Inter-rater reliability of the Brief Psychiatric Rating Scale and the Groningen Social Disabilities Schedule in a European multi-site randomized controlled trial on the effectiveness of acute psychiatric day hospitals

MATTHIAS SCHÜTZWOHL[1], JOANNA JAROSZ-NOWAK[2], JANE BRISCOE[3], KRZYSZTOF SZAJOWSKI[2], THOMAS KALLERT[1], AND THE EDEN STUDY GROUP[4]

[1]Department of Psychiatry at Dresden University of Technology, Dresden, Germany

[2]Institute of Mathematics at Wrocław University of Technology, Wrocław, Poland

[3]Unit for Social and Community Psychiatry, Barts and the London School of Medicine, Queen Mary, University of London, UK

[4]The following colleagues contributed to the EDEN Study. Dresden: Thomas Kallert, Matthias Glöckner, Christiane Matthes, Matthias Schützwohl; London: Stefan Priebe, Jane Briscoe, Rosemarie McCabe, Paul De Ponte, Alfred Okine, Maria Vidal, Donna Wright; Michalovce: Pětr Nawka, Denisa Borosová, Danica Korpová, Helena Reguliová; Prague: Jiří Raboch, Nicole Baková, Andrea Howardová, Jana Peichlová, Miroslav Sekot, Lucie Stárková; Wrocław: Andrzej Kiejna, Tomasz Adamowski, Tomasz Hadryś, Joanna Jarosz-Nowak, Krzysztof Małyszczak, Joanna Rymaszewska, Krzysztof Szajowski, Elżbieta Trypka.

ABSTRACT *The objectives of this study were to report the inter-rater reliability of the Brief Psychiatric Rating Scale (BPRS 4.0) and the Groningen Social Disabilities Schedule (GSDS-II) as assessed in a randomized controlled trial on the effectiveness of psychiatric day hospitals spanning five sites in countries of Central and Western Europe.*

Following brief training sessions, videotaped BPRS-interviews and written GSDS-vignettes were rated by clinically experienced researchers from all participating sites. Inter-rater reliability often proved to be poor for items assessing the severity of both psychopathology and social dysfunction, but findings suggest that both instruments allow for the assessment of the presence or absence of specific psychopathological symptoms or social disabilities. Inter-rater reliability at subscale level proved to be good for both instruments.

Results indicate that, with a brief training session and proper use of the instruments, psychopathology and social disabilities can be reliably assessed within cross-national research studies. The results are of particular interest given that the need to conduct cross-national multi-site studies including countries with different cultural backgrounds increases.

Key words: inter-rater reliability, multi-site psychiatric studies, EDEN study, Brief Psychiatric Rating Scale, Groningen Social Disabilities Schedule

Introduction

Strong recommendations for cross-national public mental health research have appeared in the literature within the last few years (Alonso et al., 2001; Becker and Vázquez-Barquero, 2001; Marshall et al., 2001). In particular, research on mental health service evaluation and comparison has identified the growing need for instruments that allow cross-national analyses. Currently, very few measures are available with standardized translations to ensure their cross-national applicability (Knudsen et al., 2000). In response to this, the EPSILON study (Becker et al., 2000; Leese et al., 2001) recently adapted five key instruments for the assessment of needs, service utilization and costs, informal carer involvement, quality of life, and service satisfaction, for use in five languages (Danish, Dutch, English, Italian and Spanish). Further dimensions should also be assessed in psychiatric service evaluation, however (Hansson, 2001), and future multi-site studies will increasingly involve other European languages (for example, Kallert et al., 2002; Becker et al., 2004). Therefore, the need to establish and maintain high cross-national inter-rater reliability on specific instruments will continue (cf. Shrout, 1998; Leese et al., 2001). With few resources from research grants to allow proper cross-national adaptation processes as conducted in the EPSILON study (Knudsen et al., 2000), more economical processes for successful implementation are of particular interest.

Findings on this topic are limited. According to Edson et al. (1997), inter-rater reliability across sites was possible for the assessment of psychopathology using an 18-item modified version of the original Brief Psychiatric Rating Scale (BPRS) (Overall and Gorham, 1962) and the assessment of social functioning using the Global Assessment of Functioning Scale (GAF) (American Psychiatric Association, 1994), following an initial group training session, assessment from video-taped interviews of subjects, and refresher training when necessary. All of the sites in this study were in the US, however, and thus the results cannot be generalized to cross-national and cross-cultural studies.

In this context, this article reports the inter-rater reliability for the BPRS 4.0 (Ventura et al., 1993) and the Groningen Social Disabilities Schedules (GSDS-II) (Wiersma et al., 1990) in a European multi-site randomized controlled trial on the effectiveness of psychiatric day hospitals, a study comprising researchers in Dresden (Germany), London (UK), Michalovce (Slovak Republic), Prague (Czech Republic), and Wrocław (Poland). The primary aims of this study and the participating sites have been described in detail elsewhere (Kallert et al., 2004).

The authors of the original English version of the BPRS 4.0 found good to excellent reliabilities for most of the 24 items as indicated by intra-class correlation coefficients (ICCs), both for researchers with advanced degrees (median ICC = 0.81) and for researchers with pre-doctoral degrees (median ICC = 0.83) (Ventura et al., 1993), following an initial training programme. Assessing reliabilities of the Italian version of the BPRS 4.0 following a training using videotaped BPRS interviews, Roncone and colleagues (1999) found significant differences in the median of the ICC values achieved for the 24 BPRS-items between a group of experienced clinical psychiatrists and psychologists (median ICC = 0.78), a group of psychosocial rehabilitation students (median ICC = 0.65), and a group of medical students with no previous contact to psychiatric patients (median ICC = 0.65). As in the study by Ventura et al. (1993), reliability was found to be higher for items based on the patients' self-report (ICC = 0.83 for psychiatrists and psychologists; ICC = 0.78 for psychosocial rehabilitation students, ICC = 0.79 for medical students) compared with items based on raters' observations of the patients' behaviour (ICC = 0.50 for psychiatrists and psychologists; ICC = 0.57 for psychosocial rehabilitation students, ICC = 0.38 for medical students).

To date, no international publication has addressed the reliability of the GSDS-II. When rated by an interviewer as well as an observer from a group of trained interviewers with different clinical backgrounds, good reliability of the first version of the GSDS was demonstrated, as indicated by weighted kappa coefficients ranging between 0.63 and 0.93 for the eight social roles (median = 0.78) and between 0.53 and 0.81 for the dimensions of the roles (median = 0.71) (Wiersma et al., 1988).

Because the implementation of a rigorous training programme comparable to the aforementioned studies was not feasible, reliability within the EDEN-study was expected to be somewhat lower but still acceptable on both measures.

Material and methods

Description of the instruments

The Brief Psychiatric Rating Scale, Expanded Version (BPRS 4.0).

The Brief Psychiatric Rating Scale was originally developed as a 16-item version to assess symptom change rapidly in psychiatric inpatients with various diagnoses (Overall and Gorham, 1962). It has since been expanded to a 24-item version with clear guidelines for its administration, including detailed anchor points and probe questions for each item (Lukoff et al., 1986; Ventura et al., 1993). The rating scale ranges from '1' ('not present') to '7' ('extremely severe'); 'NA' indicates symptoms not assessed. The valid ratings for each item can be added up to a total sum score and to sum scores on different subscales (see Table 2).

While the Dresden and London centres were able to use existing national versions of the BPRS 4.0, the existing scale was translated for use in the other three centres. Two or three researchers translated the instrument at each site and the results were subsequently compared and discussed. Rather than translating all of the materials, researchers in Dresden, Michalovce, Prague, and Wrocław reverted to the original English version on issues concerning the anchor points and administration manual.

The Groningen Social Disabilities Schedule, Second Version (GSDS-II)

The Groningen Social Disabilities Schedule GSDS-II (Wiersma et al., 1988; Wiersma et al., 1990) is a semi-structured interview for the assessment of social disabilities on eight different social roles, each subdivided into a number of dimensions (see Table 3). The GSDS-II defines a social role as a complex of norms and expectations prevailing within the so-called relevant reference group comprising people who in a social or other respect are of great importance to the individual.

In conducting this assessment, ratings are given for each of the eight roles (the overall role ratings), and for each dimension of the role (the dimensional ratings). For three of the roles the client must be assigned to specific categories prior to rating. For example, different rules apply for 'individuals living together' (Category I) and 'individuals living alone' (Category II) when rating the 'family role'. The rating scale ranges from '0' ('no disability') to '3' ('severe disability'). If necessary information is absent or unreliable, making an assessment impossible, a rating of '8' is given, and a rating of '9' indicates that the particular role behaviour is not applicable. In general, the overall role rating is equal to the highest of the dimensional ratings (with the exception, of course, of ratings 8 and 9), although it is permissible to subtract one point, 'whenever the interviewer has good reasons to do so' (Wiersma et al., 1990: 10). Further guidelines for making ratings for each role are given in detail in the GSDS-II manual (Wiersma et al., 1990). A sum score based on Overall Role Ratings can be computed.

The GSDS-II manual includes questions for the interview assessment and the rating form for recording all data. Because the major questions are repeated on the rating form, the use of this form during actual interviews is considered sufficient (Wiersma et al., 1990). Within the EDEN study, both the GSDS-II manual and rating form were translated from English into German, but only the rating form was translated into Czech, Polish, and Slovak. In Dresden, only one researcher (TK) translated the materials; in Prague, Wrocław and Michalovce, two or three researchers completed the translations with subsequent discussion.

Training

BPRS 4.0

A short training session on the BPRS 4.0 was conducted during the first meeting of the EDEN study group in September 2000. This followed a five-day training on the Schedules for Clinical Assessment in Neuropsychiatry (SCAN, Version 2.1; World Health Organization, 1999), a set of instruments designed to assess, measure and classify the psychopathology and behaviour associated with the major psychiatric syndromes of adult life. The initial BPRS training primarily comprised the provision of detailed information on the instrument and its usage. The total duration was 1.5 hours.

GSDS-II

A short training session on the GSDS-II was also part of the first meeting of the EDEN study group in September 2000. After an introduction of the instrument and its basic rating rules, each researcher rated five written vignettes and participated in a subsequent discussion of the ratings; the total duration was about 3 hours. Within the following 2 months, all researchers rated 20 anonymous case vignettes, written in English; a researcher from the Dresden site (MS) provided detailed written feedback on these ratings.

Reliability assessment

BPRS 4.0

To assess inter-rater reliability, BPRS 4.0 interviews were audio-visually recorded in the London site and circulated to each centre. Clients already participating in the EDEN study and receiving acute psychiatric treatment at the day hospital or on the inpatient ward at the time were approached to participate in the training. Selected clients with a range of diagnoses gave written informed consent to have the interviews recorded and used for training purposes. The local ethics committee approved the procedure. Within the first year of the EDEN study (December 2000 to November 2001), sessions with 19 patients interviewed by four different interviewers were recorded on four videotapes. Each videotape was rated by 11 to 17 of the 22 research assistants with clinical experience. Within a training and feedback session of a subsequent meeting of the EDEN study group in April 2001, ratings of the first two videotapes were discussed retroactively and consensus ratings were established. Diagnosis, according to ICD-10, was as follows: F10.2 (n = 2), F12.5, F20.0 (n = 2), F21.0, F25.0, F31.0 (n = 2), F31.2 (n = 3), F32.0 (n = 2), F33.1 (n = 2), F41.2 (n = 2), F43.2.

GSDS-II

To assess inter-rater reliability for the GSDS-II, researchers from each site sent two written anonymous case vignettes based on participants of the EDEN study to the other centres once a month. During the first year of the EDEN study, a total of 130 vignettes were each rated by 12 to 18 of the 20 researchers. The ratings were discussed retroactively during subsequent meetings of the EDEN-study group in January, April, and July 2001. The written vignettes were based on 73 women and 57 men, with an age range of 18 to 64 and a mean age of 37 years; 111 case reports referred to the last four weeks prior to admission, and 19 to the last four weeks prior to discharge from acute treatment.

Participants

The raters included the following researchers:

- Dresden: one psychiatrist and two clinical psychologists (one woman and two men), all with lengthy clinical experience.
- Prague: six psychiatrists (five women and one man), all medical specialists in psychiatry.
- London: one psychologist/mental health nurse with lengthy clinical experience and five psychologists (four women and one man) with limited clinical experience.
- Michalovce: two psychiatrists and one clinical psychologist (two women and one man), all with lengthy clinical experience.
- Wrocław: five psychiatrists (two women and three men), including three medical specialists in psychiatry and two residents in psychiatry.

Data storage

To avoid errors during data collection, a researcher from the Wrocław site (JJ-N) created specific Excel-files with defined formulas controlling entered values for both the BPRS 4.0 and the GSDS-II.

Statistical analyses

Several different statistical analysis techniques were used to assess inter-rater reliability:

- For concordance for each of the 24 items of the BPRS 4.0, a modified kappa statistic κ_M (Jarosz-Nowak, 2002) was used to assess agreement:

$$\kappa_M = \frac{p_o - p_t}{1 - p_t}$$

Here p_o denotes an observed agreement and p_t denotes a theoretical probability of agreement by chance. Thus, the modification of the classical Cohen's kappa concerns the estimation of agreement by chance. The probability of giving the same rating is not estimated from a sample. Instead it is a theoretical probability of agreement by chance treated as a parameter, which is fixed a priori and depends on the scale of the questionnaire. To calculate observed agreement, missing ratings are excluded. Modified kappa statistics are given for three cases: a) dichotomous scale (1 = absent; 2 = present with present being equivalent to a rating of ≥ 2), b) trichotomous scale (1 = absent; 2 = mild with mild being equivalent to a rating of 2, 3, or 4; and 3 = severe with severe being equivalent to a rating ≥ 5), and c) the full seven-point ordinal scale. For the dichotomous assessment, $p_t = 0.58$; for the trichotomous assessment, $p_t = 0.23$; and for the seven-point ordinal scale $p_t = 0.11$.

Furthermore, inter-rater reliability for each of the 24 BPRS-4.0-items was calculated using unbiased intra-class correlation coefficients (ICC). The ICC proposed by Bartko and Carpenter (1976) does not require the same number of raters per patient.

In order to assess the inter-rater reliability for each dimensional rating and the overall role ratings of the GSDS-II, three different methods were applied by modified kappa statistics and Cohen's kappa respectively. For all methods, the rating '8' ('the assessment is not possible') was treated as a missing value. Likewise, the rating 9 ('role behaviour is not applicable') was treated as a missing value for roles with categories where the rating '9' is imposed by the manual, because in such cases the rating was not based on a rater's decision. For roles without categories, the rating '9' was treated as a significant rating since in these cases it was a rater's decision. For roles with assignment to categories, dimensional ratings were compared only if raters assigned clients to the same category. Modified kappa statistics are given for two cases: a) the full four-point ordinal scale, and B) dichotomous assessment. For the full four-point scale, $p_t = 0.14$ for roles without an assignment to categories and $p_t = 0.11$ for roles with such an assignment. For the dichotomous assessment (0 = lack of disability; 1 = disability present with present being equivalent to a rating of 1, 2, or 3), $p_t = 0.31$ for roles without an assignment to categories and $p_t = 0.28$ for roles with such an assignment. For dichotomous assessment, a mean of Cohen's kappas was also calculated. For each pair of raters the kappa was computed. The coefficient of agreement was defined as an average of obtained kappas.

- For the assessment of agreement between raters based on mean scores of total scales and subscales of the BPRS 4.0 and the GSDS-II, unbiased ICC was applied.

Confidence intervals for kappas were evaluated by Fisher's *z*-transform constructed with the jack-knife variance (Borkowf, 2000).

According to the definition used by Roncone et al. (1999), inter-rater reliabilities ≥ +0.75 were considered to be 'good,' inter-rater reliabilities < +0.50 were considered to be 'poor,' and intermediate values were regarded as 'acceptable.' Thus, these guidelines are more stringent than those previously proposed by Cicchetti et al. (1992) or Leese et al. (2001).

Results

The Brief Psychiatric Rating Scale, Expanded Version (BPRS 4.0)

Inter-rater reliability for each of the 24 BPRS 4.0 items ranges widely in this study (Table 1). The inter-rater reliability for dichotomous assessment appears to be good for six and at least acceptable for 16 of the 24 items (median = 0.59). The inter-rater reliability for the trichotomous assessment was good for four and at least acceptable for 23 items (median = 0.69). The analyses based on the assumption of a seven-point ordinal scale showed good or at least acceptable inter-rater reliabilities for 17 items (median = 0.61). Intra-class correlation coefficients showed good inter-rater reliability for eight and at least acceptable inter-rater reliability for 14 items (median = 0.61).

Inter-rater reliability was good for self-reported symptoms (0.80) but poor for observed behaviour (0.53). At subscale level, using Ventura et al.'s (2000) four-factor solution, reliability was good for the manic/excitement and depression/anxiety domains, and acceptable for positive and negative symptom domains. Using Roncone et al.'s (1999) factor solution, reliability was good for all subscales besides the negative symptom domains. Using the five-factor solution reported by Hafkenscheid (1991) for the 18-item version of the BPRS, reliability proved to be poor for the activation domain, acceptable for the anergia domain, and good for the anxiety/depression, thought disorder and hostility domains (Table 2).

The Groningen Social Disabilities Schedule, Second Version (GSDS-II)

Inter-rater reliability proved to be good for the clients' assignment to role-specific categories – the 'family role', the 'partner role' and the 'occupational role' (Table 3).

Inter-rater reliabilities achieved for the dimensional ratings ranged from 0.36 to 0.92 (median = 0.56). On the four-point ordinal scale, inter-rater reliability appears to be good for one dimension only, acceptable for 19 dimensions and poor for two dimensions, namely for Dimension C ('active interest

Table 1. Inter-rater reliability of BPRS items

	Modified kappa for dichotomous assessment	Modified kappa for trichotomous assessment	Modified kappa for the seven-point ordinal scale	ICC
BPRS items based on patient's self-report				
1. Somatic concern	0.84 (0.81,0.87)	0.83 (0.81,0.85)	0.70 (0.64,0.76)	0.90 (0.83, 0.95)
2. Anxiety	0.80 (0.73,0.85)	0.61 (0.52,0.68)	0.36 (0.24,0.47)	0.79 (0.68, 0.89)
3. Depression	0.89 (0.86,0.92)	0.73 (0.67,0.79)	0.48 (0.36,0.59)	0.88 (0.80, 0.94)
4. Suicidality	0.90 (0.88,0.92)	0.74 (0.68,0.79)	0.60 (0.49,0.68)	0.87 (0.79, 0.94)
5. Guilt	0.76 (0.68,0.82)	0.73 (0.67,0.78)	0.55 (0.43,0.65)	0.71 (0.57, 0.84)
6. Hostility	0.59 (0.41,0.72)	0.69 (0.62,0.75)	0.47 (0.37,0.57)	0.76 (0.64, 0.87)
7. Elevated mood	0.52 (0.36,0.66)	0.63 (0.56,0.70)	0.55 (0.43,0.65)	0.60 (0.45, 0.76)
8. Grandiosity	0.77 (0.68,0.83)	0.71 (0.64,0.78)	0.64 (0.52,0.73)	0.79 (0.67, 0.89)
9. Suspiciousness	0.56 (0.41,0.67)	0.56 (0.48,0.64)	0.44 (0.32,0.55)	0.71 (0.58, 0.84)
10. Hallucinations	0.63 (0.50,0.74)	0.64 (0.55,0.71)	0.58 (0.47,0.68)	0.78 (0.66, 0.88)
11. Unusual thought content	0.66 (0.52,0.76)	0.66 (0.57,0.73)	0.58 (0.46,0.68)	0.76 (0.64, 0.87)
12. Bizarre behaviour	0.47 (0.28,0.63)	0.69 (0.61,0.76)	0.67 (0.57,0.74)	0.38 (0.25, 0.59)
13. Self-neglect	0.57 (0.41,0.70)	0.76 (0.71,0.81)	0.71 (0.64,0.77)	0.47 (0.32, 0.66)
14. Disorientation	0.57 (0.38,0.71)	0.76 (0.70,0.81)	0.77 (0.71,0.82)	0.33 (0.20, 0.53)
BPRS items based on rater's observation of patient's behaviour				
15. Conceptual disorganization	0.33 (0.14,0.50)	0.63 (0.56,0.69)	0.61 (0.53,0.69)	0.27 (0.15, 0.46)
16. Blunted affect	0.49 (0.34,0.61)	0.62 (0.55,0.68)	0.40 (0.29,0.49)	0.57 (0.42, 0.74)
17. Emotional withdrawal	0.31 (0.13,0.48)	0.59 (0.51,0.65)	0.47 (0.37,0.57)	0.43 (0.28, 0.62)
18. Motor retardation	0.36 (0.17,0.52)	0.63 (0.55,0.69)	0.54 (0.44,0.63)	0.46 (0.31, 0.65)
19. Tension	0.01 (-0.12,0.13)	0.44 (0.39,0.50)	0.33 (0.25,0.40)	0.27 (0.15, 0.46)
20. Uncooperativeness	0.58 (0.42,0.70)	0.76 (0.70,0.81)	0.78 (0.73,0.82)	0.25 (0.14, 0.44)
21. Excitement	0.52 (0.37,0.65)	0.66 (0.69,0.73)	0.63 (0.53,0.71)	0.65 (0.50, 0.80)
22. Distractibility	0.44 (0.26,0.59)	0.67 (0.61,0.73)	0.66 (0.58,0.73)	0.40 (0.26, 0.60)
23. Motor hyperactivity	0.48 (0.33,0.60)	0.66 (0.59,0.72)	0.63 (0.53,0.71)	0.51 (0.36, 0.69)
24. Mannerism/ posturing	0.54 (0.38,0.67)	0.74 (0.68,0.79)	0.76 (0.70,0.81)	0.26 (0.14, 0.45)

Note. The table shows the exact value and, in parentheses, the lower and upper confidential limits using a 95% confidence interval

in getting a partner') of the 'partner role', and Dimension B ('participation in societal groups, organizations and/or clubs') of the 'citizen role'. The inter-rater reliability for the dichotomous assessment was good for 17 dimensions, acceptable for four dimensions, and poor for Dimension B of the 'citizen role' only (median = 0.81). The mean of Cohen's kappas as achieved for the dichotomous assessment was good for nine dimensions, acceptable for eleven dimensions, and poor for two dimensional ratings ('partner role', Dim. C; 'citizen role', Dim. B; median = 0.68).

In the Overall Role Ratings, inter-rater reliability proves to be (almost) good for all social roles when calculating the modified kappa for the dichotomous assessment (median = 0.80), and to be at least acceptable when using the mean of Cohen's kappas for dichotomous assessment (median = 0.63). When calculating the modified kappa for the four-point ordinal scale, however, the standard for good inter-rater reliability was not met for any social role, and the inter-rater reliability is actually poor for the overall role rating of the family role, the partner role and the citizen role (median = 0.51).

The reliability for the GSDS-II sum score of all overall role ratings is good (ICC = 0.77 (CI 95% 0.72, 0.81)).

Discussion

This paper has sought to report the inter-rater reliability of the BPRS 4.0 and the GSDS-II within a cross-national multi-site randomized controlled trial

Table 2. Inter-rater reliability of BPRS subscales

Subscale	Numbers of items	ICC (95% CI)
BPRS 24-items global score	1–24	0.78 (0.65,0.89)
Self-report	1–14	0.80 (0.69,0.90)
Observed behaviour	15–24	0.53 (0.37,0.72)
Manic/ excitement	6, 7, 8, 21, 22, 23	0.75 (0.62,0.87)
Negative symptoms	13, 16, 17, 18	0.62 (0.47,0.79)
Positive symptoms	9, 10, 11, 12, 14	0.74 (0.61,0.87)
Depression/anxiety	2, 3, 4, 5	0.93 (0.88,0.97)
Positive symptoms	10, 11, 15	0.78 (0.66,0.89)
Negative symptoms	16, 17, 18	0.62 (0.47,0.79)
Depression	3, 4, 5	0.93 (0.87,0.96)
Psychotic disintegration scale	6, 8–11, 15–17, 20, 24	0.83 (0.73,0.92)
BPRS 18-items global score	1–3, 5, 6, 8–11, 14–21, 24	0.79 (0.68,0.90)
Anxiety/depression	1, 2 ,3, 5	0.90 (0.84,0.95)
Anergia	14, 16, 17, 18	0.63 (0.48,0.80)
Thought disorder	8, 10, 11, 15	0.83 (0.73,0.92)
Activation	19, 21, 24	0.46 (0.31,0.67)
Hostility	6, 9, 20	0.76 (0.64,0.88)

Note. The table shows the inter-rater reliability obtained by calculating ICCs for the mean score for different BPRS subscales. The subscales 'manic/ excitement', 'negative symptoms', 'positive symptoms', and 'depression/anxiety' are taken from Ventura et al. (2000); the subscales 'positive symptoms', 'negative symptoms', 'depression,' and 'psychotic disintegration scale' are consistent with those reported by Roncone et al. (1999); the subscales 'anxiety/depression', 'anergia', 'thought disorder', 'activation', and 'hostility' are consistent with those reported by Hafkenscheid (1991). It appears, however, that the composition of factors for the 24-item BPRS varies across samples (Ventura et al., 2000).

on the effectiveness of psychiatric day-hospital treatment, a study in which only economical training measures could be implemented to ensure the reliability of the assessed data.

BPRS 4.0

As expected, the inter-rater reliability of the BPRS 4.0 found within the EDEN-study was, altogether, slightly lower than the inter-rater reliabilities reported in other publications for clinicians and researchers with similar experience (Ventura et al., 1993; Roncone et al., 1999).

On item level, inter-rater reliability for each of the 24 BPRS 4.0 items ranged widely. As in previous studies on the 18-item version of the BPRS (Hedlund and Vieweg, 1980) and the 24-item version BPRS 4.0 (Ventura et al., 1993; Roncone et al., 1999), irrespective of the applied statistical method, reliabilities for items based on the patient's self-report were higher compared with those based on the rater's observation of the patient's behaviour. The latter is not surprising in this case, given the limited amount of information imparted by videotaped interviews, for example only the upper part of the body has been recorded, assessing eye contact is difficult, and so forth. Overall, findings suggest that most of the 24 BPRS 4.0 items cannot be reliably used separately in (cross-national) research studies, although the partly-low reliability coefficients might be attributable to the serious drawbacks of the ICC in analysing symptoms which are either extremely rare or extremely frequent (Anker, 1983).

In contrast, at BPRS-subscale level, findings showed largely satisfactory inter-rater reliabilities. When interpreting these findings, however, it is important to bear in mind the appropriateness of rating videotaped interviews to estimate the inter-rater reliability. Hafkenscheid (1991), for example, used repeated separate interviews by single clinicians to determine inter-rater reliability, and found lower BPRS-subscales reliabilities compared with our study or the study by Roncone et al. (1999) that also used videotaped interviews. In this respect, we would like to argue that good inter-rater reliabilities derived

Table 3. Inter-rater reliability of GSDS-II-items

Roles	Modified kappa for the four-point ordinal scale	Modified kappa for dichotomous assessment	Mean of Cohen's kappas for dichotomous assessment
The role of self-care			
dim. A 'personal care	0.61 (0.58,0.64)	0.77 (0.75,0.79)	0.71 (0.66,0.74)
dim. B 'self-presentation'	0.60 (0.56,0.63)	0.69 (0.67,0.72)	0.55 (0.49,0.61)
Overall role rating	0.60 (0.57,0.63)	0.78 (0.70,0.80)	0.68 (0.63,0.71)
Family role	0.98 (0.97,0.98)	0.98 (0.97,0.98)	0.95 (0.94,0.95)
dim. A 'contribution to atmosphere and preservation'	0.51 (0.48,0.54)	0.82 (0.81,0.84)	0.61 (0.56,0.66)
dim. B 'contribution to the economic independence'	0.52 (0.48,0.56)	0.71 (0.68,0.73)	0.61 (0.54,0.68)
dim. C 'one person household'	0.57 (0.49,0.63)	0.85 (0.83,0.87)	0.69 (0.63,0.74)
Overall role rating	0.49 (0.47,0.52)	0.83 (0.82,0.85)	0.58 (0.48,0.66)
Kinship role			
dim. A 'affective relationship with parent'	0.68 (0.65,0.71)	0.81 (0.80,0.83)	0.82 (0.80,0.84)
dim. B 'actual contacts with parents'	0.62 (0.59,0.65)	0.75 (0.73,0.77)	0.78 (0.75,0.80)
dim. C 'affective relationship and actual contacts with siblings'	0.64 (0.61,0.67)	0.78 (0.76,0.80)	0.78 (0.76,0.81)
Overall role rating	0.54 (0.51,0.57)	0.74 (0.72,0.77)	0.63 (0.58,0.68)
Partner role	0.96 (0.95,0.96)	0.96 (0.95,0.96)	0.92 (0.91,0.92)
dim. A 'affective relationship'	0.56 (0.52,0.61)	0.81 (0.78,0.83)	0.65 (0.61,0.68)
dim. B 'sexual relationship'	0.60 (0.54,0.65)	0.86 (0.84,0.88)	0.78 (0.75,0.80)
dim. C 'active interest in getting a partner'	0.36 (0.32,0.40)	0.78 (0.75,0.81)	0.36 (0.31,0.41)
Overall role rating	0.44 (0.41,0.47)	0.81 (0.79,0.82)	0.59 (0.47,0.68)
Parental role			
dim. A 'affective relationship'	0.78 (0.75,0.80)	0.86 (0.84,0.87)	0.84 (0.83,0.86)
dim. B 'actual involvement'	0.73 (0.70,0.76)	0.84 (0.82,0.85)	0.81 (0.79,0.82)
Overall role rating	0.73 (0.70,0.76)	0.85 (0.83,0.86)	0.82 (0.80,0.84)
Citizen role			
dim. A 'general interest'	0.59 (0.56,0.62)	0.86 (0.85,0.87)	0.79 (0.77,0.81)
dim. B 'participation in societal groups, organizations and/or clubs'	0.41 (0.37,0.44)	0.40 (0.36,0.45)	0.36 (0.30,0.42)
dim. C 'interest of fellow citizens'	0.63 (0.59,0.67)	0.70 (0.67,0.73)	0.57 (0.51,0.63)
Overall role rating	0.44 (0.41,0.47)	0.77 (0.75,0.78)	0.55 (0.47,0.61)
Social role			
dim. A 'quality of contacts'	0.52 (0.48,0.56)	0.70 (0.67,0.73)	0.56 (0.49,0.61)
dim. B 'frequency and extent of contacts'	0.53 (0.50,0.57)	0.87 (0.86,0.88)	0.75 (0.69,0.80)
Overall role rating	0.57 (0.54,0.60)	0.89 (0.88,0.90)	0.77 (0.73,0.81)
Occupational role	0.80 (0.78,0.81)	0.80 (0.78,0.81)	0.70 (0.67,0.73)
dim. A 'daily routine'	0.50 (0.45,0.55)	0.81 (0.78,0.83)	0.64 (0.29,0.84)
dim. B 'performance'	0.59 (0.54,0.63)	0.93 (0.92,0.93)	0.82 (0.51,0.94)
dim. C 'contacts with others'	0.56 (0.51,0.60)	0.81 (0.79,0.84)	0.69 (0.59,0.77)
dim. D '(other) daily activities'	0.51 (0.48,0.54)	0.89 (0.88,0.90)	0.61 (0.47,0.72)
Overall role rating	0.51 (0.49,0.54)	0.92 (0.91,0.93)	0.68 (0.53,0.79)

Note. The table shows the exact value and, in parentheses, the lower and upper confidential limits using a 95% confidence interval

from this procedure should not be blindly generalized to real research settings. It should also be noted, however, that the patients interviewed on videotape in the EDEN study were speaking a specific east London dialect with strong accents, but were rated by non-native speakers. Thus, language difficulties may have contributed to an underestimation of the inter-rater reliability within the EDEN study.

Finally, when interpreting these findings on the inter-rater reliability for the BPRS 4.0 within the EDEN study, we must bear in mind that researchers from different countries were rating patients from the London site only. Therefore, we do not know whether the results can be generalized to the data collection in the other sites, with patients speaking other languages, living in a range of cultural areas with diverse health care systems. Nevertheless, we do think that the findings indicate that most BPRS 4.0 subscales can be used reliably within cross-national multi-site studies involving unselected groups of psychiatric patients interviewed by experienced and reasonably trained raters.

GSDS-II

Using the original four-point ordinal scale, reliabilities achieved for the GSDS-II appeared to be merely acceptable or even poor with regard to nearly all dimensional ratings and overall ratings. This finding was discouraging as the rating of standardized written vignettes should overestimate the real reliabilities rather than underestimate them. However, ratings were often given based on sparse information with no opportunity to ask clarifying questions. Furthermore, given that GSDS-ratings must be based on norms and expectations dependent upon characteristics of the social-cultural background, cultural differences between the participating sites (Kallert et al., 2004) may additionally have led to an underestimation of the reliabilities. It became apparent that this was a particular problem for the rating of Dimension B ('participation in societal groups, organizations and/ or clubs') of the 'citizen role', and for the assignment to the 'occupational role' categories. For example, if a patient was not gainfully employed, discordances regularly occurred with respect to the assignment of patients to Category II ('people who are engaged in housekeeping') or category III ('people who are unemployed').

Using the dichotomous assessment of social disabilities, nearly all reliabilities proved to be good or at least acceptable, with the exception of the reliabilities for Dimension B of the citizen role and Dimension C of the partner role. This finding corroborates the assumption that reliabilities achieved for the first version of the GSDS (Wiersma et al., 1988) are a good indication of the reliabilities for the GSDS-II (Wiersma et al., 1990). Thus, given that researchers from different countries rated vignettes from patients from different countries, we may reasonably assume that, within cross-national research studies, the GSDS-II does not allow the reliable assessment of the degree or severity of specific social disabilities. It does allow the assessment of the presence or absence of role-specific disabilities, however. Given that the reliability for the GSDS-II sum score proved to be almost good, the GSDS-II also allows the reliable assessment of the global severity of social dysfunction.

In conclusion, the results indicate that, with a brief training session and proper use of the instruments, psychopathology and the presence or absence of social disabilities can be reliably assessed by clinically experienced raters within cross-national research studies. The results, however, also reconfirmed that cross-national studies face specific methodological problems such as inter-rater reliability across languages and cultural contexts. This study highlights that, when assessing social functioning, differences in cultural and social perception might constitute an additional challenge for transcultural mental health services research.

Acknowledgement

The multi-site research project (Acronym: EDEN-study; Web site: www.edenstudy.com) 'Psychiatric day hospital treatment: An alternative to inpatient treatment, being cost-effective and minimising post-treatment needs for care? An evaluative study in European countries with different care systems' is funded by the European Commission (Quality of Life and Management of Living Resources Programme, contract no. QLG4-CT-2000-01700). Additional national grants supporting the project have been provided by Roland-Ernst-Stiftung für Gesundheitswesen and the Faculty of Medicine at the Dresden University of Technology, the NHS Regional Research and Development Office, the Polish National Committee of Scientific Affairs, and the Slovak Ministry of Education. Pfizer Pharmaceutical Company supported travel and accommodation for the EDEN project meetings. We greatly appreciate the helpful advice we received from Dr Joseph Ventura (BPRS 4.0) and Prof. Dr Durk Wiersma (GSDS-II). Special thanks to Charlene M. Reiss for text editing.

References

Alonso J, Ferrer M, Romera B, Vilagut G, Angermeyer M, Bernert S, Brugha TS, Taub N, McColgen Z, de Girolamo G, Polidori G, Mazzi F, de Graaf R, Vollebergh WAM, Buist-Bowman MA, Demyttenaere K, Gasquet I, Haro JM, Palacín C, Autonell J, Katz SJ, Kessler RC, Kovess V, Lépine JP, Arbabzadeh-Bouchez

S, Ormel J, Bruffaerts R. The European study of the epidemiology of mental disorders (ESEMeD/ MHEDEA 2000). Project: rationale and methods. International Journal of Methods in Psychiatric Research 2001; 11: 55–67.

American Pychiatric Association. DSM-IV. Diagnostic and Statistical Manual of Mental Disorders. 4 edn. Washington DC: APA, 1994.

Bartko JJ, Carpenter WT Jr. On the methods and theory of reliability. Journal of Nervous and Mental Disease 1976; 163: 307–17.

Becker T, Knapp M, Knudsen HC, Schene AH, Tansella M, Thornicroft G, Vázquez-Barquero JL, and the Epsilon Study Group. Aims, outcome measures, study sites and patient sample. Epsilon Study 1. British Journal of Psychiatry 2000, 177 (suppl. 39): s1–s7.

Becker T, Magliano L, Priebe S, Salize HJ, Schützwohl M, Kallert T. Evidence-based mental health services research. The contribution of some recent EU-funded projects. In: Kirch W (ed.) Public Health in Europe. 10 Years EUPA. Berlin: Springer, 2004.

Becker T, Vázquez-Barquero JL. The European perspective of psychiatric reform. Acta Psychiatrica Scandinavica 2001, 104 (Suppl. 419): 8–14.

Borkowf CB. A new nonparametric method for variance estimation and confidence interval construction for Spearman's rank correlation. Computational Statistics and Data Analysis 2000; 34: 219–41.

Cicchetti DV, Volkmar F, Sparrow SS, Cohen D, Fermanian J, Rourke BP. Assessing the reliability of clinical scales when the data have both nominal and ordinal features: proposed guidelines for neuropsychological assessment. Journal of Clinical and Experimental Neuropsychology 1992; 14: 673–86.

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960; 20: 37–46.

Edson R, Lavori P, Tracy K, Adler LA, Rotrosen J and the Veterans Affairs Cooperative Study Study Group. Interrater reliability issues in multicenter trials, Part II: Statistical procedures used in Department of Veterans Affairs Cooperative Study #394. Psychopharmacology Bulletin 1997; 33: 59–67.

Hafkenscheid A. Psychometric evaluation of a standardized and expanded Brief Psychiatric Rating Scale. Acta Psychiatrica Scandinavica 1991; 84: 294–300.

Hansson L. Outcome assessment in psychiatric service evaluation. Social Psychiatry and Psychiatric Epidemiology 2001; 36: 244–8.

Hedlund JL, Vieweg BW. The Brief Psychiatric Rating Scale (BPRS): a comprehensive review. J Oper Psychiatry 1980: 11: 48–65.

Jarosz-Nowak J. Modified kappa as an alternative measure of agreement between 2 and more than 2 raters. Wrocław University of Technology, Institute of Mathematics, 2002.

Kallert T, Ganev K, Raboch J, Kastergiou A, Solomon Z, Maj M, Dembinskas A, Kiejna A, Nawka P, Torres-Gonzales F, Kjellin L, Priebe S. Aims and methods of the EUNOMIA-study (European Evaluation of Coercion in Psychiatry and Harmonisation of Best Clinical Practise). Paper presented at the Tenth Annual EUPHA meeting, Dresden, Germany, 2002.

Kallert T, Priebe S, Kiejna A, Nawka P, Raboch J. The European Day Hospital Evaluation (EDEN) Study: An example of EC-funded mental health services research. In Raboch J, Doubek P, Zrzavecká I (eds). Psychiatrie v medicíně a medicína v psychiatrii. Praha: Galen, 2002.

Kallert TW, Priebe S, Schützwohl M, Glöckner M, Briscoe J and the EDEN study group. The role of acute day hospital treatment for mental health care: research context and practical problems of carrying out the international multi-centre EDEN-study. In: Kirch W (ed.) Public Health in Europe. 10 Years EUPA. Berlin: Springer, 2004.

Knudsen HC, Vázquez-Barquero JL, Welcher B, Gaite L, Becker T, Chisholm D, Ruggeri M, Schene AH, Thornicroft G, and the Epsilon Study Group. Translation and cross-cultural adaptation of outcome measurements for schizophrenia. Epsilon Study 2. British Journal of Psychiatry 2002; 177 (suppl. 39): s8–s14.

Leese MN, White IR, Schene AH, Koeter MWJ, Ruggeri M, Gaite L. Reliability in multi-site psychiatric studies. International Journal of Methods in Psychiatric Research 2001; 10: 29–42.

Lukoff D, Liberman RP, Nuechterlein KH. Symptom monitoring in the rehabilitation of schizophrenic patients. Schizophrenia Bulletin 1986; 12: 578–602.

Marshall M, Crowther R, Almarez-Serrano A, Creed F, Sledge W, Kluiter H, Roberts C, Hill E, Wiersma D, Bond G R, Huxley P, Tyrer P. Systematic reviews of effectiveness of day care for people with severe mental disorders. (1) Acute day hospital versus admission; (2) Vocational rehabilitation; (3) Day hospital versus outpatient care. Health Technology Assessment 2001; (5) 21.

Overall JE, Gorham DR. Brief Psychiatric Rating Scale. Psychological Reports 1962; 10: 799–812.

Roncone R, Ventura J, Impallomeni M, Falloon IRH, Morosini PL, Chiaravalle E, Casacchia M. Reliability of an Italian standardized and expanded Brief Psychiatric Rating Scale (BPRS 4.0) in raters with high vs. low clinical experience. Acta Psychiatrica Scandinavica 1999; 100: 229–36.

Shrout PE. Measurement reliability and agreement in psychiatry. Statistical Methods in Medical Research 1998; 7: 301–17.

Ventura J, Green MF, Shaner A, Liberman RP. Training and quality assurance with the Brief Psychiatric Rating Scale: 'The drift busters'. International Journal of Methods in Psychiatric Research 1993; 3: 221–44.

Ventura J, Nuechterlein KH, Subotnik KL, Gutkind D, Gilbert EA. Symptom dimensions in recent-onset schizophrenia and mania: a principal components analysis of the 24-item Brief Psychiatric Rating Scale. Psychiatry Research 2000; 97: 129–35.

Wiersma D, de Jong A, Kraaijkamp HJM, Ormel J. GSDS-II. The Groningen Social Disabilities Schedule, Second Version. University of Groningen, Department of Social Psychiatry, 1990.

Wiersma D, de Jong A, Ormel J. The Groningen Social Disabilities Schedule: Development, relationship with I.C.I.D.H., and psychometric properties. International Journal of Rehabilitation Research 1988; 11: 213–24.

World Health Organization. Schedules for Clinical Assessment in Neuropsychiatry, Version 2.1. Geneva: World Health Organization, 1999.

Correspondence: Dr Matthias Schützwohl, TU Dresden, Universitätsklinikum CG Carus, Klinik und Poliklinik für Psychiatrie und Psychotherapie, Fetscherstr. 74, D – 01307 Dresden, Germany. Email: Matthias.Schuetzwohl@mailbox.tu-dresden.de.