Journal of Counseling Psychology
1975, Vol. 22, No. 4, 358-376

RESEARCH METHODOLOGY

This section reviews methodological problems in counseling psychology research and interprets current research trends and techniques to help counseling psychologists utilize developments in psychometrics, statistics, research design, and related areas.

Interrater Reliability and Agreement of Subjective Judgments

Howard E. A. Tinsley
Southern Illinois University at Carbondale

David J. Weiss
University of Minnesota

Indexes of interrater reliability and agreement are reviewed and suggestions are made regarding their use in counseling psychology research. The distinction between agreement and reliability is clarified and the relationships between these indexes and the level of measurement and type of replication are discussed. Indexes of interrater reliability appropriate for use with ordinal and interval scales are considered. The intraclass correlation as a measure of interrater reliability is discussed in terms of the treatment of between-raters variance and the appropriateness of reliability estimates based on composite or individual ratings. The advisability of optimal weighting schemes for calculating composite ratings is also considered. Measures of interrater agreement for ordinal and interval scales are described, as are measures of interrater agreement for data at the nominal level of measurement.

The rating scale is one of the most frequently used measuring instruments in counseling research. Although rating scales can assume a number of specific forms, they generally require the rater to make a judgment about some characteristic of an object by assigning it to some point on a scale defined in terms of that characteristic. In counseling psychology research, the object to be rated can be a person (e.g., a client or counselor) or a process (e.g., types of counseling or therapy). Among the characteristics of counselors that have been measured by rating scales in counseling research are accurate empathy, nonpossessive warmth, unconditional positive regard, genuineness, concreteness, and intensity of interpersonal contact; clients have been rated on degree of pathology, type of pathology, contribution to society, degree of disability, type of disability, quality of work adjustment, and type of verbal response; and counseling has been rated on its success and type of outcome. Rating scales have been used recently in studying the effectiveness of counselor-training techniques (Carkhuff & Griffin, 1970; Martin & Gazda, 1970; Myrick & Pare, 1971; Payne, Winter, & Bell, 1972; Pierce & Schauble, 1971; Truax & Lister, 1971), counseling outcomes (Garfield, Prager, & Bergin, 1971; Schuldt & Truax, 1970), the relationship of counselor verbal behaviors (McMullin, 1972) and counselor attire (Stillman & Resnick, 1972) to counseling outcomes, the content of interviews conducted by counselors of different theoretical persuasions (Wittmer, 1971), the relationship of client attributes to client employability (Tseng, 1972), and the effects of tape recording an interview on clients' self-reports (Tanney & Gelso, 1972).

Because the datum recorded on a rating scale is the subjective judgment of the rater,

While the authors assume complete responsibility for the contents of this paper, thanks are due to Joseph L. Fleiss, William G. Miller, John R. Moreland, Gordon F. Pitz, and Diane J. Tinsley for their careful reviews and many helpful comments.

Requests for reprints should be sent to Howard E. A. Tinsley, Department of Psychology, Southern Illinois University, Carbondale, Illinois 62901.

the generality of a set of ratings is always of concern. Generality is important in demonstrating that the obtained ratings are not the idiosyncratic results of one rater's subjective judgment. Knowledge of the interrater reliability and interrater agreement is crucial in evaluating the generality of a set of ratings. In almost all of the research published to date in which rating scales have been used, however, the interrater agreement of the ratings has not been reported. Failure to report the interrater reliability of ratings is also not uncommon (e.g., Carkhuff, 1971). Moreover, interrater reliability frequently has been reported as having been determined by other investigators in other research (e.g., Cannon & Carkhuff, 1969). Such practices are unacceptable.

This article differentiates between interrater reliability and interrater agreement, emphasizes the need for both types of evidence regarding a set of ratings, summarizes the various procedures suggested for determining interrater reliability and interrater agreement, presents methods for combining ratings to maximize interrater reliability, and, finally, recommends procedures for use in counseling research.

Agreement Versus Reliability

Interrater agreement represents the extent to which the different judges tend to make exactly the same judgments about the rated subject. When judgments are made on a numerical scale, interrater agreement means that the judges assigned exactly the same values when rating the same person. Interrater reliability, on the other hand, represents the degree to which the ratings of different judges are proportional when expressed as deviations from their means. In practice, this means that the relationship of one rated individual to other rated individuals is the same although the absolute numbers used to express this relationship may differ from judge to judge. Interrater reliability usually is reported in terms of correlational or analysis of variance indexes.

Table 1 shows hypothetical data (assuming interval-level measurement) in which three judges have rated 10 counselors on the Truax and Carkhuff (1967) 9-point Accurate Empathy Scale. Case 1 in Table 1 represents a set of ratings in which all three judges assign exactly the same ratings to each of the 10 counselors. These ratings have both high interrater agreement and high interrater reliability. Case 2 in Table 1 shows ratings with low interrater agreement but high interrater reliability (as indicated by the intraclass correlation). These ratings have low interrater agreement, since no two raters gave the same rating to any counselor. Low agreement among the raters is re-

TABLE 1

Hypothetical Ratings of Accurate Empathy Illustrating Different Levels of Interrater Agreement and Interrater Reliability for Interval-Scaled Data

Counselor	Case 1: High interrater agreement and high interrater reliability			Case 2: Low interrater agreement and high interrater reliability			Case 3: High interrater agreement and low interrater reliability		
	Rater			Rater			Rater		
	1	2	3	1	2	3	1	2	3
A	1	1	1	1	3	5	5	4	4
B	2	2	2	1	3	5	5	4	3
C	3	3	3	2	4	6	5	4	5
D	3	3	3	2	4	6	4	4	5
E	4	4	4	3	5	7	5	4	3
F	5	5	5	3	5	7	5	5	4
G	6	6	6	4	6	8	4	4	5
H	7	7	7	4	6	8	5	5	4
I	8	8	8	5	7	9	4	5	3
J	9	9	9	5	7	9	5	5	5
M	4.8	4.8	4.8	3.0	5.0	7.0	4.7	4.4	4.1
SD	2.7	2.7	2.7	1.5	1.5	1.5	.5	.5	.9

flected in their mean ratings of 3, 5, and 7, respectively, which translate into quite different statements about the counselors in terms of level of accurate empathy. However, the ratings assigned to the counselors by the three judges are proportional since the counselors are ordered similarly by the three judges. If the investigator is interested only in the relative ordering of the counselors by the judges, the interrater reliability is a satisfactory index of the quality of the ratings. The interrater reliability, however, fails to make evident the fact that the raters differed in their ratings of the counselors. Whenever the counselor is interested in the absolute value of the ratings, or the meaning of the ratings as defined by the points on the scale, the interrater agreement of the ratings must also be reported.

Finally, the ratings of Case 3 in Table 1 have high interrater agreement but low interrater reliability. The high interrater agreement results from the fact that the ratings assigned to the counselors by the three raters are quite similar for 7 of the 10 counselors. The interrater reliability is low primarily because of the restricted range of ratings given by the three raters. Such ratings may have occurred because the counselors were all highly similar in accurate empathy or because the judges used the rating scale improperly. Thus, when the variability of ratings is small, as is frequently the case in counseling research, interrater reliability may be low. If, however, interrater agreement is high, the possibility exists that the subjects are homogeneous on the trait of interest. This possibility can be investigated by further study of the same subjects or by having the judges rate a sample of subjects known to be heterogeneous on the trait of interest. When both interrater reliability and agreement are low, the ratings are of no value and should not be used for research or applied purposes.

In summary, high reliability is no indication that the raters agree in an absolute sense on the degree to which the ratees possess the characteristic being judged (Case 2). On the other hand, low reliability does not necessarily indicate that the raters are in disagreement (Case 3). Clearly, both types of information are important in evaluating subjective ratings.

Level of Measurement

Measurements obtained from the rating instruments used in counseling research sometimes possess the properties of nominal-level measurement. They more frequently have the properties of ordinal-level measurement but rarely can be considered to be at the interval level of measurement. Each level of measurement requires different approaches to the estimation of interrater reliability and agreement.

Nominal scales involve simple classification without order. At the nominal level of measurement the rater generally is concerned with assigning the person to be rated to one of a number of qualitatively different categories. Examples of nominal rating include clinical diagnosis (e.g., schizophrenic, psychopathic), diagnosis of physical disability (e.g., paraplegic, quadraplegic), and classification of presenting complaint (e.g., vocational concern, social–emotional concern).

Ordinal measurement implies the ability to rank individuals on some continuum, but without implying equal intervals between the ranks. Ordinal scales, in the form of the simple graphic rating scale, are probably the most frequently used scales in counseling research. Such scales may vary considerably in their appearance. The two dimensions along which these scales most commonly vary are the number of scale levels (e.g., 3-, 6-, or 11-point scales) and the manner in which the scale levels are defined. Scale levels can be defined by numerical values, adjective descriptions, or graphic intervals. Regardless of how scale points are defined, all such scales represent an ordinal level of measurement.

At the interval level of measurement, true measures of distance are possible between individuals on a continuum. While technically most counseling research instruments do not meet the formal characteristics of interval-level measurement, Baker, Hardyck, and Petrinovich (1966) have shown that statistics which assume interval-level measurement can be applied to ordinal data without distorting the sampling distribution of the statistic. Thus, the counseling psychologist may appropriately use interval-level statistics on ratings that result from only ordinal level measurement. The researcher should be aware, however, that such applications might result in measures of interrater reliability and agreement that distort the true relationship among raters, if the assumption of equal intervals is grossly inappropriate.

The distinction between interrater reliability and interrater agreement blurs when one deals with nominal scales. Since the rating categories do not differ quantitatively, the disagreements in categorization generally do not differ in their severity.[1] Therefore, at the nominal-scale level the concept of "proportionality of ratings," which is central to interrater reliability, ceases to make sense. Agreement becomes an absolute; ratings are either in agreement or disagreement. As a result, no distinction exists between interrater reliability and interrater agreement for nominal scales. The term *interrater agreement* is more appropriate, since it is more consistent with the terminology employed at the ordinal and interval levels of measurement.

Type of Replication

Investigation of the generality of a set of ratings involves replication or repetition of the obtained ratings. Most frequently, replication occurs between judges when several judges rate the same objects on the same variable, as illustrated in Table 1. Another method of replicating a set of ratings is the rate–rerate method, the analogue of the test–retest procedure for determining test reliability. In this method the judges rate a group of subjects on some characteristic, then rerate the same subjects on the same characteristic at a later date. This replication has been used recently by Carkhuff (1969), Carkhuff and Banks (1970), and Carkhuff and Griffin (1970).

The rate–rerate method is inappropriate for demonstrating interrater agreement or interrater reliability for several reasons. First, this measure of reliability does not indicate the degree to which the ratings of the different judges are in agreement. Rather, it provides a measure of the consistency or stability of the judges' ratings over time. It would be possible, for example, to achieve a high rate–rerate reliability for three raters who were in complete disagreement, as long as each rater was consistent across time. A second problem with the rate–rerate method is that over a period of time the subjects may change in terms of the characteristics being rated, thus causing the reliability or within-rater agreement of the ratings to be low. The more accurately the judges reflect this change by amending their previous ratings, the lower the reliability will appear. Rate–rerate methods may also give spuriously high reliabilities when a nonchanging data base is used (e.g., audio or video tapes of interviews) if the between-ratings time interval is short. In this situation the reratings will likely be contaminated by the rater's recall of his previous judgments. Finally, because the rate–rerate method requires that the judges rate each subject on two different occasions, this method is seldom feasible in the field situations in which much counseling research is conducted. Thus replication of ratings between judges, rather than across time, is the more appropriate method for counseling psychology research.

Interrater Reliability

Ordinal Scales

Finn (1970) has recommended the use of the following index as a measure of interrater reliability for ordinal ratings:

$$r = 1.0 - \frac{\text{Observed variance}}{\text{Chance variance}}. \qquad (1)$$

The expected or chance variance (s_c^2) of the ratings is the variance expected if the ratings were assigned at random and may be calculated as[2]

$$s_c^2 = \frac{k^2 - 1}{12}, \qquad (2)$$

[1] An exception is discussed later in the section on "weighted kappa."

[2] The authors are indebted to Joseph L. Fleiss for suggesting the use of this equation in place of a more cumbersome equation used in an earlier draft of this article.

where k = the number of scale categories. In the unusual case in which only one subject has been rated, the observed variance is simply the variance of the assigned ratings. When ratings are available for more than one subject, the observed variance is the within-subjects mean square obtained from a one-way analysis of variance (ANOVA).

The degree to which the observed variance is less than the expected variance is an indication of the amount of nonchance variance in the ratings. The ratio of the two variances gives the proportion of the chance (or random variance) present in the observed ratings. Subtracting the ratio from 1.0 results in the proportion of the total variance in the ratings that is due to nonrandom factors. An r of 1.00 indicates perfect reliability; an r of 0 indicates that the observed ratings were completely unreliable, that is, they varied as much as chance ratings. For the data in Table 1, r = 1.00, .40, and .93 for Cases 1, 2, and 3, respectively.

As we previously noted, indexes of interrater reliability are usually affected by reduced variance in the ratings. One advantage of Finn's (1970) measure of reliability is that it avoids this problem, as illustrated by the results of Case 3. Finn's r can vary from 0 to 1.0 and is not reduced by low within-judge variance.

The chance variance would be expected if an infinite population of judges assigned the ratings at random. In actual practice the counseling psychologist will be dealing with a finite and often small sample of judges. Under such circumstances the random assignment of ratings could result in an observed variance that, quite by chance, is smaller than the expected variance. Accordingly, the investigator should determine whether the observed variance is significantly less than the chance variance prior to calculating r. The hypothesis that the observed variance is equal to the chance variance can be tested with the following chi-square test:[3]

$$\chi^2 = \frac{N(K-1)S_0^2}{S_c^2}, \qquad (3)$$

with $N(K - 1)$ degrees of freedom; where N = number of subjects, K = number of raters, S_0^2 = the observed within-subjects variance and S_c^2 = the expected chance variance. Using the one-tailed chi-square test (Hays, 1963, pp. 344–45), a chi-square value lower than the lower critical value (i.e., the .99 level for a test at $p \leq .01$) would indicate that S_0^2 is significantly less than S_c^2. Finn's r should be calculated only in those instances in which the null hypothesis is rejected.

Two considerations must be kept in mind in following this procedure. First, this use of chi-square assumes that the ratings will be normally distributed and that both the subjects and raters will be randomly selected. In counseling research these assumptions will frequently be violated. Violation of the normality assumption can be quite serious when inferences are made about variances (Hays, 1963, p. 347). Accordingly, a stringent critical value (e.g., $p < .01$) is advised. Second, a significant chi-square tells the investigator only that his results are not consistent with the hypothesis of *completely* random responding.

Two problems are apparent with the use of Finn's (1970) r as an index of interrater reliability. First, computation of the chance variance requires the assumption that every rating has the same probability, because the ratings of the judges are purely random. This assumption is violated whenever the judges have a response set to avoid the extreme categories of the rating scale. Whenever this occurs, the chance variance, as calculated using Equation 2, would be greater than the "true" chance variance, thus causing r to be spuriously high.

A second cause for concern is the lack of systematic evidence available for the evaluation of this measure of reliability. This is due, in part, to its relative recency. In the comparative studies published to date (Finn, 1970, 1972), however, r and the intraclass correlation (discussed later) gave highly similar results in most instances. When the variance in the ratings is severely restricted (as in Case 3), however, r and the intraclass correlation may differ substantially.

[3] Finn (1970, p. 73) has suggested an F ratio for this purpose. This statistic appears inappropriate since it is intended for use with two sample variances (see Hays, 1963, pp. 348–355), whereas Finn's chance variance is a theoretically determined estimate of the population variance.

Interval Scales

The measure of interrater reliability most commonly used with interval data (and with ordinal scales which assume interval properties) is the intraclass correlation (R). R can be interpreted as the proportion of the total variance in the ratings due to variance in the persons being rated. Values approaching the upper limit of R (1.00) indicate a high degree of reliability, whereas an R of 0 indicates a complete lack of reliability. (Negative values of R are mathematically possible but are rarely observed in actual practice; when observed, they imply Rater × Ratee interactions.) The more R departs from 1.00, the less reliable are the ratings.

Ebel (1951) compared the product-moment correlation (applicable when only two persons are being rated), the intraclass correlation, and two other indexes of interrater reliability. He concluded that the intraclass correlation was preferable, because it permits the inclusion or exclusion of the between-rater variance as part of the error variance, because it allows an estimation of the precision of the reliability coefficient, and because it uses the familiar statistics and computational procedures of analysis of variance. At the present time, the intraclass correlation is the most appropriate measure of interrater reliability for interval scale data (Ebel, 1951; Englehart, 1959; Guilford, 1954).

More than one formula is available for the intraclass correlation. In order to make proper use of this correlation, the investigator must decide (a) whether mean differences in the ratings of the judges should be regarded as rater error, and (*b*) whether he or she is concerned with the average reliability of the individual judge or the reliability of the average rating of all the judges. The appropriate form of the intraclass correlation can then be chosen based on these decisions.

Between-Raters Variance

The data of Case 2 in Table 1 are an example in which the judges differ in their average ratings. The means of the ratings given by the three judges are 3.0, 5.0, and 7.0, respectively. These "level" differences constitute the only differences in the ratings given by the three judges. If the between-judges level differences are ignored, the interrater reliability as measured by the intraclass correlation will be 1.00. If, on the other hand, the between-judges variance is considered as error, the intraclass correlation is .18. In any case in which the mean or variance of the ratings assigned by the various judges differs, the decision regarding the between-raters variance will influence the interrater reliability of the data. If the differences in mean and/or variance are sizable, the exclusion of the between-raters variance from the error term will cause the reliability coefficient to be substantially higher than if it were included.

The desirability of removing the interjudge variance in estimating the interrater reliability depends on the way in which the ratings are to be used (Ebel, 1951). If between-judges differences in the general level of the ratings do not lead to corresponding differences in the ultimate classification of the subjects, the between-raters variance should not be included in the error term. Thus, when decisions are based on the mean of the ratings obtained from a set of observers, or on ratings which have been adjusted for rater differences (such as ranks or Z scores), the interjudge variance should not be regarded as error. On the other hand, if decisions are made by comparing subjects rated by different judges or sets of judges, or if the investigator wishes his results to be generalizable to other samples of judges using the same scale with a similar sample of subjects, the between-raters variance should be included as part of the error term.

Bartko (1966), Ebel (1951), Englehart (1959), Guilford (1954), and Silverstein (1966) have discussed the computation of the intraclass correlation. When the between-judges variance is not be to included in the error term, a two-way ANOVA procedure is employed, in which mean square for persons (MS_p), mean square for judges, mean square for error (MS_e), and total mean square are obtained following standard ANOVA computations. With the assumption of no interaction between persons and judges, the standard equation for the intraclass correlation is

$$R = \frac{MS_p - MS_e}{MS_p + MS_e(K - 1)}, \quad (4)$$

where K = the number of judges rating each person. Application of this formula to the data in Table 1 gives R = 1.00, 1.00, and −.008 for the three cases, respectively.

Use of Equation 4, based on the two-way ANOVA procedure, is justified only if the raters are regarded as "fixed" (i.e., the judges represent the *population* of judges; Bartko, 1966; Burdock, Fleiss, & Hardesty, 1963). This means the judges cannot be assumed to be a *sample* of judges from any population of judges. Accordingly, the resulting intraclass correlation represents the interrater reliability for only that set of judges and cannot be generalized to any other set of judges employing the same rating scale on the same sample of persons. In order to generalize the interrater reliability to other judges rating the same or other subjects, the investigator must satisfy two requirements not implied by the previously mentioned use of the intraclass correlation. First, the investigator must use a random sample of judges so that the judges may be considered representative of the population of judges about which generalizations are to be made. In few counseling research studies has this requirement been met. Second, the equation must incorporate some estimation of the between-judges variance in order to estimate the degree to which R will vary across samples of judges.

The standard one-way ANOVA, in which only mean square for persons (MS_p) and mean square for error (MS_e) are calculated, is the simplest method of computation when the between-judges variance is to be included in the error term (Ebel, 1951; Bartko, 1966). If one uses the one-way ANOVA (vs. the two-way ANOVA), the mean square for judges is included in the mean square for error. The intraclass correlation can then be calculated using Equation 4. Under these circumstances, the intraclass correlations for the data in Table 1 are R = 1.00, .18, and −.10, respectively. These results take into account the differences in mean level of the ratings by the raters in each of the three cases. The differences between the two estimates of reliability are largest for Case 2, where the largest mean differences among raters occurred.

In counseling research the investigator is frequently confronted with incomplete data. A problem arises in determining the appropriate value of K in Equation 4 when subjects have been rated by varying numbers of raters. When the investigator plans to use the one-way ANOVA procedure, an average value of K can be obtained using the following equation from Snedecor (1946):

$$\bar{K} = \frac{1}{N - 1}\left(\sum K - \frac{\sum K^2}{\sum K}\right), \quad (5)$$

where $\bar{K}$ = the average value of K to be inserted in Formula 4 for K, N = the number of subjects, and K = the number of judges rating each subject (this value will vary from subject to subject).

One advantage of the intraclass correlation is the possibility of determining confidence intervals around R using equations given by Ebel (1951). In most counseling research, however, the small number of raters employed will result in the lower limit of the confidence interval being below zero. Thus, calculation of confidence intervals for R will emphasize the danger of attempting to generalize interrater reliability coefficients.

Finally, the comparison of Finn's r with the two methods of computing R is informative. Apparently r is, indeed, relatively independent of the observed variance in the ratings (see Cases 1 and 3), whereas R is substantially lowered when the variance in the ratings is restricted. Moreover, Case 2 illustrates that r does include between-judges variance in the error term, as is obvious from the manner in which it is calculated.

Reliability of Composite Ratings

Average reliability or reliability of a composite. The intraclass correlation and Finn's r give the average of the reliabilities of the individual judges. As such, this measure of reliability underestimates the interrater reliability of the composite rating (e.g., a mean rating or the sum of the ratings) of the group of judges. The counseling psychologist must select the appropriate estimate of interrater reliability on the basis of the intended use of the ratings. If conclusions are typically drawn from the ratings of a single judge, the average reliability of the individual ratings (intraclass correlation, or Finn's r) is appropriate. This is true even though the ratings of several judges may be available in the experimental situation. If, on the other hand, the composite rating of a group of raters is the variate of interest, the reliability of the composite rating is more appropriate. This reliability will be higher than the reliability indicated by the intraclass correlation and can be estimated by the Spearman-Brown formula, or by the following equation (Ebel, 1951):

$$R = \frac{MS_p - MS_e}{MS_p}. \qquad (6)$$

The two equations yield identical results; studies by Clark (1935), Rosander (1936), and Smith (1935) show that the Spearman-Brown formula accurately predicts the change in the reliability of the ratings. Guilford (1954) has pointed out that this finding implies that gains in interrater reliability come from multiplying the number of raters when the initial reliability is low. Reliability increases rapidly as a result of adding the first several raters, but smaller increments in reliability occur with additional raters.

Differential weighting. The question has been raised as to whether some method of differentially weighting the ratings given a subject might increase reliability more than the simple unit weighting procedure implied in Spearman-Brown types of estimates. Kelley (1947), for example, suggested weights for combining ratings by different observers that were based on the reliability of the judges, as estimated from the intercorrelation matrix of the judges' ratings. Lawshe and Nagle (1952) reported, however, that the use of Kelley's weights offered no improvement over unit weighting and, in some instances, resulted in a composite score that was less reliable than the composite score based on unit weights.

Overall (1965) re-examined the question of differential weights and concluded that the reliabilities and variances of the individual judges were the major determinants of the efficacy of differential weights. Whenever the judges can be assumed equal in these respects, unit weights should be used, and the increase in reliability due to the use of a composite score can be estimated by the Spearman-Brown formula or Equation 6. Differential weights will yield a more reliable composite score than unit weights only when the judges' ratings differ substantially in reliability and/or variance. Overall recommended a factor-analytic procedure[4] for estimating the individual rater reliabilities and provided equations for calculating optimal weights for each judge and for estimating the reliability of the weighted composite of the ratings.[5]

The counseling psychologist must consider several factors in determining the desirability of using optimal weights as opposed to unit weights. If the judges' ratings do not differ substantially in reliability and/or variance, unit weights are recommended. Moreover, if the nature of the decision to be made does not require the greatest possible reliability, unit weights will probably suffice. Another factor to be considered is the stability of the rating situation. Whenever the rating scale is modified, the membership of the group of judges changes, or the reliability or variance of the various judges' ratings changes, a new set of optimal weights must be calculated. If raters receive consistent feedback regarding the reliability of their ratings, for example, the low-reliability raters probably will be able to improve their reliability. Thus, optimal weights cannot be assumed to be stable across rating occasions, even though the same group of judges uses the same rating scale in rating the same or a similar group of subjects.

[4] The reader not familiar with factor analysis will find helpful reviews in Weiss (1970, 1971).

[5] Krippendorff (1970) has suggested an analysis of variance procedure for estimating the reliability of the individual judges. Sufficient data for the comparison of Overall's (1965) and Krippendorff's strategies has not yet been published. Overall's procedure is conceptually simpler, however, and computer programs for factor analysis are widely available. Krippendorff's procedures require a form of computer analysis for which computer programs are not readily available.

Interrater Agreement

Only recently have statistical indexes been designed specifically to indicate interrater agreement. However, several statistical indexes designed for other purposes have been employed as measures of interrater agreement. These measures include the proportion or percentage of agreement (*P*), the pairwise correlation between judges' ratings, and various chi-square indexes.

Guttman, Spector, Sigal, Rakoff, and Epstein (1971), Kaspar, Throne, and Schulman (1968), and Mickelson and Stevic (1971) have employed the percentage or proportion of judgments in which the judges are in agreement (*P*) as a measure of interrater agreement. For the three cases in Table 1, *P* is 1.00, .00 and .10, respectively. The advantages of *P* include ease of calculation and the fact that its meaning is easily understood.

Cohen (1960) and Robinson (1957) have criticized the percentage or proportion as an index of agreement. One problem is that *P* treats interrater agreement in an absolute, all-or-none fashion. For example, the data in Case 2 of Table 1 represent perfect agreement among the judges on the relative level of each of the counselors, but the proportion of agreement index indicates that the absolute agreement is zero. Another problem is that agreement can be expected on the basis of chance alone; *P* overestimates the true absolute agreement by an amount related to the number of raters and number of points on the scale. Some adjustment in *P* that would show the proportion of nonchance agreement therefore is desirable. These difficulties have been partially avoided in more recent attempts to use *P* as a measure of interrater agreement, and, as will be shown below, *P* serves as the basis for measures of agreement in nominal ratings.

The pairwise correlation of the judges' ratings has also been used as a measure of interrater agreement (e.g., Kaspar, Throne, & Schulman, 1968). The shortcomings of pairwise correlations as a measure of interrater agreement should be obvious; correlations show proportional agreement or agreement of standardized ratings. If pairwise correlation has any merit in evaluating ratings, it is as a measure of interrater reliability, not interrater agreement. As Ebel (1951) reported, the intraclass correlation is superior to pairwise correlations as a measure of interrater reliability, since the former permits the investigator to specify the variance components included.

Finally, various indexes based on chi-square have been used as measures of interrater agreement (e.g., Taylor, 1968). Such indexes have been criticized by Cohen (1960), Lu (1971), and Robinson (1957) on the grounds that chi-square does not reflect the degree of agreement among a group of judges. Chi-square tests the hypothesis that the proportions of subjects assigned to the various rating categories by the different judges do not differ significantly. A nonsignificant chi-square indicates that the observed disagreement is not greater than the disagreement that could be expected on the basis of chance. No inferences can be made from a nonsignificant chi-square regarding the degree of agreement. A significant chi-square can occur because of a departure from chance association in the direction of greater agreement or greater disagreement. A significant chi-square may indicate that the observed disagreement is greater than the disagreement expected on the basis of chance.

Problems have been noted in the use of Kendall's coefficient of concordance as a measure of interrater agreement. This measure gives the average Spearman rank-order correlation between each pair of judges. In essence, the coefficient of concordance ignores differences in the absolute level and the dispersion of the rankings assigned by the various judges by forcing the data to take the form of ranks; it indicates the agreement among the serial orders assigned to the subjects. Lu (1971) pointed out that the frequent occurrence of tied ranks, which commonly results when the coefficient of concordance is used, makes the coefficient difficult to calculate and somewhat powerless.

Ordinal and Interval Scales

Two measures of interrater agreement have recently been formulated which merit the attention of counseling psychologists using ordinal or interval scales (Lawlis & Lu, 1972; Lu, 1971). The measures differ markedly in their concept of interrater agreement.[6] When viewed from the perspective of decision theory (Cronbach & Gleser, 1965), the two indexes offer alternative methods of quantifying the seriousness or the cost of various disagreements.

The Lawlis and Lu (1972) measure of interrater agreement allows the investigator some flexibility in selecting a criterion for agreement, thereby avoiding the necessity of treating agreement in an absolute, all-or-none fashion. For example, three raters who gave counselor ratings of 7, 8, and 9 on Truax and Carkhuff's (1967) 9-point Accurate Empathy Scale disagreed in an absolute sense, but all essentially rated the counselor as high in accurate empathy. Lawlis and Lu's (1972) index allows the option of defining agreement as identical ratings, as ratings that differ by no more than 1 point, or as ratings that differ by no more than 2 points. Agreement is tallied one subject at a time, by determining whether the total set of ratings given that subject satisfies the criterion. If agreement is defined as ratings not more than two scale categories apart, the ratings in the above example would be interpreted as in agreement.

The Lawlis and Lu (1972) approach, therefore, uses a flexible model of the seriousness of disagreements among the raters, and the investigator can distinguish between serious and unimportant disagreements. When ratings that differ by 1 scale point or less are defined as "in agreement," the intent is that a disagreement of 1 scale point is unimportant. Conversely, all disagreements that exceed the criterion (e.g., disagreements of 2 or more) are of equal seriousness.

Lawlis and Lu (1972) suggest the following nonparametric chi-square as a test of the significance of interrater agreement:

$$\chi^2 = \frac{(N_1 - NP - .5)^2}{NP} + \frac{(N_2 - N(1 - P) - .5)^2}{N(1 - P)}, \qquad (7)$$

where:

N_1 = the number of agreements,
N = the number of individuals rated,
P = the probability of chance agreement on an individual[7],
$.5$ = a correction for continuity, and
N_2 = the number of disagreements.

The statistic is distributed as chi-square with 1 degree of freedom. This test is appropriate only when the interrater agreement (N_1) is greater than the agreement expected on the basis of chance (NP).

Failure to obtain a significant chi-square means that the hypothesis of "randomly assigned ratings" cannot be rejected. Lack of significant agreement may result from some characteristics of the scale (e.g., vaguely defined or overlapping categories), the judges (e.g., carelessness, lack of proper training, inability to make the required discrimination), or the subjects (e.g., failure to emit the behavior to be rated). When a nonsignificant chi-square is obtained, the use of the scale in the manner in which the data were obtained is questionable.

A significant chi-square, on the other hand, indicates that the observed agreement is greater than the agreement that could be expected on the basis of chance. The investigator, however, also should be concerned with whether interrater agreement is high, moderate, or low, not only with whether it is better than chance. We propose the following as a measure of agreement:

$$T = \frac{N_1 - NP}{N - NP}, \qquad (8)$$

where N_1, N, and P are defined as in Equation 7. The value T, which is patterned after Cohen's (1960) κ (to be discussed later),

[6] The authors are indebted to Gordon F. Pitz for pointing out this distinction.

[7] Tables prepared by the authors are available for values of P most likely to arise in counseling psychology research. These tables, available from the first author, include P for from 2 to 10 judges on scales using 2–20 rating categories, and defining agreement as 0, 1, or 2 points discrepancy. While Lawlis and Lu (1972, p. 19) present equations for the calculation of P, their equations for 1 and 2 points discrepancy contain typographical errors. In addition, their P value for a 10-point scale with four judges ($K = 4$) and agreement defined as a discrepancy of 1 point or less ($r = 1$) should be .0136, instead of .00136, as shown in their Table 4.

should be calculated only when the hypothesis of chance agreement has been rejected. When the observed agreement is equal to the expected chance agreement, T is 0, and T is 1.0 when perfect interrater agreement is observed. Positive values of T indicate that the observed agreement is greater than chance agreement, while negative values indicate that observed agreement is less than chance agreement. The number of agreements for Case 3 of Table 1 (where agreement is defined as 0, 1, and 2 points discrepancy) is 1 (Counselor J), 7 (Counselors A, C, D, F, G, H, and J), and 10 (all counselors), respectively. The corresponding T values are .09, .68, and 1.00.

The results from Lawlis and Lu's (1972) chi-square and the associated T index are contingent upon the definition of agreement. If the definition is changed, the results will change. The counselor must study his rating scale carefully and determine the implications of the alternative definitions of agreement in the context of the research question. Only after careful study should a definition of interrater agreement be selected. That definition must, of course, be determined prior to the collection of the data. Moreover, the definition of agreement and the rationale for adopting the definition must be specified in any report on the interrater agreement of the ratings. Whenever the counseling investigator defines agreement to include a discrepancy of 1 or 2 points, he or she should also report the chi-square and the T value for agreement defined as identical ratings. This will allow the reader to evaluate the extent to which the conclusions drawn are contingent upon the definition of agreement.

One further consideration must be kept in mind when using Lawlis and Lu's (1972) chi-square. This statistic requires the assumption that every judgment has the same probability, under the hypothesis that the ratings of the judges are purely random. As was pointed out earlier, however, this assumption may not be completely applicable in some circumstances, if raters do not use the end categories of a rating scale. When this happens, the range of choice is narrowed, and the true probability of chance agreement is greater than P. P, then, is the lower limit for the unknown probability of chance agreement. This means that the probability of chance agreement is often underestimated and the significance of the observed agreement is overstated.

The careful investigator can meet this problem in two ways. He or she can adjust the required level of significance from the traditional .05 to a more stringent level or calculate the probability of chance agreement on fewer scale categories than are actually available. Thus, the investigator using a 7-point scale could employ the probability of chance agreement for a 5- or 6-point scale.

Lu (1971) also noted that agreement can vary along a continuum from absolute agreement to no agreement. He developed a coefficient of agreement which incorporates definitions of minimum agreement derived from analysis of variance and information theory. In analysis of variance terms, minimum agreement occurs when the within-subjects variance is at a maximum (i.e., when the ratings are divided equally between the two extreme categories). According to information theory, minimum agreement is achieved when the judges assign any one of the ratings to a given subject with equal likelihood.

In contrast to Lawlis and Lu (1972), Lu's (1971) approach defines all rating disagreements as important, but the seriousness of the disagreement increases as a function of the squared magnitude of the disagreement. Thus, a disagreement of 5 points is more serious than a disagreement of 4 points; a disagreement of 1 point is more serious than a disagreement of 0 points.

Lu's measure of interrater agreement is defined as[8]

$$A = \frac{S_c^2 - S_0^2}{S_c^2}, \qquad (9)$$

where S_c^2 = the expected within-subjects variance when all the ratings are equally likely, and S_0^2 = the observed within-subjects variance. A approaches 0 (reflecting

[8] While this equation is algebraically equivalent to Equation 1, Lu (1971) and Finn (1970) define both S_c^2 and S_0^2 differently.

random interrater agreement) as S_0^2 approaches S_c^2; A approaches 1.00 (reflecting perfect interrater agreement) as S_0^2 approaches 0. A can be negative, in which case it indicates that disagreement is greater than would be observed under purely chance responding. In Table 1, A equals 1.0, .34, and .60 for Cases 1, 2, and 3, respectively.

Lu's A assumes the attribute under consideration is measurable conceptually on a continuous scale. Due to practical limitations, however, attributes almost always are rated in terms of an ordered set of nonoverlapping categories. This necessitates the calculation of a set of weights to be used in determining S_c^2 and S_0^2. The weights are calculated as follows:

$$Y_m = \sum_{i=1}^{m-1} \frac{N_r + \frac{N_m}{2}}{KN}, \tag{10}$$

where:

Y_m = the weight assigned to Category m,
N_r = the number of subjects assigned to Categories 0 through $m - 1$,
N_m = the number of subjects assigned to Category m,
KN = the number of judges times the number of subjects, and
m = 1, 2, . . . , k categories.

Once calculated, the weights are placed in a Subjects × Judges matrix in place of the actual ratings. The matrix is analyzed following a two-way factorial ANOVA procedure, which yields sum of squares for between subjects, within subjects, and total. Mean square within subjects represents S_0^2. Within-subjects variance under chance responding (S_c^2) is obtained as follows:

$$S_c^2 = \frac{\sum Y_m^2}{k} - \left(\frac{\sum Y_m}{k}\right)^2, \tag{11}$$

where Y_m = the weight assigned the mth category, and k = the number of categories. As with Finn's (1970) r and the Lawlis and Lu (1972) chi-square, this definition of chance variance requires the assumption that every rating has the same probability, because the ratings of the judges are purely random.

Given the hypothesis that the ratings were assigned randomly, the expected value of S_0^2 would be S_c^2. Thus, the statistical significance of A can be determined indirectly using Equation 3. A chi-square value lower than the lower critical value (i.e., the .99 level for a test at $p \leq .01$) would indicate that S_0^2 is significantly less than S_c^2, thereby implying that A is significantly greater than 0. Again, a stringent critical value is advised because of the violation of the normality assumption.

Because of their recency, neither Lawlis and Lu's (1972) nor Lu's (1971) index has undergone rigorous, systematic investigation. Consequently, both indexes should be used with caution. The use of Lawlis and Lu's (1972) chi-square and the T index is recommended for two reasons. First, their treatment of interrater disagreements allows the counselor to more adequately distinguish rating disagreements that have no serious consequences from those that are of practical significance. An unfortunate corollary, however, is that this same flexibility may tempt the researcher to manipulate his or her definition of agreement in order to make the results reach desired levels of agreement. Second, Lu's (1971) index is more appropriately classified as a measure of interrater reliability, despite the fact that he refers to it as a measure of agreement. Lu's (1971) strategy for measuring agreement is quite similar to Finn's (1970) concept of reliability, however, and the pattern of results obtained for Cases 1, 2, and 3 in Table 1 (r = 1.00, .40, and .93; A = 1.00, .34, and .60) are quite similar. More research is needed to clarify the relationships between r and A.

Nominal Scales

Numerous writers have discussed the problem of the interrater agreement of nominal scales (Cohen, 1960, 1968; Everitt, 1968; Fleiss, 1971; Fleiss, Cohen, & Everitt, 1969; Goodman & Kruskal, 1954; Scott, 1955). All of the measures suggested by these authors are based on the percentage or proportion of agreements among judges (P). Two objections have previously been raised regarding the use of P. The fact that P

treats agreement as an absolute is an advantage rather than a disadvantage when dealing with nominal ratings, since all disagreements usually are regarded as equally serious. The problem of chance agreement remains, however. Some method of representing P as an improvement over chance agreement must be found if P is to be useful as an index of interrater agreement. Guttman et al. (1971) concluded, after a review of the literature, that there was a "tacit" consensus that 65% represented the minimum acceptable agreement. Such a standard, however, would allow the judges to be in disagreement more than one third of the time. We do not recommend acceptance of such a crude rule of thumb. The following measures of interrater agreement represent observed agreement as a function of chance agreement and provide for statistical tests of the significance of the results.

TABLE 2

Hypothetical Proportions for Categorizations of Client Interview Statements by Two Judges

Judge B	Judge A				Row total
	Negative self-reference	Positive self-reference	Request for information	Goal setting	
Negative self-reference	**.18**	.00	.02	.00	.20
Positive self-reference	.00	**.18**	.12	.00	.30
Request for information	.06	.00	**.24**	.00	.30
Goal setting	.06	.02	.02	**.10**	.20
Column total	.30	.20	.40	.10	

Note. N = 100 statements. Boldface entries indicate observed proportion of agreement for each rating category.

Two Judges

Cohen (1960) has suggested the coefficient κ as an indicator of the proportion of agreements between two raters after chance agreement has been removed from consideration. Cohen's κ can be calculated as follows:

$$\kappa = \frac{P_0 - P_c}{1 - P_c}, \qquad (12)$$

where P_0 = the proportion of ratings in which the two judges agree, and P_c = the proportion of ratings for which agreement is expected by chance.

For the data in Table 2, which shows hypothetical data for two judges who each categorized 100 interview statements, the total proportion of agreement (P_0) is .70 (i.e., .18 + .18 + .24 + .10). The expected chance agreement can be obtained as follows:

$$P_c = \sum_{j=1}^{k} \hat{P}_{jj}, \qquad (13)$$

where $\hat{P}_{jj}$ = the diagonal values obtained by finding the joint probability of the marginal proportions (i.e., $P_{jm} \cdot P_{gm}$, where j and g are the two raters and m is a category of the rating scale), and k = the number of categories. The expected chance agreement (P_c) for the data in Table 2 is .26 [i.e., (.20 × .30) + (.30 × .20) + (.30 × .40) + (.20 × .10)] and κ is .59. Computational examples and formulas for computing κ from frequency data are provided by Cohen (1960).

When the marginal frequencies are identical, κ can vary from 1.00 to -1.00. A κ of 0 indicates that the observed agreement is exactly equal to the agreement that could be expected by chance. A negative value of κ indicates the observed agreement is less than the expected chance agreement, while a κ of 1.00 indicates perfect agreement between the raters.

Others have suggested measures of interrater agreement that are identical in form to κ but differ in their definition of chance agreement. Goodman and Kruskal (1954) define P_c as the mean of the proportions in the modal (most frequently used) categories of the two judges. In Scott's (1955) coefficient of intercoder agreement, P_c is based on the assumption that the proportion of ratings assigned to each category is equal for the judges. In contrast, Cohen's (1960) κ recognizes that judges distribute their judgments differently over categories (e.g., see Table 2) and does not require the assumption of equal distribution of ratings. The only assumptions required by κ are that the subjects to be rated are independent, the judges assign their judgments independently, and the categories of the nominal scale are independent, mutually exclusive, and exhaustive. Cohen (1960) and Fleiss et al. (1969) provide formulas for the standard error of κ, for testing the significance of the difference between two κs, and for testing the hypothesis that $\kappa = 0$.

Weighted kappa. κ is based on the assumption that all disagreements in classification are equally serious. In some instances, nominal scaling notwithstanding, the counseling researcher or practitioner may consider some disagreements among judges to be more serious than others. If, for example, two counselors classified clients seen in a college counseling center as normal, neurotic, schizoid, or psychopathic, disagreement as to whether a client was normal or schizoid might be regarded as more serious than disagreement over whether the person was normal or neurotic. Cohen (1968) has developed weighted kappa (κ_w) as an index of interrater agreement for use when the investigator wishes to differentially weight disagreements among nominal ratings.

The coefficient κ may be regarded as a special case of κ_w, in which weights of 1.0 are assigned to agreements (the diagonal values in Table 2), while weights of 0 are assigned all disagreements (the off-diagonal values in Table 2). In κ_w the various types of disagreements are assigned differential weights. If, for example, the counselor has assigned weights on the basis of degree of agreement, a weight of 50 represents twice as much agreement as a weight of 25 and five times as much agreement as a weight of 10. Weights may be assigned to indicate the amount of agreement or disagreement. The wisdom of this procedure depends, of course, upon the degree to which the transformed data represent psychological reality. Computational formulas and examples have been provided by Cohen (1968), Everitt (1968), and Fleiss et al. (1969).

When the investigator wishes to order his rating categories along a unidimensional continuum, disagreement weights can be calculated using the following equation:

$$V_{jg} = (k - c)^2, \tag{14}$$

where k = the number of rating categories, and c = the number of cells in the diagonal containing Cell jg. (For Table 2, Equation 14 would yield disagreement weights of 0, 1, 4, 9; 1, 0, 1, 4; 4, 1, 0, 1; and 9, 4, 1, 0 for Rows 1 through 4, respectively.) This rating procedure transforms nominal data to data possessing the properties of ratio measurement. Accordingly, the proportionality of the ratings is again of interest. The index κ_w deals with the proportionality of the ratings and is identical to the intraclass correlation when disagreement weights are assigned as above (Fleiss & Cohen, 1973). For the reasons discussed previously, the intraclass correlation and Lawlis and Lu's (1972) chi-square and the T index are more appropriate whenever the investigator transforms his data in this manner. The rating categories may be numbered 1 through k and the subjects given the number of the category to which they are assigned.

The index κ_w will find little valid use in counseling research. The use of κ_w is appropriate only in the special case in which the counseling psychologist wishes to assign weights that are inconsistent with a unidimensional ordering of the rating categories. This would most likely occur when he or she is operating on the basis of a theory that postulates a multidimensional relationship among the rating categories. In such instances the investigator must justify the weights assigned the rating categories by explaining in detail the theory upon which the weights are predicated. The theory, of course, becomes an integral part of the hypothesis being tested. Obviously, the weights should be specified in the research report.

Variable Set of Raters

The use of κ is limited to the situation in which the same two judges rate each subject. Fleiss (1971) has formulated an extension of κ for measuring interrater agreement when subjects are rated by different sets of judges, but the number of judges per subject is constant. Using this method of measuring interrater agreement, all judges need not rate each subject. This measure of interrater agreement would be appropriate, for example, where different groups of three counselors categorized clients according to presenting complaint, such as is illustrated in Table 3.

Like Cohen (1960), Fleiss (1971) uses κ to indicate the degree to which the observed agreement exceeds the agreement expected on the basis of chance. The subscript $_\text{v}$ in the formulas below is added to indicate that the judges may *vary* from subject to subject. Fleiss' formula for κ_v is

$$\kappa_\text{v} = \frac{P_{0_\text{v}} - P_{c_\text{v}}}{1 - P_{c_\text{v}}}, \tag{15}$$

where P_{0_v} = the proportion of ratings in which the judges agree, and P_{c_v} = the proportion of ratings for which agreement is expected by chance. Observed agreement is calculated as follows:

$$P_{0_\text{v}} = \frac{\sum_{i=1}^{N} \sum_{m=1}^{k} n_{im}^2 - Nn}{Nn(n-1)}, \tag{16}$$

where:

N = the number of subjects rated,
n = the number of ratings per subject,
k = the number of categories in the rating scale,
n_{im} = the number of judges who assign Subject i to Category m,
i = 1, 2, . . . , N subjects,
m = 1, 2, . . . , k categories;

and the proportion of agreements expected on the basis of chance is

$$P_{c_\text{v}} = \sum_{m=1}^{k} P_m^2, \tag{17}$$

where

$$P_m = \frac{\sum_{i=1}^{N} n_{im}}{Nn}, \tag{18}$$

and N, n, k, n_{im}, i, and m are defined as in Equation 16. Fleiss (1971) provides a computational example. The index κ_v for the data in Table 3 is .20. Fleiss (1971) also provides formulas for the standard error of κ_v and for testing the hypothesis that the observed agreement equals chance agreement.

Interrater agreement for one subject. In special cases, the counseling researcher may wish to study the degree of agreement concerning a specific subject. Such a need may arise, for example, when the subject's diagnosis will determine the treatment or when the investigator wishes to identify subjects for whom the rating scale is inappropriate. It might be important for the counselor to know that the presenting complaint of Client 6 in Table 3, for example, is poorly categorized, while the judges agree perfectly in their categorization of Clients 3, 5, and 8. With a larger data base, the counseling researcher might be able to identify the "Client 6" type of client, for whom the scale is inappropriate.

The degree of agreement for one subject is obtained as follows:

$$P_i = \frac{\sum_{m=1}^{k} n_{im}(n_{im} - 1)}{n(n-1)}, \tag{19}$$

where n, i, m, k, and n_{im} are defined as in Equation 16. Table 3 shows the agreement of the judges for each of the 10 subjects. The value P_{0_v} is the mean of the agreements for the individual subjects (i.e., the mean of the P_is). When the P_is are of no importance,

TABLE 3

NUMBER OF AGREEMENTS AMONG THREE JUDGES ON CATEGORIZATION OF CLIENTS BY PRESENTING COMPLAINT

Client	Type of concern			Agreement on client
	Vocational	Social-emotional	Educational	
1	2		1	.33
2		1	2	.33
3		3		1.00
4	2		1	.33
5			3	1.00
6	1	1	1	.00
7	2		1	.33
8	3			1.00
9	1		2	.33
10	2		1	.33
Agreement on category	.19	.52	.03	

however, the investigator will find Equation 16 to be computationally simpler.

Interrater category agreement. Finally, the investigator may wish to determine the extent of agreement in assigning subjects to Category *m*. This will be important in studying the rating scale. A low degree of overall agreement may occur, for example, because of low agreement regarding only one or two categories. This situation may occur when some of the scale categories are poorly defined or are overlapping. Moreover, the omission of an important category from the scale may result in the inappropriate assignment of subjects to other categories by default. Whatever the reason for the lack of agreement, this information will direct attention to the categories where the greatest disagreement occurs.

Used as an index of the extent to which the observed agreement for Category *m* exceeds the expected agreement for Category *m*, κ_m is calculated as

$$\kappa_m = \frac{\sum_{i=1}^{N} n_{im}^2 - NnP_m[1 + (n-1)P_m]}{Nn(n-1)P_mQ_m}, \quad (20)$$

where n_{im}, N, n, and P_m are as defined in Equation 16 and 18, and $Q_m = 1 - P_m$. For the three categories in Table 3, κ_m is .19, .52, and .03, respectively, and is interpreted in the same manner as κ and κ_v. Fleiss (1971) provides additional computational examples as well as techniques for testing the hypothesis that the agreement in the assignment of subjects to Category *m* is no better than chance agreement.

Summary of Recommendations

General Considerations

Whenever rating scales are employed by psychologists, special attention should be paid to the interrater reliability and interrater agreement of the ratings. Evidence regarding both the reliability and agreement of the ratings is mandatory before the ratings can be accepted.

In reporting the interrater reliability and agreement of ratings, the investigator should describe the manner in which the index was calculated (e.g., "The intraclass correlation was used with between-raters variance excluded from the error term to calculate the average reliability of the individual rater"; "Lawlis and Lu's chi-square was calculated for a 5-point scale with three raters and with agreement defined as a discrepancy of 1 point or less"), and the assumptions required by the indexes employed.

Interrater reliability and agreement are functions of the subjects rated, the rating scale used, and the judges making the ratings. Therefore, the use of estimates of interrater reliability and agreement as determined from a training tape or generalized from other research or other groups of raters is undesirable. Since all measures of interrater reliability and agreement recommended for use in this article are one trial estimates, the interrater reliability and agreement should be determined appropriately for every set of rating data.

Finally, the psychologist should keep in mind that statistical tests of the hypothesis that the observed interrater reliability or agreement is equal to 0 are only preliminary to determining the degree of reliability or agreement. Of special interest to counselors would be research designed to establish minimum standards for interrater reliability and agreement and to develop statistical tests of the hypothesis that the observed reliability (or agreement) is less than or equal to the minimum standard.

Interrater Reliability

While the rate–rerate method is frequently employed as a measure of interrater reliability, ratings evaluated in this manner must be regarded as questionable, due to bias inherent in the repeated ratings procedure. The intraclass correlation (*R*, Equation 4) is recommended as the best measure of interrater reliability available for ordinal and interval level measurement. In reporting *R* for a set of ratings, the investigator must be explicit in describing the procedure employed. In order to properly interpret the results, the reader must know whether between–raters variance was included or excluded from the error term. If the between-judges variance is excluded from the error term, the investigator should point out that this precludes the possibility of generalizing his results to other situations. Generalizations are permissible only when between-raters variance is regarded as error.

In addition, the investigator should make clear whether the reported R represents the average reliability of the individual rater or the reliability of the composite rating and should describe how the ratings are to be used. Only a clear exposition of all of these factors will allow the reader to evaluate the results intelligently.

Finn's r (Equation 1) is recommended only when the within-subjects variance in the ratings is so severely restricted that the intraclass correlation is inappropriate (e.g., Case 3 in Table 1). The hypothesis that the observed within-subjects variance is equal to the chance within-subjects variance should, of course, be tested by the chi-square test (Equation 3) prior to the calculation of r. Finn's r should be interpreted cautiously, however, since chance variance may be overestimated, thereby spuriously inflating r.

Composite scores. For most purposes, psychologists will find the reliability of composite scores based on unit weights satisfactory. If the decision to be made is of such importance that the greatest possible precision of measurement is required, and if the ratings of the several judges differ in variance and/or reliability, optimal weights should be used instead of unit weights. In this case the procedure developed by Overall (1965) is appropriate. In reporting the results of research employing a composite rating based on optimal weights, the counseling psychologist should indicate the variance and the estimated reliability of each judge, the weights assigned the ratings of each judge, and the estimated reliability of the composite, based on unit weights and optimal weights. Such complete disclosure will allow the reader to evaluate thoroughly the results and to determine for himself the advantage of optimal weights in the specific instance. Finally, the counselor must keep in mind the fact that optimal weights for ratings are specific to the rating situation and cannot be assumed to generalize across rating occasions, even though the same judges may have employed the same rating scale with a similar sample of subjects on two similar occasions.

Interrater Agreement

The proportion or percentage of agreements (P), chi-square indexes, the pairwise correlation between raters, and Kendall's coefficient of concordance are inappropriate as measures of interrater agreement. The latter two indexes do not indicate interrater *agreement*, while P treats agreement as an absolute and is inflated by chance or random agreements.

We recommend the use of the Lawlis and Lu (1972) index of interrater agreement (Equation 7). The definition of agreement adopted and an indication of the extent of agreement (T index, Equation 8) should accompany the Lawlis and Lu indication of the significance of the agreement. In addition, the use of a stringent level of significance is recommended in evaluating agreement. Alternatively, the investigator should use a chance P value based on fewer scale categories than were actually available. This will partially correct for the possibility that the raters were predisposed to avoid the extreme categories of the scale, thereby causing the probability of chance agreement to be greater than the appropriate P value.

Nominal Scales

Nominally scaled data permit an analysis only of interrater agreement. The use of Cohen's κ (Equation 12) is recommended when the same two judges rate each subject. Fleiss' κ_v (Equation 15) is recommended when subjects are rated by different judges but the number of judges rating each subject is held constant. The general use of Cohen's weighted kappa (κ_w) is not recommended.

At the present time, no measure of interrater agreement for nominal scales is available for situations in which a varying number of judges rate subjects, or in which the same group of more than two judges rates each subject. The latter situation is the case most likely to occur in counseling research. Clearly, more research is needed.

REFERENCES

Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. Weak measurement vs. strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement*, 1966, *26*, 291–309.

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

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

Cannon, J., & Carkhuff, R. R. Effects of rater level of functioning and experience upon the discrimination of facilitative conditions. *Journal of Consulting and Clinical Psychology*, 1969, *33*, 189–194.

Carkhuff, R. R. Helper communication as a function of helpee affect and content. *Journal of Counseling Psychology*, 1969, *16*, 126–131.

Carkhuff, R. R. Principles of social action in training for new careers in human services. *Journal of Counseling Psychology*, 1971, *18*, 147–151.

Carkhuff, R. R., & Banks, G. Training as a preferred mode of facilitating relationship between races and generations. *Journal of Counseling Psychology*, 1970, *17*, 413–418.

Carkhuff, R. R., & Griffin, A. H. The selection and training of human relations specialists. *Journal of Counseling Psychology*, 1970, *17*, 443–450.

Clark, E. L. Spearman-Brown formula applied to ratings of personality traits. *Journal of Educational Psychology*, 1935, *26*, 552–555.

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.

Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions*. Urbana: University of Illinois Press, 1965.

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

Englehart, M. D. A method of estimating the reliability of ratings compared with certain methods of estimating the reliability of tests. *Educational and Psychological Measurement*, 1959, *19*, 579–588.

Everitt, B. S. Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 97–103.

Finn, R. H. A note on estimating the reliability of categorical data. *Educational and Psychological Measurement*, 1970, *30*, 71–76.

Finn, R. H. Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement*, 1972, *35*, 255–265.

Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 1971, *76*, 378–382.

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.

Garfield, S. L., Prager, R. A., & Bergin, A. E. Evaluation of outcome in psychotherapy. *Journal of Consulting and Clinical Psychology*, 1971, *37*, 307–313.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications. *Journal of the American Statistical Association*, 1954, *49*, 732–764.

Guilford, J. P. *Psychometric methods* (2nd ed.). New York: McGraw-Hill, 1954.

Guttman, H. A., Spector, R. M., Sigal, J. J., Rakoff, V., & Epstein, N. B. Reliability of coding affective communication in family therapy sessions: Problems of measurement and interpretation. *Journal of Consulting and Clinical Psychology*, 1971, *37*, 397–402.

Hays, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.

Kaspar, J. C., Throne, F. M., & Schulman, J. L. A study of the interjudge reliability in scoring the responses of a group of mentally retarded boys to three WISC subscales. *Educational and Psychological Measurement*, 1968, *28*, 469–477.

Kelley, T. L. *Fundamentals of statistics*. Cambridge, Mass.: Harvard University Press, 1947.

Krippendorff, K. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 1970, *30*, 61–70.

Lawlis, G. F., & Lu, E. Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin*, 1972, *78*, 17–20.

Lawshe, C. H., & Nagle, B. F. A note on the combination of ratings on the basis of reliability. *Psychological Bulletin*, 1952, *49*, 270–273.

Lu, K. H. A measure of agreement among subjective judgments. *Educational and Psychological Measurement*, 1971, *31*, 75–84.

Martin, D. G., & Gazda, G. M. A method of self-evaluation for counselor education utilizing the measurement of facilitative condition. *Counselor Education and Supervision*, 1970, *9*, 87–92.

McMullin, R. E. Effects of counselor focusing on client self-experiencing under low attitudinal conditions. *Journal of Counseling Psychology*, 1972, *19*, 282–285.

Mickelson, D. T., & Stevic, R. R. Differential effects of facilitative and nonfacilitative behavioral counselors. *Journal of Counseling Psychology*, 1971, *18*, 314–319.

Myrick, R. D., & Pare, D. D. A study of the effects of group sensitivity training with student counselor-consultants. *Counselor Education and Supervision*, 1971, *11*, 90–96.

Overall, J. E. Reliability of composite ratings. *Educational and Psychological Measurement*, 1965, *25*, 1011–1022.

Payne, P. A., Winter, D. E., & Bell, G. E. Effects of supervisor style on the learning of empathy in a supervision analogue. *Counselor Education and Supervision*, 1972, *11*, 262–269.

Pierce, R. M., & Schauble, P. G. Toward the development of facilitative counselors: The effects of practicum instruction and individual supervision. *Counselor Education and Supervision*, 1971, *11*, 83-89.

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

Rosander, A. C. The Spearman-Brown formula in attitude scale construction. *Journal of Experimental Psychology*, 1936, *19*, 486-495.

Schuldt, W. J., & Truax, C. B. Variability of outcome in psychotherapeutic research. *Journal of Counseling Psychology*, 1970, *17*, 405-408.

Scott, W. A. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 1955, *19*, 321-325.

Silverstein, A. B. A note on interjudge reliability. *Psychological Reports*, 1966, *19*, 1170.

Smith, F. F. Objectivity as a criterion for estimating the validity of questionnaire data. *Journal of Educational Psychology*, 1935, *26*, 481-496.

Snedecor, G. W. *Statistical methods* (4th ed.). Ames: Iowa State College Press, 1946.

Stillman, S., & Resnick, H. Does counselor attire matter? *Journal of Counseling Psychology*, 1972, *19*, 347-348.

Tanney, M. F., & Gelso, C. J. Effect of recording on clients. *Journal of Counseling Psychology*, 1972, *19*, 349-350.

Taylor, J. B. Rating scales as measures of clinical judgment: A method for increasing scale reliability and sensitivity. *Educational and Psychological Measurement*, 1968, *28*, 747-766.

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

Truax, C. B., & Lister, J. L. Effects of short-term training upon accurate empathy and non-possessive warmth. *Counselor Education and Supervision*, 1971, *10*, 120-125.

Tseng, M. S. Self-perception and employability: A vocational rehabilitation program. *Journal of Counseling Psychology*, 1972, *19*, 314-317.

Weiss, D. J. Factor analysis and counseling research. *Journal of Counseling Psychology*, 1970, *17*, 477-485.

Weiss, D. J. Further considerations in applications of factor analysis. *Journal of Counseling Psychology*, 1971, *18*, 85-92.

Wittmer, J. An objective scale for content analysis of the counselor's interview behavior. *Counselor Education and Supervision*, 1971, *10*, 283-290.

(Received August 19, 1974)