Brit. J. Psychiat. (1981), **139**, 138–143

Standardized Interviews in Psychiatry: Issues of Reliability

R. W. SANSON-FISHER and C. J. MARTIN

Summary: A review of studies which used standardized interviews as a means of data collection indicated that the issue of reliability was frequently neglected in the psychiatric literature. It is suggested that this may be due to a misinterpretation of the concept of reliability which is apparently perceived as a permanent feature of an interview instrument. The importance of reliability is discussed, and recommendations are made as to how a methodologically adequate assessment of reliability might be undertaken.

The current widespread use of standardized interviews in psychiatric research raises a number of questions about the reliability with which such instruments are applied. An instrument can be said to have been used reliably when there are small errors of measurement and when there is stability, consistency and dependability of scores for individuals on the trait, characteristic or behaviour which is being assessed (Mitchell, 1979). Commonly, reliability refers to the extent to which measures by one observer agree with the independent judgments of another (inter-rater reliability) or the same observer recording the same sequences at two different times (intra-rater reliability) (Kazdin, 1977). As reliability is a necessary, although not sufficient, condition for validity, the issue is obviously of considerable importance (Bartko and Carpenter, 1976; Maxwell, 1977).

Individual studies have shown that interview instruments such as the Present State Examination (Wing *et al*, 1974) can be used with a high degree of inter-rater reliability (Wing *et al*, 1977; Cooper *et al*, 1977; Luria and Berry, 1979). However, this should not permit other researchers to assume that such findings will be automatically duplicated in their studies. Yet, a review of the 73 studies which used interview methods and were published in the British Journal of Psychiatry from 1978–80, showed that reliability was not assessed in 77 per cent of them.

In this paper an attempt will be made to elaborate issues relating to reliability, its assessment, methods of calculation, and its monitoring. Although much of the present data relating to the assessment of reliability comes from observational research, the similarity with interview procedures appears clear. Both approaches use observers to record the behaviour of subjects and employ a predetermined scoring system to organize and categorize the basic material; both seek to obtain, maintain and express reliability in a way which ensures that the interviewer or observer is not distorting the information he receives and codes.

Components of reliability

Of the 73 studies which used an interview procedure, 23 per cent provided data on reliability. All of these presented the agreement with which the scoring of subjects' responses was achieved by two independent raters. There are, however, at least two levels of potential reliability variation—that of interview administration and interview scoring.

(i) *The reliability of interview administration*

While variation in the administration of an interview is reduced by the use of structured questionnaires, research suggests that interviewers may not always use the specified format in the same way as a function of boredom and other variables (Cannell *et al*, 1977). Moreover, most interviews which are used in psychiatric research can be argued to be only semi-structured. For example, while the PSE comprises a large number of standard questions, it also includes several defined cut off points which are determined by a subject's responses. At these points there is a possibility of interviewer unreliability.

It might be assumed that disagreements in interview administration would invariably be reflected in scoring disagreements. However, four combinations of agreement in administration and scoring are possible. One of these is that, while interviewers may agree in the scoring of responses, they may disagree in the administration of the interview. If this occurred and the reliability of interviewing was not checked, investigators would be led into the

https://doi.org/10.1192/bjp.139.2.138 Published online by Cambridge University Press

spurious assumption that satisfactory reliability had been achieved. The reliability of the interview administration, while problematic to assess (Cooper *et al*, 1972), might be undertaken by the audio-recording of the interview and then having a second interviewer comparing the style of questioning with the standard. Monitoring of interview administration also ensures that the interview is not shaping or encouraging specific respondent responses (Cannel *et al*, 1977).

(ii) *The reliability of interview scoring*

The responses to interview questions are normally scored on a number of rating scales according to a set of predetermined codes or criteria. For example, in some PSE questions, the symptom can be rated on a three-point scale (0, 1, 2). A rating of 0 represents the absence of a symptom, and a rating of 2 represents its presence in a severe or marked form. Assuming that the questionnaire has been administered in a reliable way, any rating disagreements which result are likely to be due to a differential application of the scoring criteria.

That 77 per cent of the studies reviewed did not appear to have carried out any form of reliability assessment suggests that the issue is perceived as a constant or structural feature of an interview instrument. However, there appears to be no logical or experimental justification for this assumption. Even when reliability was tested, only interview scoring was considered to be an area requiring assessment. However, since such agreement could be the result of the interaction between the adequacy of the interview administration and scoring agreement, there is a need to ensure that both components are tested for reliability.

In the following sections, some of the methods which can be employed to assess reliability and the means by which this can be undertaken are outlined and discussed.

When reliability should be assessed

Of the seventeen studies reviewed which presented reliability figures, in only eight studies was reliability assessed throughout the period of the investigation. In seven studies, it appears that reliability was calculated at the beginning of the study. In the latter research, the assumption appears to have been, that once a satisfactory level of agreement had been achieved, it was maintained throughout the investigation. However, one of at least three things could happen to obtained reliability after the initial training:

(i) Reliability may be maintained with raters applying definitions in the same manner. In order to test this possibility, it would obviously be necessary to assess reliability more than once in the course of an investigation. In those studies where reliability has been checked over time, the results indicate that agreement does not remain stable (O'Leary *et al*, 1973; Reid, 1970).

(ii) A second possibility is that reliability between raters may be maintained while 'accuracy' declines where 'accuracy' refers to the extent to which ratings accord with the criteria defined at the inception of a study (Kazdin, 1977). Thus, raters may consensually change the manner in which they apply codes and definitions over time. O'Leary *et al* (1973) found that observers who frequently discussed and compared their ratings developed their own, rather than the investigators' interpretation of codes. While raters still displayed high agreement they had as a group 'drifted' away from the standards applied at the beginning of the study and were therefore making 'inaccurate' but reliable observations.

(iii) A third possibility is that reliability over time will not be maintained, with one or more raters developing an idiosyncratic style of scoring. For example, in a study by Reid (1970) it was found that inter-rater agreement fell by 25 per cent when observers believed that their ratings were no longer being checked for reliability.

Such findings indicate that initial levels of agreement between raters may not be maintained over time. Therefore, continuous monitoring of reliability over the period of a study appears advisable. However, such assessment alone may not reveal the phenomenon of 'consensual drift', with the agreement between raters remains high while accuracy declines (O'Leary *et al*, 1973). Such drift may be controlled by continually training interviewers together, and by periodically introducing newly trained raters to permit 'between', rather than 'within' comparisons (Kazdin, 1977). The establishment of sets of data with agreed upon ratings against which later ratings may be compared, may also serve an important function in maintaining interviewer accuracy.

Interviewer training and reliability

Intuitively, it seems obvious that some interviews will be easier to administer and score reliably because they are less complex. Indeed, several studies indicate that reliability is lower for complex interactions, where the number of discriminations required of an observer

https://doi.org/10.1192/bjp.139.2.138 Published online by Cambridge University Press

are greater than for interactions of lesser complexity (Taplin and Reid, 1973; Mash and McElwee, 1974). Skindrud (1972) found that the increasing levels of inter-rater agreement obtained over a three week training period (47 per cent to 68 per cent) was not only a function of increasing observer proficiency, but also of the less complex interactions which occurred in the second and third weeks of training. Similarly, observers who have been trained to rate 'predictable' behaviour (i.e., conversations with redundant information) were less reliable than when they later coded 'unpredictable' behaviors. Observers trained with sequence which were unpredictable did not display a decline in reliability (Mash and McElwee, 1974).

Thus, if observers are trained on data of low complexity, assessments of reliability based on ratings from such material may be spuriously high and may not generalize to the experimental phase, where the complexity of interactions is greater. In interview studies, interviewers may be trained to administer and score a standardized psychiatric interview reliably with a control or community sample. In these subjects, the extent and severity of symptomatology, and hence the complexity of the data, may be lower when compared to an in-patient sample. In the latter group, the discriminations required for the administration and scoring of the interview are likely to be of greater complexity.

As a result of the interaction between complexity and reliability, it is recommended that interviewers should be trained on material which reflects a similar level of complexity that will be encountered in the investigative stages of the study.

The reactivity of reliability assessment

Research by Taplin and Reid (1973) indicates that if observers know that reliability is being assessed, they apply the rating standards more stringently in an effort to achieve the required level of agreement. When raters were unaware that reliability was being checked, the level of agreement dropped when compared to overt reliability assessment. On the days when spot checks were expected, the level of agreement was found to increase. (Reid, 1970; Romanczyk *et al*, 1973; Kent *et al*, 1977).

Any changes in rater performance over time may, therefore, not be identified by overt assessment, suggesting that covert reliability checks might be advisable. One way in which this could be achieved would be by audio-taping all interviews. This would allow a second rater to select randomly interviews for re-rating throughout an investigation.

A further source of observer reactivity which may bias reliability estimates has also been identified. O'Leary *et al* (1975) found that if observers were informed of what to expect in a video-taped interaction (e.g., a decrease in disruptive behaviours) and were given feedback by the investigator in the form of approval or disapproval of their results, observations were found to be biased in the perceived desired direction. Expectancies alone, however, do not appear to be capable of influencing data (Kent *et al*, 1974; Redfield and Paul, 1976). Rather, it appears that it is the specific effects of expectations combined with feedback from an influential figure which may cause bias.

As a result, while it may be difficult to fully control for interviewer expectancies, it is recommended that the principal investigator should give few indications as to the desirability of the obtained results. Potential biases may also be controlled by having independent raters periodically check the data (Kazdin, 1977).

The unit of data for calculating reliability

Data from standardized interviews can usually be analysed at more than one level. For example, PSE scores can be analysed at the level of the individual symptom (item), or scores for related symptoms may be grouped together to yield syndrome scores (e.g., general anxiety). Alternatively, an overall category such as psychiatric case vs. non-case, based on the number and severity of certain symptoms, can be calculated. The appropriate unit of data on which reliability is calculated should be dependent upon the unit of data used in the analysis. Ideally, agreement should be calculated on the unit of data which is to be used in subsequent analyses (Hartmann, 1977).

This would mean that if a later analysis is concerned with assessing the extent and severity of psychiatric symptomatology, reliability should be calculated on the full range of possible scores for each symptom-item. Assessing agreements merely for ratings of symptom presence or absence (e.g., in the PSE—0 vs. 1, 2) could be considered inadequate if the extent or severity is of interest in the analysis stage. In the same way, the use of syndrome scores or derived categories may give an inaccurate representation of the actual level of agreement between raters. Raters may show agreement in their section scores or in their assignment of a category, but disagreement in item scores, on which the derived scores are based.

An example (Table) in which simulated PSE scores are provided may illustrate these points.

Three units of data can be identified in this example, each of which would yield a different estimate of inter-rater agreement if used as the base for reliability calculation. The first unit is a syndrome score which is the sum of certain symptom scores. For example, a rating of 2 for the syndrome 'general anxiety' requires

https://doi.org/10.1192/bjp.139.2.138 Published online by Cambridge University Press

TABLE

Simulated scores for section: 'Autonomic anxiety' (PSE: 9th Edition)

	PSE scores		
	Raters		
Symptom	1	2	
Free-floating anxiety	1	1	*†
Anxious foreboding	2	1	†
Panic attacks	0	0	*†
Situational anxiety	1	2	†
Anxiety on meeting people	2	1	†
Specific phobias	1	0	
Anxiety avoidance	0	1	
General anxiety (syndrome score)	2	2	

Rating scale: 0 = absence of symptom.
1 = presence of symptom (moderate variety).
2 = presence of symptom (marked severity).

* Agreement for severity rating (trichotomised scores).

† Agreement for presence/absence rating (dichotomised scores).

a rating for the symptom 'free floating anxiety' plus a rating for either 'specific phobia' or 'anxiety avoidance'. Thus, the two raters in the presented example show complete agreement but have based their rating on the presence of different symptoms. The dichotomized item agreement rate (0 vs. 1 or 2) is 71 per cent, whereas agreement for the trichotomized item scores (0 vs. 1 vs. 2) is only 29 per cent. Given this variation, it seems reasonable that reliability calculations should be carried out at the level of the individual item, and that the full range of the rating scales be utilized.

Methods of calculating reliability

The statistical method employed for calculating reliability is a further determinant of the magnitude of the reliability coefficient obtained (Kent and Foster, 1977). Even when applied to the same data, different statistical techniques may yield different agreement values (Hartmann, 1977). Hall (1974) suggests that a statistical test for calculating reliability with rating scales should be distribution-free, allow credit for partial rater agreement, correct for differences in rater mean scores, make use of individual items in the rating scale, and correct for rater agreements due to chance alone. Just as raters may agree because they are applying definitions or codes in the same way, they may also agree simply by chance.

Two statistical techniques which are commonly used to calculate reliability, but which are open to criticism on the above grounds, are percentage agreement and product-moment correlation.

(i) *Percentage agreement*

In this method, the total number of agreements between raters is divided by the sum of agreements plus disagreements and multiplied by a hundred, making it a relatively simple means of calculating agreement. However, percentage agreement has a number of drawbacks: it has no established lower level of acceptability (Hartmann, 1977), although estimates of 70–80 per cent are usually considered satisfactory (Kazdin, 1977). Furthermore, it does not take account of chance agreement between raters. As a result of these criticisms, it has been suggested that only a low level of percentage agreement serves any useful function in informing about the unreliability of data (Kent and Foster, 1977).

(ii) *Product-moment correlation*

This is a well-known estimate of concordance based on paired scores, a coefficient of +1.00 reflecting perfect agreement between raters. Although there exist established levels of significance dependent on the total number of paired ratings, the product-moment correlation does not explicitly control for chance agreements. As a result, it can also be argued to be a less than robust test of reliability.

Of the seventeen studies which presented reliability data, fourteen used either product moment correlations or percentage agreement.

(iii) *Kappa*

One statistic which fulfils all the criteria outlined by Hall (1974) is the weighted kappa (Cohen, 1968). Kappa is represented by (Po–Pc)/(1–Pc), where Po is the proportion of observed agreements and Pc is the proportion of chance or expected agreements. It is a correlation-like measurement which can be tested for significance for different probability values. Despite these advantages, kappa was used by only three studies reviewed. If kappa is used to calculate reliability, the number of points on the rating scale has implications for the sample size or number of trials across which agreement should be assessed. Thus, a 3-point rating scale would require a sample size of at least twenty while a 7-point scale would need a sample of around one hundred (Ciccheti, 1976).

https://doi.org/10.1192/bjp.139.2.138 Published online by Cambridge University Press

(iv) *Analysis of variance*

Another method of calculating reliability is that described by generalisability theory which uses the statistical technique of analysis of variance (Cronbach *et al*, 1972). Generalisability theory assumes that there are a number of sources of variation in any measurement, each of which is called a facet. Each facet is treated as a separate condition in the analysis, and might include interviewers, coders, subjects and occasions. By systematically sampling conditions for each facet, and carrying out an analysis of variance across all scores, it is possible to obtain a generalisability coefficient which would reflect the size of the contribution of each facet to the observed scores (Mitchell, 1979). Thus, if measurements were shown to vary as a function of the scorers, this would indicate one source of variance or instability in the data (Jones *et al*, 1975).

However, although it is a powerful statistical tool, this form of analysis could be considered to be more complex and rigorous than required by all but the most sophisticated research endeavours.

As a result of such concerns and given the limitations of interview studies in psychiatry, it is suggested that kappa currently represents the most appropriate statistic for calculating the reliability of interview administration and scoring. A fuller discussion of available statistical methods for calculating reliability are discussed by Bartko and Carpenter (1976) and Hartmann (1977).

Conclusions

It has been argued that reliability is not a constant feature of an interview instrument and should be estimated whenever a particular schedule is employed. A survey of studies using interview methods indicated that this does not appear to be current practice; only 23 per cent of the reviewed studies carried out any any form of reliability assessment. Furthermore, none of these studies examined the reliability of interview administration. Given that this may affect the nature of the results, it is suggested that it be monitored whenever possible. The reliability of interview administration could be checked by having a second interviewer listen to the audio-taped interviews and scoring their agreement or disagreement with the administration of the questionnaire.

When assessing administration or scoring reliability the following procedures are recommended:

(i) In order to ensure adequate levels of agreement over time, reliability should be continually monitored throughout an investigation. Ideally, interviewers should be trained together with newly trained interviewers periodically introduced to permit between, rather than within, group comparisons. Sets of data with agreed ratings should also be utilised so that the accuracy of rater performance can be assessed. These procedures should help to prevent changes in reliability and consensual drift by the raters.

(ii) Audio-taped interviews should be selected on a random basis for assessment in an effort to reduce rater reactivity to reliability checks and hence potentially spuriously high agreement figures.

(iii) Given the impact of complexity on reliability, raters should be trained to satisfactory levels of agreement using material similar to that which they are likely to encounter in the investigative stage of the study.

(iv) Since feedback by the principal investigator about the desirability of obtained ratings may influence reliability, comments by the researchers should be limited to the accuracy of ratings. No comments should be made which may indicate the direction in which it is hoped the results will go.

(v) Because of the variation in levels of agreement which can be obtained using different units of data as the base for reliability assessment, it is recommended that agreement should be calculated on the smallest unit of data which is to be used in subsequent analyses.

(vi) As a result of the spurious influence of sample size and scale range on reliability estimates, chance-corrected statistics such as kappa should be used whenever there is a probability of chance agreements.

While it is acknowledged that the offered recommendations will lead to an increase in research time, this may be a necessary outcome if a high scientific standard is to be maintained in psychiatric research.

References

BARTKO, J. J. & CARPENTER, W. T. (1976) On the methods and theory of reliability. *The Journal of Nervous and Mental Disease*, **163**, 307–17.

CANNELL, C. F., MARQUIS, K. H. & LAURENT, A. (1977) *A summary of studies of interviewing methodologies*. DHEW No HRA 77-1343.

CICCHETTI, D. V. (1976) Assessing inter-rater reliability for rating scales: resolving some basic issues. *British Journal of Psychiatry*, **129**, 453–6.

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

https://doi.org/10.1192/bjp.139.2.138 Published online by Cambridge University Press

COOPER, J. E., COPELAND, J. R. M., BROWN, G. W., HARRIS, T. & GOURLEY, A. J. (1977) Further studies in interviewer training and inter-rater reliability of the Present State Examination (PSE). *Psychological Medicine*, **7**, 517–23.

—— KENDELL, R. E., GURLAND, B. J., SHARPE, L., COPELAND, J. R. M. & SIMON, R. (1972) *Psychiatric diagnosis in New York and London.* Institute of Psychiatry Monographs, No 20, Oxford University Press, London.

CRONBACH, L. J., GLESER, G. C., NANDA, H. & RAJARATNAM, N. (1972) *The Dependability of Behavioral Measurements: Theory of Generalisability for Scores and Profiles.* New York: Wiley.

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

HARTMANN, D. P. (1977) Considerations in the choice of inter-observer reliability estimates. *Journal of Applied Behaviour Analysis*, **10**, 103–16.

JONES, R. R., REID, J. B. & PATTERSON, G. R. (1975) Naturalistic observation in clinical assessment. In *Advances in Psychological Assessment* Vol 3, (ed. P. McReynolds). San Francisco: Jossey Bass.

KAZDIN, A. E. (1977) Artifact, bias, and complexity of assessment: the ABC's of reliability. *Journal of Applied Behavioural Analysis*, **10**, 141–50.

KENT, R. N., O'LEARY, K. D., DIAMENT, C. & DIETZ, A. (1974) Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology*, **42**, 774–80.

—— & FOSTER, S. L. (1977) Direct observational procedures: methodological issues in naturalistic settings. In *Handbook of Behavioural Assessment*, (eds.) A. R. Ciminero, K. S. Calhoun and H. E. Adams. John Wiley and Sons, New York.

—— KANOWITZ, J., O'LEARY, K. D. & CHEIKEN, M. (1977) Observer reliability as a function of circumstances of assessment. *Journal of Applied Behavior Analysis*, **10**, 317–24.

LURIA, R. E. & BERRY, R. (1979) Reliability and descriptive validity of PSE syndromes. *Archives of General Psychiatry*, **36**, 1187–95.

MASH, E. J. & MCELWEE, J. (1974) Situational effects on observer accuracy: behavioural predictability, prior experience, and complexity of coding categories. *Child Development*, **45**, 367–77.

MAXWELL, A. E. (1977) Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry*, **130**, 79–83.

MITCHELL, S. K. (1979) Interobserver agreement, reliability, and generalisability of data collected in observational studies. *Psychological Bulletin*, **86**, 376–90.

O'LEARY, C. D. & KENT, R. N. (1973) Behavior modification for social action: research tactics and problems. In *Critical Issues in Research and Practice*, (eds.) L. A. Hamerlynck, P. O. Davidson and L. E. Acker. Research Press, Illinois.

—— —— & KANOWITZ, J. (1975) Shaping data collection congruent with experimental hypotheses. *Journal of Applied Behaviour Analysis*, **8**, 43–51.

REDFIELD, J. & PAUL, G. L. (1976) Bias in behavioural observation as a function of observer familiarity with subjects and typicality of behavior. *Journal of Consulting and Clinical Psychology*, **44**, 156.

REID, J. B. (1970) Reliability assessment of observation data: a possible methodological problem. *Child Development*, **41**, 1143–50.

ROMANCZYK, R. G., KENT, R. N., DIAMENT, C. & O'LEARY, K. D. (1973) Measuring the reliability of observational data: a reactive process. *Journal of Applied Behavior Analysis*, **6**, 175–84.

SKINDRUD, K. (1972) An evaluation of observer bias in experimental field studies interaction. Unpublished doctoral dissertation, quoted *In* Jones, R. R. *et al* (op. cit.).

TAPLIN, P. S. & REID, J. B. (1973) Effects of instructional set and experimenter influence on observer reliability. *Child Development*, **44**, 547–54.

WING, J. K., COOPER, J. E. & SARTORIUS, N. (1974) *The Measurement and Classification of Psychiatric Symptoms.* Cambridge University Press.

—— NIXON, J. M., MANN, S. A. & LEFF, J. P. (1977) Reliability of the PSE (9th Edition used in a population survey). *Psychological Medicine*, **7**, 505–16.

* Robert W. Sanson-Fisher, PH.D., *Professor of Behavioural Science in Medicine, Faculty of Medicine, Newcastle University, New South Wales, Australia*

Claudia J. Martin, M.A., *Research Assistant, Department of Psychiatry, University of Manchester, The University Hospital of South Manchester, West Didsbury, Manchester M20 8LR*

* Correspondence to the senior author who was an Australian NH and MRC Research Fellow when the paper was completed.

(*Received 13 October; revised 23 December 1980*)

https://doi.org/10.1192/bjp.139.2.138 Published online by Cambridge University Press