The validity of the German version of the Health of the Nation Outcome Scales (HoNOS-D): a clinician-rating for the differential assessment of the severity of mental disorders

SYLKE ANDREAS,¹ TIMO HARFST,² SVEN RABUNG,¹ ROBERT MESTEL,³ HENNING SCHAUENBURG,⁴ MARIA HAUSBERG,¹ STEPHAN KAWSKI,¹ UWE KOCH¹ & HOLGER SCHULZ¹

1 Department of Medical Psychology, Centre of Psychosocial Medicine, University Medical Centre Hamburg-Eppendorf, Germany

2 Federal Chamber of Psychotherapists (Bundespsychotherapeutenkammer) Berlin, Germany

3 Clinic for Psychosomatic Medicine, Bad Groenenbach, Germany

4 Department of Psychosomatic Medicine, University of Heidelberg, Germany

Key words

HoNOS-D, validity, mental disorders, outcome

Correspondence

Sylke Andreas, Institute and Clinic for Medical Psychology, University Medical Centre, Hamburg-Eppendorf, Centre for Psychosocial Medicine, Martinistr. 52, House W 26, D-20246 Hamburg, Germany.
Telephone +49-(0)40-7410-57705
Fax: +49-(0)40-7410-54940
Email: sandreas@uke.uni-hamburg.de

Received 10 June 2008;
revised 10 April 2009;
accepted 27 July 2009

Abstract

The Health of the Nation Outcome Scales (HoNOS) is an internationally-established clinician-rating instrument for the differential assessment of the severity of patients with mental disorders. The aim of this study was to examine the validity of the German version of the HoNOS (HoNOS-D). Evaluation of validity, including factor validity, convergent and discriminant validity and sensitivity to change, was conducted on a large, virtually representative, clinical sample of patients with mental disorders in inpatient psychotherapy (Study 1, $N = 3169$). Additional assessment of criterion-based validity was completed using another clinical sample of patients with mental disorders (Study 2, $N = 55$). Although factor validity of the HoNOS-D and its total score could not be confirmed as expected, the majority of single items of the HoNOS-D proved to be valid in terms of convergent validity, criterion-based validity and sensitivity to change. Hence, single items, rather than the total score of the HoNOS-D, can be recommended for obtaining a picture of the impairment of patients with mental disorders in the clinical setting. *Copyright © 2010 John Wiley & Sons, Ltd.*

Introduction

A central issue in outcomes research, especially in the area of mental health care, is the lack of a gold standard for the measurement of improvement in health. For example, even if a patient's manic episodes decrease during treatment, the patient will not necessarily experience this change to be beneficial (Slade, 2002). One possible approach to address this issue is to assess outcomes combining two different perspectives: self-reports and clinician-rated measures (Slade, 2002).

The Health of the Nation Outcome Scales (HoNOS, Wing *et al.*, 1998) represents a particularly relevant and internationally-established, clinician-rated instrument for the assessment of severity and outcomes in patients with mental disorders. Today, this instrument is routinely applied in the treatment of patients with mental disorders in many countries (e.g. Denmark, Nova Scotia, Italy and Australia) (Bech *et al.*, 2003; Kisely *et al.*, 2007; Lora *et al.*, 2001; Trauer *et al.*, 1999).

Given its common use in other countries, and the absence of other suitable clinician-rated instruments for the differential assessment of severity, the HoNOS has been translated into German with authorization from the original instrument developers. The translation process and empirical results of the instrument regarding its feasibility and retest-reliability have already been described in a previous study (Andreas *et al.*, 2007). In a study assessing interrater reliability in a sample of 123 patients with mental disorders we found satisfactory coefficients for all of the 12 items after training of the raters (ICCs > 0.6, Andreas *et al.*, 2009; Fleiss and Cohen, 1973).

The aim of the development of the HoNOS was to construct an instrument for routine use assessing the severity of psychosocial problems in 12 independent areas irrespective of the patients' diagnosis (Wing *et al.*, 1998). A differentiated glossary is available for the 12 problem areas providing reference examples of individual ratings (zero to four).

By now, a number of studies have examined the psychometric properties of the HoNOS (Bech *et al.*, 2003; Bech *et al.*, 2006; Gigantesco *et al.*, 2007; Kisely *et al.*, 2007; Lora *et al.*, 2001; Pirkis *et al.*, 2005; Trauer, 1999; Trauer *et al.*, 1999; Wing *et al.*, 1998).

The factor structure of the instrument is considered to be a central to its validity. Wing *et al.* (1998), on the basis of an unreported factor analysis, proposed the following four subscales: (1) subscale 'Behaviour' (item 1 'Overactive, aggressive, disruptive and agitated behaviour', item 2 'Non-accidental self-injury' and item 3 'Problem drinking or drug-taking'), (2) subscale 'Impairment' (item 4

'Cognitive problems' and item 5 'Physical disability'), (3) subscale 'Symptoms' (item 6 'Hallucinations and delusions', item 7 'Depressed mood' and item 8 'Other mental and behavioural problems') and (4) subscale 'Social' (item 9 'Problems with relationships', item 10 'Problems with activities of daily living', item 11 'Problems with living conditions' and item 12 'Problems with occupation and activities').

So far, studies investigating the original factor structure have revealed no satisfactory results; therefore, alternative models have been suggested (Lauzon *et al.*, 2001; Trauer, 1999). The study by Trauer (1999) identified a five-scale solution using an exploratory factor analysis in a sample of 2137 patients from five different psychiatric services in Australia, with over 50% schizophrenia as a main diagnosis. The structure was replicated in a study by Eagar *et al.* (2005) using confirmatory factor analysis. The five-scale solution of the HoNOS contains a 'Social' subscale (items 9, 10–12), a 'Behaviour' scale (items 1 and 3), a 'Depression' scale (items 2, 7–9), an 'Impairment' scale (items 4 and 5) and a 'Hallucinations/delusions' scale consisting of one corresponding item only (item 6).

The aim of the present study was to examine the German Version of the HoNOS (HoNOS-D) with regards to its factor structure, with the intent that empirical findings will contribute to the ongoing debate on psychometric properties of the HoNOS. We sought to identify overall scale structure solutions as well as to assess the validity of single items. In order to assess the applicability of the HoNOS in routine clinical settings, the focus of the evaluation was on criterion-based validity and sensitivity to change.

Method

Design and setting

Examination of the validity of the HoNOS-D was conducted on a large, representative, clinical sample for inpatients in psychotherapy (Study 1). Additional evaluations of validity (e.g. construct validity or differential validity) were conducted in another clinical sample (Study 2).

Study 1: Within the framework of a comprehensive quality assurance programme, samples of consecutive patients ($N = 3169$) were obtained both at the beginning and at the end of inpatient treatment from 19 clinics over a period of 5 years, from June 2002 to July 2007. A total of 3743 patients were included in the study. Of those, 301 patients did not participate in the study for a number of reasons (e.g. refusal) and for 273 patients only incomplete datasets were available. Dropout analyses comparing clinically relevant aspects of study participants ($n = 3169$)

and patients who did not participate in the study ($n = 574$) showed a small number of significant mean differences. The sample of study participants can thus be considered as virtually representative. The evaluation measures (e.g. HoNOS-D) were administered by clinicians following their initial contact with the patients during the first week and the last week of inpatient treatment, just before the patients' discharge. Ecob *et al.* (2004) pointed out that investigations of HoNOS scores must take possible inter-rater variation into account. In order to control for this aspect, the same clinician completed all of the clinician-rated instruments for one specific patient both at admission and discharge. The raters were not specially trained, as evidence suggests that training is not an indispensable precondition for sufficiently reliable ratings of the HoNOS (Rock and Preston, 2001). However, clinicians received detailed instructions containing a list with the most important points to be considered in the use of the HoNOS-D.

Study 2: Secondly, HoNOS-D assessments were carried out in a sample of consecutive patients ($N = 55$) from the Clinic of Psychosomatic Medicine at Bad Groenenbach (Germany) on the basis of video recordings, which had been recorded during the first week of inpatient treatment. Sixty-five inpatients were asked to participate in the study, from whom a total of 10 inpatients were excluded for various reasons (e.g. refusal to participate). There were no significant differences between study participants and study drop-out patients regarding socio-demographic and clinical characteristics (e.g. psychotherapy motivation).

Additionally, DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, fourth edition, American Psychiatric Association, 1994) diagnoses of patients were obtained using standardized and structured clinical interviews [Diagnostic interview for patients with mental disorders – short version, Mini-DIPS (Margraf, 1994) and Structured Clinical Interview for DSM-IV SCID-II (Fydrich *et al.*, 1995)]. These diagnoses were assessed by a clinical experienced diagnostician (RM), who was blinded to the HoNOS-D ratings.

Sample

Study 1: Data was obtained from a total of 3169 patients for both admission and discharge. The patients were 72% female and the mean age was 45 years ($SD = 14$ years; range: 13–96 years). At admission, 51% were in full-time employment, 11% were unemployed and 6% had retired. Forty-four per cent were married and 29% were single. The most frequent primary International Classification of

Diseases (ICD-10) treatment diagnoses (WHO, 1992) were depressive disorders (ICD-10: F32–33, F34.1, 35%), adjustment disorders and neurasthenias (F43.2–9, F48, 16%), anxiety disorders (F40–41, 9%), somatoform disorders (F45, 6%) and personality disorders (F60, 4%). On average, patients exhibited a total of 2.7 ICD-10 diagnoses ($SD = 1.4$). The study sample can be regarded as virtually representative for inpatients in psychotherapy. The average treatment duration in the participating clinics was 43 days ($SD = 17$ days, range: 1–146 days).

Study 2: The second sample consisted of 55 females from a mental health clinic. The mean age was 39 years ($SD = 9$ years, ranging from 19 to 54 years). All participants were married; 21% lived with their first marriage spouse. Forty-five per cent of the patients had 10 years of school education (O-level equivalent). Twenty-nine per cent of participants were in full-time employment, while 29% were on sick leave, unemployed or not working. The average Global Severity Index of the Symptom-Check-List-90-R (SCL-90-R, Derogatis, 1983) was 1.29 ranging from 0.14 to 2.75. DSM-IV diagnoses (American Psychiatric Association, 1994) were assigned on the basis of standardized and structured clinical interviews (Mini-DIPS, Margraf, 1994). The most frequent diagnoses at admission were personality disorders (DSM-IV: 301.9, 301.83, 301.00 and 301.82, 59%) and 36% exhibited borderline personality disorder (DSM-IV: 301.9). Sixteen per cent suffered from affective disorders (DSM-IV: 296.00–296.89, 296.20, 296.3, 296.90–301.13, 296.90), followed by eating disorders (DSM-IV: 307.1–307.51, 9%) and pain disorders (DSM-IV: 307.80, 4%). The remaining six patients were each diagnosed with a substance disorder, agoraphobia, obsessive-compulsive disorder, post-traumatic stress disorder, adjustment disorder and a dissociative disorder.

Instruments

Study 1: In addition to the HoNOS-D, additional self-evaluation instruments such as the short version of the Symptom Checklist (SCL, Harfst *et al.*, 2002), the short version of the Inventory of Interpersonal Problems (IIP, Harfst *et al.*, 2004) and the SF-8 Health Survey (SF-8, Ware *et al.*, 2000) were used. The SF-8-F Health Survey (SF-8-F, Ware *et al.*, 2000) clinician-rated evaluation measures were also administered.

Study 2: In Study 2 alongside the HoNOS-D, the SCL (Derogatis, 1977), the IIP (Horowitz *et al.*, 1988), the Beck Depression Inventory (BDI, Beck *et al.*, 1961), the Global Assessment of Functioning Scale (GAF, Endicott *et al.*, 1976) and the German version of a symptom severity

measurement (degree of impairment score, German: Beeinträchtigungs-Schwere-Score, BSS, Schepank, 1995) were also deployed. While the self-report instruments are widely used and possess good psychometric properties, the psychometric properties of the stated clinician-rated instruments have not yet unequivocally been confirmed.

Statistical analyses

Statistical analyses were performed using SPSS 15.0 (2001) statistical package.

Based on the scale structures used by the test authors (Wing *et al.*, 1998), the factor structure was assessed using confirmatory factor analysis. Due to the skewness of the HoNOS items, analyses were carried out on the basis of polychoric correlation matrices using the computer program EQS (Structural Equations Program, Bentler and Wu, 2004).

Additionally, the factor structure identified by Trauer (1999) and Lauzon *et al.* (2001) was also assessed. The goodness-of-fit of the data was assessed using the fit indices recommended by Hu and Bentler (1999) (standardized-root-mean-square-residual, SRMR, Bentler, 1995) and the root-mean-square-error-of-approximation (RMSEA, Steiger, 1990) as well as the 'Comparative Fit Index' (CFI, Bentler, 1990). The recommended cut-off values are: SRMR ≤ 0.08 (or 0.10) and RMSEA ≤ 0.06 and CFI ~ 0.95 (Hu and Bentler, 1999). The intercorrelation of scales will be used as another measure of the factorial validity.

Convergent and discriminant validity was assessed using correlations (Spearman-Rank correlation) between HoNOS scales and single items, each with the scores of the other scales.

Study 1: short version of the SCL (Harfst *et al.*, 2002), short version of the IIP (Harfst *et al.*, 2004), SF-8 (Schulz *et al.*, 2009; Ware *et al.*, 2000);
Study 2: BDI (Hautzinger *et al.*, 1995), BSS (Schepank, 2002) and GAF (Endicott *et al.*, 1976).

It was expected that the HoNOS scales and single items would individually be substantially correlated with convergent self-assessed scales (with at least a medium effect size $r > 0.3$) (Cohen, 1988) (e.g. item 7 'Depressed mood' with the depression scale of the SCL or item 9 'Problems with relationships' with the total score of the IIP). Furthermore, it is expected that single items of the HoNOS-D will substantially correlate with at least a large effect size ($r > 0.5$) with convergent scales of clinician-rated instruments (e.g. item 5 'Problems with physical disability' with

the physical component score of the SF-8-F). Smaller correlations are expected with divergent scales (e.g. item 1 'Overactive, aggressive, disruptive behaviour' with the physical component score of the SF-8). In order to test the hypotheses average Fisher's- Z correlations were calculated, which resulted in moderate convergent and divergent validity.

In order to assess the criterion-based validity, differences in the severity of individual HoNOS-D items across different DSM-IV diagnosis groups were investigated using the Mann Whitney U test. In order to adjust for the alpha error a Bonferroni-Holm correction (Holm, 1979) was carried out. The diagnoses groups were generated as follows: affective disorders (DSM-IV: 296.20, 296.3, 300.4), substance-related disorders (DSM-IV: 303.00–305.90), borderline personality disorder (DSM-IV: 301.83) and somatoform disorders (DSM-IV: 300.11, 300.81–300.82, 307.42–307.44, 307.80). Cohen's d (Cohen, 1988) was used as a measure of effect size ($0.20 < d < 0.50$ was considered small effect size, $0.50 < d < 0.80$, medium effect size and $d > 0.80$, large effect size). In line with empirical results from previous studies, it was expected that patients with a specific disorder (e.g. affective disorder) would score higher on the index item (e.g. HoNOS-D item 7 for affective disorders) in comparison to patients with no affective disorders.

The pre-post data of the HoNOS from Study 1 was used to investigate sensitivity to change. In order to test for significance the Wilcoxon test was applied. Using the formula from Kazis *et al.* (1989), a standardized pre-post effect size (ES) was calculated in order to determine sensitivity to change: $(M_1 - M_2)/SD_1$ (M_1 = mean at the beginning of treatment, M_2 = mean at the end of treatment, SD_1 = standard deviation at the beginning of treatment). The values of ES were interpreted in line with the recommendations of Cohen (1988) with 0.2, 0.5 and 0.8 being small, medium and large, respectively. For comparing the ES of the HoNOS-D, the ES of the three subscales of the SCL (depression, phobic anxiety and somatisation) and the mental component score of the SF-8 clinician- and self-rating assessments were used.

Results

Descriptive item analyses

The distribution of the HoNOS-D items in Study 1 is displayed in Table 1. A lower item mean score (< 1.00) was found for five of 12 HoNOS-D items (item 1 'Overactive, aggressive, disruptive and agitated behaviour', item 2 'Non-accidental self-injury', item 3 'Problem drinking or drug taking', item 6 'Problems with hallucinations and

Table 1 Distribution of HoNOS items in the first sample ($N = 3169$)

HoNOS items	Mean (SD)	Floor (%)	Ceiling (%)	Skewness	Kurtosis	rit
AGGR	0.86 (1.43)	61.8	3.0	2.53	9.56	0.30
SI	0.30 (1.18)	88.1	1.1	5.36	33.02	0.34
ADD	0.38 (1.27)	84.7	0.9	4.85	27.21	0.28
COG	1.16 (1.35)	47.0	4.9	1.15	2.25	0.28
PhysDis	1.36 (1.42)	42.9	6.8	0.72	0.47	0.12
Delusions	0.17 (0.95)	92.6	0.3	7.62	63.96	0.27
DEP	2.83 (0.95)	3.1	21.9	-0.60	2.48	0.29
Other	3.17 (0.94)	3.2	36.9	-0.93	4.86	0.27
REL	2.46 (1.31)	12.7	17.9	-0.12	1.79	0.28
ACT	1.81 (1.20)	20.6	4.5	0.05	0.53	0.36
LIV	0.76 (1.50)	66.6	1.9	3.11	12.70	0.26
OCC	1.98 (1.60)	26.5	11.0	0.85	2.76	0.27

Note: rit, discriminatory power (corrected); SD, standard deviation; AGGR, item 1 'Aggression'; SI, item 2 'Non-accidental self-injury'; ADD, item 3 'Problem drinking or drug taking'; COG, item 4 'Cognitive problems'; PhysDis, item 5 'Physical illness or disability problems'; delusions, item 6 'Problems associated with hallucinations and delusions'; DEP, item 7 'Depressed mood'; Other, item 8 'Other problems'; REL, item 9 'Problems with relationships'; ACT, item 10 'Problems with activities of daily living'; LIV, item 11 'Problems with living situation'; OCC, item 12 'Problems with occupation and activities'.

delusions' and item 11 'Problems with living situation' whereas item 7 'Depressed mood', item 8 'Other problems' and item 9 'Problems with relationships' possessed the highest item mean coefficients. There are distinct floor effects (i.e. answer category 0 = no problem > 20%) for seven items: item 1 'Overactive, aggressive, disruptive and agitated behaviour', item 2 'Non-accidental self-injury', item 3 'Problem drinking or drug taking', item 4 'Cognitive problems', item 5 'Physical illness or disability problems', item 11 'Problems with living situation', and item 6 'Problems with hallucinations and delusions' (with the highest effects of 93%). In contrast, ceiling effects (i.e. answer category 4 = severe problem > 20%) were only found for item 8 'Other problems' with 37% and for item 7 'Depressed mood' with 22%. Furthermore, the values below 1.0 showed acceptable skewness indices with almost normal distribution for item 5 'Physical illness or disability problems', item 7 'Depressed mood', item 8 'Other problems', item 9 'Problems with relationships', item 10 'Problems with daily activities' and for item 12 'Problems with occupation and activities' (Table 1). The majority of HoNOS-D items are right-skewed distributed (Table 1). Only item 7 'Depressed mood', item 8 'Other problems' and item 9 'Problems with relationships' display a left-skewed distribution. With the exception of item 5 'Physical illness or disability problems', all other items of the HoNOS-D had medium discriminatory power indices ($0.30 < r < 0.50$) (Table 1).

Factorial validity

First an intercorrelation matrix of single items was calculated to form a basis for the calculation of confirmatory factor analysis. Of 66 possible correlations between the 12 HoNOS-D items only five correlations with a medium ES of $r = 0.30$ were found.

Using confirmatory factor analysis based on polychoric correlation matrices the following values emerged in the examination of the original structure (Wing *et al.*, 1998): $\chi^2 = 455.69$ ($df = 48$, $p = 0.001$), RMSEA = 0.12, SRMR = 0.08 and CFI = 0.69. Thus, values of the postulated four-subscale-solution lay outside the cut-off limit recommended by Hu and Bentler (1999), and the CFI = 0.69 does not reach the recommended value of CFI ~0.95.

The examination of the five-scale-solution by Trauer (1999) in Study 1 resulted in the following: $\chi^2 = 1557.93$ ($df = 41$, $p = 0.001$), RMSEA = 0.12, SRMR = 0.07 and CFI = 0.73, whereby all values lay outside the cut-off limit recommended by Hu and Bentler (1999).

Lauzon *et al.* (2001) examined a one-dimensional model of the HoNOS-D using confirmatory factor analysis. In line with the methodological procedure by Lauzon *et al.* (2001) two one-factor-models were examined using Study 1. The first one-dimensional model included items 1 to 10, the second one-dimensional model included items 1 to 12. Items 11 and 12 were excluded from model

examination in the first analysis, as these had shown low reliability in a number of previous studies (Andreas *et al.*, 2007). The one-dimensional model with 10 items presented the following fit indices values: $\chi^2 = 394.11$ ($df = 35$, $p = 0.001$), RMSEA = 0.13, SRMR = 0.09 and CFI = 0.63, all lying outside the recommended limit. The one-dimensional model with 12 items showed a similar picture: $\chi^2 = 549.42$ ($df = 54$, $p = 0.001$), RMSEA = 0.12, SRMR = 0.08 and CFI = 0.64.

The calculation of the factor structure using confirmatory factor analysis of the HoNOS-D showed no satisfactory results. The reliability of the total scores using Cronbach's α , was 0.60 for the HoNOS-D total score across 10 items, and was 0.62 across all 12 items; thus, the total score demonstrated moderate internal consistency. Because results indicated that the 12 single items of the HoNOS-D were relatively independent scales, validity assessments were carried out on the single item level. In addition to the validation of single items we decided to report the total score (12 items) of the HoNOS for contentwise reasons.

Convergent and discriminant validity

As displayed in Tables 2 and 3, three of four assessed convergent correlations between single items of the HoNOS-D and the total score, respectively, with scales of self-rating instruments were confirmed: item 5 'Physical illness or disability problems' correlates with contentwise corresponding scales (SF-8- physical sum scale) at $r = 0.31$ and item 7 'Depressive mood' correlates with two depression scales of the SCL-90 at $r = 0.43$. For the HoNOS total scale the expected correlation was also confirmed ($r = 0.37$) (see Tables 2 and 3). The average convergent correlation of the HoNOS with contentwise corresponding scales of self-rating instruments lies at $r = 0.34$, thereby confirming the convergent validity.

Four convergent correlations of single HoNOS items and the total score, respectively, with scales of clinician-rated instruments were assessed. Thereby correlations were confirmed for the total score ($r = 0.54$) and for item 5 'Physical problems' with corresponding scales (SF-8-physical total scale and the BSS physical total scale) ($r = 0.60$). Unexpectedly no correlations were found for item 9 'Problems with relationships' and the corresponding BSS scale social communication and for item 10 'Problems with daily activities' with the GAF scale (see Table 3). The average convergent correlation of the HoNOS with corresponding scales of clinician-rated instruments is $r = 0.44$, thereby not confirming the convergent validity.

The validity of the German version of the HoNOS-D

The assessed divergent correlations of single HoNOS-D items with self- and clinician-rated, non-corresponding scales are expectedly low ($r = 0.02$ – 0.21), thereby confirming the divergent validity.

Other significant interactions between single items of the HoNOS-D and scales of self-report instruments are shown in Tables 2 and 3.

Criterion-based validity

Study 2 investigated the extent to which index items of the HoNOS-D discriminate between patients with and without a specific diagnosis using Mann–Whitney U tests. As expected, the largest significant difference was found for the values of item 7 'Depressed mood' between patients with an affective disorder and without affective disorder with at least medium ES ($d = 0.65$) as displayed in Figure 1 (Mann–Whitney U test, $p < 0.05$). As expected the largest significant difference in patients with substance-related disorders and without substance-related disorders was in item 3 'Problem drinking or drug taking' ($d = 1.30$, Mann–Whitney U test, $p < 0.001$). As can be seen in Figure 1, patients with somatoform disorders differ from patients without somatoform disorders on item 5 'Physical illness or disability problems' with the largest significant difference as expected ($d = 0.93$, Mann–Whitney U test, $p < 0.001$). Furthermore, there is a significant difference for item 2 'Non-accidental self-injury' ($d = 1.00$, Mann–Whitney U test, $p < 0.001$) between patients with BPS and no BPS.

Sensitivity to change

With regard to sensitivity to change between the two time points of admission and discharge, at least small effects ($d > 0.20$) can be found for nine of 12 items of the HoNOS-D, and at least medium to large ES values can be found for four of 12 items (item 7 'Depressed mood', item 8 'Other problems', item 9 'Problems with relationships' and item 10 'Problems with daily activities') (Table 4). No effects were found for item 2 'Non-accidental self-injury', item 6 'Problems with hallucinations and delusions' and item 11 'Problems with living situations', most likely because these items had low values at the beginning of treatment (Table 4).

The ES of item 5 'Physical illness or disability problems' was lower than the ES for the corresponding physical component scores of the SF-8-S and SF-8-F (self- and clinician-rating) whereas the ES of item 7 'Depressed mood' was higher than the ES of the corresponding depression scale of the SCL.

Table 2 Spearman's rank correlation coefficients with confidence intervals (95% CI) between the 12 scales of the HoNOS and scales of other clinician-rating and self-report instruments in Study 1 ($N_{\min} = 2950$)

HoNOS-D	Self-rated						Clinician-rated		
	SCL			SF8-S			IIP		
	DEP	SOM	PHOB	GSI	SF8-MCS	SF8-PCS	GS	SF8-MCS	SF8-F
AGGR	0.09 (0.05–0.12)	0.04 (0.00–0.08)	0.02 (–0.02–0.06)	0.07 (0.03–0.10)	0.08 (0.05–0.12)	0.02 (–0.01–0.06)	0.06 (0.03–0.10)	0.15 (0.01–0.29)	0.14 (0.00–0.28)
SI	0.19 (0.16–0.23)	0.00 (–0.04–0.03)	0.11 (0.07–0.14)	0.16 (0.12–0.19)	0.11 (0.08–0.15)	–0.02 (–0.05–0.02)	0.14 (0.10–0.17)	0.20 (0.07–0.33)	–0.04 (–0.18–0.10)
ADD	0.08 (0.04–0.11)	–0.02 (–0.05–0.02)	0.02 (–0.02–0.05)	0.05 (0.01–0.09)	0.03 (0.00–0.07)	–0.03 (–0.07–0.01)	0.04 (0.00–0.08)	–0.04 (–0.18–0.11)	–0.08 (–0.22–0.06)
COG	0.12 (0.08–0.15)	0.00 (0.01–0.08)	0.09 (0.06–0.13)	0.11 (0.08–0.15)	0.11 (0.07–0.14)	0.04 (0.00–0.07)	0.08 (0.05–0.12)	0.37 (0.19–0.52)	0.21 (0.07–0.34)
PhysDis	–0.04 (–0.08–0.01)	0.25 (0.21–0.28)	0.08 (0.04–0.11)	0.08 (0.04–0.11)	0.03 (0.00–0.07)	0.31 (0.28–0.34)	–0.05 (–0.09–0.02)	0.14 (–0.01–0.27)	0.50 (0.38–0.60)
Delusions	0.05 (0.01–0.08)	–0.02 (–0.06–0.01)	0.05 (0.01–0.08)	0.03 (0.00–0.07)	0.04 (0.01–0.08)	0.00 (–0.04–0.03)	0.03 (0.00–0.07)	0.08 (–0.06–0.22)	0.00 (–0.14–0.14)
DEP	0.37 (0.34–0.40)	0.12 (0.09–0.16)	0.15 (0.11–0.18)	0.31 (0.28–0.34)	0.30 (0.27–0.34)	0.15 (0.11–0.18)	0.23 (0.19–0.26)	0.52 (0.40–0.61)	0.27 (0.14–0.40)
Other	0.17 (0.14–0.21)	0.09 (0.06–0.13)	0.12 (0.08–0.15)	0.18 (0.14–0.21)	0.16 (0.12–0.19)	0.08 (0.05–0.12)	0.10 (0.06–0.13)	0.32 (0.18–0.44)	0.27 (0.13–0.40)
REL	0.28 (0.25–0.32)	0.01 (–0.03–0.04)	0.06 (0.03–0.10)	0.20 (0.16–0.23)	0.19 (0.16–0.22)	0.00 (–0.03–0.04)	0.23 (0.19–0.26)	0.36 (0.23–0.47)	0.06 (–0.08–0.20)
ACT	0.26 (0.23–0.30)	0.15 (0.11–0.18)	0.20 (0.17–0.24)	0.28 (0.25–0.31)	0.25 (0.22–0.29)	0.19 (0.15–0.22)	0.18 (0.14–0.21)	0.52 (0.41–0.62)	0.31 (0.17–0.43)
LIV	0.19 (0.16–0.22)	0.01 (–0.02–0.05)	0.08 (0.04–0.11)	0.15 (0.11–0.18)	0.12 (0.09–0.16)	0.01 (–0.02–0.05)	0.10 (0.07–0.14)	0.06 (–0.09–0.20)	0.02 (–0.13–0.16)
OCC	0.13 (0.10–0.17)	0.03 (0.00–0.07)	0.03 (0.00–0.07)	0.10 (0.07–0.14)	0.16 (0.12–0.19)	0.05 (0.02–0.09)	0.10 (0.06–0.13)	0.17 (0.03–0.31)	0.00 (–0.15–0.14)
Tot12	0.33 (0.30–0.36)	0.13 (0.10–0.17)	0.16 (0.12–0.20)	0.30 (0.26–0.33)	0.28 (0.24–0.31)	0.17 (0.13–0.20)	0.21 (0.18–0.25)	0.54 (0.39–0.67)	0.35 (0.21–0.47)

Note: $r > 0.30$ for self-report scales and $r > 0.50$ for clinician-rated scales; bold, convergent validity; italic, divergent validity; cv, average convergent correlation; dr, average divergent correlation; S, self-rated; C, clinician-rated; AGGR, item 1 'Aggression'; SI, item 2 'Non-accidental self-injury'; ADD, item 3 'Problem drinking or drug taking'; COG, item 4 'Cognitive problems'; PhysDis, item 5 'Physical illness or disability problems'; delusions, item 6 'Problems associated with hallucinations and delusions'; DEP, item 7 'Depressed mood'; Other, item 8 'Other problems'; REL, item 9 'Problems with relationships'; ACT, item 10 'Problems with activities of daily living'; LIV, item 11 'Problems with living situation'; OCC, item 12 'Problems with occupation and activities'; Tot12, total score (12 items) of HoNOS; SCL, Symptom-Check-List-90-R; SCL-DEP, depression scale of the SCL; SCL-GSI, Global Severity Index of the SCL; SF-8-F, Short-Form-8 clinician-rating; SF-8-S, Short-Form-8 self-report; SF8-MCS, Short-Form-8-Mental Component Score; SF8-PCS, Short-Form-8-Physical Component Score; IIP, Inventory of Interpersonal Problems; IIP-GS, Global Score of the IIP; BSS, Beeinträchtigungsschwerescore (impairment severity score); BSS-Ment, mental score; BSS-Phys, physical score; BSS-SOC, social-communicative score; BSS-GS, global score of the impairment severity score; BDI, Beck Depression Inventory; BDI-GS, global score of the BDI; GAF, Global Assessment of Functioning Scale; GAF-GS, global score of the GAF.

Table 3 Spearman's rank correlation coefficients with confidence intervals (95% CI) between the 12 scales of the HoNOS and scales of other clinician-rated and self-report instruments in Study 2 ($N_{\min} = 48$)

HoNOS-D	Self-rated						Clinician-rated					
	SCL						BSS					
	DEP	SOM	PHOB	GSI	BDI	IIP	Phys	Soc	Ment	GS	GAF	dr
AGGR	0.20 (-0.07-0.44)	0.31 (0.05-0.53)	0.28 (0.02-0.51)	0.28 (0.01-0.51)	0.26 (-0.01-0.49)	0.22 (-0.04-0.46)	0.33 (0.06-0.55)	0.26 (-0.01-0.49)	0.32 (0.05-0.54)	0.44 (0.20-0.63)	-0.36 (-0.58-(-0.10))	0.16 _s , 0.15 _c
SI	0.39 (0.13-0.60)	0.30 (0.03-0.53)	0.16 (-0.12-0.42)	0.41 (0.16-0.61)	0.55 (0.33-0.72)	0.42 (0.16-0.62)	0.27 (-0.01-0.50)	0.34 (0.08-0.57)	0.41 (0.15-0.62)	0.46 (0.21-0.65)	-0.33 (-0.56-(-0.05))	0.16 _s
ADD	0.34 (0.08-0.56)	0.33 (0.07-0.55)	0.42 (0.17-0.62)	0.45 (0.21-0.64)	0.35 (0.09-0.57)	0.25 (-0.02-0.49)	0.41 (0.16-0.61)	0.32 (0.06-0.54)	0.37 (0.11-0.58)	0.55 (0.32-0.71)	-0.26 (-0.50-0.01)	0.21 _s
COG	0.15 (-0.12-0.40)	0.30 (0.04-0.53)	0.17 (-0.10-0.42)	0.36 (0.11-0.57)	0.24 (-0.03-0.47)	0.35 (0.09-0.56)	0.13 (-0.15-0.38)	0.31 (0.05-0.53)	0.23 (-0.05-0.46)	0.30 (0.04-0.53)	-0.21 (-0.46-0.07)	0.17 _s
PhysDis	-0.13 (-0.39-0.14)	0.34 (0.08-0.55)	0.17 (-0.10-0.41)	0.00 (-0.27-0.26)	0.07 (-0.19-0.33)	-0.14 (-0.39-0.13)	0.68 (0.50-0.80)	-0.09 (-0.35-0.19)	0.22 (-0.05-0.46)	0.50 (0.26-0.67)	0.07 (-0.21-0.33)	0.31 _s , 0.60 _c
Delusions	0.26 (0.00-0.49)	0.27 (0.00-0.50)	0.18 (-0.09-0.42)	0.35 (0.10-0.57)	0.28 (0.01-0.50)	0.37 (0.11-0.58)	0.30 (0.03-0.52)	0.35 (0.09-0.56)	0.37 (0.12-0.58)	0.48 (0.25-0.66)	-0.36 (-0.58-(-0.10))	0.21 _s
DEP	0.39 (0.13-0.59)	0.28 (0.01-0.50)	0.21 (-0.06-0.45)	0.38 (0.12-0.58)	0.53 (0.31-0.70)	0.19 (-0.08-0.43)	0.07 (-0.20-0.33)	0.11 (-0.17-0.36)	0.43 (0.19-0.63)	0.30 (0.04-0.53)	-0.09 (-0.36-0.19)	0.43 _s , 0.19 _s
Other	0.32 (0.05-0.54)	0.24 (-0.03-0.48)	0.28 (0.02-0.51)	0.37 (0.11-0.58)	0.48 (0.24-0.66)	0.27 (0.00-0.50)	0.45 (0.20-0.64)	0.08 (-0.20-0.34)	0.40 (0.14-0.60)	0.50 (0.26-0.68)	-0.18 (-0.44-0.10)	0.23 _s , 0.14 _s
REL	0.15 (-0.13-0.40)	0.34 (0.08-0.55)	0.15 (-0.12-0.40)	0.35 (0.09-0.56)	0.25 (-0.02-0.48)	0.24 (-0.03-0.48)	0.24 (-0.04-0.48)	0.29 (0.02-0.52)	0.16 (-0.12-0.41)	0.30 (0.03-0.53)	-0.25 (-0.49-0.03)	0.29 _c , 0.15 _c
ACT	-0.18 (-0.44-0.11)	0.12 (-0.17-0.39)	-0.03 (-0.31-0.26)	-0.07 (-0.35-0.21)	0.16 (-0.13-0.42)	0.06 (-0.23-0.34)	0.17 (-0.12-0.44)	0.08 (-0.21-0.36)	0.12 (-0.18-0.39)	0.15 (-0.14-0.42)	-0.29 (-0.54-(-0.00))	-0.29 _c , 0.12 _s
LIV	0.24 (-0.04-0.49)	0.20 (-0.09-0.46)	0.07 (-0.22-0.34)	0.22 (-0.06-0.48)	0.24 (-0.05-0.49)	0.34 (0.07-0.57)	0.13 (-0.16-0.40)	0.40 (0.13-0.62)	0.25 (-0.04-0.50)	0.35 (0.07-0.58)	-0.33 (-0.57-(-0.04))	0.23 _s , 0.14 _s
OCC	0.03 (-0.29-0.35)	0.20 (-0.13-0.49)	0.21 (-0.12-0.49)	0.23 (-0.10-0.51)	0.19 (-0.13-0.48)	0.10 (-0.22-0.41)	0.28 (-0.05-0.55)	0.15 (-0.18-0.45)	0.35 (0.03-0.61)	0.36 (0.05-0.62)	-0.16 (-0.46-0.17)	0.29 _c , 0.15 _c
Tot12	0.38 (0.12-0.60)	0.54 (0.31-0.71)	0.42 (0.16-0.62)	0.55 (0.32-0.72)	0.57 (0.34-0.73)	0.43 (0.17-0.63)	0.60 (0.38-0.75)	0.38 (0.11-0.60)	0.57 (0.35-0.74)	0.75 (0.59-0.85)	-0.42 (-0.63-(-0.15))	0.37 _s , 0.54 _c

Note: $r > 0.30$ for self-report scales and $r > 0.50$ for clinician-rated scales; bold, convergent validity; italic, divergent validity; cv, average convergent correlation; dr, average divergent correlation; S, self-rated; C, clinician-rated; AGGR, item 1 'Aggression'; SI, item 2 'Non-accidental self-injury'; ADD, item 3 'Problem drinking or drug taking'; COG, item 4 'Cognitive problems'; PhysDis, item 5 'Physical illness or disability problems'; Delusions, item 6 'Problems associated with hallucinations and delusions'; DEP, item 7 'Depressed mood'; Other, item 8 'Other problems'; REL, item 9 'Problems with relationships'; ACT, item 10 'Problems with activities of daily living'; LIV, item 11 'Problems with living situation'; OCC, item 12 'Problems with occupation and activities'; Tot12, total score (12 items) of HoNOS; SCL, Symptom-Check-List-90-R; SCL-DEP, depression scale of the SCL; SCL-GSI, Global Severity Index of the SCL; SF-8-F, Short-Form-8 self-report; SF-8-MCS, Short-Form-8-Mental Component Score; SF-8-PCS, Short-Form-8-Physical Component Score; IIP, Inventory of Interpersonal Problems; IIP-GS, Global Score of the IIP; BSS, Beeinträchtigungs-Schwerescore (impairment severity score); BSS-Ment, mental score; BSS-Phys, physical score; BSS-SOC, social-communicative score; BSS-GS, global score of the impairment severity score; BDI, Beck Depression Inventory; BDI-GS, global score of the BDI; GAF, Global Assessment of Functioning Scale; GAF-GS, global score of the GAF.

Figure 1 Mean values of the HoNOS-D items for the different diagnosis groups in study 2 ($n_{\min} = 10$).

Analyses with $n < 10$ were excluded from figures; AGGR = Item 1 'Aggression'; SI, Item 2 'Non-accidental self-injury'; ADD, Item 3 'Problem drinking or drug taking'; COG, Item 4 'Cognitive problems'; PhysDis, Item 5 'Physical illness or disability problems'; delusions, Item 6 'Problems associated with hallucinations and delusions'; DEP, Item 7 'Depressed mood'; Other, Item 8 'Other problems'; REL, Item 9 'Problems with relationships'; ACT, Item 10 'Problems with activities of daily living'; LIV, Item 11 'Problems with living situation'; OCC, Item 12 'Problems with occupation and activities'; * $p < 0.05$; ** $p < 0.01$ and *** $p < 0.001$.

Discussion

It was the aim of the present study to examine the validity of HoNOS-D on a large clinical, representative sample for the first time (Study 1, $N = 3169$), and to assess additional aspects of validity in a second study based on video recordings of structured, clinical interviews ($N = 55$). Differing distribution patterns emerged for the 12 HoNOS-D items. While item 7 'Depressed mood', item 8 'Other problems', item 9 'Problems with relationships' and item 10 'Problems with daily activities' were satisfactorily distributed, several other items, particularly item 6 'Problems with hallucinations and delusions', showed distinct ceiling phenomena. These results are in line with the fact that the sample contained virtually no patients with disorders corresponding to the items. Furthermore, it can be assumed that the distribution of a number of items (e.g.

item 3 'Problem drinking or drug taking') depend on the extent to which patients with substance-related disorders have been included in the sample (Audin *et al.*, 2001).

Consistent with findings from other studies (Lauzon *et al.*, 2001; McClelland *et al.*, 2000; Trauer, 1999), when the factor structure of the HoNOS (Wing *et al.*, 1998) was examined, the present German sample was unable to confirm the factor validity of the HoNOS via the formation of subscales or a total score. Alternative models by Trauer (1999) and Lauzon *et al.* (2001) on the factor structure of the HoNOS were also not confirmed. An analysis of item intercorrelations showed that only five of 66 possible correlations showed at least medium values. Hence, the internal consistency of the total score did not yield satisfactory results, which again is consistent with results of other studies (Lauzon *et al.*, 2001; Orrell *et al.*, 1999; Page *et al.*, 2001).

Table 4 Sensitivity to change of the HoNOS-D items and the SCL and SF-8 (self-report and clinician-rating)

	Pre-treatment M_1 (SD ₁)	Post-treatment M_2 (SD ₂)	p^*		ES
<i>HoNOS-D</i>					
AGGR	0.76 (1.16)	0.37 (0.76)	0.001		0.35
SI	0.20 (0.72)	0.10 (0.44)	0.001		0.15
ADD	0.25 (0.76)	0.09 (0.44)	0.001		0.21
COG	1.13 (1.27)	0.69 (1.00)	0.001		0.35
PhysDis	1.33 (1.37)	0.90 (1.14)	0.001		0.32
Delusions	0.09 (0.47)	0.05 (0.35)	0.001		0.08
DEP	2.83 (0.93)	1.57 (1.00)	0.001		1.34
Other	3.17 (0.90)	2.14 (0.96)	0.001		1.14
REL	2.44 (1.23)	1.81 (1.16)	0.001		0.50
ACT	1.80 (1.18)	0.94 (1.02)	0.001		0.73
LIV	0.62 (1.07)	0.46 (0.91)	0.001		0.15
OCC	1.89 (1.39)	1.34 (1.24)	0.001		0.40
Tot12	16.99 (6.31)	10.86 (6.02)	0.001		0.97
<i>SCL</i>					
DEP	1.77 (1.04)	1.10 (0.95)	43.24	3168	0.65
PHOB	0.53 (0.79)	0.31 (0.61)	22.50	3168	0.29
SOM	1.23 (0.99)	0.84 (0.89)	27.72	3168	0.39
GSI	1.26 (0.74)	0.80 (0.69)	44.03	3168	0.63
<i>SF-8</i>					
MCS-S	2.23 (0.88)	1.47 (0.93)	47.73	3168	0.86
MCS-F	2.60 (0.66)	1.36 (0.73)	88.26	3044	1.88
PCS-S	1.86 (0.92)	1.30 (0.88)	39.69	3168	0.60
PCS-F	1.91 (0.81)	1.06 (0.68)	66.66	3037	1.04

Note: AGGR, item 1 'Aggression'; SI, item 2 'Non-accidental self-injury'; ADD, item 3 'Problem drinking or drug taking'; COG, item 4 'Cognitive problems'; PhysDIS, item 5 'Physical illness or disability problems'; delusions, item 6 'Problems associated with hallucinations and delusions'; DEP, item 7 'Depressed mood'; Other, item 8 'Other problems'; REL, item 9 'Problems with relationships'; ACT, item 10 'Problems with activities of daily living'; LIV, item 11 'Problems with living situation'; OCC, item 12 'Problems with occupation and activities'; Tot12, total score (12 items) of HoNOS; SCL, Symptom-Check-List-90-R; SCL-DEP, depression scale of the SCL; SCL-PHOB, phobia scale; SCL-SOM, somatoform scale; SCL-GSI, Global Severity Index of the SCL; SF-8, Short-Form-8; SF8-MCS-S, Short-Form-8-Mental Component Score-self rating; SF8-PCS-S, Short-Form-8-Physical Component Score-self rating; SF8-MCS-F, Short-Form-8-Mental Component Score-clinician rating; SF8-PCS-F, Short-Form-8-Physical Component Score-clinician-rating; ES, Cohen's d ($d < 0.20$ = small effect, $d = 0.20$ – 0.50 medium effect and $d > 0.80$ large effect).

*Wilcoxon rank sum tests; all t -values are significant ($p < 0.001$).

On the basis of these findings, additional assessments of the differential validity, criterion-based validity and sensitivity to change of the HoNOS-D were carried out on single item level and additionally with the total score, whereby HoNOS-D items suitable for the identification of patients with mental disorders in inpatient treatment were identified. To our knowledge, this is the first study to comprehensively evaluate the validity of single items of the HoNOS.

The convergent validity of the single items of the HoNOS-D was demonstrated for the items where

convergent instruments had been applied (e.g. HoNOS item 7 'Depressed mood' and depression scale of the SCL or HoNOS item 5 'Physical illness or disability problems' and the physical component score of the SF-8-S/F and of the BSS). The range of the correlations was also in line with empirical findings of other studies (Adams *et al.*, 2000; Brooks, 2000; Page *et al.*, 2001), and Study 2 even exceeded prior findings. However, it needs to be pointed out that empirical findings of previous studies were based on analyses of the total score.

Unexpectedly, no significant correlations with at least medium effect size were found between the HoNOS item 9 'Problems with relationships' and the self-report total score of the IIP. One reason for this may be that item 9 refers to difficulties with maintaining interpersonal relationships in general, whereas the IIP asks for specific interpersonal behaviours (e.g. too dominant, too insecure). Contrary to empirical findings of other studies (Browne *et al.*, 2000; McClelland *et al.*, 2000; Parker *et al.*, 2002), this study was able to show that none of the 12 HoNOS-D items correlated with the clinician-rating of the GAF at sufficiently large effect size. This could be due to the fact that the information required for the assessment of social functioning was only partly provided in the video-based recording of the standardized and structured interviews in Study 2.

Overall, criterion-based validity was confirmed for individual index items of the corresponding mental disorders. For example, it was confirmed that item 7 'Depressed mood' discriminates between the presence and absence of depression in patients with affective disorders. This result was also found for the index item 3 'Problem drinking or drug taking' in patients with and without substance-related disorders. Furthermore, it needs to be pointed out that, as expected, further items of the HoNOS-D (e.g. item 8 'Other problems' or item 2 'Non-accidental self-injury') also differentiate between patients with and without substance-related disorders, as these patients often exhibit comorbid disorders and multi-faceted problem areas (Ilgen *et al.*, 2005).

The majority of HoNOS-D items and the total score proved to be moderate to sufficiently sensitive to change. Items 7, 8, 9 and 10 showed medium to large values of ES within the range of $ES = 0.50$ – 1.34 . Only three out of 12 items showed no change during the course of treatment (item 2, item 6 and item 11). Yet, item 2 'Non-accidental self-injury' and item 6 'Problems with hallucinations and delusions' can be regarded as being of minor relevance in inpatient psychotherapeutic treatment of the investigated patients with mental disorders due to their distribution patterns. Item 11 'Problems with living situation' refers to the assessment of the living situation, thus, if at all, only a minor change in this item over the course of inpatient treatment is expected. Future examinations are required in additional studies.

Some limitations of the study need to be pointed out. First, the presented results are derived from a sample where patients with psychotic disorders (e.g. schizophrenia, F2-diagnosis according to ICD-10) are under-represented; therefore, future studies need to examine validity aspects of the HoNOS-D and particularly of

important single items (e.g. item 6 'Problems with hallucinations and delusions') with more representation of such patients. Second, this study partly lacks specific external criteria for the validation of single items (e.g. Brief Psychiatric Rating Scale, BPRS, Overall and Gorham, 1976), hence additional studies are required. A further limitation poses the fact that although raters in Study 1 were provided with structured material for the assessment, they did not receive specific training. A following study will investigate whether a HoNOS-D training for raters is necessary. Third, it needs to be pointed out that another limitation of this study is the small sample size of Study 2. Future studies with larger samples including patients with diagnosis based on standardized and structured interviews according to DSM-IV are required for the validity assessment of the HoNOS-D.

Practical implications

The results of the present study show that single items of the HoNOS-D can be appropriately applied to get a picture of the problems in patients with mental disorders. Despite the fact that the factor structure of the HoNOS could not be confirmed in the present study, the majority of examined individual items have proven to be applicable to differentially assess the degree of mental problems and the severity in patients with mental disorders in inpatient treatment. Additionally the total score offers a global measure of impairment in a number of different problem areas. For the patients with mental disorders in inpatient treatment, 4 of 12 items appeared particularly suited to differentially assess sensitivity to change (item 7 'Depressed mood', item 8 'Other problems', item 9 'Problems with relationships' and item 10 'Problems with daily activities'). In addition, an earlier study confirmed that the HoNOS-D is a practical instrument with good inter-rater and re-test reliability (Andreas *et al.*, 2007). Therefore, single items and the total score of the HoNOS-D can be recommended for clinical use to evaluate the impairments of patients with mental disorders and for future research including outcome monitoring.

Acknowledgements

The authors would like to thank all participants of the study. This study was funded by the German Research Foundation (GZ: AN 382/3-1). The authors would also like to thank the anonymous reviewers for their very valuable comments.

Declaration of interest statement

The authors have no competing interests.

References

- Adams M., Palmer A., O'Brien J.T., Crook W. (2000) Health of the nation outcome scales for psychiatry: Are they valid? *Journal of Mental Health*, **9**, 193–198.
- American Psychiatric Association. (1994) *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition, American Psychiatric Association.
- Andreas S., Harfst T., Dirmaier J., Kowski S., Koch U., Schulz H. (2007) A psychometric evaluation of the German version of the 'Health of the Nation Outcome Scales, HoNOS-D': On the feasibility and reliability of a clinician-rated measure of severity in patients with mental disorders. *Psychopathology*, **40**, 116–125.
- Andreas S., Rabung S., Horstmann S., Schauenburg H., Harris-Hedder K., Schwenk W., Koch U., Schulz H. (2009) Ist eine Schulung von Ratern zur differenzierten Beurteilung des Schweregrades bei Patienten mit psychischen Störungen in der Routineversorgung notwendig? [Training of mental health care professionals in routine outcome assessment – necessary or dispensable?] *Psychologische Medizin*, **20**, 117.
- Audin K., Marginson F.R., Clark J.M., Barkham M. (2001) Value of HoNOS in assessing patient change in NHS psychotherapy and psychological treatment services. *The British Journal of Psychiatry*, **178**, 561–566.
- Bech P., Bille J., Schutze T., Sondergaard S., Wiese M., Waerst S. (2003) Health of the Nation Outcome Scales (HoNOS): Implementability, subscale structure and responsiveness in the daily psychiatric hospital routine over the first 18 months. *Nordic Journal of Psychiatry*, **57**, 285–290.
- Bech P., Bille J., Waerst S., Wiese M., Borberg L., Treufeldt P., Kessing L. (2006) Validity of HoNOS in identifying frequently hospitalized patients with ICD-10 mental disorders. *Acta Psychiatrica Scandinavica*, **113**, 485–491.
- Beck A.T., Ward C.H., Mendelson M., Mock J.E., Erbaugh J.T. (1961) An inventory for measuring depression. *Archives of General Psychiatry*, **4**, 561–571.
- Bentler P.M. (1990) Comparative fit indexes in structural models. *Psychological Bulletin*, **107**, 238–246.
- Bentler P.M. (1995) *EQS: A Structural Equations Program Manual*, Multivariate Software, Inc.
- Bentler P.M., Wu E. (2004) *EQS 6.1 for Windows*, Multivariate Software Inc.
- Brooks R. (2000) The reliability and validity of the Health of the Nation Outcome Scales: Validation in relation to patient derived measures. *Australian and New Zealand Journal of Psychiatry*, **34**, 504–511.
- Browne S., Doran M., McGauran S. (2000) Health of the Nation Outcome Scales (HoNOS): Use in an Irish psychiatric outpatient population. *Irish Journal of Psychological Medicine*, **17**, 17–19.
- Cohen J. (1988) *Statistical Power and Analysis for the Behavioral Sciences*, second edition, Lawrence.
- Derogatis L. (1983) *SCL-90-R, Administration, Scoring & Procedures Manual for the R(evised) Version*, Clin Psychometric Research.
- Derogatis L. (ed.). (1977) *SCL-90-R, Administration, Scoring & Procedures Manual-I for the R(evised) Version*, John Hopkins University School Medicine.
- Eagar K., Trauer T., Mellsop G. (2005) Performance of routine outcome measures in adult mental health care. *Australian and New Zealand Journal of Psychiatry*, **39**, 713–718.
- Ecob R., Croudace T.J., White I.R., Evans J.E., Harrison G.L., Sharp D., Jones P.B. (2004) Multilevel investigation of variation in HoNOS ratings by mental health professionals: a naturalistic study of consecutive referrals. *International Journal of Methods in Psychiatric Research*, **13**, 152–164.
- Endicott J., Spitzer R.L., Fleiss J.L., Cohen J. (1976) The Global Assessment Scale: A procedure for measuring overall severity of psychiatry disturbance. *Archives of General Psychiatry*, **33**, 766–771.
- Fleiss J.L., Cohen J. (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, **33**, 613–619.
- Fydrich T., Schmitz B., Bodem M. (1995) *SKID-II. Strukturiertes Klinisches Interview für DSM-IV, Achse II [Structured Clinical Interview for DSM-IV Personality Disorders, (SCID-II) German version]*, Beltz-Test.
- Gigantesco A., Picardi A., de Girolamo G., Morosini P. (2007) Discriminant ability and criterion validity of the HoNOS in Italian psychiatric residential facilities. *Psychopathology*, **40**, 111–115.
- Harfst T., Dirmaier J., Koch U., Schulz H. (2004) Entwicklung und psychometrische Überprüfung einer multidimensionalen Kurzform des Inventars zur Erfassung interpersonaler Probleme [Development and evaluation of a short version of the Inventory of Interpersonal Problems]. *DRV-Schriften*, **52**, 53–55.
- Harfst T., Koch U., Kurtz von Aschoff C., Nutzinger D.O., Rüddel H., Schulz H. (2002). Entwicklung und Validierung einer Kurzform der Symptom Checklist-90-R [Development and evaluation of a short version of the Symptom Checklist-90-R]. *DRV-Schriften*, **33**, 71–73.
- Hautzinger M., Bailer M., Worrall H., Keller F. (1995) *Das Beck Depressionsinventar (BDI) [Beck Depression Inventory – German version]*, Huber.
- Holm S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.
- Horowitz L.M., Rosenberg S.E., Baer B.A., Ureño G., Villaseñor V.S. (1988) Inventory of Interpersonal Problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology*, **56**, 885–892.

- Hu L., Bentler P.M. (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, **6**, 1–55.
- Ilgen M.A., Quyen T., Finney J.W., Harris A.H. (2005) Recent suicide attempt and the effectiveness of inpatient and outpatient substance use disorder treatment. *Alcoholism, Clinical and Experimental Research*, **29**, 1664–1671.
- Kazis L.E., Anderson J.J., Meenan R.F. (1989) Effect sizes for interpreting changes in health status. *Medical Care*, **27**(3 Suppl), S178–S189.
- Kisely S., Campbell L.A., Crossman D., Gleich S., Campbell J. (2007) Are the Health of the Nation Outcome Scales a valid and practical instrument to measure outcomes in North America? A three-site evaluation across Nova Scotia. *Community Mental Health Journal*, **43**, 91–107.
- Lauzon S., Corbiere M., Bonin J.P., Bonsack C., Lesage A.D., Ricard N. (2001) [Validation of the French version of the Health of the Nation Outcome Scales (HoNOS-F)]. *Canadian Journal of Psychiatry*, **46**, 841–846.
- Lora A., Bai G., Bianchi S., Bolongaro G., Civeni G., Erlicher A., Maresca G., Monzani E., Panetta B., von Morgen B., Rossi F., Torri V., Morosini P. (2001) The Italian version of HoNOS (Health of the Nation Outcome Scales), a scale for evaluating the outcome and severity in mental health services. *Epidemiologia e psichiatria sociale*, **10**, 198–212.
- Margraf J. (1994) *Mini-DIPS: Diagnostisches Kurz-Interview bei psychischen Störungen* [Mini-DIPS: Short Diagnostic Interview for Mental Disorders], Springer.
- McClelland R., Trimble P., Fox M.L., Stevenson M.R., Bell B. (2000) Validation of an outcome scale for use in adult psychiatric practice. *Quality in Health Care*, **9**, 98–105.
- Orrell M., Yard P., Handysides J., Schapira R. (1999) Validity and reliability of the Health of the Nation Outcome Scales in psychiatric patients in the community. *British Journal of Psychiatry*, **174**, 409–412.
- Overall J.E., Gorham D.R. (1976) BPRS. Brief Psychiatric Rating Scale. In: *ECDEU Assessment Manual for Psychopharmacology* (ed. Guy W), pp. 157–169, National Institute of Mental Health.
- Page A.C., Hooke G.R., Rutherford E.M. (2001) Measuring mental health outcomes in a private psychiatric clinic: Health of the Nation Outcome Scales and Medical Outcomes Short Form SF-36. *Australian and New Zealand Journal of Psychiatry*, **35**, 377–381.
- Parker G., O'Donnell M., Hadzi-Pavlovic D., Proberts M. (2002) Assessing outcome in community mental health patients: A comparative analysis of measures. *International Journal of Social Psychiatry*, **48**, 11–19.
- Pirkis J.E., Burgess P.M., Kirk P.K., Dodson S., Coombs T., Williamson M.K. (2005) A review of the psychometric properties of the Health of the Nation Outcome Scales (HoNOS) family of measures. *Health and Quality of Life Outcomes*, **3**, 1–12.
- Rock D., Preston N. (2001) HoNOS: Is there any point in training clinicians? *Journal of Psychiatric and Mental Health Nursing*, **8**, 405–409.
- Schepank H. (1995) *Der Beeinträchtigungs-Schwere-Score (BSS)* [Degree of Impairment Score – German version], Beltz Test.
- Schepank H. (2002) BSS: Der Beeinträchtigungs-Schwere-Score [Degree of impairment score – German version]. In: *Diagnostische Verfahren in der Psychotherapie* (eds Brähler E, Strauß B), pp. 68–72, Hogrefe-Verlag.
- Schulz H., Harfst T., Andreas S., Kowski S., Koch U., Rabung S. (2009) Zusammenhang zwischen Patienten- und Therapeuteinschätzungen der Ergebnisqualität in der Rehabilitation von Patienten mit psychischen/psychosomatischen Erkrankungen am Beispiel des SF-8. [Relationship between patient and clinician rated outcome measures of patients with mental disorders]. *Rehabilitation*, **48**, 270–276.
- Slade M. (2002) Routine outcome assessment in mental health services. *Psychological Medicine*, **32**, 1339–1343.
- SPSS. (2001) *SPSS Version 15.0 for Windows*, SPSS.
- Steiger J.H. (1990) Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, **25**, 173–180.
- Trauer T. (1999) The subscale structure of the Health of the Nation Outcome Scales (HoNOS). *Journal of Mental Health*, **8**, 499–509.
- Trauer T., Callaly T., Hantz P., Little J., Shields R., Smith J. (1999) Health of the Nation Outcome Scales: Results of the Victorian field trial. *British Journal of Psychiatry*, **174**, 380–388.
- Ware J.E., Kosinski M., Dewey J.E., Gandek B. (2000) *How to Score and Interpret Single-item Health Status Measures: A Manual for Users of the SF-8 Health Survey*, Quality Metric.
- World Health Organization (WHO). (1992) *The ICD-10 Classification of Mental and Behavioural Disorders. Clinical Descriptions and Diagnostic Guidelines*, WHO.
- Wing J.K., Beevor A.S., Curtis R.H., Park S.B., Hadden S., Burns A. (1998) Health of the Nation Outcome Scales (HoNOS): Research and development. *British Journal of Psychiatry*, **172**, 11–18.