Arch Psychiatr Nervenkr (1982) 232:395–406

Archiv für Psychiatrie und Nervenkrankheiten
Archives of Psychiatry and Neurological Sciences

Some Problems Concerning the Reliability and Structure of the Scales in the Inpatient Multidimensional Psychiatric Scale (IMPS)*

Victoria Cairns, Toni Faltermaier, Hans-Ulrich Wittchen, Horst Dilling, Werner Mombour, and Detlev von Zerssen
Max-Planck-Institute for Psychiatry, Kraepelinstrasse 10, D-8000 München 40, Federal Republic of Germany

Summary. Two problems concerning the IMPS are revealed. First, the importance of an appropriate reliability measure is demonstrated in a sample of 124 inpatients. Widely divergent results are found using three different intraclass correlation coefficients. The one producing the highest results was used by Klett and McNair (1966) and by Behrends et al. (1971). However, the results from this measure only apply to certain types of investigation. A measure more applicable to most investigations in psychiatric research (the intraclass correlation coefficient of individual ratings with between-rater variance included) produced results that were clearly lower, although still acceptable, with an average correlation of 0.72 between the 12 scales. Secondly, the multiplication of item scores by 2 or 8 when forming the scale scores is shown to lead to unusual and undesirable scale distributions in a sample of 1932 inpatients. The resulting implications for assessing change scores on some of the IMPS scales are discussed.

Key words: Psychiatric rating scales – Inter-rater reliability – Intraclass correlation coefficient

Introduction

In recent years there has been an increase in the number of structured and standardized clinical interviews as well as lists of self-rated symptoms (National Institute of Mental Health 1978; von Zerssen and Moeller 1980; Mombour 1972). Their primary aims are to improve the reliability and validity of psychiatric diagnoses and to increase comparability between national and international studies.

Offprints requests to: V. Cairns, GSF-MEDIS-Institut, Ingolstädter Landstrasse 1, D-8042 München-Neuherberg, Federal Republic of Germany

* This investigation was supported in part by a grant from the Deutsche Forschungsgemeinschaft

These instruments make it possible, through fairly structured rating procedures, to arrive at syndrome diagnoses that are relatively reliable and that can be replicated in many studies. One such instrument is the Inpatient Multidimensional Psychiatric Scale (IMPS, Lorr and Klett 1967). The particular advantages of the IMPS are: (a) the formulation of the psychiatric symptoms in everyday language, (b) its relative simplicity compared to more structured interviews such as the PSE (Wing et al. 1974), and (c) the presumed high inter-rater agreement on symptoms and syndromes.

In 1966 Klett and McNair performed a reliability study on the IMPS. They reported high reliability scores for both scales and items. These values were unusually high for assessments of psychiatric disorders. Three other reliability studies have been performed on German versions of the IMPS. As well as the study by Behrends et al. (1971)[1], which only gives the scale reliabilities, there are two unpublished dissertations. One, by Jacobi (1974), discusses in detail the strengths and weaknesses of a German version of the IMPS, but his study is based on different sets of rating scales for the individual items and different combinations of items into scales ("syndromes"), so few comparisons are possible with his work. The other study, by Bender (1976), also contains certain differences. One is that his sample under study, a sample of 311 inpatients, was a highly selected group, selected in particular to have very clearly defined diagnoses. He aimed at better discrimination between particular diagnostic groups on the basis of the IMPS items. This selection procedure, however, can lead to deceptively high reliability scores.

Klett and McNair (1966), on the other hand, had for their reliability study a large sample of 822 newly admitted inpatients that were selected as manifesting functional psychoses. However, their chosen reliability measure (the intraclass correlation coefficient for average ratings with between-rater variance excluded) is one that only applies to certain types of investigation (see below). This measure will be shown to produce particularly high results.

Mariotto and Farrell (1979) also assessed inter-rater reliability in a sample of 10 patients using 11 raters. They found very good reliabilities for ranked scores, as did Klett and McNair (1966), but also found that differences in levels of rating did exist between the raters, over the whole instrument as well as between the syndrome scales.

Although the practical usefulness of the IMPS can hardly be doubted, a closer look at the methodological criteria, in particular reliability, has revealed some problems. The following study involves an additional reliability study and compares the results with those found from previous reliability studies. Particular problem areas are identified and suggestions for improvement are given.

In addition, problems relating to the structure of the scales are examined. The items of the IMPS are rated on three different rating scales: the first 45 on a 9-point scale (0–8), the next 13 on a 5-point scale (0–4), and the remaining 32 on a 2-point scale (0–1). The scores from particular items are added to produce the 12 "syndrome" scores. To ensure that all the items receive the same weight in a "syndrome" scale, scores on all those items on a 5-point scale are multiplied by

1 The study by Behrends et al. (1971) was part of an international study. These reliability results have also been reported by Lorr and Klett (1968)

2 to put them in the range of 0 to 8, and scores on all those items rated on a 2-point scale are multiplied by 8 also to put them in the range of 0 to 8. This procedure, however, leads to unusual "syndrome" scale distributions, as will be shown.

As a result of this multiplication of item scores by a scalar, serious difficulties arise when looking at change scores on the IMPS scales. Mariotto and Paul (1974) said that "extreme caution should be excercised in any usage of the IMPS where absolute level differences are required (e.g., comparisons across patient groups, assessing changes in functioning, etc.)" (p. 507). Suggestions are made here that may help solve this problem.

Thus, the following questions were examined:
(a) What are the appropriate reliability indices for this instrument? How different are the results from different reliability indices applied to the same data set?
(b) How do the results from this data set compare with earlier studies? Which items and scales appear to have particularly low reliability? Do the reliability scores on the 9-point items change if the ratings are condensed into fewer categories?
(c) What effect does the multiplication of scores from the 2-point and 5-point items have on the distribution of the scale scores? How will this multiplication of scores effect further analyses?

Method

Subjects

The 124 inpatients for the reliability study came from three Bavarian mental hospitals (Haar, Kaufbeuren and Günzburg), from the psychiatric University hospital in Munich and from a neurological hospital in Munich (Hirnverletztenheim)[2]. A large proportion of the sample were long-stay patients. Of the 124 patients 23 were diagnosed psychotic with brain trauma or unspecified cerebral condition (I.C.D. 293.5 and 293.9) (I.C.D. 8th revision, Degkwitz et al. 1971), and 77 were diagnosed schizophrenic (2 with I.C.D. 295.3 and 295.6, and 75 with I.C.D. 295.9). Diagnoses of an affective psychosis (either I.C.D. 296.0 or 296.2) were given to 14, 9 were given diagnoses of a neurosis (I.C.D. 300.0, 300.4 or 300.5), and 1 was given a diagnosis of an unspecified personality disorder (I.C.D. 301.9).

The second sample included 1932 newly admitted inpatients of the hospital of the Max-Planck-Institute of Psychiatry, Munich. They were interviewed between the years 1970 and 1979 using the IMPS; 46% were men, 54% women, and their ages ranged from 13 to 85 years with a median of 31. Their diagnoses covered a broad spectrum and included 47% diagnosed psychotic (I.C.D. 290 to 299) and 50% diagnosed non-psychotic (I.C.D. 300 to 309), of which 55% were diagnosed neurotic (I.C.D. 300). The remaining 3% received other diagnoses. From this particularly large sample, the distributions of the IMPS scale scores could be seen.

Procedure

In our reliability study, each patient was seen by two raters, one as interviewer and one as observer. The 9 raters were all psychiatrists, although not especially trained in the use of the IMPS. They alternated in their roles of leading the interview and observing. The rating followed an open interview that lasted from 30 to 40 min. Each rater saw between 15 and

2 We are indebted to the directors of these institutions and their co-workers for their kind support of our investigation. The data were collected by H. Dilling, B. Lueth, D. Mattke, B. Schaeufelen, M. Wolf, K. Fliege, M. Adler, L. Barth, and G. Rose

40 patients. After the interview, the raters filled out the IMPS independently. In our second sample, patients admitted to the hospital were interviewed by a psychiatrist who also filled out the IMPS after the interview, which usually took place between 2 and 6 days after admission. Approximately 100 psychiatrists were involved in interviewing these patients.

Instrument

The IMPS contains 90 items and the scores on selected items are added to form the scale scores. The translation by Flegel and Mader (Düsseldorf) of Lorr's (1966) original version of the IMPS containing 75 items, along with the 15 additional items from the revised version, was employed in this study.

Data Analysis

With psychiatric scales one common form of measuring reliability is by looking at the agreement between different examiners of a patient[3]. Usually two examiners participate in the same interview and their following independent ratings are compared. This method, however, only takes into account the agreement between raters observing the same material. If raters made separate interviews with the same patient, further sources of variance would be added; the psychopathology observed would be different. Everyday psychiatric practice is generally characterized by independent interviews, so the reliability of results will not in general be as high as the reported reliabilities. Furthermore, it is important to calculate individual item reliabilities as well as the scale reliabilities, as it is well known that the magnitude of the reliability of a scale is a direct function of the number of items forming the scale (Sarris and Lienert 1974).

Recently, reliability studies have been criticized for employing inadequate statistical methods to calculate reliability. It is important that the reliability index is chosen to match the data structure (Bartko and Carpenter 1976), that chance agreement is taken into account (Spitzer and Fleiss 1974; Helzer et al. 1977; Bartko and Carpenter 1976), and that the chosen methods make it possible to compare the results with other studies. We therefore calculated the intraclass correlation coefficient (ICC) (Bartko 1966), weighted kappa (Cohen 1968) and, for comparison, the product moment correlation (although the product moment correlation coefficient is not to be recommended as it can produce high values even when differences do exist between the raters (Bartko and Carpenter 1976)).

However, there are several different formulae that may be used when calculating the ICC, a fact that is not always made quite clear. One must first decide whether one wants the "reliability of individual ratings" or the "reliability of average ratings". The latter is used when, in practice, two or more ratings are made on each individual and the results are averaged. If, however, individuals usually receive only one rating, then the "reliability of individual ratings" should be calculated (Ebel 1951).

Two different formulae are used to calculate these reliabilities. One need only use two raters per patient. (It is also possible to use more raters per patient but without averaging their ratings when calculating reliability.) Then from the one-way analysis of variance table the two ICC's may be calculated:

$$\text{ICC1} = (\text{MSB} - \text{MSW})/(\text{MSB} + (k\text{-}1)\ \text{MSW})$$

$$\text{ICC2} = (\text{MSB} - \text{MSW})/\text{MSB}$$

where: MSB = mean sum of squares between patients
MSW = mean sum of squares within patients
k = number of ratings per patient.

ICC1 assesses the reliability of individual ratings and ICC2 assesses the reliability of average ratings.

3 Other measures of reliability include the internal consistency of a scale and the stability of scores over time (test-retest). The latter tends to be low for psychiatric scales as psychiatric symptoms are often unstable over time. Results for the internal consistencies of the IMPS scales differed from the inter-rater reliabilities presented here. These will be discussed in another paper

As Bartko and Carpenter (1976) say: "One interpretation of ICC (E9) (ICC2) is: if another random sample of raters rated the same patients then ICC (E9) (ICC2) is approximately the correlation between the averaged ratings from the two sets of raters" (p. 317). That is, theoretically, if another set of raters were to rate the same set of patients, ICC2 would approximate the correlation between the averaged ratings. One does not need to use four raters per patient in a reliability study in order to assess the reliability of average ratings. Using only two raters per patient, ICC2 gives us the reliability of average ratings. These results then apply to the situation when in practice the ratings from two raters are averaged.

One must also decide whether to include or exclude the between-rater variance. If each rater's ratings are later standardized or ranked, then the differences in levels of rating between the raters should be removed, i.e., the between-rater variance should be excluded. Otherwise it should be included (Ebel 1951; Cohen 1968). We can calculate: ICC3 = (MSB − MSW)/MSB, where this time the within-patient sum of squares would be calculated as the total SS minus the between-patient SS and minus the between-rater SS, rather than just the total SS minus the between-patient SS. It is more common in psychiatric research to use the raw data in analyses, i.e., data that have not been standardized for each rater. Thus, in general, the between-variance should be included as part of the error.

Results

Scale Reliability

Three different ICCs and the product moment correlations between the ratings for the 12 scales are shown in Table 1, along with the results from Klett and McNair (1966)[4], Behrends et al. (1971) and Bender (1976). For 2 of the scales, Disorientation (DIS) and Obsessive-Phobic (OBS), all the patients in our study received low scores, and 94% and 88% of them, respectively, received a zero from both raters. Thus, these reliability scores were only based on the lower range of the scales, where one would not expect very high reliability.

What is particularly striking in Table 1 is the rather large differences between the three ICCs: ICC1, ICC2 and ICC3. (These are, respectively, the ICC of first individual and then average ratings with the inter-rater variance included, and finally the ICC of average ratings with the inter-rater variance excluded.) These results were calculated on the same set of data and yet produced differences between ICC1 and ICC3 of up to 0.25 (RTD) and between ICC1 and ICC2 of up to 0.17 (in DIS and OBS). The mean difference over the 12 scales between ICC1 and ICC2 was 0.10 and between ICC1 and ICC3 was 0.17. The importance of choosing the correct index of reliability is thus clearly demonstrated. The results from these three ICC indices apply to different research situations, that for ICC1 being perhaps the most common. There was good agreement between our (high) ICC3 results and those of Klett and McNair (1966) and Behrends et al. (1971) except on the 2 scales, DIS and OBS. Bender (1976), who calculated product moment correlations, tended to find higher values than we did, particularly in the scales ANX, RTD, DIS and OBS.

4 Klett and McNair's (1966) results from the analysis for "individual raters" are presented here. In their table of reliabilities, "combined raters" means four raters see one patient and the averages of pairs of ratings are compared. Our analysis is comparable to the case with "individual raters", in which the ratings of two individual raters are compared

Table 1. Reliability of the IMPS scales. Intraclass correlations (ICC) and product moment correlations (PM)

		Ours (N=124)				Klett and McNair (N=822)	Behrends et al. (N=201)	Bender (N=311)
		PM	ICC1	ICC2	ICC3	ICC3	ICC3	PM
Excitement	(EXC)	0.72	0.78	0.87	0.97	0.90	0.95	0.92
Hostility	(HOS)	0.82	0.81	0.89	1.01[a]	0.91	0.94	0.85
Paranoid projection	(PAR)	0.91	0.90	0.94	0.99	0.93	0.96	0.87
Grandiose expansiveness	(GRN)	0.94	0.94	0.97	1.00	0.92	0.96	0.88
Perceptual distortion	(PCP)	0.89	0.88	0.94	0.95	0.93	0.94	0.87
Anxious intropunitiveness	(ANX)	0.70	0.68	0.81	0.85	0.91	0.96	0.88
Retardation and apathy	(RTD)	0.62	0.62	0.77	0.87	0.90	0.95	0.87
Disorientation	(DIS)	0.45	0.45	0.62	0.68	0.96	0.75	0.92
Motor disturbances	(MTR)	0.75	0.78	0.88	0.94	0.82	0.92	0.88
Conceptual disorganization	(CNP)	0.76	0.76	0.87	0.90	0.89	0.91	0.88
Impaired functioning	(IMF)	0.75	0.73	0.85	0.94	—	0.95	0.86
Obsessive-phobic	(OBS)	0.35	0.35	0.52	0.58	—	0.88	0.81
Item range:								
Minimum		−0.01	−0.01	−0.02	0.04	—	—	0.43
Maximum		1.00	1.00	1.00	1.02[a]	—	—	0.91
Median item score		(0.63)	(0.59)	(0.75)	(0.82)	—	—	(0.77)

ICC1: Intraclass correlation coefficient of individual ratings with inter-rater variance included.
ICC2: Intraclass correlation coefficient of average ratings with inter-rater variance included.
ICC3: Intraclass correlation coefficient of average ratings with inter-rater variance excluded.

[a] As each rater did not see each patient, the sum of squares for raters must be estimated. This can sometimes give an inflated figure which leads to a negative sum of squares and thus an ICC greater than 1!

For comparison, Klett and McNair's and our ICC results on the 1st 10 scales are based on the original, standard form of the IMPS with 75 items. Both sets of PM results and Behrends et al.'s ICC results are based on the revised version of the IMPS with 90 items. However, only minor differences occur between the two.

Behrends et al. called the 11th scale "Depressive Mood (DPR)", rather than "Impaired Functioning (IMF)", as the items in this scale are related to depressive symptoms.

Table 2. Reliability of the IMPS items. Number of items with weighted kappa less than 0.4, between 0.4 and 0.6, and greather than 0.6. (Those items for which the raters agreed on the ratings of more than 120 of the cases were excluded from this table.)

	<0.4	0.4 to 0.6	>0.6
9-point (not at all → extremely)	17	25	3
5-point (not at all → very often)	3	1	7
2-point (no, yes)	2	9	10

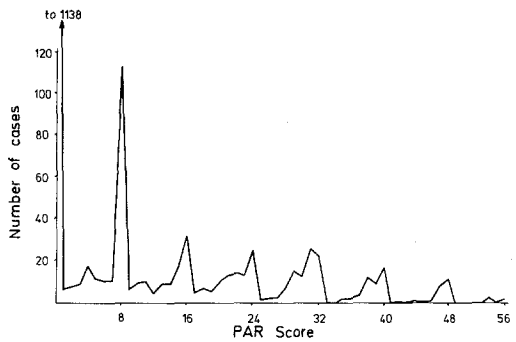

Fig. 1. Frequencies of scores on the Paranoid Projection Scale (PAR) in a sample of 1932 inpatients. *Note:* PAR is the sum of 7 items: 1 on a 9-point scale (scored 0 to 8) and 6 on 2-point scales (scored 0 or 8)

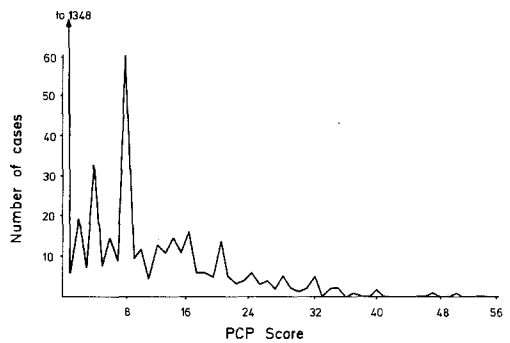

Fig. 2. Frequencies of scores on the Perceptual Distortion Scale (PCP) in a sample of 1932 inpatients. *Note:* PCP is the sum of 7 items: 1 on a 9-point scale (scored 0 to 8), 5 on a 5-point scale (scored 0, 2, 4, 6 or 8), and on 1 a 2-point scale (scored 0 or 8)

Item Reliability

An appropriate index for calculating the item reliabilities is weighted kappa. We calculated this for each item using as weights the differences between the row and column numbers, i.e. a linear weighting. For the 90 items, weighted kappa ranged between 0.06 and 0.76 with a median of 0.44 in our sample.

The number of items with low, medium and high values of weighted kappa for the items with the three different rating ranges appear in Table 2. The values of 0.4 and 0.6 have been chosen as cut-off points to identify low, medium and high values. Those items with values under 0.4 are listed below and those with values under 0.2 (particularly low) are marked by an asterisk:

9-point: Slowed speech, verbally hostile, posturing, contemptuous, fixed faces, blames self*, slowed movements, anxious (specific), speech blocking*, apathy, apprehensive, self-depreciating*, blames others, slovenly, guilt*, failure to answer, obsessive.

Table 3. Change scores on the Paranoid Projection Scale (PAR) between admission and discharge in a sample of 1103 inpatients. These included all the inpatients with full data on the 7 items making up the scale, both on admission and at discharge. Only part of the table is included here, as the whole scale ranges from 0 to 56

PAR on admission	PAR on discharge										
	0	1	2	3	4	5	6	7	8	9	10
0	696	1	0	2	1	1	0	1	8	1	1
1	4	1	0	0	0	0	0	0	0	0	0
2	4	0	0	1	0	0	0	0	0	0	0
3	7	0	0	0	0	0	0	0	0	0	0
4	8	0	0	0	0	0	0	0	1	0	0
5	5	2	0	0	0	1	0	0	0	0	0
6	7	2	0	0	1	0	0	0	0	0	0
7	8	0	0	0	0	0	0	0	0	0	0
8	52	1	1	0	0	0	0	0	9	0	0
9	2	0	0	0	0	0	0	0	0	0	0
10	5	0	0	0	0	0	0	0	0	0	0

5-point: Grimacing, talks to self, voices threaten.
2-point: Hopeless, gastro-intestinal symptoms.

Only 3 of the 45 items on a 9-point scale produced good reliability scores of over 0.6, while 17 had scores that were unsatisfactorily below 0.4. Of these 17 items with weighted kappa less than 0.4 12 were in the 2 scales ANX and RTD.

Bender (1976) lists his results for the product moment correlations for the individual items, so comparisons could be made between his and our studies. He found a minimum correlation of 0.43 and a median of 0.77, whereas we found a minimum of −0.01 and a median of 0.63 (see Table 1). Our results were lower than his primarily in those items making up the 2 scales ANX and RTD.

Of our ICC3 item scores 21 were over 0.9 compared to 1 of the ICC1 item scores. There was a difference of 0.24 between the ICC1 and ICC3 median item scores. The lower ICC scores tended to be from items in the scales ANX and DIS.

A large proportion of the items with low reliability scores were from the group of items rated from 0 to 8. These ratings are labelled: "not at all, very slightly, a little, mildly, moderately, quite a bit, distinctly, markedly, extremely". The ratings on these items were re-grouped into 0–1, 2–3, 4–5, 6–8 to correspond to the labels "none, mild, moderate, severe". Reliability coefficients were recalculated and the mean weighted kappa score remained at 0.42.

The Structure of the Scales

The combination of items with different rating scales leads to unusual scale distributions. See, for example, Figs. 1 and 2. In Fig. 1 the frequencies of the different scores for the Paranoid Projection scale (PAR) are plotted. This scale is made up of 6 items scored 0 or 8 and 1 item scored 0 to 8. Thus, the multiples of 8 on the scale are frequent. If the 9-point item (delusional) is scored 0 or 8 then

the score on PAR must be a multiple of 8. A score of 8 is seen to be particularly common. This occurs when 1 of the 2-point items is given a positive rating and all the other 6 items receive a zero, something which is quite likely, or when the 9-point item is given a rating of 8 (extremely) and all the other items receive a zero, something which is less likely.

In Fig. 2, the frequencies of the different scores for the Perceptual Distortions scale (PCP) are plotted. This scale is made up of 5 items on a 5-point scale, 1 on a 2-point scale and 1 on a 9-point scale. The 1 on a 2-point scale creates a peak at 8 and the 5 on 5-point scales cause the up and down zigzag of the distribution so that even numbered scores are more frequent. Odd numbered scores may only occur when there is an odd numbered rating on the item on a 9-point scale (hears voices). The same sort of phenomenon can be seen in the frequencies of scores on the 3 scales GRN, ANX and OBS.

Such odd-shaped distributions are not desirable if further analyses on the data are planned. For example, particular care must be taken when considering change scores on these scales. Table 3 gives a section of a contingency table for the change scores between admission and discharge on PAR in a sample of 1103 inpatients (a subsample of those in Fig. 1). Note the relatively large proportions of cases with changes from 8 to 0 and from 0 to 8.

A change from 8 to 0 is quite reasonably seen as a greater change than one from 7 to 0 or 6 to 0. However, when the 52 cases with PAR scores changing from 8 to 0 are looked at in more detail, this assumption turns out to be not quite so reasonable. On admission 11 of them had extremely severe delusions while the remaining 41 had one paranoid symptom present. Under this system, a patient with, for example, mild ideas of persecution which later disappear will be seen to have improved more than a patient who had marked delusions initially which also disappear.

Similar problems will arise with the analysis of some of the other scales. The scale PCP, ANX and OBS also all have a noticably higher proportion of cases with scores of 8, due to this multiplication of item scores by 8.

Discussion

The first point to be gathered from this study is the importance of an appropriate reliability measure. It must suit the data structure and, above all, produce results that are applicable to general situations in psychiatric research. The wide variability of different measures has been demonstrated here through the large differences produced in the results on the same set of data by the three different ICCs: ICC1, ICC2 and ICC3. Even between ICC1 and ICC2, two fairly realistic measures, differences of up to 0.17 (on DIS and OBS) were found in this sample. ICC1 is a reliability measure producing results that apply when, in practice, each patient is seen by only one rater. ICC2 produces results that apply when, in practice, each patient is seen by 2 raters. The latter situation is more desirable as the average ratings of two raters will be more reliable, but it is frequently not possible due to limited manpower, so ICC1 is then the appropriate measure of reliability. (Generally, the measure ICC1 is meant when the ICC is mentioned.)

The unusually high reliability coefficients found by Klett and McNair (1966) and Behrends et al. (1971) can now be explained. They calculated the ICC for the case when, in practice, each patient is seen by two raters and, in addition, the differences in the levels of rating between the raters is removed (thus, ICC3). This coefficient usually produced values of over 0.9 on these scales in both their studies as well as in our study. These results, however, only apply to the rather unusual situation when each rater's ratings are ranked or standardized. In general it is the raw scores given by the raters that are used in any further analyses.

On the whole, when comparing results using like measures, our results were as high as those found by Klett and McNair (1966) and by Behrends et al. (1971). Our item reliabilities were seen to be lower than Bender's (1976), particularly in those items making up the scales ANX and RTD and, moreover, in many of the 9-point items. His generally higher item reliabilities could be explained by the fact that all his patients were chosen to have very clear diagnoses, and therefore possibly clearer symptoms. It may be that those items with 9-point ratings are more difficult to rate, being more subjective, and thus, when the patients have less clear symptoms, the reliabilities on these particular items will drop. Some training of the interviewers in the use of the IMPS and a review of some of the items would probably lead to more reliable results.

In our study, those items on a 9-point scale tended to have lower reliabilities than those on 5-point or 2-point scales. Whether this is a function of the number of points in the scale or the nature of the items themselves is not clear. However, reducing the number of categories in the scale to 4 did not change the mean weighted kappa score for these 45 items. This result is in accordance with the results from other authors who have shown that reliability is not dramatically affected by the number of points in a scale, particularly above 5-points (Remington et al. 1979; Mattel and Jacobi 1971; Lissitz and Green 1975). It appears that other criteria should be used to select the number of categories for a scale.

Another problem identified in this study is that caused by adding item scores from items with different rating scales. The item scores are multiplied by 1, 2 or 8 to give all items the same possible maximum score of 8. Thus, the 2-point items are scored 0 or 8 and the 5-point items are scored 0, 2, 4, 6, or 8. This procedure has been shown to lead to unusual and undesirable scale distributions that could distort the results in any further analyses.

In order to avoid this, the same rating scale should be used for all items. Furthermore, this rating scale should be limited to 4 or 5 points with clearly defined psychological meaning, as some symptoms may be difficult to rate on a finer scale. There may still be some difficulty with items which may only be rated on a present-absent scale.

One solution for data that have already been gathered using the three different rating scales is to reduce those on 9- and 5-point scales to a 4-point scale (0 to 3) and to rate the 2-point items with scores of 0 or 2. One possible reduction for the rating scales is: for the 9-point items to group the scores 0 and 1 together, 2 and 3 together, 4 and 5 together and 6, 7 and 8 together and for the 5-point items to group the scores 3 and 4 together. These new item scores ranging from 0 to 3 may then be summed as before to produce the syndrome scores. Data that have already been collected may be reevaluated using this system. This does not

entirely remove the problem, as peaks will still occur for particular values in the distributions, but the distortions caused by the 2-point items are somewhat reduced.

References

Bartko JJ (1966) The intraclass correlation coefficient as a measure of reliability. Psycholog Rep 19 : 3-11

Bartko JJ, Carpenter WT (1976) On the methods and theory of reliability. J Ner Ment Dis 163 : 307-317

Behrends K, Flegel H, Helmchen H, Hippius H, Höffken KD, Schacht L, Schulte PW (1971) Quantifizierung psychotischer Symptome unter transkulturellen Aspekten. Soc Psychiatr 6 : 66-72

Bender W (1976) Studie zur Reliabilität und differentiellen Validität der Lorr-Skala (IMPS). Unpublished Dissertation, Hamburg

Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70 : 213-220

Degckwitz R, Helmchen H, Kockott G, Mombour W (1971) Diagnosenschlüssel und Glossar Psychiatrischer Krankheiten: ICD (International Classification of Diseases). 8th Revision. Deutsche Ausgabe der Internationalen Klassifikation der WHO und des Internationalen Glossars. Springer, Berlin Heidelberg New York

Ebel RL (1951) Estimation of the reliability of ratings. Psychometrika 16 : 407-424

Helzer JE, Robins LN, Taibleson M, Woodruff RA, Reich T, Wish ED (1977) Reliability of psychiatric diagnosis: I. A methodological review. Arch Gen Psychiatr 34 : 129-133

Jacobi P (1974) Untersuchungen zur Faktorenstruktur, Zuverlässigkeit und Gültigkeit einer deutschen Bearbeitung der IMPS nach Lorr. Unpublished Dissertation, Hamburg

Klett CJ, McNair DM (1966) Reliability of the acute psychotic types. In: Lorr M (ed) Explorations in typing psychotics. Pergamon Press, Oxford London New York, pp -

Lissitz RW, Green SB (1975) Effect of the number of scale points on reliability: A Monte Carlo approach. J Appl Psychol 60 : 10-13

Lorr M (ed) (1966) Explorations in typing psychotics. Pergamon Press, Oxford London New York

Lorr M, Klett CJ (1967) Manual for the inpatient multidimensional psychiatric scale (IMPS); revised. Consulting Psychologists Press, Palo Alto, Calif.

Lorr M, Klett CJ (1968) Major psychotic disorders. Arch Gen Psychiatr 19 : 652-658

Mattel MS, Jacobi J (1971) Is there an optimal number of alternatives for Likert scale items; Study I: Reliability and validity. Educ Psychol Measurement 31 : 657-674

Mariotto MJ, Farrell AD (1979) Comparability of the absolute level of ratings on the Inpatient Multidimensional Psychiatric Scale within a homogeneous group of raters. J Consult Clin Psychol 47 : 59-64

Mariotto MJ, Paul GL (1974) A multimethod validation of the Inpatient Multidimensional Psychiatric Scale with chronically institutionalized patients. J Consult Clin Psychol 42 : 497-508

Mombour W (1972) Verfahren zur Standardisierung des psychopathologischen Befundes. Teile 1 und 2. Psychiatria Clin 5 : 73-120; 5: 137-157

National Institute of Mental Health (1978) Handbook of psychiatric scales, (2nd ed). Preston & Assoc, Inc, Rockville, Maryland

Remington M, Tyrer PJ, Newson-Smith J, Cicchetti DV (1979) Comparative reliability of categorical and analogue rating scales in the assessment of psychiatric symptomatology. Psychol Med 9 : 765-770

Sarris V, Lienert GA (1974) Konstruktion und Bewährung von klinisch-psychologischen Testverfahren. In: Schrant WJ, Baumann H (Hrsg) Klinische Psychologie. Huber, Bern, S 286-351

Spitzer RL, Fleiss JL (1974) A re-analysis of the reliability of psychiatric diagnosis. Br J Psychiatr 125 : 341–347

Wing JK, Cooper JE, Sartorius N (1974) Measurement and classification of psychiatric symptoms. An instruction manual for the PSE and Catego program. University Press, Cambridge

Zerssen D von, Möller H-J (1980) Psychopathometrische Verfahren in der psychiatrischen Therapieforschung. In: Biefang S (Hrsg) Evaluationsforschung in der Psychiatrie: Fragestellungen und Methoden. Enke, Stuttgart, S 121–166

Received April 29, 1982