Psychiatry Research 281 (2019) 112587

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Reliable and clinically significant change based on the Health of the Nation Outcome Scales

Albert E. Boon[a,b,*], Sjoukje B.B. de Boer[a], Melissa van Dorp[a,c], Yolanda A.M. Nijssen[a,d]

[a] *Parnassia Psychiatric Institute. Youz: child and adolescent psychiatry, Netherlands*
[b] *Curium-LUMC: child and adolescent psychiatry, Leiden University, Leiden, Netherlands*
[c] *Academische Werkplaats Risicojeugd/Intermetzo, Netherlands*
[d] *Tranzo, Tilburg University, Tilburg, Netherlands*

ARTICLE INFO

Keywords:
Outcome study
HoNOSCA
Child and adolescent psychiatry
reliable and clinical significant change

SUMMARY

To evaluate treatment outcomes of individual patients based on clinician-rated instruments, the assessment of reliable and clinically significant change (RCSC) is essential. In heterogeneous samples, RCSC underestimates treatment outcome. Therefore, the Reliable Change Index (RCI) was adjusted by a stratification into subsamples.

This method was tested on the Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA) ratings (n = 12,547) at admission and discharge of youths (age 6–18 years) treated in ten psychiatric institutions. Based on the test-retest reliability of a subsample (n = 397), the RCI was calculated for three severity groups (“very severe,” “moderately severe,” and “subclinical/mild”). Individuals who accomplished reliable change during treatment and moved to a subclinical score were classified as recovered.

Using the traditional RCSC calculation, the large majority (75.1%) of the sample would be considered as unchanged, 2.9% as deteriorated, 18.9% as improved, and 3.1% as recovered. Using RCI cutoff points based on the severity ratings at admission results in a more representative distribution of outcome groups, where 54.6% of the sample was stable, 7.5% worsened, 21.6% improved, and 16.3% recovered.

This methodological framework for calculating RCSC for heterogeneous populations is applicable for all HoNOS instruments, making it very useful for mental health professionals.

1. Introduction

Outcome research in psychotherapy generally relies on statistical significance tests. These methods are useful for identifying group effects but do not provide information on the meaningfulness of the observed changes for an individual. Methods to calculate individual change exist and are generally accepted and are based on the statistics of the total sample to which the individual belongs. As a consequence of this statistical procedure, individuals who are part of a heterogeneous patient population (e.g., general mental health care) are at a disadvantage compared to patients from a more homogeneous patient population (e.g., specialized mental health care). The aim of the present study was to adjust the method of calculating reliable and clinically significant change (RCSC) for the Health of the Nation Outcome Scales (HoNOS) group of outcome instruments in order to overcome this disadvantage for samples with a large dispersion in scores.

In order to monitor progress and evaluate treatment, the assessment of individual clinically meaningful change is essential (Eisen et al., 2007). Jacobson and Truax (1991) proposed a method to determine RCSC. This concept ensures reliable change (RC), meaning that the change an individual experiences goes beyond what could be attributed to measurement errors or chance. This leads to three categories: “improved”, “no reliable change” and “deteriorated”. When the score of an individual reliably improves and moves into the range of the functional population, it constitutes clinically significant change (CSC) and is considered as “recovered”. Normative data from a functional population to calculate CSC are crucial and are available for widely used outcome scales in adult psychiatry, such as the Hamilton Rating Scale for Depression (Grundy et al., 1996), the Symptom Checklist-90-R (Schauenburg and Strack, 1999), and the Brief Psychiatric Rating Scale (Hafkenscheid, 2000). For child and adolescent psychiatry, normative data are also available for instruments such as the Child Behavior Checklist (Achenbach and Rescorla, 2001) and the Strength and Difficulties Questionnaire (Goodman et al., 2000).

For the different versions of the widely used HoNOS (Wing et al., 1996), developed by the British Royal College of Psychiatrists, such

[*] Corresponding author.
E-mail address: a.boon@youz.nl (A.E. Boon).

https://doi.org/10.1016/j.psychres.2019.112587
Received 12 April 2019; Received in revised form 24 September 2019; Accepted 26 September 2019
Available online 27 September 2019

normative data are lacking. The HoNOS group of outcome measures consists of the HoNOS (working age adults), the HoNOS 65+ (older adults), the HoNOSCA (children and adolescents), the HoNOS-Secure (forensic services), the HoNOS-LD (learning disabilities), and the HoNOS ABI (acquired brain injury). Versions of the HoNOS are translated into Croatian, Danish, Dutch, Finnish, French, German, Italian, Korean, Norwegian, Spanish, and Thai and are used in several European countries, Australia, New Zealand, Malaysia, and India (Delaffon et al., 2012; James et al., 2018). The instruments are included in compulsory routine outcome monitoring (ROM) in Canada, Denmark, Holland, Germany, Italy, Norway, Spain, Australia, and New Zealand (Delaffon et al., 2012).

Outcome based on the total score of the HoNOS instruments is ambiguous. Because each item covers a separate issue, dramatic reductions in item scores can occur, reflecting important clinical change, while the overall score can remain more or less the same if small changes in the opposite direction occur on other items (MacDonald, 2002). This problem can be solved by reporting changes on each of the thirteen items separately (e.g., Brann and Coleman, 2010), which gives real insight into the changes that occur on each during treatment. This method is preferable because of its completeness, but clinicians and other stakeholders like health insurance companies and government agencies often prefer single outcome indicators. This is probably why, although the HoNOS does not measure a single, underlying construct of mental health status (Trauer et al., 1999), most researchers have concentrated on differences in the total score (and sometimes section scores) as an indicator of change (Gowers et al., 2000, 1999; Harnett et al., 2005; Manderson and McCune, 2003) or described change through the percentage of patients whose difference in total score increased, decreased, or remained unaltered (Kisely et al., 2007).

Besides the problem of the ambiguity of outcome based on the total score, attempts to calculate RCSC for these instruments confront the researcher with several problems. First, data from a functional population are lacking. To collect normative data from a functional or nonclinical population, trained professionals (psychologist and psychiatrists) would have to assess a large sample of the general population. This is a (far too) expensive undertaking. Second, the instrument is often used in ROM for general mental health institutions with heterogeneous populations (Delaffon et al., 2012). Heterogenic scores affect the size of the standard deviation, which is an essential part of the Reliable Change Index (RCI) formula. As a consequence, patients from heterogenic populations have to change a very high number of points in order to accomplish RC. This explains why outcome studies found that 70% (Brann and Coleman, 2010) to 92% (Parabiaghi et al., 2005) of the sample remains stable (show no improvement or deterioration). These outcomes can lead to discouragement of professionals because only a small proportion of their patients are considered significantly improved. This matter was discussed by Brann and Coleman (2010), who assumed that the 95% confidence level that is a common premise of statisticians is much higher than what characterizes clinical decision making. Others have proposed considering everyone who moved half of the standard deviation in the desired direction on the HoNOS as improved (Nugter et al., 2012).

In all, the aims of the present study were to adjust the RCI for heterogeneous samples by a stratification in subsamples and to propose an alternative for a functional sample to determine whether an individual has returned to normal functioning.

2. Methods

2.1. Setting

In the Netherlands, mental health institutions are obliged to perform ROM (de Beurs et al., 2011). In child and adolescent psychiatry, the HoNOSCA (Gowers et al., 1999) is one of the instruments that can be used for this purpose. Ten institutions throughout the Netherlands that provided child and adolescent mental health services (CAMHS) had collected data for a number of years. The current research used the data collected in these institutions.

2.2. Procedure

The research was in accordance with Dutch medical ethical research regulations, and the anonymized data were handled according to the Personal Data Protection regulation. The therapists of the cooperating CAMHS had been trained in the use of the authorized Dutch version of the HoNOSCA (Staring et al., 2003), ensuring that they would apply the instrument reliably. According to the nationally agreed requirements for ROM, attending therapists were asked to fill in a HoNOSCA at the start (T1) of treatment, then annually and subsequently at termination (T2) of treatment. Because of the demands of ROM at the annual review, the HoNOSCA was sometimes administered several times during therapy. Therefore, it was possible to select a dataset of patients for whom the HoNOSCA was filled in twice within five to ten days during ongoing treatment. Data from this subsample (n = 397) were used to calculate the test-retest reliability.

2.3. Sample

At admission (T1), the HoNOSCA (n = 17,761) was routinely completed for each individual admitted to one of the collaborating ten child mental health facilities from March 2010 to December 2016 (de Beurs et al., 2011). Of this sample, 70.6% (n = 12,547) were reassessed at termination of treatment (T2). The group with only one (T1) measurement and the group with both measurements (T1 and T2) differed significantly in age (M = 11.77 years, SD = 4.01 versus M = 11.33 years, SD = 3.73) (t(17,759) = 7.01, p < .001, d = 0.11) but not on sex (χ^2 (1, n = 17,761) = 0.039, p = .844) or treatment duration (t(17,759) = 0.23901, p = .811). The included sample (T1 + T2) comprised 7700 males (61.4%) and 4847 females (38.6%). The mean age of these youths at admission was 11.35 years (SD = 3.57, range 6 –18). The mean treatment duration was 273 days (SD = 162.08, range 7–1378 days). The ethnic background, based on the country of birth of the patients' parents (the birth country of the mother was used when the parents were born in different foreign countries), of 79.4% of the sample was known. The composition of this group was Native Dutch (76.9%), non-Western (mostly children of immigrants from Turkey, Morocco, or former Dutch colonies) (17.1%), and Western (mostly from European countries) (6%). The main clinical diagnoses (DSM classifications) made by the attending psychiatrists or psychologists were attention deficit hyperactivity disorder (30.2%), disorders of infancy, childhood, or adolescence not otherwise specified (19.6%), autism spectrum disorders (18.5%), anxiety disorders (11.6%), mood disorders (7.4%), conduct disorders (4.0%), V codes (2.1%), personality disorders (0.8%), and other disorders (5.8%).

2.4. Health of the nation outcome scales for children and adolescents

The HoNOSCA is the child and adolescent version of the HoNOS (Wing et al., 1996). The instrument was developed as a standardized assessment tool for routine use in mental health services. The HoNOSCA is a clinician-rated instrument (see Box 1) composed of 13 symptom or function items and 2 optional items concerning the parents' lack of knowledge about services and about difficulties. Most studies do not report on the last two items, and it has been suggested that they be omitted (Brann and Coleman, 2010; Garralda and Yates, 2000). The items can be scored using a five-point Likert scale: 0 "No problem", 1 "Minor problem requiring no action", 2 "Mild problem but definitely present", 3 "Moderately severe problem" and 4 "Severe to very severe problem". The glossary specifies that a rating of 2 or higher indicates a clinically significant symptom, worthy of clinical attention, such as further assessment, treatment, monitoring, or documentation; a rating of 0 or 1 indicates no clinical problem. The HoNOSCA scores are commonly reported as the total score ("overall severity of physical, personal and social problems associated with mental illness"), which is the summed score of the first 13 items (range 0–52) (Gowers et al., 1998). Multiple studies have supported acceptable validity, reliability, sensitivity to clinical change, and feasibility of the HoNOSCA (Bilenberg, 2003; Brann and Coleman, 2010; Garralda et al., 2000; Gowers et al., 1999b). The convergent validity has been found to be satisfactory (Garralda et al., 2000; Harnett et al., 2005; Urben et al., 2014). Although it was stated that the "key component sections" were supported by a principal component analysis (Gowers et al., 2000a), later research could not replicate this finding (Brann, Unpublished in Pirkis et al., 2005; Tiffin and Rolling, 2012). The interrater reliability is good, with intraclass correlations greater than 0.8 for most of the items (Hanssen-Bauer et al., 2007; Hunt and Wheatley, 2009; Yates et al., 2006). Several studies have established the instrument's concurrent validity (Hanssen-Bauer et al., 2010) and reliability (Gowers et al., 1999; Harnett et al., 2005). The test-retest reliability is moderate, and the total score is moderately sensitive to change (Garralda et al., 2000; Gowers et al., 2002; Pirkis et al., 2005).

2.5. Classification of severity

Because of the ambiguity of the total score of the HoNOSCA a more advanced method for classifying severity was applied. This method was proposed for the HoNOS (Lelliott, 1999) and the HoNOSCA (Gowers et al., 2000) and later elaborated for the HoNOS (Parabiaghi et al., 2005). All attempts to make a severity index are based on the fact that the answer categories for each item are severity scales in themselves. Ranging from 0 "no problem" to 4 "very severe problem", and scores of 2 or higher are assumed to indicate clinically significant problems. No research is known to support these assumptions, but based on the content of the answer categories a severity index makes sense. Therefore severity was defined as follows: "very severe" patients with a score of ≥3 on at least two items (item 6, physical illness, excluded), "moderately severe" patients with a score of ≥3 on one item (item 6, physical illness, excluded), "mild" patients with at least one item's score of 2, and "subclinical" patients with no scores of 2 or higher. Table 1 shows the division of the sample into severity groups at the start and end of treatment.

2.5.Statistics

2.5.1. Reliable change and clinically significant change

The advice of Tingey et al. (1996) to use multiple clinical groups based on severity of symptoms and to determine cutoff points for each group was followed. Because at T1 only a very small proportion (2.4%) of the sample was classified as "subclinical," these individuals were added to the "mild" group, resulting in three subgroups based on severity ("subclinical/mild," "moderately severe," and "very severe"). For each group, RC was calculated using the formula of Jacobson and Truax (1991) to compute the RCI for the total and section scores. The cutoff points (the number of points a subject has to change between two measurements to have reliably changed) were calculated using the test-retest reliability (r_{xx}) of the subsample (n = 397) and the standard deviations of the three severity groups at T1. The standard error of measurement (SEM) was calculated (s√1-r_{xx}), DIFF = √2(SEM2), RCI = xt1-xt2/SDIFF. The 95% criterion was used, so change greater than this would only occur by unreliability of measurement alone in less than 5% of the cases. The resulting categories were "deteriorated" (a significant increase in score between T1 and T2), "no reliable change" (no significant increase or decrease in score between T1 and T2), and "improved" (a significant decrease in score between T1 and T2). Consequently, CSC was calculated in two ways: (a) by following the suggestion of Jacobson and Truax (1991) to categorize patients who improved according to the RCI and also moved more than two standard deviations under the mean of the sample they belonged to at pre-treatment as "recovered," and (b) by categorizing patients who improved and also moved into the subclinical range of severity as "recovered" (no clinical problems). The last method implies that only patients with full remission are considered to be recovered. The cutoff points were calculated for the total score and the sections (behavioral problems, impairment problems, symptomatic problems, and social problems). For the section scores, only the last method was followed, so only patients with no scores of 2 or higher on the associated items were considered to be recovered.

Table 1
Division into severity categories.

	T1		T2	
	n	%	n	%
Very severe	6345	50.6	3099	4.7
Moderately severe	3821	30.5	2720	21.7
Mild	2079	16.6	3988	31.8
Subclinical	302	2.4	2740	21.8
Total	12,547	100.0	12,547	100.0

2.5.2. Statistical analyses

All analyses were performed using SPSS, version 25.0 (IBM, 2017). Descriptive statistics were used to calculate the means and standard deviations. The test-retest reliability was calculated using correlations. Differences in scores between T1 and T2 were tested using the paired samples *t*-test (one-sided). Effect sizes using Cohen's *d* were calculated.

3. Results

The descriptive statistics were calculated at T1 and T2 for the total score and the sections. The HoNOSCA scores of the sample at admission (T1) and discharge (T2) were compared (paired *t*-test) and showed a significant decrease in the total and all section scores but with small effect sizes (Cohen's *d*) (Table 2).

3.1.Test-retest reliability

A sample of 397 patients for which the HoNOSCA was completed twice within ten days during treatment (M = 7.41 days, range 5–10 days, SD = 1.46) was used to calculate the test-retest reliability for the total score (r_{xx} = 0.812) and the sections on behavioral problems (r_{xx} = 0.783), impairment problems (r_{xx} = 0.651), symptomatic problems (r_{xx} = 0.803), and social problems (r_{xx} = 0.765).

3.2. Reliable change

First, the test-retest values (r_{xx}) of the subsample with a repeated measurement within 10 days and SD at T1 of the total sample were used in the formula of Jacobson and Truax (1991) to calculate the RCI for the total and section scores. The cutoff point for the total score was 8,

Table 2
Comparison of HoNOSCA scores between T1 and T2 (n = 12,547).

	T1		T2				
	M	*SD*	*M*	*SD*	*t*	*p*	ES
Behavior	3.42	2.67	2.40	2.04	54.88	< 0.001	.48
Impairment	1.48	1.51	1.05	1.38	31.37	< 0.001	.29
Symptoms	2.70	2.12	1.77	1.82	50.69	< 0.001	.47
Social	4.77	2.91	3.62	2.93	47.28	< 0.001	.39
Total score	12.37	5.86	8.84	6.21	66.96	< 0.001	.59

T1 = admission, T2 = discharge; *M* = mean; SD = standard deviation; ES = effect size (Cohen's *d*).

meaning that a patient whose HoNOSCA score decreased 8 points or more on the total score was classified as "improved," and an increase of 8 points or more was classified as "deteriorated"; patients within the range of −7 to +7 were classified as "no reliable change." The sample had a wide range in scores (0–51), and about a fifth (20.4%) had an initial score of 7 or lower. This meant that a large minority could not achieve a clinically significant improvement because their initial score was not high enough to decrease 8 points. By stratifying the sample into severity categories, resulting in a lower cutoff, reaching clinically meaningful change also became possible for individuals with an initial low score.

The ranges of the total scores for the severity groups were calculated: "subclinical" (1–9), "mild" (2–31), "moderately severe" (3–34), and "very severe" (6–51). This demonstrates the advantage of using the severity categorization over a categorization based on an simple addition of the item scores, because an individual with a total score between 6 and 9 can be found in each severity group but is categorized from "subclinical" when there are no clinical problems to "very severe" when there are two or more clinically significant symptoms for which treatment is necessary.

Because at T1, only a small group (2.4%) was classified as "subclinical" and could not be considered as dysfunctional and thus could not reach RC, these individuals were added to the "mild" group to calculate the cutoff points for the RCI for the groups based on severity. The descriptive statistics for the severity groups were calculated for the total score and the section scores. Based on these statistics, the cutoff points for the RCI were calculated (see Table 3).

As Table 3 shows, the RCI cutoff point for the total score of the "very severe" group (7) is only one point lower than that of the total population. For the other severity groups, however, the number of points that someone has to change was considerably lower. Based on this sample, the cutoff points of the total score for the "mild" and the "moderately severe" groups are the same, and only one difference in the section scores is found.

Based on the cutoff points in Table 3, the RCSC was calculated (see Table 4). The "improved" group was divided into two groups: "improved" and "recovered." The "improved" group referred to patients who had a reliable improvement but had not reached the cutoff point for the severity category "subclinical." The "recovered" group in Table 4 were those patients who reliably improved and at the end of treatment reached the cutoff point for the severity category "subclinical," thus showing full remission.

Another method to calculate recovery is to use the Jacobson and Truax (1991) criterion identifying persons who had reliably improved and moved two standard deviations below the mean at T1 (J&T criterion). For the total score, this would mean that 2138 patients (17%) would be classified as recovered. Although the percentages (16.3–17.0%) for both methods are about equal, 2642 patients were classified as recovered according to the J&T criterion or subclinical criterion. Of this group 1535 (58.1%) were considered recovered according to both criteria, 600 (22.7%) met the J&T criterion but did not reach the subclinical level, and 504 (19.2%) were subclinical but were not recovered according to the J&T criterion.

To test whether the improvement based on the RCSC criteria could be supported by group statistics, a paired sample *t*-test was performed for the four RCSC categories (see Table 5).

Table 5 makes clear that all groups underwent significant change between T1 and T2. A small effect (0.22 SD) is found in the "no change" group, while all other groups had large effects, up to a change of 2.68 standard deviations for the recovered group. Compared to the RCSC based on the undivided sample (with J&T criterion of recovery) the two distributions differed significantly (χ^2(9, *N* = 12,547) = 28,612,46, $p < 0.001$). The numbers and effect sizes are "recovered" (*N* = 395; 3.1%) ES = 4.00, "improved" (*N* = 2374; 18.9%) ES = 2.14, "no change" (*N* = 9420; 75.1%) ES = 0.32, and "deteriorated" (*N* = 358; 2.9%) ES = 1.89. Because the criteria based on the undivided sample are stricter, the effect sizes are higher, but as a consequence, less than 20% of the recovered group according to our method would be considered as such. Compared to the conventional RCSC method, in our proposed method, a minority (4.7%) of patients have a less desirable outcome (587 moved from "no reliable change" to "deteriorated"), and 2815 (22.4%) have a more desirable outcome (1547 moved from "no reliable change" to "improved" and 827 from "improved" to "recovered").

4. Discussion

Clinicians need methods to monitor the progress or decline of individual patients during treatment. A very useful method is that of RCSC. The formula for RC is highly dependent on the dispersion of the scores of the sample under study. Hence, institutions that offer general psychiatric care and, as a consequence, serve heterogeneous populations with often broad dispersion of scores have a disadvantage compared to specialized mental health services with, in most cases, a much smaller coefficient of variation (CV). Especially in very heterogeneous samples, the results of this method are often perceived as unrealistic and disappointing and "undermine everyone's motivation for continuing treatment" (Delaffon et al., 2012) p. 1102). Therefore, the aim of the present study was to adjust the method to calculate RCSC in such a way that the disadvantage for samples with a large dispersion in scores would be corrected. A number of steps were proposed and applied to the youth version of the HoNOS (HoNOSCA).

Many studies using the RCI have rather homogeneous populations with a CV between 0.1 and 0.2. To introduce their method, Jacobson and colleagues used an imaginary sample with a CV of 0.19 (*M* = 40, SD = 7.5) (Jacobson and Truax, 1991). In other studies, the CV was 0.15 (Bauer et al., 2004). The sample of the present study was very heterogeneous (*M* = 12.37, SD = 5.85, CV = 0.47), and because the standard deviation is an essential part of the RCI formula, only a very small proportion of the sample would have been classified as significantly changed if this criterion had been applied. This phenomenon can result in outcome studies in which more than 90% of the patients remain unchanged during treatment. Therefore, the standard deviations of more homogeneous groups based on pretreatment severity were used to determine the cutoff points.

Although it has been stated that the total score of the HoNOSCA provides a good quantitative measure of clinical severity when compared with a severity index (Gowers et al., 2000), this only holds true

Table 3
Mean, standard deviations, and cutoff points based severity at T1 for total and section scores (*n* = 12,547).

Severity at T1	Mild			Moderately severe			Very severe		
	M	*SD*	Cutoff	*M*	*SD*	Cutoff	M	SD	Cutoff
Behavior	1.99	1.34	2	2.94	1.18	2	4.24	1.68	3
Impairment	0.80	0.94	2	1.09	1.19	2	1.97	1.70	3
Symptoms	1.39	1.25	2	2.06	1.74	3	3.52	2.23	3
Social	2.69	1.68	3	3.38	1.91	3	6.50	2.89	4
Total score	6.86	3.31	4	9.47	3.20	4	16.23	5.47	7

T1 = admission.

Table 4
RCSC based on severity at T1.

	Total score		Behavior		Impairment		Symptoms		Social	
	n	%	*n*	%	*n*	%	*n*	%	n	%
Recovered	2042	*16.3*	2500	*20.0*	1559	*12.4*	1578	*12.6*	1789	*14.3*
Improved	2715	*21.6*	1048	*8.4*	99	*0.8*	308	*2.5*	854	*6.8*
No reliable change	6845	*54.6*	8211	*65.4*	10,269	*81.8*	10,220	*81.5*	9151	*72.9*
Deteriorated	945	*7.5*	778	*6.2*	620	*4.9*	441	*3.5*	753	*6.0*
	12,547	*100.0*	12,547	*100.0*	12,547	*100.0*	12,547	*100.0*	12,547	*100.0*

T1 = admission.

Table 5
Paired sample *t*-test for HoNOSCA total score by RCSC (n = 12,547).

		T1		T2				
	n	*M*	SD	*M*	SD	*t*	*p*	ES
Recovered	2042	11.51	4.55	2.14	1.92	94.31	< 0.001	2.68
Improved	2715	16.36	6.12	7.60	4.22	119.04	< 0.001	1.67
No change	6845	11.51	5.47	10.27	5.66	38.40	< 0.001	0.22
Deteriorated	945	8.98	4.86	16.50	6.83	− 62.07	< 0.001	1.27

T1 = admission, T2 = discharge; *M* = mean; SD = standard deviation; ES = effect size (Cohen's *d*).

Box 1
Items and sections of the HoNOSCA.

Items	*Sections*
1 *Disruptive, antisocial or aggressive behaviour*	*Behaviour*
2 *Over-activity, attention or concentration*	
3 *Non-accidental self-injury*	
4 *Alcohol, substance/solvent misuse*	
5 *Scholastic or language skills*	*Impairment*
6 *Physical illness or disability problems*	
7 *Hallucinations, delusions*	*Symptoms*
8 *Non-organic somatic symptoms*	
9 *Emotional and related symptoms*	
10 *Peer relationships*	*Social*
11 *Self-care and independence*	
12 *Family life and relationships*	
13 *Poor school attendance*	

Example: Instruction for scale 9: Problems with emotional and related symptoms
***Rate** only the most severe clinical problem not considered previously*
***Include** depression, anxiety, worries, fears, phobias, obsessions or compulsions, arising from any clinical condition including eating disorders.*
***Do not include** aggressive, destructive or overactive behaviours attributed to fears, phobias, rated at scale 1.*
***Do not include** physical complications of psychological disorders, such as severe weight loss, rated at scale 6.*

when severity classifications are compared with group means. On an individual level, the range of 6–9 for total score can represent a severity range of individuals with no clinically relevant problems to individuals with very severe problems. To overcome this ambiguity of the total score of the HoNOSCA, a division into four severity categories was made, based on the item scores. These categories, “very severe,” “moderately severe,” “mild,” and “subclinical,” are considered a more reliable way to assess the severity of an individual than a division based solely on the total score. To calculate the RCI, three severity subgroups at T1 were used (“subclinical” and “mild” were combined).

Calculating CSC data from a functional population is the gold standard for deciding whether a patient has moved past the criterion of the functional group. When data from a normal population are not available, patients who fall outside the range of the dysfunctional population, defined as extending to two standard deviations beyond (in the direction of functionality) the mean for that population, can be considered as functional (Jacobson and Truax, 1991). We propose however using the “subclinical” category of the severity index to decide whether a patient who has clinically significantly improved is recovered, because patients who are labeled recovered according to the J&T criterion can still have very severe problems. Our method implies that only patients with full remission are considered as functional. Because no real data from the functional population are available, whether the Jacobson and Truax statistical approach is closer to the truth than ours is up for debate.

Because the HoNOSCA is a clinician-rated instrument, the results could be inflated (e.g., by giving lower scores at the end of treatment). The outcome of the group statistics of the total group (paired *t*-test), however, showed a medium effect size (d = 0.59). This was comparable to or lower than results found in meta-analyses of psychotherapy for children and adolescents, in which effect sizes ranged from 0.69 (Weisz et al., 2006) to 0.71 for general CAMHS populations (Weisz et al., 1995) and 0.88 in a nationwide Australian study (Network, 2005). Therefore, it can be concluded that the therapists in the present study were probably realistic in their assessments of the changes that occurred in the course of treatment.

Looking at the results of the present study, we think that the RCI method that we proposed is a more realistic assessment, because about half of the sample was stable between the beginning and end of treatment. When the RCI is calculated for the undivided sample, the outcome is far less favorable and about three quarters of the sample would be considered unchanged.

A small proportion of the sample was considered as “moderately severe” or “very severe” at the start of treatment but had a total score of respectively ≤ 7 or ≤ 4 and thus could not reach the cutoff point for RC and had no opportunity to improve. In the sample of the present study, 267 individuals (2.2%) met this condition. Of this group, 101 individuals had reached the “subclinical” category of severity at the end of treatment (0.8%). A consideration is to add these individuals to the recovered group, because it seems illogical to classify a person with one or two (very) severe problems at the start of treatment who moves to the subclinical level as unchanged.

When test-retest reliability is not available, internal consistency (Cohen's alpha) is used as the reliability in the formula of Jacobson and Truax. However, this should be considered the second-best option, since SEM should be based on the agreement between the results of successive measurements (Jacobson and Truax, 1991; Parabiaghi et al., 2005). Because test-retest reliability was calculated under almost optimal conditions (M = 7 days, range 5–10 days), this reliability can be classified as good and equal (Brann and Coleman, 2010; Harnett et al., 2005) or higher (Garralda et al., 2000) than in other studies that used a longer test-retest time period. The use of Cohen's *alpha* for the present sample (0.629 at T1 and 0.774 at T2) would have led to less favorable results.

The HoNOSCAs in our sample were collected as part of ROM. One of the goals of ROM is to perform benchmark studies and compare outcome scores between mental health organizations. Often, there are significant differences between mental health services in the severity of their patient groups. Some institutions offer help to less severe patients and exclude patients who need, and subsequently receive, more specialized treatment at other institutions. The proposed method can support more realistic and fair comparisons when the results of groups based on severity within institutions are compared between institutions.

We also applied our method to the key component sections of the HoNOSCA, although we are aware of the disputed status of these sections. Although at the introduction of the HoNOSCA, it was stated that there was statistical evidence that the sections represented an underlying factor, no other research has confirmed this claim. Neither does a principal component analysis on our sample support this division into sections. This shows the problematic status of the naming of the HoNOS instruments. The items are incorrectly called scales, because in test theory, a scale is a set of questions that measure a latent variable (e.g., Michell, 1990), and the component sections do not represent statistically sound components. The reason we chose to use the behavior, impairment, symptoms, and social sections is that they have proven to be acceptable to clinicians from a range of disciplines and services (Gowers et al., 1999) and have a high face validity (Gowers et al., 2000), meaning that they make sense for clinicians working with children and adolescents as distinctive domains related to mental health problems. Although we advocate to report outcome results based the 13 separate items, we realize that in many occasions a single outcome indicator is desired. An outcome report that adds the results on the four sections to this single outcome indicator becomes more insightful.

The method we propose can be implemented for all versions of the HoNOS. Because the answer categories for all versions are identical, a division into subgroups based on severity can be accomplished and the RCI can be calculated by using the SDs of these subgroups. It is advisable to perform a test-retest study (preferable with an interval of a week). When this is not an option, the Cohen's *Alpha* reliability test can be used, with the caveat that this last option is less reliable, and in the present study, it would have led to less preferable outcome results. When the SD and reliability are known, the RCI can be easily calculated with an RCI calculator on the internet (e.g., https://www.psyctc.org/stats/rcsc1.htm).

An alternative way of measuring outcome can be to compare the severity score at the beginning and termination of treatment. However, this can lead to an improvement to a less severe category based on a reduction of the total score of one point. Therefore, the combination of the severity categories and the RCI index must be considered superior.

4.1. Limitations

The data on which the cutoff points for the RCI were calculated were collected in the Netherlands. Although research has found that psychiatric problems of children and adolescents in Western countries are comparable (Achenbach et al., 2008), further research on the international generalizability is needed.

Our method is based on the classification of severity that has been proposed by several authors, but these severity groups have never been substantiated. Future research should be aimed at testing the external validity of the severity criteria used.

Declaration of Competing Interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this paper.

Acknowledgments

The authors would like to thank the Routine Outcome Monitoring Consortium Kinder- en Adolescenten Psychiatrie (ROMCKAP) of the Dutch Knowledge Centre of Child and Adolescents Psychiatry that provided the data from their members for research purposes. Special thanks go to their information analysts, Pieter Koppelaar and Jan Berend Dusink, for their efforts to aggregate the data.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2019.112587.

Bibliography

A.M.H.a.O.C.c. Network, 2005. Child and adolescent national outcomes and casemix collection standard reports, version 1.1. In: Australia, G.o.S. (Ed.), Brisbane, Queensland.

Achenbach, T., Rescorla, L., 2001. Manual For the ASEBA School-Age Forms & Profiles. University of Vermont, Research Center for Children, Youth, & Families, Burlington, VT.

Achenbach, T.M., Becker, A., Döpfner, M., Heiervang, E., Roessner, V., Steinhausen, H., Rothenberger, A., 2008. Multicultural assessment of child and adolescent psychopathology with ASEBA and SDQ instruments: research findings, applications, and future directions. J. Child Psychol. Psychiatry 49 (3), 251–275.

Bauer, S., Lambert, M., Nielsen, S., 2004. Clinical significance methods: a comparison of statistical techniques. J. Personal. Assess. 82 (1), 60–70.

Bilenberg, N., 2003. Health of the nation outcome scales for children and adolescents (HoNOSCA)–results of a Danish field trial. Eur. Child Adolesc. Psychiatry 12 (6), 298–302.

Brann, P., Coleman, G., 2010. On the meaning of change in a clinician's routine measure of outcome: HoNOSCA. Aust. N Z J Psychiatry 44 (12), 1097–1104.

Brann, P., Unpublished. Routine Outcome Measurement in Child/Adolescent Mental Health: HoNOSCA - Valid Enough? Feasible Enough?Melbourne, Monash University.

de Beurs, E., den Hollander-Gijsman, M., van Rood, Y., van der Wee, N., Giltay, E., van Noorden, M., van der Lem, R., van Fenema, E., Zitman, F., 2011. Routine outcome monitoring in the Netherlands: practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. Clin. Psychol. Psychother. 18, 1–12.

Delaffon, V., Anwar, Z., Noushad, F., Ahmed, A.S., Brugha, T.S., 2012. Use of health of the nation outcome scales in psychiatry. Adv. Psychiatric Treatment 18 (3), 173–179.

Eisen, S., Ranganathan, G., Pradipta, S., Spiro, A., 2007. Measuring clinically meaningful change following mental health treatment. J. Behav. Health Serv. Res. 34, 272–289.

Garralda, E., Yates, P., 2000. HoNOSCA: uses and limitations. Child Psychol. Psychiatry Rev. 5 (3), 131–132.

Garralda, M., Yates, P., Higginson, I., 2000. Child and adolescent mental health service use. HoNOSCA as an outcome measure. Br. J. Psychiatry 177 (1), 52–58.

Goodman, R., Ford, T., Simmons, H., Gatward, R., Meltzer, H., 2000. Using the strengths and difficulties questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. Br. J. Psychiatry 177 (6), 534–539.

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

Gowers, S., Harrington, R., Whitton, A., Beevor, A., Lelliott, P., Jezzard, R., Wing, J., 1998. Health of the nation outcome scales for children and adolescents (HoNOSCA). Glossary for HoNOSCA score sheet. Br. J. Psychiatry 174, 428–431.

Gowers, S., Harrington, R., Whitton, A., Lelliott, P., Beevor, A., Wing, J., Jezzard, R., 1999. Brief scale for measuring the outcomes of emotional and behavioural disorders in children. Health of the Nation Outcome Scales for children and Adolescents (HoNOSCA). Br. J. Psychiatry 174, 413–416.

Gowers, S., Levine, W., Bailey-Rogers, S., Shore, A., Burhouse, E., 2002. Use of a routine, self-report outcome measure (HoNOSCA-SR) in two adolescent mental health services. Health of the Nation Outcome Scale for Children and Adolescents. Br. J. Psychiatry 180, 266–269.

Grundy, C.T., Lambert, M.J., Grundy, E.M., 1996. Assessing clinical significance: application to the Hamilton rating scale for depression. J. Mental Health 5, 25–33.

Hafkenscheid, A., 2000. Psychometric measures of individual change: an empirical comparison with the brief psychiatric rating scale (BPRS). Acta Psychiatr. Scand. 101, 235–242.

Hanssen-Bauer, K., Aalen, O., Ruud, T., Heyerdahl, S., 2007. Inter-rater reliability of clinician-rated outcome measures in child and adolescent mental health services. Admin. Policy Ment. Health Ment. Health Serv. Res. 34 (6), 504–512.

Hanssen-Bauer, K., Langsrud, Ø., Kvernmo, S., Heyerdahl, S., 2010. Clinician-rated mental health in outpatient child and adolescent mental health services: associations with parent, teacher and adolescent ratings. Child Adolescent Psychiatry Mental Health 4, 29.

Harnett, P., Loxton, N., Sadler, T., Hides, L., Baldwin, A., 2005. The health of the nation outcome scales for children and adolescents in an adolescent in-patient sample. Aust. N. Z. J. Psychiatry 39 (3), 129–135.

Hunt, J., Wheatley, M., 2009. Preliminary findings on the health of the nation outcome scales for children and adolescents in an inpatient secure adolescent unit. Child Care Pract. 15 (1), 49–56.

IBM, 2017. SPSS Statistics for Windows, 25.0 ed. IBM Corp., Armonk, NY.

Jacobson, N., Truax, P., 1991. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J. Consult. Clin. Psychol. 59 (1), 12–19.

James, M., Painter, J., Buckingham, B., Stewart, M.W., 2018. A review and update of the health of the nation outcome scales (HoNOS). Br. J. Psychiatry Bull. 42 (2), 63–68.

Kisely, S., Campbell, L., 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. Commun. Ment Health J. 43 (2), 91–107.

Lelliott, P., 1999. Definition of severe mental illness. In: Charlwood, P., Mason, A., Goldacre, M., Cleary, R., Wilkinson, E. (Eds.), Health Outcome indicators: Severe Mental illness. Report of a Working Group to the Department of Health. National Centre for Health Outcomes Development, Oxford, pp. 87–93.

MacDonald, A.J.D., 2002. The usefulness of aggregate routine clinical outcomes data. The example of HoNOS65+. J. Ment. Health 11, 645–656.

Manderson, J., McCune, N., 2003. The use of honosca in a child and adolescent mental health service. Irish J. Psychol. Med. 20 (2), 52–55.

Michell, J., 1990. An Introduction to the Logic of Psychological Measurement. Lawrences Erlbaum Associates, Hillsdales, NJ.

Nugter, M.A., Buwalda, V.J.A., Dhondt, A.D.F., Draisma, S., 2012. The use of Honos in the treatment of patients. Tijdschrift voor Psychiatrie 54 (2), 153–159.

Parabiaghi, A., Barbato, A., D'Avanzo, B., Erlicher, A., Lora, S., 2005. Assessing reliable and clinically significant change on health of the nation outcome scales: method for displaying longitudinal data. Austr. N. Z. J. Psychiatry 39, 719–725.

Pirkis, J., Burgess, P., Kirk, P., Dodson, S., Coombs, T., Williamson, M., 2005. A review of the psychometric properties of the health of the nation outcome scales (HoNOS) family of measures. Health Q. Life Outcomes 3 (76), 1–12.

Schauenburg, H., Strack, M., 1999. Measuring psychotherapeutic change with the symptom checklist SCL-90-R. Psychother. Psychosom. 68, 199–206.

Staring, T., Hofman, E., Mulder, N., 2003. Health of the Nation Outcome Scales for Children and Adolescents (Dutch version). Trimbos instituut, Utrecht.

Tiffin, P.A., Rolling, K., 2012. Structure of the health of the nation outcome scales for children and adolescents: An ordinal factor analysis of clinician ratings of a sample of young people referred to community mental health services. Psychiatry Res. 197 (1-2), 154–162.

Tingey, R., Lambert, M., Burlingame, G., Hansen, N., 1996. Assessing clinical significance: proposed extensions to method. Psychother. Res. 6, 109–123.

Trauer, T., Callaly, T., Hantz, P., Little, J., Shields, R.B., Smith, J., 1999. Health of the nation outcome scales: Results of the Victorian field trial. Br. J. Psychiatry 174, 380–388.

Urben, S., Baier, V., Mantzouranis, G., Schwery, J., Mahi, C., Courosse, S., Guignet, B., Halfon, O., Holzer, L., 2014. The french adaptation of the health of the nation outcome scale for children and adolescents self-rated form (F-HoNOSCA-SR): validation and clinical routine use. Psychiatry Res 218 (1-2), 229–235.

Weisz, J., Weiss, B., Han, S., Granger, D., Morton, T., 1995. Effects of psychotherapy with children and adolescents revisited: a meta-analysis of treatment outcome studies. Psychol. Bulletin 117, 450–468.

Weisz, J., McCarty, C., Valeri, S., 2006. Effects of Psychotherapy for Depression in Children and Adolescents: A Meta-Analysis. Psychol. Bulletin 132 (1), 132–149 Copyright 2006 by the American Psychological Association.

Wing, J., Curtis, R., Beevor, A., 1996. HoNOS: Health of the Nation Outcome Scales. Trainers' Guide. College Research Unit, London.

Yates, P., Kramer, T., Garralda, M., 2006. Use of a routine mental health measure in an adolescent secure unit. Br. J. Psychiatry 188, 583–584.