Adm Policy Ment Health (2007) 34:504–512
DOI 10.1007/s10488-007-0134-y

ORIGINAL PAPER

Inter-rater Reliability of Clinician-rated Outcome Measures in Child and Adolescent Mental Health Services

Ketil Hanssen-Bauer · Odd O. Aalen · Torleif Ruud · Sonja Heyerdahl

Published online: 11 September 2007

Abstract This study investigated the inter-rater reliability when 169 out of 171 clinicians working in 10 Norwegian child and adolescent mental health services rated 20 written vignettes using the following outcome measures: Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA), Children's Global Assessment Scale (CGAS) and Global Assessment of Psychosocial Disability (GAPD). Three clinicians rated both patients and vignettes. On vignettes the intraclass correlation coefficient (ICC) for the HoNOSCA total score was 0.81 (single scales 0.47–0.96), for the CGAS 0.61 and for the GAPD 0.60. The reliability was not lower on patients. The rater's profession, experience or clinic did not have effect on the scores.

Keywords Children · Adolescent · Mental health services · Reliability · Outcome

K. Hanssen-Bauer (✉) · S. Heyerdahl
Centre for Child and Adolescent Mental Health, Eastern and Southern Norway, P.O. Box 4623 Nydalen, NO-0405 Oslo, Norway
e-mail: ketil.hanssen-bauer@r-bup.no

O. O. Aalen
Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway

T. Ruud
Division of Mental Health Services, Akershus University Hospital, Lørenskog, Norway

T. Ruud
Faculty Division Akershus University Hospital, Faculty of Medicine, University of Oslo, Oslo, Norway

Introduction

There has been an increased focus on outcome of child and adolescent mental health services (CAMHS) (Kazdin 2003; Williams and Kerfoot 2005; Epstein et al. 2005), and measuring outcome as part of routine clinical practice is recommended (Weiss 1998). Information about different domains (Hoagwood et al. 1996; Jensen et al. 1996; Fonagy 2002) and from different perspectives are relevant (Wolpert et al. 2005). The referred child's symptom severity and level of functioning scored by clinical staff is considered a valuable part of comprehensive outcome measurement systems (National Mental Health Strategy 2004; MH-SMART 2007; CAMHS Outcome Research Consortium 2007).

There are different broad staff rated measures of symptoms and level of functioning in use as routine in CAMHS, both multidimensional and unidimensional measures. In the United States and Canada the Child and Adolescent Functional Assessment Scale (CAFAS) developed by Hodges (Hodges and Gust 1995; Hodges et al. 1998; Hodges and Wotring 2004) is widely used. In several states in the US and in one province of Canada CAFAS is mandatory as a routine outcome measure (Bates 2001; CAFAS in Ontario 2005). This multidimensional measure of level of functioning was developed for the Fort Bragg study (Bickman et al. 1995). The second version of the CAFAS consists of eight scales for rating the child/youth: School/work, Home, Community (reflect delinquent acts), Behavior towards others, Moods/emotions, Self-harmful behavior, Substance use, and Thinking. It is also possible to score the caregiver's resources on two optional scales: Material needs and Family/social support. All scales are rated on a four point scale (severe, moderate, mild or minimal/no impairment) which gives a scored profile and a summed total score.

In Australia, New Zealand and Denmark, another and not very different multidimensional scale, the Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA) is established as a mandatory routine outcome measure in CAMHS (National Mental Health Strategy 2004; Australian Mental Health Outcomes and Classification Network 2005; MH-SMART 2007; Bilenberg, 2003). It is also widely used in the United Kingdom (Johnston and Gowers 2005; Ford et al. 2006) where it was developed as part of a family of broad scales to measure health and social functioning (Gowers et al. 2000). The HoNOSCA consists of 13 clinical scales and two scales on the caregiver's need for information and knowledge giving both a clinical profile and a total severity score (Gowers et al. 1999a; Gowers et al. 1999b; Yates et al. 1999; Garralda et al. 2000; Brann et al. 2001; Bilenberg 2003; Pirkis et al. 2005).

Unidimensional scales of level of functioning (Schorre and Vandvik 2004; Winters et al. 2005) are also widely used as routine clinical outcome measures. The Global Assessment of Functioning (GAF) Scale constitute the Axis V of the DSM-IV-TR (American Psychiatric Association 2000) and is used for both children, adolescents and adults. The Children's Global Assessment Scale (CGAS) and the Global Assessment of Psychosocial Disability (GAPD) are specific for the age group below 18 and widely used in CAMHS. The CGAS (Shaffer et al. 1983) is investigated in several studies, well established and widely used. The GAPD was introduced a decade ago by the World Health Organization (WHO) as Axis Six in their Multiaxial Classification of Child and Adolescent Psychiatric Disorders (World Health Organization 1996), based on ICD-10 (World Health Organization 1994). GAPD is a poorly evaluated measure (Schorre and Vandvik 2004).

Inter-rater reliability (IRR) or agreement is often reported between motivated and well-trained raters in research settings. Results from such studies should not be generalized to clinical settings where patient population and rater motivation and competence can be different (Vatnaland et al. 2007). Since routine outcome measurement is implemented in CAMHS in several countries with large clinical databases to be analyzed, the psychometric properties in ordinary clinical settings are essential.

There are several reports on the IRR of CGAS in clinical settings indicating fair (we categorize Intraclass Correlation Coefficients (ICC) as Shrout (1998) suggested) to substantial reliability and one report on the GAPD indicating moderate reliability (Green et al. 1994; Rey et al. 1995; Dyrborg et al. 2000; Schorre and Vandvik 2004). There is one report of the IRR of HoNOSCA in a clinical setting and this indicates fair reliability of the total score (Brann et al. 2001).

The aim of this study was to investigate the inter-rater reliability (IRR) when clinical staff in CAMHS use the HoNOSCA, GAPD and CGAS. We also investigated whether ratings were related to clinic, profession or clinical experience; if clinicians agree more after written feedback on their ratings and if IRR differs when rating patients compared to written vignettes.

Method

Sample of Raters

To get a broad sample of clinicians, 10 out of a total of 66 Norwegian CAMHS treating outpatients were recruited through personal contact and poster announcements. The clinic could join the project if they had multi-professional teams and agreed to let all professionals participate. No clinicians had used the HoNOSCA previously, some had used the CGAS and all routinely used the GAPD. From December 2002 until June 2003 the clinicians rated written vignettes (which took 4–6 weeks to complete for each clinic).

In the 10 CAMHS, 169 out of 171 clinicians rated ten written vignettes each. The two missing came both from the same clinic. The number of raters from each clinic was between 8 and 28 (mean 16.7, SD 6.7). Clinicians were not included if on leave of absence or sick leave at the time of rating. Raters included 55 psychologists (31 clinical specialists), 24 physicians (11 child and adolescent psychiatrists), 37 social workers and 53 with other bachelor degrees. The professions' distribution was not different from national data on full time equivalent (FTE) at Dec. 31st 2002 (Pearson χ^2 test, $P = 0.56$). (It was not possible to correct for our sample counting number of persons and the national data counting FTE.) At that time employed professionals in outpatient CAMHS in Norway amounted to 1153 FTE (Sitter and Hagen 2004).

To investigate if written feedback improves IRR, clinicians at two of the 10 clinics rated another set of 10 vignettes after they had received a personal letter with detailed feedback on their first ratings according to a consensus-based standard. Thirty-nine out of the 45 clinicians at the chosen clinics, who had rated the first time, completed this second rating. They used the same instruments as the first time.

To investigate if IRR between raters is different in a patient-design than a vignette-design, three clinicians from another clinic with 33 employed clinicians first assessed patients and later written vignettes. They used the HoNOSCA and the GAPD, not the CGAS. The raters were one male and two females, one psychologist, one educational therapist and one social worker.

Sample of Written Vignettes and Patients

Twenty written vignettes, each one page long (number of words: mean 446, SD 71), were developed for this study. They were based on anonymous clinical descriptions from experienced clinicians working in CAMHS in different countries. All clinical descriptions where changed to make them untraceable to and unrecognizable by any original patient. Symptoms and problems from different main parts of DSM-IV and ICD-10 (chapter V) where included with a normal distribution of severity measured by the CGAS. The vignettes are on average more severe (Table 1) than cases from studies of outpatients (Yates et al. 1999; Gowers et al. 1999b) since we included a wide range of problems and symptoms in a restricted number of cases.

Each vignette had a heading specifying age and a made up name, followed by an introduction describing the reason for referral to an outpatient CAMHS and some background information. The main text was presented with the subheadings "Symptoms and behavioral problems", "Social problems", "Developmental disorders and somatic problems" and "Lack of information and knowledge". The text was in common clinical words, and did not contain descriptions specific for HoNOSCA, GAPD or CGAS. A variety of problems and symptoms were included. The 20 vignettes were divided in two groups of ten. We used the CGAS scores to check that each group of vignettes was normally distributed and contained a variation in severity of problems and symptoms. The patients in the vignette sample were from 4–17 years old (mean age 10.6 years, SD 4.26, 10 boys and 10 girls).

The three clinicians, who rated both patients and written vignettes, rated a sample of 30 consecutively referred patients. Each clinician rated 20 patients and each patient was rated by two clinicians. The patients were 4.1–12.5 years old (mean age 8.84, SD 2.24), 19 (63.3%) boys and 11 (36.7%) girls.

The study was approved by the Regional Committee for Medical Research Ethics, Southern Norway, and the Norwegian Data Inspectorate.

Outcome Measures

The Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA) (Gowers et al. 1999a) was developed in United Kingdom to measure mental health and outcome in clinical settings. The HoNOSCA focuses on clinically significant problems and symptoms and consists of 15 scales, each rated from 0 (no problem) to 4 (severe to very severe problem). The first 13 scales are summarized to a total score indicating severity of mental health problems. The HoNOSCA was translated from English to Norwegian by the first author (K.H-B.) and a reference group of four professionals commented on drafts. A complete Norwegian version was independently back-translated into English by a bilingual psychologist and compared to the original English version by Simon Gowers in the HoNOSCA Project (HoNOSCA Project 2003). The translation was accepted following minor changes of the Norwegian text. HoNOSCA has been evaluated in several studies and found to be easy to use, reliable, valid and sensitive to change (Gowers et al. 1999b; Brann et al. 2001; Garralda et al. 2000; Yates et al. 1999; Bilenberg 2003).

The Children's Global Assessment scale (CGAS) is a well-known unidimensional scale measuring global functioning, rated from 1 (lowest functioning) to 100 (excellent functioning). CGAS has been evaluated in several studies and is widely used to assess outcome (Shaffer et al. 1983; Rey et al. 1995).

The World Health Organization has introduced the Global Assessment of Psychosocial Disability (GAPD) as Axis Six in their Multiaxial Classification of Child and Adolescent Psychiatric Disorders, based on ICD-10 (World Health Organization 1996). The GAPD is rated from 8 (lowest functioning) to 0 (excellent functioning). Schorre and Vandvik (2004) reviewed the research literature and found only one study evaluating GAPD (Dyrborg et al. 2000).

Procedures

The HoNOSCA, the CGAS and the GAPD were implemented in 10 Norwegian outpatient CAMHS and training given to all clinicians working there. They started to use the instruments as part of a larger project to validate the HoNOSCA.

Training

The first author gave standardized training to all participating raters. One hour focused on the CGAS and the GAPD, providing general information about the scales and their use followed by a discussion on how to rate five short written vignettes. The HoNOSCA training lasted 2½ h as the scales are more complex. The training included general information about the HoNOSCA and its use, and the clinicians rated and discussed two clinical cases presented as video taped reconstructions of clinical interviews used as training materials available from the HoNOSCA Project (HoNOSCA Project 2003). The interviews were in English and a written translation to Norwegian was provided.

Rating Written Vignettes

To reduce their work load, each clinician rated 10 of the 20 vignettes. The clinicians were randomly allocated by a computer program (SPSS 11.5, http://www.spss.com/) to either use the CGAS or the GAPD and either rate vignette no 1–10 (85 clinicians of whom 40 used the CGAS and 45 the GAPD) or vignette no 11–20 (84 clinicians of whom 38 used the CGAS and 46 the GAPD). There was an exception to the randomization in one clinic where the clinic leader did not want the clinicians to use the CGAS as the GAPD was already routinely being employed. Those 14 clinicians used the GAPD. This resulted in slightly uneven numbers of raters using the CGAS (78 clinicians) and the GAPD (91 clinicians). All 169 clinicians used the HoNOSCA (15 scales). The vignette order was systematically changed in eight different ways in each vignette group to avoid systematic effects from experience and fatigue. For all three measures the instruction was to rate the previous 2-week period.

Feedback on Ratings

To evaluate if IRR improves after feedback on individual ratings we asked two of the 10 clinics to rate a second time using different vignettes. They received feedback on their first ratings and 39 of the 45 clinicians (86.7%) rated the other 10 vignettes 4–6 months later, giving a sub-sample of 39 (23.1%) out of the total of 169 clinicians. The feedback was given as a table showing how their ratings differed from a consensus-based standard, including mean and standard deviation of the differences on every scale. The consensus-based standard was made by six professionals involved in introducing the HoNOSCA as a routine clinical outcome measure. Some of the clinicians used the rating scales in their clinical work between the vignette ratings. They rated the other group of vignettes using the same instruments as the first time (the HoNOSCA, and either the CGAS or the GAPD).

Rating Patients

Three clinicians from another clinic rated 20 patients each as part of their intake procedure. In the first consultation two clinicians met both the patient and his/her parent(s). At the next consultation one of the two clinicians talked with the parents and the other with the patient separately. Then, the two clinicians met and shared information before the third meeting, where everyone participated. The clinicians subsequently rated the case independently. They did not receive any feedback or discuss the ratings of the patients. Afterwards they all rated the 20 written vignettes.

Data Analyses

All statistical analyses were conducted using the SPSS 13.0 for Windows (http://www.spss.com/), except bootstrap analysis which were conducted using the NLME package in the program R for Windows (www.r-project.org).

To analyze inter-rater reliability we computed the intraclass correlation coefficient (ICC) from variance components (Dunn 1989) using the General Linear Model based on restricted maximum likelihood. Both raters and vignettes were considered as random factors and we used an absolute ICC-model (Shrout and Fleiss 1979) with variance component from raters contributing to the total error variance in the denominator of ICC. Vignette group and the rater's gender were built into the model as covariates. We used Shrout's standards for reliability results (virtually none: 0.00–0.10; slight: 0.11–0.40; fair: 0.41–0.60; moderate: 0.61–0.80; and substantial 0.81–1.0) (Shrout 1998).

To investigate whether profession (psychologist, medical doctor, social worker or other bachelor level professions), experience (specialist or not) or the rater's workplace had a systematic effect on ratings, we computed the mean score for the HoNOSCA total score, the CGAS and the GAPD across the vignettes for every rater. These values were used as dependent variables in univariate ANOVA. Since vignette group (raters were randomized to rate one of two fixed vignette groups) had a significant effect on the mean HoNOSCA total score, vignette group was included in the model as a covariate, together with the clinician's gender.

Standard error (SE) of measurement was computed as the square root of the sum of the error variance components. A clinician's rating of a vignette will with a probability of 95% be within 1.96 × the SE of measurement of that vignette's "true" rating defined as the mean rating given by many raters (Anastasi and Urbina 1997).

To test for statistical significance of difference between the reliability (ICC-values) of two measures used by same raters and the same vignettes we conducted an analysis of 1,000 bootstrap samples to take care of the dependency (Efron and Tibshirani 1993).

SE of measurement is the most appropriate index to compare the IRR in a vignette design with the IRR in a patient design, since patient variability can be different in the two samples and this index is independent of patient variability (Anastasi and Urbina 1997).

Results

The HoNOSCA total score (sum of the first 13 scales), the CGAS and the GAPD were strongly correlated for the sample of the 20 written vignettes. Using the mean of all raters the Pearson r for the HoNOSCA total score/CGAS was 0.84; for the HoNOSCA total score/GAPD 0.82; and for the CGAS/GAPD 0.97. For the sample of the 30 patients the Pearson r for the HoNOSCA total score and the GAPD were 0.44, when using mean of the two raters as the score.

Overall Inter-rater Reliability on Vignette Ratings

The overall inter-rater reliability (IRR) results for the vignette ratings are presented in Fig. 1 and Table 1. The ICC was 0.81 for the HoNOSCA total score, indicating substantial reliability (Shrout 1998), 0.61 for the CGAS and 0.60 for the GAPD indicating moderate/fair reliability. The ICC was significantly higher (bootstrap analysis) for the HoNOSCA total score than for both the CGAS ($P < 0.001$) and the GAPD ($P = 0.001$). For the 15 HoNOSCA scales the reliability was substantial for four scales, moderate for six scales and fair for five scales. The least reliable scales were "Physical illness or disability problem" (Scale 6, ICC = 0.47) and "Lack of information about treatment" (Scale 15, ICC = 0.48).

The associated SE of measurement was 3.51 for the HoNOSCA total score, 9.92 for the CGAS and 1.09 for the GAPD. These results indicate that for a mean outpatient (Gowers et al. 1999b; Yates et al. 1999) with a true score of 11 on the HoNOSCA total score, there is a 95% probability that the actual clinical rating will be between 4 and 18, and 68% probability that it will be between 7.5 and 14.5.

The raters' profession, experience or which clinic they worked at did not significantly influence the mean score of the vignettes for any of the three measures (univariate ANOVA). The effects on the HoNOSCA total score from profession was $F(3, 153) = 1.20$ ($P = 0.31$), from experience $F(1, 153) = 0.18$ ($P = 0.67$) and from clinic $F(9, 153) = 1.91$ ($P = 0.055$). The effects on the CGAS score from profession was $F(3, 63) = 1.47$ ($P = 0.23$), from experience $F(1, 63) = 1.67$ ($P = 0.20$) and from clinic $F(8, 63) = 1.30$ ($P = 0.26$). The effects on the GAPD score from profession was $F(3, 75) = 0.50$ ($P = 0.69$), from experience $F(1, 75) = 0.89$ ($P = 0.35$) and from clinic $F(9, 75) = 1.70$ ($P = 0.11$).

The group of clinicians who repeated vignette ratings after feedback on their first ratings had very similar IRRs on both occasions (Fig. 2).

Inter-rater Reliability on Vignette Ratings versus Patients

Three clinicians first rated 30 patients (20 patients each) after intake assessment and later rated the 20 written vignettes. Their results are shown in Table 2. The SE of measurement for HoNOSCA total score was 2.59 with

Table 1 Inter-rater reliability for HoNOSCA, CGAS and GAPD (rating of 20 vignettes)

		n (raters)	Mean	SE of measurement	ICC	95% CI for ICC
HoNOSCA scales						
1	Disruptive, aggressive problem	169	1.51	0.62	0.82	0.73–0.92
2	Overactive, attention difficulty	169	2.11	0.87	0.69	0.55–0.83
3	Self injury	169	0.66	0.41	0.90	0.84–0.96
4	Alcohol, drug misuse	169	0.32	0.20	0.96	0.94–0.99
5	Scholastic or language skills problem	169	1.90	0.99	0.60	0.44–0.76
6	Physical illness, disability problem	169	0.77	0.90	0.47	0.31–0.64
7	Hallucinations, delusions	169	0.87	0.82	0.70	0.56–0.84
8	Psychosomatic problem	169	1.07	0.92	0.59	0.43–0.75
9	Emotional symptom	169	2.39	0.91	0.63	0.48–0.78
10	Peer relationship problem	169	2.65	0.68	0.74	0.61–0.86
11	Self care, independence problem	169	1.62	0.93	0.62	0.47–0.78
12	Family problem	169	1.39	0.84	0.60	0.44–0.76
13	Poor school attendance	169	1.23	0.49	0.91	0.86–0.97
14	Lack of knowledge about difficulties	169	2.32	0.86	0.66	0.51–0.81
15	Lack of information about treatment	169	2.23	1.05	0.48	0.32–0.65
HoNOSCA total score (*sum scale 1–13*)		169	18.34	3.51	0.81	0.70–0.91
CGAS		78	44.90	9.92	0.61	0.45–0.77
GAPD		91	4.53	1.09	0.60	0.44–0.76

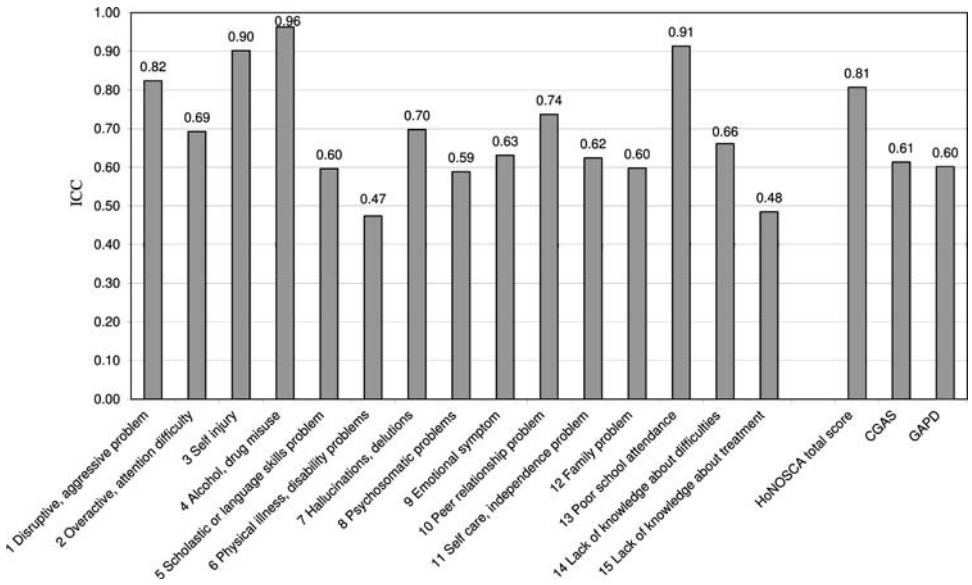

Fig. 1 Inter-rater reliability (ICC) from Table 1 for HoNOSCA items, HoNOSCA total score, CGAS and GAPD

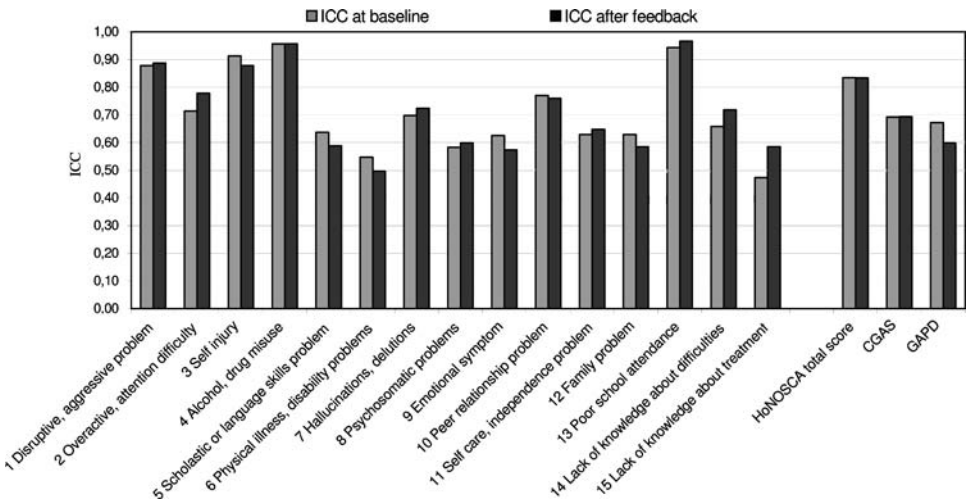

Fig. 2 Change of inter-rater reliability (ICC) after feedback

patients and 2.93 with vignettes and the SE of measurement for the GAPD was 0.68 with patients and 0.89 with vignettes. The SE of measurement was lower in the patient design for 10 of the 15 scales of HoNOSCA.

Discussion

We examined and compared the IRR of different outcome measures (the HoNOSCA, the CGAS and the GAPD) when rated by a large sample of clinicians at several CAMHS using a design with written vignettes. Our results support the HoNOSCA total score as a measure with substantial agreement (measured by ICC) and the CGAS and the GAPD with moderate/fair agreement between clinicians in routine clinical settings both across the actual professional disciplines, professional experiences and clinics. We found better IRR (ICC values) for the HoNOSCA total score than for the CGAS and the GAPD. The IRR of the single scales of the HoNOSCA ranged from ICC 0.47 (physical illness, disability problem) to ICC 0.96 (alcohol, drug misuse).

We found CGAS to be less reliable (ICC 0.61) than Shaffer et al. (1983) reported (ICC 0.84). They studied five second-year child psychiatry fellows who rated 19 written case vignettes. Our results were closer to what Rey et al. (1995) found (ICC 0.53 for inpatients and 0.63 for outpatients) when they studied 20 experienced clinicians with different professional background who rated 162 children and adolescents.

Dyrborg et al. (2000) investigated both the CGAS and the GAPD in a clinical setting with different subgroups of patients and raters. Each rater used both instruments. A group of five raters (three experienced clinicians and two trainees), scored 28 patients each from case notes and achieved an ICC of 0.59 on the CGAS, and an ICC of 0.74 on the GAPD. They found higher ICC-values for a

Table 2 Inter-rater reliability for HoNOSCA, CGAS and GAPD (ratings of 30 patients versus 20 vignettes by three raters)

		Patients (n = 30)				Vignettes (n = 20)			
		Mean	SE of measurement	ICC	95% CI for ICC	Mean	SE of measurement	ICC	95% CI for ICC
HoNOSCA scales									
1	Disruptive, aggressive problem	2.17	0.55	0.81	0.69–0.93	1.73	0.59	0.88	0.78–0.97
2	Overactive, attention difficulty	2.77	0.63	0.70	0.52–0.89	2.42	0.43	0.91	0.85–0.98
3	Self injury	0.33	0.26	0.90	0.83–0.97	0.87	0.34	0.94	0.90–0.99
4	Alcohol, drug misuse	0.00	0.00	1.00	–	0.37	0.18	0.97	0.95–0.99
5	Scholastic or language skills problem	2.40	0.52	0.79	0.65–0.92	2.07	0.73	0.76	0.61–0.92
6	Physical illness, disability problem	0.57	0.68	0.58	0.34–0.82	0.90	0.98	0.40	0.12–0.68
7	Hallucinations, delusions	0.67	0.67	0.66	0.45–0.87	0.83	0.72	0.75	0.60–0.91
8	Psychosomatic problem	1.30	0.60	0.77	0.63–0.92	1.38	0.89	0.65	0.45–0.86
9	Emotional symptom	2.25	0.85	0.49	0.22–0.76	2.50	0.85	0.67	0.47–0.87
10	Peer relationship problem	2.67	0.75	0.24	0.00–0.56	2.68	0.65	0.78	0.64–0.93
11	Self care, independence problem	1.58	0.81	0.54	0.29–0.79	1.90	1.01	0.55	0.39–0.71
12	Family problem	1.23	0.41	0.89	0.81–0.96	1.15	0.52	0.84	0.74–0.95
13	Poor school attendance	0.73	0.00	1.00	–	1.25	0.22	0.98	0.97–1.00
14	Lack of knowledge about difficulties	1.71	0.84	0.39	0.08–0.70	2.50	0.65	0.79	0.63–0.96
15	Lack of information about treatment	1.98	0.88	0.46	0.18–0.75	2.40	0.87	0.62	0.39–0.85
HoNOSCA total score (*sum scale 1–13*)		18.67	2.59	0.74	0.57–0.90	20.05	2.93	0.88	0.80–0.97
GAPD		4.13	0.68	0.55	0.30–0.80	4.53	0.89	0.71	0.53–0.89

subgroup of the three more experienced raters using both measures on 95 cases (test of differences not reported; CGAS 0.86 and GAPD 0.88).

Different researchers have evaluated the IRR of the HoNOSCA. Gowers et al. (1999b) studied three unspecified raters who rated 20 cases simultaneously. They found an ICC greater than 0.80 for eight out of the 12 main scales for which an ICC value could be computed. They did not report the IRR of the total score. Garralda et al. (2000) established IRR for 15 unspecified case vignettes assessed by three unspecified raters achieving ICC = 0.42–0.82 for different subgroups of scales, not reporting on the total score.

These studies had different weaknesses. The studies of Gowers et al. and Garralda et al. were both restricted to three raters, making it difficult to generalize to clinicians in general. Brann et al.'s study was restricted to three case vignettes making it difficult to generalize to a population of patients in clinics. None of them report confidence intervals for their ICC.

Considering the confidence intervals in our study for the ICC of the CGAS, the GAPD and the HoNOSCA total score, our results do not differ from the other investigations of these measures in clinical settings (Rey et al. 1995; Dyrborg et al. 2000; Brann et al. 2001). However, we found somewhat different ICC compared to Brann et al. on some of the HoNOSCA scales, which can be due to different number of vignettes studied.

Our results on SE of measurement are adding information to the results on ICC by indicating the uncertainty of a rating given to an individual patient. About 95% of ratings for that patient would be within $\pm 1.96 \times$ SE of measurement from the "true" rating, i.e., the mean rating given by a large number of raters. We found that the SE of measurement for the CGAS was 9.9 while Shaffer et al. (1983) found 8.6.

Our study extends the reliability studies of the HoNOSCA, the CGAS and the GAPD further into clinical settings by including a larger and more clinical representative sample of raters with all clinicians at several outpatient CAMHS rating 20 different vignettes. We do not know studies examining the IRR of the HoNOSCA, the CGAS and the GAPD on such a large and representative sample of clinicians.

We did not find any effect on IRR from the detailed feedback on ratings, which we expected would improve the IRR. This could be due to the quality or the appropriateness of the feedback or to the design which was an uncontrolled intervention design.

This study is based on information in written vignettes. It may be less variance in the information accessible to the raters when reading written vignettes than when interviewing patients in a clinical situation. To address whether IRR may be lower in clinical settings than we found in the vignette study, a group of three clinicians first rated real patients and then rated the written vignettes using the

HoNOSCA and the GAPD. We did not find reduced agreement when they rated real patients than when they rated the vignettes. Since only three clinicians participated in rating patients it is more difficult to generalize these results on written vignette versus real patients to a larger population of clinicians. However, this is in accordance with the results from a recent study by Peabody et al. (2004) who found that vignettes are a valid tool for measuring the quality of clinical practice by physicians as also commented in an editorial (Norcini 2004).

Although the vignettes were chosen as being representative of children seen in CAMHS they are a selection and therefore may not fully capture the wealth of clinical complexity seen in clinics giving another limitation in the vignette design.

Further studies of additional aspects of reliability, validity and sensitivity for change of these outcome measures are needed.

Clinical Implications

In ordinary CAMHS the HoNOSCA total score can be used with substantial and the CGAS and the GAPD with moderate/fair inter-rater reliability. The HoNOSCA total score had higher inter-rater reliability than the CGAS and the GAPD. To evaluate outcome of ordinary clinical practice reliability issues should be considered when choosing outcome measures and analyzing such data in CAMHS. In clinical practice the data from these measures should be interpreted cautiously for an individual patient.

Limitations

To make such a large study feasible we could not randomly select clinics and we used written case vignettes as proxy for patients.

Acknowledgements The study was financially supported by The Research Council of Norway, and by Centre for Child and Adolescent Mental Health, Eastern and Southern Norway. The authors thank all the raters. *Disclosure*: The authors report no conflicts of interests.

References

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., Text Revision ed.). Washington, DC: American Psychiatric Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall International.

Australian Mental Health Outcomes and Classification Network (2005). *Child & Adolescent National Outcomes & Casemix Collection Standard Reports*, Version 1.1. Brisbane, Queensland.

Bates, M. P. (2001). The child and adolescent functional assessment scale (CAFAS): Review and current status. *Clinical Child & Family Psychology Review, 4*(1), 63–84.

Bickman, L., Guthrie, P. R., Foster, E. M., Lambert, E. W., Summerfelt, W., Breda, C. S., et al. (1995). *Evaluating managed mental health services: The Fort Bragg experiment*. New York, NY, US: Plenum Press.

Bilenberg, N. (2003). Health of the nation outcome scales for children and adolescents (HoNOSCA). Results of a Danish field trial. *European Child & Adolescent Psychiatry, 12*(6), 298–302.

Brann, P., Coleman, G., & Luk, E. (2001). Routine outcome measurement in a child and adolescent mental health service: An evaluation of HoNOSCA. *The Australian and New Zealand Journal of Psychiatry, 35*(3), 370–376.

CAFAS in Ontario (2005). *Child and adolescent functional assessment scale*. http://www.cafasinontario.ca/html/home.asp, access date 12.4.2007.

CAMHS Outcome Research Consortium (2007). *CORC homepage*. http://www.corc.uk.net/, access date 19.3.2007.

Dunn, G. (1989). *Design and analysis of reliability studies: The statistical evaluation of measurement errors*. New York: Oxford University Press.

Dyrborg, J., Larsen, F. W., Nielsen, S., Byman, J., Nielsen, B. B., & Gautre-Delay, F. (2000). The children's global assessment scale (CGAS) and global assessment of psychosocial disability (GAPD) in clinical practice—substance and reliability as judged by intraclass correlations. *European Child & Adolescent Psychiatry, 9*(3), 195–201.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Epstein, M. H., Kutash, K., & Duchnowski, A. J. (2005). *Outcomes for children and youth with emotional and behavioral disorders and their families: Programs and evaluation best practices* (2nd ed.). Austin, TX: Pro-Ed.

Fonagy, P. (2002). Outcome measurement in children and adolescents. In W. W. IsHak, T. Burt, & L. I. Sederer (Eds.), *Outcome measurement in psychiatry: a critical review* (pp. 59–75). Washington, DC: American Psychiatric Publishing, Inc.

Ford, T., Tingay, K., Wolpert, M., & the CORC Steering Group. (2006). CORC's survey of routine outcome monitoring and national CAMHS dataset developments: A response to Johnston and Gowers. *Child and Adolescent Mental Health, 11*(1), 50–52.

Garralda, M. E., Yates, P., & Higginson, I. (2000). Child and adolescent mental health service use. HoNOSCA as an outcome measure. *British Journal of Psychiatry, 177*, 52–58.

Gowers, S. G., Harrington, R. C., Whitton, A., Beevor, A., Lelliott, P., Jezzard, R., et al. (1999a). Health of the nation outcome scales for children and adolescents (HoNOSCA). Glossary for HoNOSCA score sheet. *British Journal of Psychiatry, 174*, 428–431.

Gowers, S. G., Harrington, R. C., Whitton, A., Lelliott, P., Beevor, A., Wing, J., et al. (1999b). Brief scale for measuring the outcomes of emotional and behavioural disorders in children. Health of the nation outcome scales for children and adolescents (HoNOSCA). *British Journal of Psychiatry, 174*, 413–416.

Gowers, S., Bailey-Rogers, S. J., Shore, A., & Levine, W. (2000). The health of the nation outcome scales for child & adolescent mental health (HoNOSCA). *Child Psychology & Psychiatry Review, 5*(2), 50–56.

Green, B., Shirk, S., Hanze, D., & Wanstrath, J. (1994). The children's global assessment scale in clinical practice: An empirical evaluation. *Journal of the American Academy of Child and Adolescent Psychiatry, 33*(8), 1158–1164.

Hoagwood, K., Jensen, P. S., Petti, T., & Burns, B. J. (1996). Outcomes of mental health care for children and adolescents: I. A comprehensive conceptual model. *Journal of the American Academy of Child and Adolescent Psychiatry, 35*(8), 1055–1063.

Hodges, K., & Gust, J. (1995). Measures of impairment for children and adolescents. *Journal of Mental Health Administration, 22*(4), 403–413.

Hodges, K., Wong, M. M., & Latessa, M. (1998) Use of the child and adolescent functional assessment scale (CAFAS) as an outcome measure in clinical settings. *Journal of Behavioral Health Services & Research, 25*(3), 325–336.

Hodges, K., & Wotring, J. (2004) The role of monitoring outcomes in initiating implementation of evidence-based treatments at the state level. *Psychiatric Services, 55*(4), 396–400.

HoNOSCA Project (2003). *Health of the nation outcome scales for children and adolescents.* http://www.liv.ac.uk/honosca/Home.htm, access date 16.4.2007.

Jensen, P. S., Hoagwood, K., & Petti, T. (1996). Outcomes of mental health care for children and adolescents: II. Literature review and application of a comprehensive model. *Journal of the American Academy of Child and Adolescent Psychiatry, 35*(8), 1064–1077.

Johnston, C., & Gowers, S. (2005). Routine outcome measurement: A survey of UK child and adolescent mental health services. *Child and Adolescent Mental Health, 10*(3), 133–139.

Kazdin, A. E. (2003). Psychotherapy for children and adolescents. *Annual Review of Psychology, 54*, 253–276.

MH-SMART (2007). *The mental health standard measures of assessment and recovery summary.* http://www.tepou.co.nz/page/tepou_11.php, access date 19.3.2007.

National Mental Health Strategy. (2004). *Australian mental health outcomes and classification network.* http://www.mhnocc.org/amhocn/, access date 14.3.2007.

Norcini, J. (2004) Back to the future: clinical vignettes and the measurement of physician performance. *Annals of Internal Meicine*, *141*(10), 813–814.

Peabody, J. W., Luck, J., Glassman, P., Jain, S., Hansen, J., Spell, M., et al. (2004). Measuring the quality of physician practice by using clinical vignettes: A prospective validation study. *Annals of Internal Medicine*, *141*(10), 771–780.

Pirkis, J. E., Burgess, P. M., Kirk, P. K., Dodson, S., Coombs, T. J., & 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*, 76.

Rey, J. M., Starling, J., Wever, C., Dossetor, D. R., & Plapp, J. M. (1995). Inter-rater reliability of global assessment of functioning in a clinical setting. *Journal of Child Psychology and Psychiatry, 36*(5), 787–792.

Schorre, B. E., & Vandvik, I. H. (2004). Global assessment of psychosocial functioning in child and adolescent psychiatry. A review of three unidimensional scales (CGAS, GAF, GAPD). *European Child & Adolescent Psychiatry, 13*(5), 273–286.

Shaffer, D., Gould, M. S., Brasic, J., Ambrosini, P., Fisher, P., Bird, H., et al. (1983). A children's global assessment scale (CGAS). *Archives of General Psychiatry, 40*(11), 1228–1231.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research, 7*(3), 301–317.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Sitter, M., & Hagen, H. (2004). Opptrappingsplanens mål: Status psykisk helsevern for barn og unge. In V. Halsteinli (Ed.), *Samdata psykisk helsevern. Sektorrapport 2003. Sammenligningsdata for psykisk helsevern*. Sintef Helse.

Vatnaland, T., Vatnaland, J., Friis, S., & Opjordsmoen, S. (2007). Are GAF scores reliable in routine clinical use? *Acta Psychiatrica Scandinavica, 115*(4), 326–330.

Weiss, B. (1998). Annotation: Routine monitoring of the effectiveness of child psychotherapy. *Journal of Child Psychology and Psychiatry, 39*(7), 943–950.

Williams, R., & Kerfoot, M. (2005). *Child and adolescent mental health services. Strategy, planning, delivery and evaluation*. Oxford University Press.

Winters, N. C., Collett, B. R., & Myers, K. M. (2005). Ten-year review of rating scales, VII: Scales assessing functional impairment. *Journal of the American Academy of Child and Adolescent Psychiatry, 44*(4), 309–338.

Wolpert, M., Thompson, M., & Tingay, K. (2005). Data collection, clinical audit, and measuring outcomes. In R. Williams & M. Kerfoot (Eds.), *Child and adolescent mental health services. Strategy, planning, delivery and evaluation* (pp. 519–533). Oxford University Press.

World Health Organization (1994). *International statistical classification of diseases and related health problems 10th revision (ICD-10)*.

World Health Organization (1996). *Multiaxial classification of child and adolescent psychiatric disorders*. Cambridge University Press.

Yates, P., Garralda, M. E., & Higginson, I. (1999). Paddington complexity scale and health of the nation outcome scales for children and adolescents. *British Journal of Psychiatry, 174*, 417–423.