The HoNOS Family of Measures:

A technical review of their psychometric properties

The NATIONAL CENTRE *of* MENTAL HEALTH RESEARCH, INFORMATION *and* WORKFORCE DEVELOPMENT

www.tepou.co.nz

Te Pou
o Te Whakaaro Nui

MENTAL HEALTH PROGRAMMES LIMITED TRADING AS TE POU

Version 2, Published in November 2012 by Te Pou o Te Whakaaro Nui
The National Centre of Mental Health Research, Information and Workforce Development.
PO Box 108-244, Symonds Street, Auckland, New Zealand.

Web: www.tepou.co.nz
Email: info@tepou.co.nz

Acknowledgements

This report was produced by Aleksandra Antevska (Research Lead) and Mark Smith (Clinical Lead) with other members of the research and evaluation team and delivery team within Te Pou.

The authors thank Dr Malcolm Stewart (Counties Manukau District Health Board) and Professor Graham Mellsop (Waikato District Health Board) for their reviews of the report.

Citation: Te Pou. (2012). *The HoNOS Family of Measures: A technical review of their psychometric properties.* Auckland: Te Pou o Te Whakaaro Nui.

Contents

List of tables

Executive summary

The Health of the Nation Outcome Scale (HoNOS) measures the symptom severity and social functioning of those experiencing mental health issues across time. The original adult measure has been adapted into other measures, five of which (as of July 1st 2012) will be mandated for collection in the New Zealand clinical context. The other four measures are: HoNOSCA, which assesses outcomes for children and adolescents, HoNOS 65+, for older individuals, HoNOS-LD for those with learning difficulties and HoNOS secure, for those in a secure setting. There is anecdotal evidence, feedback from training workshops and from previous reviews (Pirkis, Burgess, Kirk, Dodson, Coombs & Williamson, 2005, Deady 2010) which suggests that some clinicians and researchers question the validity and reliability of the HoNOS family of measures. Given that these measures are mandated for collection in New Zealand it is important to address these concerns in this technical report. Te Pou has initiated this report in order to address these concerns. The audience for the report are interested mental health practitioners, policy makers and stakeholders.

This technical report updates an earlier review of research on the psychometric properties of the HoNOS family of measures Pirkis et al 2005. Pirkis et al., (2005) summarised the evidence on psychometric properties on the HoNOS, HoNOSCA and HoNOS65+ from a range of published works. They concluded that these three measures have mostly good or adequate reliability, validity and sensitivity to therapeutic change and are able to be used to measure outcomes for different groups on a range of mental health related constructs. However, some of the properties were under-investigated and warranted additional research, these included the content validity of the HoNOSCA and most of the validity and reliability constructs of the HoNOS65+.
This report includes research published since 2005 and research on the HoNOS-LD and HoNOS- secure as these were not included in the previous review. This report provides an update on recent research for New Zealand practitioners, policy makers and stakeholders.

Method

Search strategy

A search of the literature was conducted to identify all relevant articles published since 2005. See page 12 for full search strategy details.

Establishing validity and reliability

The psychometric properties of instruments and measures are used to determine their quality and usefulness in the required setting. These properties are split into reliability, validity and sensitivity to therapeutic change. Reliability refers to the consistency of a set of items or a measure. Reliability is the extent to which we can be sure that the score received on a test is consistent over time and across conditions. It is used to describe how good the test is at eliminating confounding error. Validity refers to whether the test actually measures what it is intended to measure. Validity testing is concerned with what the test measures, and how well it does this.

Sensitivity to therapeutic change is the measures ability to measure change across time. Feasibility is the degree to which the measure is acceptable to stakeholders or in this case useful in clinical practice. Feasibility is covered in training for the use of the measures in New Zealand and is not included in this review (See pages 10 and 11 in the Method section for further explanation of these properties).

Based on the literature published since 2005 this technical report will consider the reliability tests (which are test-retest reliability, inter-rater reliability, internal consistency), the validity tests (which are content validity, criterion validity, construct validity) and the sensitivity to therapeutic change of the five instruments: HoNOS, HoNOS 65+, HoNOSCA, HoNOS- secure and HoNOS- LD. See page 10-11 for more details of these reliability and validity tests.

Results

There is strong evidence about some of the psychometric properties for the HoNOS family of measures, but there are also some substantial gaps and limitations. See Table 1 on page 27 for more details.

The overall conclusion which can be drawn from this review is that, while there are still areas in which further evaluation of the psychometric properties of these measures could be undertaken, the preponderance of evidence suggests that all the HoNOS family of measures do have adequate reliability, validity, and sensitivity to therapeutic change to be useful in clinical settings in New Zealand. All the measures have shown some evidence of reliability and validity, and some evidence suggestive of sensitivity to therapeutic change. This supports the adoption of these measures by the New Zealand mental health services.

More research is needed to determine if the HoNOS-secure and the HoNOS-LD have sound psychometric properties.

Background

The Health of the Nation Outcome Scale (HoNOS) was developed by Wing and colleagues in response to a mental health target to improve the health and social functioning for those with experience of mental illness (Wing, Beevor, Curtis, Park, Hadden & Burns, 1998; Wing, Lelliott & Beevor, 2000). The measure was designed to assess symptom severity and social functioning of people experiencing mental health issues across time and is widely used in mental health services around the world at admission, during treatment and at follow up. Once data is collected it can be used to measure clinical outcomes at various information levels (individually, team, DHB, nationally), as well as to monitor, evaluate and improve services. In the United Kingdom, Australia and New Zealand various members of the HoNOS family have been adopted as outcome tools. They are widely used in surveys or as a research tool in several European countries (Lovagilo & Monzani, 2011).

The HoNOS family of measures are used to collect information for the New Zealand mental health database (PRIMHD-programme for the integration of mental health data) which informs service planning and funding at a national level and also at DHB and team level. This involves HoNOS, HoNOSCA and HoNOS 65+ presently and – as of July 1st 2012—it will include HoNOS-secure and HoNOS-LD. These will be mandated for collection using the same information collection protocol as the existing mandated measures. However, there is some anecdotal evidence which questions the psychometric soundness of the measures. This has been expressed in both the clinical setting and previous research reviews (Pirkis, Burgess, Kirk, Dodson, Coombs & Williamson, 2005; Deady, 2010).

A review published in 2005 compiled the research evidence for the psychometric properties of the HoNOS, HoNOS65+ and the HoNOSCA. Pirkis et al., (2005) concluded that the three measures had adequate (or better) psychometric properties and can therefore be regarded as appropriate for routinely monitoring consumer outcomes. However at the time of publication not all of the psychometric properties of the HoNOSCA and HoNOS 65+ had been sufficiently researched and no evidence for the soundness of the HoNOS-LD and HoNOS-secure was considered.

The aims of this report are to:

- Produce a review of the recent research evidence to update and extend the work of Pirkis et al., (2005) by:
    - including research published on the psychometric properties on the HoNOS, HoNOSCA and HoNOS 65+ since 2005
    - including all research on the HoNOS-LD and HoNOS-secure
- Update interested mental health practitioners, policy makers and stakeholders on the recent research evidence in an accessible way and be relevant to the New Zealand context

Method

Establishing validity and reliability

In any service delivery or clinical setting the treatment or rehabilitation process begins with an assessment of the service users' interests, abilities, capacities and needs. This information can also be used for further analysis to monitor and develop services. Given the importance of these two goals, standardized, valid and reliable assessment tools must be used. In the case of outcome measurement in mental health, it is important that clinicians understand what the measures are assessing and are confident that they are indeed measuring this construct in a consistent way. Psychometric properties of instruments and measures are used to determine their quality and appropriateness in the required setting. Psychometric properties are split into reliability and validity.

Validity refers to whether the test is actually measuring what it is intending to measure. Validity testing is concerned with what the test measures, and how well it does this. The three forms of validity are construct, content and criterion validity. All three are important but which form is emphasized depends on the type of test. Reliability is the consistency of a set of items or a measure. Reliability is the extent to which we can be sure that the score received on a test is consistent over time and across conditions. It is used to describe how good the test is at eliminating confounding error. Forms of reliability include; inter-rater reliability, test-retest reliability, and internal consistency. Sensitivity to therapeutic change refers to the ability of the measure to reflect change over time. Reliability and validity are related concepts, a test must reliably measure a construct to be valid, and a test can be reliable but not valid, therefore reliability is needed but not sufficient for validity. A reliable test must be valid to allow inferences to be made of the score and therefore be useful in a clinical, funding or planning setting. Psychometric properties are demonstrated through various tests and analyses which have certain criteria that need to be met if the measure is to be considered valid or reliable.

The following psychometric properties (as modified from Pirkis et al., 2005; McGoey, Cowan, Rumrill and LaVogue, 2010) were used as criteria for critical appraisal.

- Content validity refers to the degree to which a measure reflects the intended area or domain of content. For an instrument to be content valid, its items must be representative of a set of traits the instrument is intent on measuring. Content validity is commonly assessed by having stakeholders or experts on the topic decide on how accurate the items are in representing the topic or construct. Face validity is often thought of as the starting point of content validity. Face validity is a non-statistical judgement on whether or not the test appears to be valid to the test administrator.
- Criterion validity is used to show the accuracy of a measure by comparing it with other measures and standards that reflect the same variable. There are two forms of criterion validity; concurrent and predictive validity. Concurrent validity compares the current measure to an existing measure which has established validity or is currently the 'gold standard'. The predictive validity accesses the measures

ability to measure a future outcome, by looking at the strength and direction of the relationship between the score on the current measure and a score measured sometime in the future on another relevant measure.

- Construct validity is the extent to which a measure is said to accurately and thoroughly measure the construct or trait. It involves understanding the constructs which are to be assessed by the measure, their dimensions, attributes, internal structures of the instrument, and the relationship between different test items. Establishing construct validity can also take into account concurrent and predictive validity as well convergent and discriminant validity. Convergent validity is degree to which the measure correlates with measures it is theoretically predicted to correlate with. Discriminant validity is the degree to which the measure does not correlate with other measures it is predicted to not correlate with.
- Test-retest reliability is a measure of the consistency of an individual's scores across two separate administrations of the test. It is used to demonstrate how stable the construct is over time. It is a useful statistic for reliability for a relatively stable construct. Test-retest reliability can be affected by learning factors, practice and fatigue.
- Inter-rater reliability assesses the consistency between two or more testers, scores or observers that administer the measure. High agreement between the scores obtained by different raters suggests that the measure provides a valid representation of the person's trait that is being measured. Low scores of inter-rater reliability suggest subjectivity with scoring and interpreting results. A correlation of >0.80 is considered to show sufficient reliability.
- Internal consistency measures the similarity of scores of all the items in the measure. It is often employed when an alternate form (giving an analogous measure to correlate scores with) is unavailable. It can be measured by taking the split-half reliability coefficient which is simply spitting the measure in half or inter-item reliability which looks at the extent to which all items are related to each other. Cronbach's Alpha Coefficient can be calculated to demonstrate internal consistency.
- Sensitivity to therapeutic change refers to the ability of an instrument to detect change in the subject or the constructs over time. This is measured by comparing the change detected by the measure in question with another measure that has an established sensitivity.
- Feasibility is the degree to which the measure is acceptable to stakeholders or in this case useful in clinical practice. This is covered in training for the use of the measures in New Zealand and is not included in the review. Te Pou conducted a national survey of the views of clinicians and others towards outcome measures in late 2011. The results of this survey were fed back to individual District health boards (DHB's) and late in 2012 it is anticipated a national version of the survey will be made available. The survey results indicate a high level of clinician acceptance and approval with outcome measures and their collection. Since 2008 New Zealand has been collecting outcome scores in the PRIMHD data warehouse. Te Pou and the Ministry of Health have made available aggregated reports at national, District health board (DHB) and team level which show how outcome measurement can be useful in workforce planning, research, quality initiatives and team planning. There is evidence that outcome measurement can also be useful at the individual level with service users for planning care and allocating caseloads with clinicians and the like (From Data to Information, 2009). In particular the concept of 'clinical significance' in rating the HoNOS family of measures (anything rating 2 or more on the

individual items of the HoNOS family of measures and anything rating 1 or more on the secure ratings for HoNOS secure) can help clinicians to identify clinical priorities.

Procedure

Technical update

This report updates the research review published by Pirkis et al., (2005) by qualitatively summarising the findings on the HoNOS, HONOSCA and HoNOS 65+ that have been published since 2005 and all those on the HoNOS-LD and HoNOS-secure. Overall the methodology employed was similar to that of Pirkis et al., in the sense that it summarised results rather than statistically combining them and used the same criteria to present the evidence for the measures.

Data collection

We used the following strategy to identify studies on the reliability and validity of the HoNOS family of measures.

The following key words were used to identify relevant work: 'HoNOS family measures', 'HoNOS', 'HoNOSCA', 'HoNOS65+', 'HoNOS-LD' and 'HoNOS-secure' were combined with key search terms relating to reliability and validity such as 'psychometric properties', 'psychometric characteristics', 'reliability', 'validity'.

The following databases were searched: Google Scholar, Medline, PsychINFO, PsychEXTRA, Pubmed, Scopus, and Cochrane Databases of Systematic Reviews.

All articles were downloaded and considered unless they were in a different language and no translation was available. Studies on self-report variations of the HoNOS were not considered as this is not the way the measure is used in the New Zealand context; however one was included as it used New Zealand data.

The search produced 20 articles that explicitly assessed psychometric properties of the HoNOS (11), HoNOSCA (2) and HoNOS65+ (4) and all works on the HoNOS-LD (2) and HoNOS-secure (1). Studies were not appraised on their quality but were assumed to have gone through some kind of academic check as they were published in peer-reviewed journals. This was in line with the data collection process that Pirkis et al., (2005) employed.

Results

We report the results separately for each HoNOS measure. For HoNOS, HoNOS 65+, HoNOSCA, HoNOS-secure, HoNOS- LD in the following results section we will cover content, criterion and construct validity; test-retest reliability, inter-rater reliability, internal consistency and sensitivity to change.

HoNOS

Content Validity

As summarised by Pirkis et al., (2005) early studies explored the content validity of the HoNOS by asking service users, carer advocacy groups and mental health professionals to comment on whether items reflected areas of concern for them. Overall, responses were positive suggesting this was an appropriate, well designed measure (Shergill et al., 1999; Orrell et al., 1999; McClelland et al., 2000 as cited in Pirkis et al., 2005).
However, the restriction to indicate only one problem on item number 8 (Other mental health and behavioural problems) was a concern as often there was more than one problem that could have been listed. There was also concern about item number 6's (Problems associated with hallucinations and delusions) ability to accurately describe the symptoms of a person with schizophrenia (Orrell et al, 1999, as cited in Pirkis et al., 2005). The social items (10, 11, 12) were said to be problematic because of the complexity of the information needed to rate them (Orrell et al, 1999; McClelland et al, 2000 as cited in Pirkis et al., 2005). There was a problem of subjectivity in some of the terminology used; this meant that raters may have difficulties knowing which items to use when rating symptoms. Shergill et al., (1999) and Orrell et al., (1999) also observed failure to "take into account factors such as culture, poverty, abuse, safety and risk, bereavement, medication and compliance" (Pirkis et al., 2005, p. 4). McClelland et al, (2000) added that the HoNOS was open to human error and misinterpretation (as cited in Pirkis et al., 2005).

In a recent study, Kisely et al., (2007) tested the HoNOS in routine clinical practice in the US across three pilot sites with both inpatient and outpatients referrals, over a period of four months. They used the HoNOS, the HoNOSCA, and the HoNOS65+ tools as appropriate for different age groups. About 80% of the 5620 patients had at least one rating and overall 60% of patients had more than one rating and were included in the analysis. This study was the first to look at the performance of the HoNOS across all age groups in a way that resembled normal clinical practice. They asked clinicians about their experience with the HoNOS measure using a 5-point Likert scale. The median scores indicated that the HoNOS was easy to use but 80% of clinicians indicated that the item 'Other Symptoms' was hard to use. In terms of usefulness, relevance and sensitivity to change, clinicians found the HoNOSCA glossary less useful compared to the adult version as well as the HONOSCA being less useful in terms of relevant information, sensitivity to change and their social items. This provides some support for the content validly of the HoNOS but suggests that the HoNOSCA is slightly more complicated.

Burgess et al., (2009) carried out an anonymous web-based survey which was completed by 94 outcome measurement experts, most with clinical experience. They were asked to rate the clinical significance of the items, their relative importance in clinical settings and which items they would expect to show change with improvement. Results showed that all items are viewed as clinically significant by these experts and each item was considered by experts to show evidence of a clinically significant problem that requires attention. Overall these findings provide support for the content validity and clinical utility of the HoNOS/65+/CA.

Criterion Validity

Pirkis et al., (2005) list the numerous studies which have assessed the performance of the HoNOS against more established instruments in order to consider the concurrent validity. To summarise, the HoNOS performs well against the following clinician administered instruments; the Role Functioning Scale, Brief Psychiatric Rating Scale, Global Assessment Scale, Life Skills Profile, Manchester Audit Tool, Clifton Assessment Procedures for the Ellery – Behaviour Rating Scale, Clinical Dementia Rating, Mini-Mental State Examination, Schedules for Clinical Assessment in Neuropsychiatry, Broad Rating Schedule, Disability Assessment Schedule, Social Adjustment Scale, Location of Community Support Scale, Social Behavioural Schedule, Hamilton Rating Scale for Depression and the positive and Negative Symptoms Scale (see Pirkis et al., 2005). Low correlations were found between the HoNOS and the Brief Psychiatric Rating Scale in a single study and the Beck Depression Inventory in another. This gives support to the criterion validity of the HoNOS.

The HoNOS showed poor or mixed performance with self or service user-rated instruments such as the Symptom Check List 90-Revised, the Social Adjustment Scale, Medical Outcomes Study Short Form 36, Camberwell Assessment of Need Sort Appraisal Schedule, Quality of Life Scale, Avon Mental Health Measure, Outcome of Problems of Users of Services, an instrument adapted from the Quality of Life Index for Mental Health and the self-rating HoNOS. Exceptions include some moderate correlations between the HoNOS and the Camberwell Assessment of Need Short Appraisal Schedule, Medical Outcomes Study Short Form 36, General Health Questionnaire, and Comprehensive Quality of Life Scale. However these were generally lower than clinician rated measures (see Pirkis et al., 2005). Pirkis et al., (2005) assert that these findings are to be expected considering that observers or test administrators have access to different information than the clinicians. This means that although the HoNOS showed poor performance against these measures, it does not detract from its criterion validity.

More recent studies found mostly moderate correlations of the HoNOS and other measures. Phuaphanprasert, Srisurapanont, Pannarunothai & Geater (2007) report a good correlation ($r>0.80$) with other standard measures. The Thai HoNOS had the highest correlation with the Brief Psychiatric Rating Scale (BPRS), followed by Global Assessment of Functioning (GAF) and Clinical Global Impression (CGI). The Thai HoNOS/GAF and Thai HoNOS/BPRS correlations were higher than McClelland's study where the correlation of HoNOS/BPRS was 0 .72 ($P<.001$, $n=149$). This suggests high concurrent validity. Gigantesco et al., (2007) found that the HoNOS had only moderate correlations with other measures such as the GAF, the Physical Health Index (PHI) and the Life Skills Profile (LSP). Oiesvold, Bakkejord and Sexton (2011) took HoNOS scores from newly admitted patients

and showed that the HoNOS and the SCL-90-R had only moderate correlations. They concluded that the HoNOS and the SCL-90-R measured different phenomena and may be complementary to each other.
The ability of the HoNOS to discriminate between consumer groups was used to test its concurrent validity. According to Pirkis and colleagues (2005) several studies have found high total scores to be associated with various psychopathological and addiction diagnoses. Studies have also shown the HoNOS to discriminate between consumers' levels of need and disability, as indicated by the level of treatment they receive. More recently the HoNOS has been found to discriminate between frequently and non-frequently hospitalized patients as well as severity of diagnosis (as ranked by ICD-10) (Bech et al., 2006). Phuaphanprasert et al., (2007) also found that the Thai HoNOS was able to discriminate between acute patients which had significantly higher scores and non-acute patients (Wilcoxon W = 293, Z= -5.548, $p<0.001$) giving support to the criterion validity.

Gigantesco et al., (2007) carried out a discrimination function analysis on 707 service users which were diagnosed with schizophrenia and assessed by the HoNOS, GAF and the PHI. They were then categorized based on the presence or absence of positive symptoms (categories were; without symptoms, with symptoms, psychological disability but no symptoms, physical disability and no symptoms). All three discriminant functions were significant and over all the classification procedure correctly classified 55.7% of the patients. This suggests that the HoNOS lacks discriminatory ability for some patient groups and is insufficient for routine clinical evaluation. Further, the HoNOS was found to measure the severity of mental illness in the consultation liaison psychiatry setting but less than half of the change in scores was related to the consumers' mental health care. This questions the use of the HoNOS measure in this setting (Duke, 2010). Another recent study showed that the German version of the HoNOS was able to discriminate between patients that had a specific diagnosis and those that did not. Item 7 (Depressed mood) had the biggest difference for patients with an affective disorder and those without, item 3 (Problem drinking or drug taking) in patients with or without substance disorders, item 5 (Physical illness and disability problems) for patients with and without somatoform disorders, item 2 (Non-accidental self-injury) for patients with and without personality disorders (Andreas et al., 2010). This speaks to a larger issue in outcome measurement where by the 'outcome' being measured is dependent on a number of factors and how much is due to each cannot be assessed (From Data to Information, 2009). New Zealand research also found that three versions of the HoNOS (clinician rated, significant other rated, service user rated) compared well to GAF scores (Stewart, 2009)

Pirkis et al., (2005) also found that most studies that looked at predicative validity, found it to be reasonably good. The measure explained a significant proportion of variance in treatment use (e.g., service contact, length of stay) and treatment outcome (readmission rate, retention in community, treatment response). However some studies found limited correspondence between the total scores of the HoNOS and resource use (Goldney, Fisher & Walamsley, 1998; Boot & Andrews, 1997 as cited in Pirkis et al., 2005).

Kisely et al., (2007) showed that the HoNOS was able to differentiate between groups of inpatients and outpatients, with inpatients having significantly larger total scores. They also examined patterns of scores for three age groups (<39, 40-49, >50) and whether they reflected the expected prevalence of symptoms. In their sample, males had significantly higher scores on 'Alcohol and Drugs', by contrast, females had higher ratings on

'Depression'. Consumers over 50 were less likely to use alcohol and drugs than the other groups, but more likely to have higher scores on 'Cognitive Impairment' and 'Physical Problems'. Patients with schizophrenia and other non-affective psychoses scored significantly higher for 'Delusions/Hallucinations' while depressed patients had the highest scores for 'Deliberate Self Harm', 'Depression' and 'Other symptoms'. As expected patients with anxiety scored the highest in 'Other symptoms', where anxiety scores are usually recorded, and patients with personality disorders, had high overall scores, in particular on problems with 'Aggression' and 'Relationships'. This gives support for the predictive validity of the HoNOS.

A follow up study by Kisely, Campbell, Cartwright, Cox & Campbell (2010) obtained at least one rating from 4620 patients in an outpatient facility. They found that the HoNOS and HoNOSCA had satisfactory sensitivity to therapeutic change and predictive validity for routine clinical use. They obtained at least one rating from 4620 patients with either the HoNOS or HoNOSCA. Using global and individual scores they found good predictive validity in terms of service contacts, lengths of stay, readmission rates and retention in the community.
They found that the scores were related to service use which provides support for the predictive validity of the measures. For example, the HoNOS scores showed significant associations with health service use and those with high scores were likely to be readmitted within the year.

Construct Validity

The original factor structure proposed by Wing et al., (1998) of the HoNOS contained the following four subscales: (1) subscale 'Behaviour' (Items 1 'Overactive, aggressive, disruptive and agitated behaviour', 2 'Non-accidental self-injury', 3 'Problem drinking or drug taking), (2) subscale 'Impairment' (items 4 'Cognitive problems' and 5 'Physical Disability'), (3) subscale 'Symptoms' (items 6 'Hallucinations and delusions', 7 'Depressed mood' and 8 'Other mental and behavioural problems'), and (4) subscale 'Social' (items 9 'Problems with relationships 11 'Problems with living conditions', 12 (Problems with occupation and activities'). There have been several studies done but no support has been found for this factor structure.

Examining the subscale structure of the HoNOS, Preston (2000, as cited in Pirkis et al., 2005) found that the original model with four factors was a good fit, however, the contribution of individual items to their subscales varied in two different mental health services. This suggested that the construct was interpreted in different ways across the sector. Trauer's (1999, as cited in Pirkis et al., 2005) examination of the subscale structure had a poorer fit than Preston's, which led him to propose an alternative five-factor structure. This was supported by Eagar et al., (2005 as cited in Pirkis et al., 2005). McClelland et al., also identified alternative factors (as cited in Pirkis et al., 2005).

Consistent with results of previous studies the German version of the HoNOS was unable to confirm the original four factor structure of the measure, via subscales or the total scores (Andreas et al., 2010). Alternative models that have been proposed by other researchers but these have also been contested. A five factor structure proposed by Eagar, Trauer & Mellsop (2005 as cited in Pirkis et at., 2005) has been supported by others and can be used in the reporting of the PRIMHD data base in New Zealand. Lovalglio and Monzani (2011) looked at the factor structure of the Italian HoNOS and concluded that it did not measure a single, underlying construct of

mental health status. This suggests that the instrument is more multi-dimensional and might not be a clinically valid tool. However, scores suggested sufficient reliability and precision which may make it a good research tool. Overall there is still debate over the factor structure of the HoNOS. However, as total scores, individual scores and sub-scale scores are used to obtain a representation of the complicated and multi-dimensional construct of mental health status, this does diminish the overall validity of the HoNOS.

Convergent validity of single items was shown in a study of the German version of the HoNOS (Andreas et al., 2010a). This was the first study to look at single items as well as total scores. They showed that the average convergent correlations with other clinician administered scales were only 0.44 and therefore not supporting convergent validity. However there was support for divergent validity with self and clinician administered, non-corresponding scales producing low correlations between 0.02-0.21. They showed through a regression analysis that the single items of the HoNOS were the most important predictors of service utilization.

Test-retest Reliability

Pirkis et al., (2005) reported that the limited data on test-retest reliability shows 'fair to moderate' overall reliability scores. Items 1 (Over reactive, aggressive, disruptive and agitated behaviour), 3 (Problem drinking or drug taking), 7 (Problems with depressed mood) and 10 (Problems with activities of daily living) have been reported to be most problematic as the scores for these items have shown inconsistencies between administrations.

Inter-rater Reliability

Most studies have shown moderate agreement between raters, according to Pirkis et al., (2005). However, items 4 (Cognitive problems), 7 (Problems with depressed mood), 8 (Other mental health and behavioural problems), 9 (Problems with relationships), 11 (Problems with living conditions), and 12 (Problems with occupation and activities) have been named in some studies as most problematic as the scores for these items have shown changes across raters.

More recently, inter-rater reliability was found to be satisfactory to excellent as the study had interclass correlation coefficients for inter-rater reliability between 0.75 and 0.98. The raters were two nurses who conducted the assessment together but completed individual scoring (Phuaphanprasert et al., 2007). Kisely et al., (2007) also looked the inter-rater reliability of the HoNOS with 68 clinical staff as the raters. All the participating staff had training and an assessment of inter-rater reliability showed good inter-rater reliability coefficients (ICC) values with the use of case-vignettes instead of live service users.

Andreas et al., (2010a) were able to show that inter-rater reliability was able to be improved by training. After training, all items showed satisfactory ICC values (ICC > 0.6), improving on the low values obtained for some items prior to training. This is contradictory to other studies such as Rock and Preston (2001) who found that all four test groups showed adequate inter-rater reliability. These were mental health nurses, with standard or modified training, and those with no clinical experience or experience with other scales but not patients.

Internal Consistency

Studies assessing the internal consistency of the HoNOS items, as summarised by Pirkis and colleagues (2005), produced Cronbach's Alpha values ranging from 0.59-0.76. This indicates moderate internal consistency and low item redundancy. (Wing et al, 2000; Shergill, 1999;Orrell et al., 1999; McClelland et al., 2000; Stedman, et al., 1997; Trauer, 1999; Page et al., 2001; Eagar et al., 2005 as cited in Pirkis et al., 2005). However, Trauer (1999 as cited in Pirkis et al., 2005) has argued that the HoNOS does not measure a single underlying construct of mental health status and proposed an alternative subscale structure.

McClelland (1999) found that Items 7 (Depressed mood), 8 (Other mental and behavioural problems) and 9 (Problems with relationships) had the biggest contribution to the total score (15%, 19% and 14% respectively). Items 11 (Living conditions) and 12 (Occupation and activities) had the least contribution with only a 3% contribution for each. More recently Phuaphanprasert and colleagues (2007) report a Cronbach's alpha coefficient of 0.68 which falls just under the 0.70 threshold for acceptable internal consistency. Andreas et al., (2010b) produced a value of 0.62 and Oiesvold et al (2011) showed that the internal consistency of the HoNOS sum score and the SCL-90-R-GSI was satisfactory for both testing times giving support to this psychometric property.

Sensitivity to Therapeutic Change

A number of studies have examined the direction and movement of the HoNOS scores over time. Simple studies examined the change in HoNOS scores over time in a given sitting, hypothesizing that there should be a decrease in severity as a consumer reaches the end of an episode. These studies have found the greatest decrease in magnitude in inpatient settings and lesser magnitude in community settings (McClelland et al., 2000 cited in Pirkis et al., 2005). There is some evidence that there may be an interaction between setting, diagnosis and severity, and the HoNOS may be able to detect change in community settings for those with depression and anxiety (Adams et al., 2000 cited in Pirkis et al., 2005) and those with high scores at the start of the episode (Parabiaghi et al., 2005 as cited in Pirkis et al., 2005). Some items have also been found to interact with the setting. Example item 11 has been found to not decrease with the rest of the items in a community setting (McClelland et al., 2000 cited in Pirkis et al., 2005). Another study found that Items 7 (Problems with depressed mood), 8 (Other mental health and behaviour problems), and 9 (Problems with relationships) were the only ones that had relevance and variability over time in the same setting (Audin et al., 2001 cited in Pirkis et al., 2005).

Sensitivity has also been tested by using clinician or consumer judgement about whether change has occurred and to see if this has been reflected in the scores of the HoNOS. Three studies found correlations between clinical judgement or self-report rating of condition improvement, stability or deterioration and the HoNOS scores between initial and repeat ratings (Taylor &Wilkinson, 1997; Gallagher & Teesson, 2000; Hunter et al., 2004 cited in Pirkis et al., 2005).

Studies have also explored the capacity of the HoNOS to detect change compared to more established outcome measures. The HoNOS has been found to perform commensurately with the Global Assessment Scale, the Brief Psychiatric Rating Scale (McClelland et al., 2002 cited in Pirkis et al., 2005), the Modified Clinical Global Impressions Scale (Sharma et al., 1999 cited in Pirkis et al., 2005), the Clifton Assessment of Strengths, Interests and Goals amongst other quality of life scales designed for geriatric patients (Ashaye et al., 1999 cited in Pirkis et al., 2005). By contrast, it was found to perform less well compared to the Schedules for Clinical Assessment in Neuropsychiatry and the Social behaviour Schedule (Bebbington et al., 1999 cited in Pirkis et al., 2005).

Another way that sensitivity was tested was to assess if improvement reflected in the HoNOS scores was associated with consumers receiving evidence-based therapies which have been shown to reduce symptom severity. Bech and colleagues (2003 cited in Pirkis et al., 2005) correctly hypothesized that patients who receive lithium and/or ECT would show greater improvement on the HoNOS than ones who did not.

A new study by Kisely and colleagues (2010) found that the HoNOS and HoNOSCA had satisfactory sensitivity and predictive validity for routine clinical use. They obtained follow up ratings from 599 patients with the HoNOS in an outpatient setting. The findings for sensitivity of change are consistent with findings from previous studies where changes in global score are more consistent than with the sub-scores. The authors suggest that the HoNOS may be better able to detect changes in consumers with higher baseline scores, older adults, inpatients, and adult outpatients with depression and anxiety. It appeared less useful in detecting change in psychotherapy outpatients.

Kisely et al., (2007) had previously examined the sensitivity to change of the HoNOS by obtaining scores on two or more occasions, four months apart. There were significant decreases in the total HoNOS scores for inpatients and outpatients reflecting improvement across time. Significant change in individual items also showed that items measuring 'Deliberate self-harm', 'Depression' and 'Other symptoms'. For HoNOSCA, there were changes in items measuring 'Aggression', 'Over activity', 'Alcohol/Drug use', 'Non-Organic Somatic Symptoms', 'Emotional Disorders' and 'School attendance'. Andreas et al., (2010b) also found sensitivity to be moderate to sufficient, with items 7, 8, 9 and 10 showing the most change and items 2 ('Non-accidental self-injury'), 6 ('Problems with hallucinations and delusions') and 11 ('Problems with living situation') showed no change.

HoNOS65+

Content Validity

During initial development of the HoNOS65+, Burns and colleagues (1999) asked mental health professionals working with older consumers to review the content of the HoNOS. Some modifications to the Glossary were made in order to improve the comprehensiveness for older consumers. Since then, issues have been noted and refinements made (Burns et al., 1999; Allen et al., 1999; Macdonald, 1999).

The HoNOS65+ measure was also successfully used in a retrospective analysis which showed effective treatment interventions in an inpatient psychogeriatric unit showing it is suitable and appropriate for this setting (Cheung & Strachan, 2007). There was also some support provided by the Burgess et al., (2009) study (See HoNOS Content Validity)

Criterion Validity

Pirkis et al., (2005) reported that there were no studies available for predictive validity. In a recent study, Canuto et al., (2009) have assessed the sensitivity to treatment of the HoNOS65+F (French version) compared to five routinely used scales. Thirty elderly patients with ICD-10 diagnosis of dementia and depression were evaluated at admission and discharge using paired sample t-tests. The 'depressive moods' item from the Brief Psychiatric Rating Scale (BPRS) was used as a gold standard and a receiver operating characteristics curve assessed the HoNOS65+F 'depressive symptoms' item score changes. Clinical improvement at discharge was reflected in significant changes to total scores for the HoNOS65+F, BPRS and Global Assessment Functioning Scale. BPRS has previously shown to have good reliability and validity, supporting the use of the HoNOS65+F. The Geriatric Depression Scale (GDS), Mini Mental Health Examination and Activities of Daily Living scores did not change significantly from admission to discharge. This fits with previous research which suggests that the GDS ability to detect depression decreases with moderate dementia.

Of the HoNOS65+F items, 'Behavioural disturbance', 'depressive symptoms', 'activities of daily life' and 'drug management' showed high and significant changes between admission and discharge. The 'depressive symptoms' item was also shown to classify 93% of the cases with good sensitivity (0.95) and specificity (0.88). This data suggests that the 'depressive symptoms' items of the HoNOS65+F may provide a valid assessment of depressive symptoms in dementia patients and overall provides support for the discriminatory power of the HoNOS65+F.

Several studies have examined correlations between the HoNOS65+ and other clinician administered measures in similar domains. Reasonable correlations have been found between the HoNOS65+ and the Mini-Mental State Examination, Crighton Royal Behavioural Rating Scale and Barthel Activities of daily living. Several items have been found to have stronger correlations with specific scales. There have been conflicting results with the correlation of the Geriatric Depression Scale and the HoNOS65+ (Pirkis et al., 2005).

There is little evidence of the HoNOS65+ having ability to discriminate between different consumer groups. However some studies have shown that the HoNOS65+ can discriminate between those with dementia and those with functional psychiatric disorders, scoring higher on items 1 (Behavioural Disturbance), 4 (Cognitive Problems) and 10 (Problems with activities of daily living) and generally having higher total scores than those with mood disorders but lower scores on the symptoms subscale (Pirkis et al., 2005).

Construct Validity

The only evidence about the construct validity of the HoNOS65+ is from the original pilot work by Burns et al.,(1999 as cited in Pirkis et al., 2005) where factor analysis showed that a four factor model accounted for 57.4% of the variance in the item scores. More work needs to be done to demonstrate construct validity.

Test-retest Reliability

There is currently no research that explores this psychometric property.

Inter-rater Reliability

According to Pirkis and et al., (2005), two studies found inter-rater reliability to be good to very good, with only items 2, 10, 11 and 12 not performing well consistently in one study and items 4, 5 and 9 in the other (Burns et al, 1999; Spear et al., 2002). A third study found a broader range of items had inter-rater reliability problems, and thought this was related to difficulties in interpretation (Allen et al., 1999 as cited in Pirkis et al., 2005).

Internal Consistency

There is currently no research that explores this psychometric property.

Sensitivity to Therapeutic Change

Spear et al., (2002) found that consumers improved on HoNOS65+ subscales and total scores between assessment and discharge from inpatient and community services. The total score and change in scores on the HoNOS65+ showed moderate but significant correlations with the Clinicians Interview Based Impression of Change Scale (Pirkis et al., 2005).

HoNOSCA

Content Validity

No studies were available at the time of the Pirkis et al., (2005) review but there was also some support provided by the Burgess (2009) study and Kisely et al., (2007) (See HoNOS Content Validity).

Criterion Validity

According to Pirkis et al., (2005) several studies have compared the HoNOSCA's performance against other instruments in an attempt to establish concurrent validity. In general, correlation of total scores between the HoNOSCA and other clinician-rated measures are moderate ($r < 0.6$). Some examples were the Children's Global Assessment Scale, the Paddington Complexity and the Global Assessment of Psychosocial Disability.

Studies that have compared the HoNOSCA against parent and child/adolescent-rated instruments have usually produced weaker correlations. Yates et al., (1999 as cited in Pirkis et al., 2005) found modest correlations between the HoNOSCA and the Behaviour Check List, Strength and Difficulties Questionnaire, Child Health Related Quality of Life Questionnaire and Modified Harter Self-Esteem Questionnaire. Further, low agreement was found between the HoNOSCA and a consumer rated version (Gowers et al., 2002 as cited in Pirkis et al., 2005). According to Pirkis et al., (2005) this was to be expected (and therefore adds to criterion validity), as outcome measures rely on different classes of informants and are unlikely to correlate when their informants are from a different class.

Studies have assessed the ability of the HoNOSCA to discriminate between groups of service users based on their clinical profile. The HoNOSCA can distinguish between child and adolescents that were either in inpatient and outpatient settings or between those presenting to clinics with different areas of focus (Gowers et al., 1999; 2000; Yates et al., 1999 as cited in Pirkis et al., 2005). The HoNOSCA total scores were associated with the number of critical incidents consumers experienced (Harnett et al, 2005 as cited in Pirkis et al., 2005) and reflected a number of gender and age differences. For example boys scored higher than girls on item 1 (Problems with disruptive, antisocial or aggressive behaviour), but lower on item 9 (Problems with emotional and related symptoms), younger children scores higher than older children on Item 5 (Problems with scholastic and language skills) but lower on Item 3 (Non-accidental self-injury). Overall the HoNOSCA results intuitively reflected diagnoses, for example; conduct and attention deficit disorders scored highest on Items 1 and 2 (Problems with disruptive, antisocial or aggressive behaviour and Problems with over-activity or attention or concentration). Overall high scores were found to be associated with comorbid diagnoses.

HoNOSCA total scores at community assessment could discriminate between adolescents who later received more intensive care from those who progressed to other forms of community care, providing support for the predictive validity of the measure (Brann, 2005 as cited in Pirkis et al., 2005). The patterns of scores for the HoNOSCA showed a significant difference between the sexes on 'Emotional Disorders' including higher scores for anxiety and depression for females than males. Also, consumers over 12 years old were more likely than the younger patients to have self-harmed or used drugs and alcohol. This gives some support for the predictive validity of the HoNOSCA (Kisely et al., 2007).

A new study by Kisely et al., (2010) found that the HoNOS and HoNOSCA had satisfactory sensitivity and predictive validity for routine clinical use. They obtained at least one rating from 4620 patients with either the

HoNOS or HoNOSCA. They found good predictive validity in terms of service contacts, lengths of stay, readmission rates and retention in the community.

Construct Validity

Development studies on the HoNOSCA examined the subscale structure of the HoNOSCA considering individual items and subscales. Correlations between individual items were found to be low, which was taken as evidence that each item carried independent weight (Gowers et al., 1999; 200; Harnett et al., 2005 as cited in Pirkis et al., 2005). The factor structure of the HoNOSCA was also found to mirror the subscales. Another study however, examined the factor structure and found some evidence for a different set of factors (Brann, 2005 as cited in Pirkis et al., 2005).

These studies tested the extent to which the HoNOSCA total score reflected clinical severity. They found the total score increased as a linear function of high individual item scores (Gowers et al., 1999; 2000; Brann et al., 2005 as cited in Pirkis et al., 2005).

Test-retest Reliability

At the time of the Pirkis review there was little published work on the test-retest reliability of the HoNOSCA, but there was some evidence from studies which tested sensitivity to change or lack of change of the instrument. One study looked at change over a 6 month period for a group of consumers which clinicians reported no change on a global rating scale and reported a correlation of 0.69 (Garralda et al., 2000). Similarly, another study reported correlations of 0.80 over 3 months and 0.76 over five-months (Brann, Forthcoming as cited by Pirkis et al., 2005). These should not really be considered as evidence for test-retest reliability as the time interval is too long and we would expect change to occur during this time. Another study reported a correlation of 0.80 between initial and follow up scores over a 2-4 week period where adolescents were admitted as inpatients and their scores were expected to remain stable during a settling in period (Harnett et al., 2005).

Inter-rater Reliability

Pirkis et al., (2005) report that inter-rater reliability has been good to very good for the majority of section A. However there is little agreement about which items perform poorly. One study noted a low intra-class correlation (0.06) for item 10, and another reported a moderate to high correlation (0.77). As for section B there is debate for its inter-rater reliability, some reporting high intra-class correlations for the two items and later studies reporting much lower figures.

The Kisely et al., (2007) study provides some support for the Inter-rater reliability of the HoNOSCA as this measure was used for their younger participants and over-all the study (which utilised HoNOS, HoNOSCA and HoNOS 65+) found support for inter-rater reliability.

Internal Consistency

There is currently no research that explores this psychometric property.

Sensitivity to Therapeutic Change

From several attempts to determine the sensitivity of the HoNOSCA the weakest has been simply determining if the total scores change over time, with no reference to whether or not this reflects real change. Several studies observed reductions in scores (Gowers et al., 1999; 2000; Manderson & McCune, 2003; Harnett et al., 2005 as cited in Pirkis et al., 2005).

Other studies have also demonstrated that the HoNOSCA total scores had similar changes in direction and magnitude compared to other clinician-rated measures such as the Children's Global Assessment Scale and the Global Assessment of Psychosocial Disability but to a lesser extent to parent and/or service user rated measures such as the HoNOSCA-SR, and the Behaviour Check list and the Strengths and Difficulties Questionnaire (Pirkis et al., 2005).

Another approach was to compare global outcome judgments, where a clinician or parent or other referrer judges whether the clients has improved, remained stable or deteriorated and compared this to HoNOSCA scores of the clients. Several studies have found correlations between change reported on this global measure and the HoNOSCA (Pirkis et al., 2005).

A new study by Kisely et al., (2010) obtained follow up ratings from 209 service users with either the HoNOS or HoNOSCA in an outpatient setting. The findings for sensitivity of change are more modest than the HoNOS but still showed significant change even with the small numbers. The results reflect previous findings where change in HoNOSCA scores was greater in the items which measure symptoms and behaviour rather than social impairment.

HoNOS-secure

Content Validity

There is only one empirical study on the HoNOS-secure and therefore limited information on which to draw conclusions. The single study suggests that the HoNOS-secure has sufficient reliability by looking at the inter-rater reliability and internal consistency. This is a good start in validating the measure and calls for more research (Dickens, Sugarman & Walker, 2007).

Criterion Validity

Another study took data collected by the HoNOS and HoNOS-secure to introduce two performance indicators of clinical effectiveness. Results show consistent 90-day improvement rates and increasing stability over time. The results appear valid as changes were in the predicted direction. This study shows that using the HoNOS and HoNOS-secure as an outcome measure to demonstrate clinical effectiveness is appropriate (Sugarman, Walker & Dickens, 2009).

Construct Validity

There is currently no research that explores this psychometric property.

Test-retest Reliability

There is currently no research that explores this psychometric property.

Inter-rater Reliability

60 inpatients were rated independently by two clinicians; there were 34 different raters in total. The ICCs for six of the seven security items indicated at least a moderate agreement, and the one item fair agreement. The ICC for the 12 HoNOS items indicated fair to substantial consistency between raters (Dickens et al., 2007).

Internal Consistency

The Cronbach's alpha values were 0.73 for the security scale and 0.79 for the HoNOS scale, indicating acceptable internal reliability (Dickens et al., 2007).

Sensitivity to Therapeutic Change

There is currently no research that explores this psychometric property.

HoNOS-LD

Content Validity

The HoNOS-LD was developed so that those with learning disabilities, who are also subject to the same range of mental illnesses and psychological ill health as the general population, have a suitable outcome measure to improve service effectiveness. Initial development work was done by Roy et al., (2002) and their findings support the reliability and validity of the measure. They piloted the HoNOS-LD at 26-sites and with 372 consumers.

Acceptability studies conducted by Roy et al., (2002) showed that clinicians at their pilot sites agreed that the language used was appropriate and the instrument was easy to understand and use. Some suggested items could be added, few thought the measure should be reduced. This was seen as support towards content and construct validity.

Criterion Validity

There is currently no research that explores this psychometric property.

Construct Validity

Roy et al., (2002) showed that the HoNOS-LD has a good correlation with the well-established instrument, the Aberrant Behavioural Checklist. This supports convergent validity as they had raters complete both the HoNOS and Aberrant Behaviour Checklist on the same clients and found high correlations at two different times.

Test-retest Reliability

There is currently no research that explores this psychometric property.

Inter-rater Reliability

364 raters assessed 327 consumers. The raters were of different professions in mental health care, these were; clinical psychologist, nurse, occupational therapist, psychiatrist, speech and language therapist and support worker. Peasons correlations showed a high degree of correlation between raters and Cohen's kappa calculations also showed high levels of significant correlation providing support for the inter-rater reliability of the HoNOS-LD. The inter-rater reliability was found to be high with high correlations between the different groups of professionals. However the authors suggest that an informant who knows the client well is necessary when administering the HoNOS-LD and completing the measure purely from reviewing case notes was shown to be less reliable (Roy et al., 2002).

Internal Consistency

There is currently no research to support this psychometric property.

Sensitivity to Therapeutic Change

Roy and colleagues (2002) conducted a paired sample T-Test and shows that scores between time one and two as measures by the HoNOS-LD had reduced as expected. Therefore the instrument is useful in measuring change.

Results Summary

In the following table the 'Adequate' means that there is sufficient research evidence to suggest that the at least minimal requirements (statistical or otherwise) have been met for that particular psychometric property. 'Good' means that criteria were met and exceeded. Insufficient evidence means there is a gap in the research evidence for this psychometric property and more research needs to be done. There was no evidence to suggest that any psychometric property was violated.

Table 1. *Summary table of Psychometric properties for the HoNOS family of measures*

	HoNOS	HoNOS65+	HoNOSCA	HoNOS-LD	HoNOS-Secure
Content Validity	Good	Adequate*	Adequate	Adequate*	Adequate*
Criterion Validity	Good	Adequate	Good	Insufficient evidence	Insufficient evidence
Construct Validity	Good but issues with factor structure	Adequate	Adequate	Adequate*	Insufficient evidence
Test-Retest reliability	Adequate	Insufficient Evidence	Adequate	Insufficient evidence	Insufficient evidence
Inter-Rater Reliability	Adequate	Adequate	Good	Adequate*	Adequate*
Internal Consistency	Adequate	Insufficient evidence	Insufficient evidence	Insufficient evidence	Adequate*
Sensitivity to change	Adequate	Adequate	Good	Adequate	Insufficient evidence

*Emerging evidence, based on a small amount of research, more is needed

Discussion

There is strong evidence about some of the psychometric properties for the HoNOS family of measures (designated by good and adequate in the above table) but there are also some substantial gaps. The summary table shows where the evidence supports the use of the measure and where the gaps are.

Overall, the evidence for content validity of the measures is strengthening. As part of the criterion validity, concurrent validity has shown good correlations with similar measures and the HoNOS has been shown to be able to discriminate between different patient groups. Recently, studies evaluating predictive validity have given evidence further supporting criterion validly of the HoNOS. This is the case for all the five measures, although the HoNOS-secure and the HoNOS-LD only have development research to rely on currently.

The construct validity of the HoNOS has come into question with several researchers examining the original factor structure and proposing variations. The HoNOS may be multi-dimensional rather than simply measuring one underlying construct of 'mental health status'. A sub-scale structure has not been agreed upon although there is support for both four-factor and five-factor models. However comparing the psychometric properties of single items has proved useful suggesting single item scores rather than total score might be useful in obtaining a picture of mental health status. This conclusion is the same for the other measures and is a characteristic for the whole family of HoNOS measures that have had their factor structure looked at (HoNOS, HoNOSCA, HoNOS65+). Reflecting this, internal consistency was found to be moderate with more research needed. The little research that was conducted on the convergent and divergent validity of the HoNOS gave support for these psychometric properties. These issues with construct validity are to be expected as the underlying construct of mental health and social functioning is complicated and difficult to define and measure.

Over all test-retest and inter-rater reliability was found to be satisfactory. However there was uncertainty in the literature of whether training was needed to improve inter-rater reliability, with studies showing different results. There was difficulty with interpretation of some items of the HoNOS65+ which could affect inter-rater reliability. According to the evidence, the HoNOS family of measures are sensitive to change.

Conclusion

In conclusion, there has been a slow but valuable trickle of evidence regarding the psychometric properties of the HoNOS family of measures published since Pirkus et al.'s 2005 review. This evidence extends the earlier evidence that has largely supported earlier findings that the HoNOS outcome measure is a reliable and valid tool that can be used to assess various aspects of mental health status. However, more research is needed to determine the factor structure of the HoNOS and therefore how best to interpret total score or individual item scores. The HoNOSCA and the HoNOS65+ measures have less research than the HoNOS measure but they have performed well in the majority of the studies, both prior to and following 2005. They share the same concern for their factor structure, just like the HONOS. However, further research on these measures would be valuable. In particular, the test-retest reliability of the HoNOS65+ needs more research evidence as does the internal consistency of the HoNOSCA. More research is particularly needed to determine if the HoNOS-secure and the HoNOS-LD have sound psychometric properties. The current research is indicative of reasonable reliability and validity, but more evidence, particularly obtained in settings independent of the development processes for these measures would be useful.

Despite these limitations, the overall conclusion which can be drawn from this review is that, while there are still areas in which further evaluation of the psychometric properties of these measures could be undertaken, the preponderance of evidence suggests that all the HoNOS family of measures do have adequate reliability, validity, and sensitivity to therapeutic change to be useful in clinical settings in New Zealand. All the measures have shown some evidence of reliability and validity, and some evidence suggestive of sensitivity to therapeutic change. This supports the adoption of these measures by the New Zealand mental health services.

Throughout training with the HoNOS measures, and in many of the studies of the psychometric properties of these measures, the issue of maintaining fidelity to the use of the rating criteria as presented in the Rating Glossaries is frequently emphasised. Irrespective of how good the psychometric properties of the HoNOS family (or any other) measures are, if staff who are using them do not rate according to the established criteria the validity and utility of the resulting data will not be maintained. It is therefore important to ensure that steps to facilitate fidelity with rating to standard criteria are undertaken. Strategies to facilitate this include training (and refresher training) that emphasises use of the criteria and glossaries, ensuring that glossary information is readily available for review when people are completing the measures, discussion of the results of the measures as part of routine clinical review, and feeding back meaningful analyses of team aggregated data to the clinicians (See From Data to Information, 2009).

To date, very little psychometric evaluation of the HoNOS family of measures has been undertaken in New Zealand. While the international literature gives a solid foundation for accepting the reliability and validity of these measures, it would be valuable to also accumulate local evidence, including evidence regarding the reliability and validity of these measures with Maori, Pacific people, and other ethnic groups common in New Zealand. As a large amount of HoNOS family of measures data is already being collected in New Zealand, there is good scope for conducting this kind of research here. With the introduction of the HoNOS-secure and

HoNOS-LD there is a particular opportunity to augment the relatively scant evidence currently available internationally. Services in New Zealand may want to consider establishing projects to undertake such psychometric evaluations.

References

Andreas, S., Harries-Hedder, K., Schwenk, W., Hausberg, M., Koch, U. & Schulz, H. (2010a). Is the Health of the Nation Outcome Scales appropriate for assessment of symptom severity in patients with substance-related disorders? *Journal of Substance Abuse Treatment, 39*, 32-40.

Andreas, S., Harfst, T., Rabung, S., Mestel, R., Schauenburg, H., Hausberg, M., Kawski, S., Koch, U. & Schulz, H. (2010b). 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. *Int J Methods Psychiatr Res, 19*(1), 50-62.

Bech, P. Bille, J. Waarst, S. Wiese, M. Borberg, L. Treufeld, P. & Kessing, L. (2006) Validity of HoNOS in identifying frequently hospitalised patients with ICD-10 mental disorders. *Acta Psychiatr Scand.* 113, 485-491.

Brann, P. & Coleman, G. (2010) On the meaning of change in a clinician's routine measure of outcome: HoNOSCA. *Australian and New Zealand Journal of Psychiatry*, 44, 1097-1104.

Burgess, P., Trauer, T., Coombs, T., McKay, R. & Pirkis, J. (2009) What does 'clinical significance' mean in the context of the Health of the Nation Outcome Scales? *Australas Psychiatry, 17*(2), 141-8.

Canuto, A., Rudhard-Thomazic, V., Herrmann, F.R., Delaloye, C., Giannakopoulos, P. & Weber, K. (2009) Assessing depression outcome in patients with moderate dementia: sensitivity of the HoNOS65+ scale. *J Neurol Sci, 283*(1-2), 69-72.

Cheung, G. & Strachan, J. (2007) Routine 'Health of the Nation Outcome Scales for elderly people' (HoNOS65+) collection in an acute psychogeriatric inpatient unit in New Zealand. *N Z Med J,* 120(1259):U2660.

Deady, M. (2009) *A review of Screening Assessment and Outcome measures for Drugs and Alcohol Settings.* Network of Alcohol & Other Drugs Agencies.

Dickens, G., Supgarman, P. & Walker, L. (2007) HoNOS-secure: A reliable outcome measure for users of secure and forensic mental health services. *The Journal of Forensic Psychiatry & Psychology, 18*(4), 507-514.

Duke, B. (2010) HoNOS in the consultation liaison psychiatry setting: is it valid? *Australas Psychiatry, 18*(6), 547-50.

Gigantesco, A., Picardi, A. de Girolama, G. & Morosini, P. (2007) Discriminant Ability and Criterion Validity of the HoNOS in Italian Psychiatric Residential Facilities. *Psychopathology*, 40, 111-115.

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. *Community Mental Health Journal*, 43, 91-107.

Kisely, S., Campbell, L.A., Cartwright, J., Cox, M., & Campbell J. (2010) Do the Health of the Nation Outcome Scales measure outcome? *Canadian Journal of Psychiatry, 55*(7), 431-9.

Lovaglio, P. G. & Monzani, E. (2011) Validation aspects of the health of the nation outcome scales. International *Journal of Mental Health Systems, 5*(20), 1-11.

McGoey, K. E., Cowan, R. J., Rumrill, P. P. & LaVogue, C. (2010) Understanding the psychometric properties of reliability and validity in assessment. *Work,* 36, 105-111.

Oiesvold, T., Bakkejord, T. & Sexton, J. A. (2011) Concurrent validity of the Health of the Nation Outcome Scales compared with a patient-derived measure, the Symptom Checklist-90-Revised in out-patient clinics. *Psychiatry Research*, 187, 297-300.

Preston, N.J. (2000) The Health of the Nation Outcome Scales: Validating factorial structure and invariance across two health services. *Australian & New Zealand Journal of Psychiatry, 34*(3), 512-519.

Phuaphanprasert, B., Srisurapanont, M., Silpakit, C., Pannarunothai, S., Udomratn, P., Geater, A. & Prapaphomrn, P. (2007) Reliability and validity of the Thai version of the Health of the Nation Outcome Scales (HoNOS). *J Med Assoc Thai, 90*(11), 2487-93.

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 Qual *Life Outcomes, 28*(3), 76.

Roy, A., Matthews, H., Clifford, P., Fowler, V., & Martin, D. (2002) Health of the Nation Outcome Scales for People with Learning Disabilities (HoNOS-LD). *British Journal of Psychiatry*, 180, 61-66.

Stewart, M. (2009). Service user and significant other versions of the Health of the Nation Outcome Scales. *Australian Psychiatry, 17*(2).

Sugarman, P., Lorraine, W. & Dickens, G. (2009) managing outcome performance in mental health using HoNOS: Experience at St Andrew's Healthcare. *Psychiatric Bulletin,* 33, 285-288.

Te Pou (2009). From Data to Information: Data use guidelines for standard measures collected in the New Zealand mental health system. Auckland

Wing, J.K., Beevor, A.S., Curtis, R.H., Park, S.B.G., Hadden, S. & Burns, A. (1998) Health of the Nation Outcome Scales (HoNOS): Research and development. *British Journal of Psychiatry*, 172, 11-18.

Wing, J.K., Lelliott, P., & Beevor, A.S. (2000) Progress on HoNOS. *British Journal of Psychiatry*, 176, 392-393.

Glossary

HoNOS: Health of the Nation Outcome Scales

HoNOSCA: HoNOS for use with children and adolescents

HoNOS 65+: HoNOS for use with adults 65 and over

HoNOS-secure: For use with individuals in a secure setting

HoNOS LD: The HoNOS outcome measure for use by those with learning disabilities

AUCKLAND
Level 2, 8 Nugent Street (B), Grafton
PO Box 108-244, Symonds Street
Auckland 1150, NEW ZEALAND
T +64 (9) 373 2125 F +64 (9) 373 2127

HAMILTON
293 Grey Street, Hamilton East
PO Box 219, Waikato Mail Centre
Hamilton 3240, NEW ZEALAND
T +64 (7) 857 1202 F +64 (7) 857 1297

WELLINGTON
Level 3, 147 Tory Street
PO Box 6169, Marion Square
Wellington 6141, NEW ZEALAND
T +64 (4) 237 6424 F +64 (4) 238 2016

CHRISTCHURCH
21 Birmingham Drive, Middleton
PO Box 22105, High Street,
Christchurch 8142, NEW ZEALAND
T +64 (3) 339 3782 F +64 (3) 339 3783

The NATIONAL CENTRE *of* MENTAL HEALTH RESEARCH, INFORMATION *and* WORKFORCE DEVELOPMENT

www.tepou.co.nz

Te Pou o Te Whakaaro Nui

MENTAL HEALTH PROGRAMMES LIMITED TRADING AS TE POU