*Brit. J. Psychiat.* (1976), **129**, 452-6

# Assessing Inter-Rater Reliability for Rating Scales: Resolving some Basic Issues

By DOMENIC V. CICCHETTI

**Summary.** This paper extends the recent work of Hall (1974) by presenting the minimal sample sizes and the specific linear agreement weights required for assessing the reliability of rating scales commonly used in neuropsychiatric and other clinico-medical settings. The weights are shown to vary as a function of (a) whether or not the rating scale contains a point of 'absence', and (b) the number of ordinal points on the scale.

## Introduction

In a recent article, Hall (1974) argues convincingly that a number of assumptions underlying the use of standard statistical tests of reliability may not be valid when applied to rating scales. He states, further, that the test of choice for calculating reliability with rating scales should: (1) be distribution free; (2) allow credit for partial rater agreement; (3) correct for rater agreement due to chance alone; (4) make use of *individual items* in the rating scale; and (5) correct for differences in rater mean scores. One method which appears to meet these criteria satisfactorily is weighted kappa, introduced by Cohen (1968). Its revised standard error has been given by Fleiss, Cohen and Everitt (1969).

However, the research investigator, even after selecting weighted kappa as the statistic of choice, has two further problems to consider: (1) which agreement weights should be used to determine complete rater agreement and each level of partial rater agreement? and (2) what is the minimal sample size required for the valid application of weighted kappa? It is the purpose of this brief report to extend the work of Hall by providing answers to these basic questions. This information should be especially valuable to the many clinical investigators who use rating scales in neuropsychiatric and other medical settings.

### *Selecting rater agreement weights*

Hall (1974, p 250) argues persuasively that *linear* weights should be used with weighted kappa when the data are presumably ordinal. He shows, in fact, that other systems (e.g., quadratic weights) can weight equally probable rater errors rather unequally. Thus, in the absence of information about score distributions, Hall makes the reasonable assumption that it is prudent to adopt linear or equal weight intervals. Hall also gives the linear weights for the 3-, 4-, and 5-point ordinal scales. It will be demonstrated here that these weights vary as a function of (a) the type of ordinal scale under consideration; and (b) the number of points on the ordinal scale. General formulae for selecting appropriate sets of linear agreement weights will be given.

### *Types of ordinal rating scales*

There are, basically, two types of ordinal rating scales possible, each with differing implications for rater agreement weights. The distinction hinges simply upon whether the specific items on the ordinal scale do or do not contain a point of 'absence'. When there is no point of 'absence' the scale can be described as having three or more points of 'presence' (e.g., a patient's level of activity might be described as 'low', 'moderate', or 'high'). Rater confusions between adjacent sets of



https://doi.org/10.1192/bjp.129.5.452 Published online by Cambridge University Press

categories on such rating scales appear clinically equivalent. Thus, rater confusion between 'moderate' and 'low' is no more serious clinically than the confusion between 'moderate' and 'high'.

We will compare this situation to the one of assessing rater reliability on a scale containing a point of 'absence' and two or more degrees of 'presence'. Important clinical examples of this type of ordinal scale include health rating scales in general, or neuropsychiatric (NP) rating scales in particular.

*Continuous-ordinal scales*

The first type of ordinal scale, which contains no point of 'absence' can be described as 'continuous-ordinal' (C-O). The number of weights for such scales is determined, simply, by $\kappa$, the number of ordinal points on the scale. Each weight corresponds to a category of agreement. For example, on a 3-point ordinal scale, there will be 3 linear weights: the first for rater pairings in complete agreement (1–1, 2–2, 3–3); the second for rater pairings in partial agreement (1–2, 2–1, 2–3, 3–2); and, the third for ratings in complete disagreement (1–3 and 3–1 pairings, which represent maximal disagreement). The actual weights, for any size $\kappa$ ordinal scale, follow the formula given by Cicchetti (1972) and Cicchetti and Allison (1973):

$$\text{C-O (Linear weights)} = \frac{\kappa-1}{\kappa-1}; \frac{\kappa-2}{\kappa-1}; \frac{\kappa-3}{\kappa-1}; \ldots \frac{\kappa-\kappa}{\kappa-1} \qquad \text{(i)}$$

in which $\kappa$ refers simply to the number of points on a given 'continuous-ordinal' scale. These appear below for the 3-, 4-, 5-, 6-, and 7-point scales Table I. This encompasses the range of ordinal scales commonly used in clinical or research setting.

The assessment of rater reliability for behaviours of the type measured on health rating scales (e.g., the neuropsychiatric (NP) scales developed by Lorr, Klett, McNair and Lasky (1962); Overall and Gorham (1962); and Overall and Klett (1972)), presents a special problem to clinical investigators. Unlike the 'continuous-ordinal' rating scales discussed above, such scales contain a point of 'absence' as well as two or more points of 'presence'. Because of this, the distances between category intervals are *not* always clinically equivalent, as is true of C-O scales. As early as 1962 Overall and Gorham were aware of the implications this phenomenon has for the valid assessment of rater reliability. Following their reasoning, consider, as an example, an NP scale which categorizes 'delusions' as 'absent', 'mild', 'moderate', or 'severe'. Rater confusion between 'mild' and 'absent' is much more serious, clinically, than that between 'mild' and 'moderate', in spite of the fact that the category 'mild' appears midway between 'absent' and 'moderate'. The crucial clinical implication of this is that rater confusions between 'mild' and 'moderate' should receive higher agreement weights than rater confusions between 'mild' and 'absent'. The linear weighting system, given in (i) does not accomplish this. On our 4-point 'delusions' scale ('absent'; 'mild'; 'moderate'; 'severe'), the weights according to (i) are 1, ·667, ·333, and 0 (or the same weights given by Hall

TABLE I

*Linear agreement weights for continuous-ordinal rating scales*

| Number of $\kappa$ ordinal points on rating scale | Categories of rater agreement and their corresponding linear agreement weights | | | | | | |
|---|---|---|---|---|---|---|---|
| | Complete agreement | 1 Scale point apart | 2 Scale points apart | 3 Scale points apart | 4 Scale points apart | 5 Scale points apart | 6 Scale points apart |
| 3 | 1 | ·50 | 0 | | | | |
| 4 | 1 | ·67 | ·33 | 0 | | | |
| 5 | 1 | ·75 | ·50 | ·25 | 0 | | |
| 6 | 1 | ·80 | ·60 | ·40 | ·20 | 0 | |
| 7 | 1 | ·83 | ·67 | ·50 | ·33 | ·17 | 0 |

https://doi.org/10.1192/bjp.129.5.452 Published online by Cambridge University Press

(1974, p 251) for a 4-point rating scale). Since both rater confusions ('mild' with 'absent' and 'mild' with 'moderate') represent a one ordinal scale point discrepancy, they would both receive an agreement weight of ·67. This occurs in spite of the fact that the confusion between 'mild' and 'absent' is far more serious clinically than that between 'mild' and 'moderate'. It is the purpose here to demonstrate that this serious problem, which is peculiar to rating scales containing an 'absence' point, can be resolved by deriving an appropriate set of linear agreement weights for such scales. In the present discussion, these scales and their linear agreement weights will be referred to as 'dichotomous-ordinal' (D-O).

*Dichotomous-ordinal scales*

To arrive at a simple mathematical formula for deriving these weights, one has first to determine how many of the possible $\kappa$ categories of rater agreement are affected by the 'absence' category in a manner that clearly calls for adjusting the linear weighting system one would otherwise utilize by applying formula (i). Using our 4-point 'delusions' scale, the categories of agreement would be 'complete agreement', 'agreement within 1 scale point', 'agreement within 2 scale points', and 'complete disagreement'. It becomes readily apparent that neither the 'complete rater agreement' category nor the 'complete rater disagreement' category changes in clinical meaning. Thus, when both raters score 'absence', this belongs in the same category (complete agreement) as do the 'mild-mild', 'moderate-moderate', and 'severe-severe' rater pairings. Such pairings, using an agreement weighting system varying between 0 and 1 must, of course, receive an agreement weight of 1. Further, rater confusions between 'absence' of delusions and 'severe' delusions represent maximal disagreement and therefore receive the lowest possible linear agreement weight of 0. These weights are the same as those for comparable pairings on C-O rating scales. However, it is when rater scoring of 'absence' is paired with the *remaining* middle or $\kappa$–2 categories that the problems occur with the C-O linear weighting system. In our 'delusions' scale, the two $\kappa$–2 or middle categories are 'agreement within 1 scale point' and 'agreement within 2 scale points'. As noted previously, if such a scale were treated as a C-O scale, the linear agreement weights would become ·67 for the first category and ·33 for the second category. The problem is that *each* $\kappa$–2 category on a D-O scale contains two rater pairings which confuse 'absence' with some degree of 'presence'. The remaining pairings in the category represent rater discrepancies between degrees of 'presence' only.

If we wish to assign higher linear agreement weights to the latter than to the former types of rater confusion, it is clear that we must double our original $\kappa$–2 middle categories of rater agreement. This produces $2(\kappa-2)$ middle categories for any given D-O scale. If we add to this the 'complete agreement' and 'complete disagreement' rater categories, we require $2(\kappa-2)+2$ linear agreement weights (W) for any given D-O scale. This formula can be simplified, mathematically, as:

$$\underset{\text{(Number of D-O weights)}}{\mathrm{W}} = 2(\kappa-1) \qquad \text{(ii)}$$

in which $\kappa$ again refers to the number of points on the ordinal scale. Once W is obtained, for a given D-O rating scale, its value can be substituted into the formula below for determining the appropriate *set* of linear agreement weights:

$$\underset{\text{(Linear weights)}}{\text{D-O}} = \frac{\mathrm{W}-1}{\mathrm{W}-1}; \frac{\mathrm{W}-2}{\mathrm{W}-1}; \frac{\mathrm{W}-3}{\mathrm{W}-1}; \ \ldots \ \frac{\mathrm{W}-\mathrm{W}}{\mathrm{W}-1} \qquad \text{(iii)}$$

Applying formulae (ii) and (iii), we obtain the linear agreement weights for D-O rating scales. These appear in Table II for scales once again ranging between 3 and 7 ordinal points. It should be noted that the $2(\kappa-2)$ middle categories of agreement, for a given D-O scale, have the property that the first of each successive pair of rater categories always contains rater discrepancies between degrees of 'presence'. The second contains the confusions between 'absence' and some degree of 'presence'. Finally, both members of each successive pair of agreement categories contain rater pairings which are the

https://doi.org/10.1192/bjp.129.5.452 Published online by Cambridge University Press

same number of scale points apart. Formula (iii), above, then, always makes it possible to assign a higher agreement weight to the first member in each pair (the case with no 'absence-presence' confusion) than to the second one (always denoting 'absence-presence' confusion of one type or another). As given in Table II, the D-O linear agreement weights provided by this formula should obviate the aforementioned rater reliability problem which occurs with scales containing a point of 'absence'. In the context of our 'delusions' scale, the confusion between 'mild' and 'moderate' will receive a weight of ·8, while that between 'mild' and 'absent' will receive one of ·60. This clinically meaningful distinction contrasts with the weight of ·67 which would be assigned to *both* types of rater confusion if the scale were erroneously treated as 'continuous-ordinal'.

In effect, then, simply knowing the number of points on a rating scale and whether or not it contains a point of absence, we can refer to either Table I or II to obtain the appropriate linear weighting system for assessing inter-rater reliability. Specifically, these weights should be used in the calculation of weighted kappa which involves a statistical comparison between observed rater agreement and that expected on the basis of chance alone. The steps required to accomplish this are clearly delineated in Fleiss, *et al* (1969); and, in the context of ward rating scales, by Hall (1974). As such they need not be repeated here.

*Minimal number of cases*

The last question we seek to answer is what is the minimal sample size required for assessing the reliability of rating scales? A conservative, untested, large sample size estimate of $\geq 200$ was given by Fleiss, Cohen and Everitt (1969). However, a very recent Monte Carlo study (Cicchetti and Fleiss, 1975) of rating scales between $3 < \kappa < 7$ ordinal points shows that even when raters are extremely different in their application of a given rating scale, weighted kappa produces valid results. This holds true even in the rare instances in which raters confuse the meanings of the end points on a given ordinal scale. The minimal sample size for which this holds is approximately $N \geq 2\kappa^2$. This produces Ns ranging between about 20 for the 3-point rating scale to about 100 for the 7-point scale. The 4-, 5-, and 6-point scales, then, require minimal sample sizes of about 30, 50, and 75, respectively. Given the difficulty of obtaining sample sizes of $\geq 200$, this finding should be of some comfort to investigators using these statistical approaches.*

* The reader interested in the details of this investigation can obtain a copy of this study by writing to the author.

TABLE II

*Linear agreement weights for dichotomous-ordinal rating scales*

| Number of $\kappa$ ordinal points on rating scale | Value of W (D-O) | Complete agreement | Categories of rater agreement and their corresponding linear agreement weights | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 Scale point apart | | 2 Scale points apart | | 3 Scale points apart | | 4 Scale points apart | | 5 Scale points apart | | 6 Scale points apart | |
| | | | Absence-Presence Confusion: | | Absence-Presence Confusion: | | Absence-Presence Confusion: | | Absence-Presence Confusion: | | Absence-Presence Confusion: | | Absence-Presence Confusion: | |
| | | | No | Yes | No | Yes | No | Yes | No | Yes | No | Yes | No | Yes |
| 3 | 4 | 1 | ·67 | ·33 | – | 0 | | | | | | | | |
| 4 | 6 | 1 | ·80 | ·60 | ·40 | ·20 | – | 0 | | | | | | |
| 5 | 8 | 1 | ·86 | ·71 | ·57 | ·43 | ·29 | ·14 | – | 0 | | | | |
| 6 | 10 | 1 | ·89 | ·78 | ·67 | ·56 | ·44 | ·33 | ·22 | ·11 | – | 0 | | |
| 7 | 12 | 1 | ·91 | ·82 | ·73 | ·64 | ·45 | ·45 | ·36 | ·27 | ·18 | ·09 | – | 0 |

https://doi.org/10.1192/bjp.129.5.452 Published online by Cambridge University Press

ACKNOWLEDGEMENTS

The author expresses appreciation to Joseph Vitale and Sandra L. Aivano who developed the necessary computer programs to assess rater agreement with both 'continuous-ordinal' and 'dichotomous-ordinal' rating scales. Copies of these programs, which will be described in future publications, are available by writing to the author.

REFERENCES

CICCHETTI, D. V. (1972) A new measure of agreement between rank ordered variables. *Proceedings of the American Psychological Association*, **7**, 17–18.

—— & ALLISON, T. (1973) Assessing the reliability of scoring EEG sleep records: an improved method. *Proceedings and Journal of the Electro-physiological Technologists' Association*, **20**, 92–102.

CICCHETTI, D. V., & FLEISS, J. L. (1975) 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.

COHEN, J. (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213–20.

FLEISS, J. L., COHEN, J. & EVERITT, B. S. (1969) Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, **72**, 323–7.

HALL, J. N. (1974) Inter-rater reliability of ward rating scales. *British Journal of Psychiatry*, **125**, 248–55.

LORR, M., KLETT, J., MCNAIR, D. M. & LASKY, J. J. (1962) *Inpatient Multidimensional Psychiatric Scale (IMPS)*. Palo Alto: Consulting Psychologists Press.

OVERALL, J. E. & GORHAM, D. R. (1962) The brief psychiatric rating scale. *Psychological Reports*, **10**, 799–812.

—— & KLETT, C. J. (1972) *Applied Multivariate Analysis*. New York: McGraw-Hill (Introductory chapter).

Domenic V. Cicchetti, PH.D., *Research Psychologist and Biostatistician, V. A. Hospital, West Haven, Connecticut 06516; Associate Professor of Psychology, Department of Psychiatry, Yale University.*

This research was supported by the West Haven V. A. Hospital (MRIS 1416). Use V. A. address for reprint requests.

(*Received 29 January 1976*)

https://doi.org/10.1192/bjp.129.5.452 Published online by Cambridge University Press