# Assessing the content validity of the revised Health of the Nation Outcome Scales (HoNOS 2018)

**Meredith Harris[1], Caley Tapp[1], Urska Arnautovska[1], Tim Coombs[1], Rosemary Dickson[1], Mick James[2], Jon Painter[3], Mark Smith[4], Angela Jury[4], Jennifer Lai[4], Philip Burgess[1]**

[1] Australian Mental Health Outcomes Classification Network (AMHOCN), Australia
[2] Royal College of Psychiatrists, United Kingdom
[3] Sheffield Hallam University, United Kingdom
[4] Te Pou, New Zealand

**April 2021**

# Acknowledgements

The authors gratefully acknowledge the time and effort contributed by the experts who participated in this study. We are grateful to the National Mental Health Information Development Expert Advisory Panel (NMHIDEAP) in Australia for their advice and review over previous iterations of the study methodology and survey instrument. We thank the NMHIDEAP and other colleagues for assisting with the nomination of experts in Australia. We thank the General Adult Faculty of the Royal College of Psychiatrists (RCPsych), the College's Centre for Advanced Learning and Conferences and other colleagues for assisting with the nomination of experts in England. We acknowledge the national stakeholder group for the Programme for the Integration of Mental Health Data (PRIMHD) for assisting with participant recruitment in New Zealand.

This project was led by the Australian Mental Health Outcomes and Classification Network (AMHOCN). AMHOCN managed the collection of data in Australia. AMHOCN received funding from the Australian Government Department of Health to support the implementation, training and public reporting of the National Outcomes and Casemix collection, which includes the HoNOS. In England, this work was conducted by the RCPsych in partnership with Sheffield Hallam University. Te Pou managed the collection of New Zealand data. Te Pou is currently funded by the Ministry of Health to deliver HoNOS training and reporting of HoNOS data in New Zealand.

# Abbreviations

| | |
|---|---|
| AD | Average deviation |
| AMHOCN | Australian Mental Health Outcomes and Classification Network |
| COSMIN | COnsensus-based Standards for the selection of health Measurement INstruments |
| HoNOS | Health of the Nation Outcome Scales |
| HoNOSCA | Health of the Nation Outcome Scales for Children and Adolescents |
| HoNOS OA | Health of the Nation Outcome Scales for Older Adults |
| HoNOS 2018 | Health of the Nation Outcome Scales (2018 version) |
| HoNOS 65+ | Health of the Nation Outcome Scales for people aged 65 years and over |
| I-CVI | Item-level content validity index |
| ISPOR | International Society for Pharmacoeconomics and Outcomes Research |
| M | Mean |
| MHCT | Mental Health Clustering Tool |
| N, n | Number (sample size) |
| NMHIDEAP | National Mental Health Information Development Expert Advisory Panel |
| PRIMHD | Programme for the Integration of Mental Health Data |
| SD | Standard deviation |

# Contents

# Executive Summary

## About the HoNOS 2018

The Health of the Nation Outcome Scales (HoNOS) were developed in the 1990s as a means for clinicians to measure the outcomes of working-age adults in contact with specialised mental health services. The HoNOS comprises 12 scales that cover the kinds of problems that may be experienced by this group. In 2014, a collaborative project was commenced to review the HoNOS. This project was led by the Royal College of Psychiatrists, as the copyright holder, with the participation of representatives from Australia and New Zealand. As a result of the review, an updated version was published in 2018 and is known as the HoNOS 2018. The revisions were intended to reduce ambiguity and inconsistency in the glossary, and to promote rating consistency and clinical utility, without changing the measure's structure. Whether these benefits have been achieved is unknown.

## Context of this study

The HoNOS 2018 has already been taken up in some mental health services in England. In Australia and New Zealand, empirical evidence regarding its measurement properties and utility is needed to inform decisions about implementation. To this end, the Australian Mental Health Outcomes and Classification Network (AMHOCN) was tasked by the Australian Government Department of Health to investigate key measurement properties of the HoNOS 2018. Content validity was identified as a priority because deficits in content validity can affect all other measurement properties. With guidance from the National Mental Health Information Development Expert Advisory Panel (NMHIDEAP) and input from colleagues in England and New Zealand, AMHOCN designed a study to evaluate the content validity of the HoNOS 2018 scales. England and New Zealand expressed an interest in undertaking the study locally; AMHOCN supported the content validity studies in each country by providing relevant study documentation.

## Method

This descriptive study involved the completion of an anonymous, web-based survey by HoNOS experts in Australia, England and New Zealand. At least 10 participants were sought from each country with expertise in: making or supervising HoNOS ratings, psychometric or clinical effectiveness research involving the HoNOS, or use of HoNOS ratings at a macro level (e.g., staff training, monitoring service quality). Experts were identified through nomination by national bodies, bibliographic database searches, and professional networks.

The identified experts were emailed an invitation to complete the survey. The survey gathered basic information about their professional backgrounds. Six 'core' questions were developed to measure the relevance, comprehensiveness and comprehensibility of each HoNOS 2018 scale (giving a total of 72 'core' questions). In response to these questions, experts indicated their opinion on a 4-point ordinal Likert scale ranging from negative to positive (e.g., 1=Not important, 2=Somewhat important, 3=Important, 4=Very important). Experts were asked to elaborate on their reasons for any 'negative' ratings they made. At the end of the survey, they were invited to provide additional comments about the content of the HoNOS 2018.

An item-level content validity index (I-CVI) was calculated from experts' ratings on each core question. An I-CVI value of ≥0.75 indicated 'excellent' content validity. An average deviation (AD) index was calculated to show the dispersion of responses around the median. An AD value of ≤0.68 indicated 'acceptable and statistically significant agreement' between experts. Open-ended comments were analysed thematically.

## Findings

Of 43 invited experts, 32 completed the survey (74% response rate). Experts comprised a mix of professional groups, although psychiatrists (52%) and nurses (23%) accounted for the majority. Experts reported a mean of 15 years (SD 5 years) working with the HoNOS. Few (9%) had used the HoNOS 2018 in their work.

The I-CVI values show that 'positive' ratings were made by at least 50% (i.e., I-CVI ≥0.5) of experts on all but one of the 72 core questions. The number of scales that met the *a priori* criterion for excellent content validity (I-CVI ≥0.75) varied according to the question asked. For example, on the question assessing *importance for determining overall clinical significance* (an indicator of relevance), 11 of the 12 scales met the criterion. In contrast, on the question assessing *coverage of problems typically seen among adult mental health service consumers/patients* (an indicator of comprehensiveness), 5 scales met the criterion.

Several scales met the criterion for excellent content validity on all questions; these were Scale 6 (Problems associated with hallucinations and /or delusions), Scale 7 (Problems with depressed mood), and Scale 9 (Problems with relationships). In contrast, Scale 5 (Physical illness or disability problems) only met the criterion on the question assessing *importance for determining overall clinical significance*.

Almost all AD index values were equal to or below the critical 0.68 threshold, indicating acceptable and statistically significant agreement between experts.

Thematic analysis of experts' concerns provided insights into the variability in ratings for different aspects of content validity across scales. For example, one concern was that some scales combine multiple phenomena, which may have resulted in ambiguity in item wording or inadequate descriptions of severity levels, in turn creating challenges for raters. Another concern was a perceived lack of fit between the intention of the ratings and usual clinical thinking about certain types of problems (e.g., the desire to rate future physical health risks). Other comments pointed to areas of clarification that could be a focus for training and support materials (e.g., incorporating cultural and contextual factors into ratings).

In their final comments, several experts said they expected the revisions to result in improved reliability, validity and sensitivity to change. Conversely, others perceived lack of clinical utility as a greater concern, regardless of any benefits due to the revisions.

## Conclusion

Findings indicate that the HoNOS 2018 scales remain important for determining clinical severity of adults in contact with specialised mental health services, and that the revisions have not altered this core aspect of content validity. Although evidence on other aspects of content validity was more variable, the majority of experts who participated in this study rated the relevance, comprehensiveness and comprehensibility of the HoNOS 2018 scales positively. Finding from this study have the potential to inform the refinement of training and support materials in contexts where the HoNOS 2018 has already been implemented, and to inform decisions about the implementation of the HoNOS 2018 in contexts where this is being considered. They may also assist in the interpretation of results from future studies of the measurement properties of the HoNOS 2018. Given the breadth of content covered by the HoNOS 2018, training and support materials remain critical for ensuring the scales are rated as intended. Progression to testing of inter-rater reliability, utility and other measurement properties is now indicated.

# 1. Introduction

## 1.1 About the HoNOS 2018

In mental health services, routinely collected measures of clinical status and functioning are necessary tools for monitoring individual consumer/patient[a] progress and evaluating service effectiveness. The Health of the Nation Outcome Scales (HoNOS)[1] was developed as a means for clinicians to measure the outcomes of adults in contact with specialised mental health services. Since its development in the 1990s, the HoNOS has become one of the most widely implemented clinician-rated outcome measures in mental healthcare. It forms part of a coordinated national approach to outcome measurement for adults (usually, those aged 18 through 64 or 65 years) in several countries. For example, in Australia, the HoNOS has been mandated for collection in all specialised public sector mental health services as part of the National Outcomes and Casemix Collection (NOCC) which was implemented from 2001. The HoNOS is also used to monitor outcomes in private hospitals with psychiatric beds.[2] In England, the collection of HoNOS data was initially mandated for all National Health Service (NHS) funded specialist mental health care services in 2003 as part of the Mental Health Minimum Data Set (MHMDS).[3, 4] In New Zealand, the HoNOS has been mandated for collection by mental health services since 2008 and is part of the Programme for the Integration of Mental Health Data (PRIMHD) national data collection.[5] The HoNOS is also used extensively in service evaluation studies.[6]

In addition to measuring and monitoring outcomes at the individual and service level, the HoNOS is used in classification models for funding services. In Australia, the HoNOS is an important component of the Australian Mental Health Care Classification,[7] which will eventually be used for activity-based funding in the public mental health service sector. In England, the HoNOS is currently used within the Mental Health Clustering Tool (MHCT) as part of on-going development of the National Tariff Payment system. The MHCT is used to allocate consumers/patients to a cluster which can then be used to allocate a fixed price for that consumer/patient's care, for a set period of time.[8]

There is now 25 years of accumulated evidence about the measurement properties of the HoNOS. Over this time several reviews have found the HoNOS to have acceptable reliability, validity, sensitivity to change, clinical utility and interpretability.[9-11] However, the glossary had not been updated to reflect clinicians' experiences or advances in mental health service provision.[12] To that end, a collaborative international review of the HoNOS[b] commenced in 2014, led by the Royal College of Psychiatrists in the United Kingdom (the copyright holder). An advisory board was established, comprising members from England, Australia and New Zealand experienced in using the HoNOS for staff training, clinical practice, service monitoring and governance purposes. The advisory board, in turn, sought the opinions of clinicians' in their networks regarding aspects of the HoNOS that required refinement.[12] As a result of the review, the glossary was revised with the aim of reducing ambiguity and inconsistency in the glossary, thereby promoting rating consistency and clinical utility, while maintaining the fundamental structure. Most of the 12 HoNOS scales underwent a degree of revision. The nature of the revisions varied across scales and included: linguistic changes to existing scale wording and/or rating descriptions to improve clarity or relevance to the target population; the inclusion of new examples in the descriptors; and

[a] The term 'consumer' is more commonly used in Australia and New Zealand, while 'patient' is more commonly used in England.

[b] The HoNOS for people aged 65 years and over (HoNOS 65+) was also revised through this process, resulting in the HoNOS for Older Adults (HoNOS OA). The Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA) was not in scope for revision.

changes to the scope of what is to be rated or included. However, the overall structure of the measure was not changed. The revised measure was published in 2018 and is known as the HoNOS 2018.[13]

Like its predecessor, the HoNOS 2018 consists of 12 scales that cover the types of problems experienced by adults in contact with specialised mental health services (see Table 1). Each scale is rated on 1 of 5 severity levels (from 0 = no problem to 4 = severe to very severe problem), representing the maximum severity over the rating period, usually the previous two weeks. In assigning ratings, the clinician makes use of a glossary that provides summary rating instructions (general guidance that applies to all scales) as well as scale-specific guidance about what to include when making ratings, and descriptors that explain the meaning of each rating level. Once the clinician is trained and familiar with the HoNOS glossary, ratings take approximately 5 minutes to complete. No special interviews or procedures are required. Rather, the clinician should draw on all available information (e.g., case notes, interviews with the consumer/patient and family, and team meetings).

**Table 1.** The HoNOS/HoNOS 2018 scales

| Scale titles | Range of scale scores [a] |
|---|---|
| 1. Overactive or aggressive or disruptive or agitated behaviour [b] | 0 – 4 |
| 2. Non-accidental self-injury | 0 – 4 |
| 3. Problem drinking or drug-taking | 0 – 4 |
| 4. Cognitive problems | 0 – 4 |
| 5. Physical illness or disability problems | 0 – 4 |
| 6. Problems associated with hallucinations and /or delusions [c] | 0 – 4 |
| 7. Problems with depressed mood | 0 – 4 |
| 8. Other mental and behavioural problems | 0 – 4 |
| 9. Problems with relationships | 0 – 4 |
| 10. Problems with activities of daily living | 0 – 4 |
| 11. Problems with housing and living conditions [d] | 0 – 4 |
| 12. Problems with occupation and activities | 0 – 4 |

**Notes**. [a] Scales are rated on a 5-point scale: 0 = no problem; 1 = minor problem requiring no action; 2 = mild problem but definitely present; 3 = moderately severe problem; 4 = severe to very severe problem. [b] In the original HoNOS, the title for Scale 1 is 'Overactive, aggressive, disruptive or agitated behaviour'. [c] In the original HoNOS, the title for Scale 6 is 'Problems associated with hallucinations and delusions'. [d] In the original HoNOS, the title for Scale 11 is 'Problems with living conditions'.

## 1.2 Context of this study

The HoNOS 2018 has been taken up in some mental health services in England; in Australia and New Zealand, it was identified that empirical evidence regarding its measurement properties and utility is needed to inform decisions about implementation. To this end, the Australian Mental Health Outcomes and Classification Network (AMHOCN) was tasked by the Australian Government Department of Health to investigate key measurement properties of the HoNOS 2018. Content validity was identified as a priority for investigation, as it can affect all other measurement properties.[14, 15] With guidance from the National Mental Health Information Development Expert Advisory Panel (NMHIDEAP), AMHOCN designed a study to evaluate the content validity of the HoNOS 2018 scales. England and New Zealand expressed an interest in undertaking the study locally; AMHOCN supported the content validity studies in each country by providing relevant study documentation.

## 1.3 Assessing content validity

According to the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) initiative, when a measure is modified its measurement properties must be re-assessed.[15] The

assessment of content validity (i.e., the degree to which the content of a measure adequately reflects the construct(s) of interest) is a priority because deficits in content validity can affect all other measurement properties.[14, 15] The assessment of content validity should take into account the construct(s) being assessed, target population and context of use.[15] Importantly, for multi-dimensional measures such as the HoNOS 2018, each scale should be considered separately. COSMIN provides a set of 10 criteria for evaluating 'good' content validity, covering the 3 aspects of:

- relevance (the item(s) are consistent with and specific to the construct(s) of interest);
- comprehensiveness (the item(s) measure all facets of the construct(s) of interest); and
- comprehensibility (the item(s) can be understood as intended).[15]

The assessment of content validity requires the collection of information from individuals deemed to hold 'expert' knowledge in relation to the constructs being measured and their application.[15] There are 2 key considerations relating to experts; how they are chosen, and the task they are asked to complete. Experts should have training, experience, and qualifications relevant to the construct. Other types of expertise also include clinical expertise and experience conducting research on the phenomenon of interest.[16, 17] Careful specification of criteria to guide the selection of experts is a crucial aspect of the study design, as bias can easily be introduced through incorrect expert selection.[16, 18] Once experts are selected, the experts should be provided with guidance to ensure their task is clear. For example, it is critical that experts are provided with both the conceptual and operational definitions of the construct and, in survey methods, the use of bolding and underlining can help to emphasise important parts of the survey instructions and focus experts on key parts of the questions.[7, 17] If experts are unclear about the concepts or what is to be rated, this can in turn lead to unclear results in the data analysis phase.[17]

Information about content validity can be gathered using qualitative and/or quantitative methods. On one hand, qualitative methods such as interviews and focus groups may enable more in-depth information to be obtained via inter-personal interaction, and saturation can often be reached with relatively few experts.[14, 19] On the other hand, quantitative methods are frequently used, in part due to the comparative ease by which data can be gathered. In addition, statistics summarising the level of interrater agreement among experts can be calculated taking into account agreement occurring due to chance, and these can be interpreted against established thresholds. This provides a standardised way of determining whether an excellent level of content validity has been reached.[20, 21]

## 1.4 Aim of this study

With these requirements in mind, the purpose of this study was to gather empirical evidence regarding the content validity of the 12 HoNOS 2018 scales.

# 2. Method

## 2.1 Design and participants

This study used a descriptive design involving the completion of an anonymous, web-based survey by individuals with HoNOS-related expertise from 3 countries – Australia, England and New Zealand. In order to obtain survey data from a range of contexts in which the HoNOS is used, we sought at least 10 experts from each country with expertise in one or more of the following: making or supervising HoNOS ratings; psychometric or clinical effectiveness research involving the HoNOS; or HoNOS training or use of HoNOS ratings at a macro level (e.g., to monitor service quality).

Each site received approval to conduct the study and to pool data for analysis - Australia (University of Queensland Medicine, Low & Negligible Risk Ethics Sub-Committee, 2019/HE002824; Research Ethics and Integrity, 2021/HE000113), England (Sheffield Hallam University Research Ethics Committee, ID ER21666298) and New Zealand (ethics review not required; Ministry of Health, Health and Disability Ethics Committees, 20/STH/109). Written (online) informed consent was obtained from all participants.

## 2.2 Survey instrument

A purpose-designed, web-based content validity survey was developed for this study. The survey gathered basic information about experts' professional backgrounds and areas of HoNOS expertise. A series of 'pages' presented each section of the HoNOS 2018 along with a corresponding set of content validity questions. This meant that the questions could be answered even if the expert was not familiar with the HoNOS 2018.

Six 'core' questions were developed to measure the relevance, comprehensiveness and comprehensibility of each HoNOS 2018 scale (giving a total of 72 'core' questions). The questions focused on aspects of content validity potentially impacted by the revisions (see Appendix Table A.1). The content of the questions was informed by the COSMIN criteria for content validity[14, 15] and other relevant literature including the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) guidance on assessing content validity.[19, 22, 23] The 'core' questions were:

1. How important is this scale for determining overall clinical severity for adult mental health service patients? *(relevance)*
2. How likely are repeat ratings on this scale to capture change in [scale-specific problems] during a period of mental health care? *(relevance)*
3. How well do the descriptors for each rating of 0-4 cover the range of [scale-specific problems] typically seen among adult mental health service patients? *(comprehensiveness)*
4. How helpful is the glossary for determining what to include when rating [scale-specific problems]? *(comprehensibility)*
5. How well do the descriptors for each rating of 0-4 correspond to the different levels of severity of [scale-specific problems]? *(comprehensibility)*
6. How consistent is the wording of the glossary with language used in contemporary mental health practice? *(comprehensibility)*

Additional questions focused on the summary of rating instructions (4 questions) and about scale-specific changes to the glossary (1 question each for Scales 3, 4, 7, 8, 11, 12). As per best practice recommendations, key phrases within each question were underlined to focus experts' attention on the rating task and, as necessary, the population and context of use were explicitly stated in the question.[7, 17]

In response to these questions, experts indicated their opinion on a 4-point Likert scale (e.g., 1=Not important, 2=Somewhat important, 3=Important, 4=Very important). A 4-point ordinal scale was used as it is recommended there are no ambivalent middle ratings when calculating agreement among experts.[20, 24] For each scale an open-ended question invited experts to elaborate on their reasons for any 'negative' ratings (i.e., ratings of 1 or 2). At the end of the survey, experts were invited to make final comments about the content of the HoNOS 2018.

The survey content for each country was identical except that locally-relevant response options were provided when asking experts about the service settings in which they had used the HoNOS.

## 2.3 Procedures

Potential participants (experts) were identified through multiple methods including nomination by national bodies, bibliographic database searches, and professional networks. In some cases, experts were initially contacted by telephone to confirm their eligibility or, because of the anonymous nature of the survey responses, to confirm their correct email contact details. Experts were invited to participate via an email which contained a link to the survey. The survey commenced with a 'page' displaying the information sheet/consent form. Upon providing informed consent, participants entered the survey (described above), which took approximately 30 minutes to complete. Each country identified and recruited local experts and hosted the survey on their own platform.

## 2.4 Data analysis

The main body of this report presents results for the total sample. Country-level results are provided in Appendix Tables A.2 through A.9.

An item-level content validity index (I-CVI)[20, 25] shows the proportion of experts who rated each scale positively on each core question. The I-CVI was calculated as the sum of the number of 'positive' ratings (i.e., ratings of 3 or 4), divided by the number of raters. At the 5% significance level, an I-CVI value ≥0.75 indicates 'excellent' content validity when there are ≥16 raters.[25] This method of determining the I-CVI threshold takes sample size into account, which addresses concerns about the inflation of agreement merely by chance.[7, 24]

We also calculated an average deviation (AD) index. The AD index measures the dispersion of responses around the median, with lower values indicating less dispersion and therefore better agreement.[26] It is calculated by summing the absolute differences in individual ratings from the median and dividing by the number of experts. The AD index value is then compared to a threshold for acceptable and statistically significant agreement, determined by the number of response categories and number of experts. At the 5% significance level with a 4-point response scale, AD index values ≤0.68 indicate 'acceptable and statistically significant agreement' when there ≥15 raters.[26, 27] Statistical analyses were conducted in Stata version 16.0 (StataCorp, College Station, TX, USA).

Open-ended comments were analysed thematically using NVivo 12 plus (QSR International, 2018).

# 3. Results

## 3.1 Sample characteristics

Of 43 invited experts, 32 completed the survey (74% response rate).[c] Table 2 presents the characteristics of the participating experts. Three-quarters were psychiatrists or nurses, the remaining one-quarter represented a range of disciplines including psychology and social work. Most experts reported having HoNOS experience in at least 2 areas – more than 80% in making or reviewing HoNOS ratings, more than 60% in research involving the HoNOS and more than 60% in macro-level use of the HoNOS. Collectively, the experts had used the HoNOS across a mix of clinical settings as well as in non-clinical settings. On average, they had worked in mental health for 28 years and with the HoNOS for 15 years. More than half knew of the HoNOS 2018, but few (9%) had used it in their work.

**Table 2.** Characteristics of experts who completed the survey (N = 32)

| | n | % |
|---|---|---|
| **Main professional background** [a] | | |
| Nurse | 7 | 23 |
| Psychologist | 3 | 10 |
| Clinical psychologist | 2 | 6 |
| Social worker | 1 | 3 |
| Psychiatrist | 16 | 52 |
| Psychiatric registrar | 0 | 0 |
| Occupational therapist | 0 | 0 |
| Other | 2 [b] | 6 |
| **Expertise with HoNOS** [c] | | |
| Rating HoNOS or reviewing HoNOS ratings made by others | 27 | 84 |
| Research in the measurement properties of the HoNOS and/or measuring clinical effectiveness | 20 | 63 |
| HoNOS staff training and/or using HoNOS results at a macro level | 20 | 63 |
| **Mental health settings worked with HoNOS** [c] | | |
| Inpatient | 22 | 69 |
| Residential [d] | 7 | 22 |
| Community | 30 | 94 |
| Other, non-clinical setting | 4 | 13 |
| **Aware of HoNOS 2018 prior to survey** | | |
| No, I was not aware of the HoNOS 2018 at all | 12 | 38 |
| Yes, I was aware of the HoNOS 2018, but have not used it in my work | 16 | 50 |
| Yes, I have used the HoNOS 2018 in my work | 3 | 9 |
| Not sure | 0 | 0 |
| Other | 1 [e] | 3 |
| | **M (SD)** | **Range** |
| **Years worked in mental health** [f] | 28 (9) | 10-43 |
| **Years worked with the HoNOS** | 15 (5) | 5-28 |

HoNOS, Health of the Nation Outcome Scales. M, mean. SD, standard deviation.

[a] Missing data for one respondent (n = 31). [b] "Clinical epidemiologist", "Consumer and Family Leader". [c] Categories not mutually exclusive. [d] 'Residential' category included only in the Australian version of the survey. [e] "I have not seen the revised HoNOS 2018". [f] Missing data for two respondents (n = 30).

[c] Response rates were 87% (13/15) in Australia, 83% (10/12) in England, and 56% (9/16) in New Zealand.

## 3.2 Experts' ratings

Tables 3 and 4 show summary statistics including the I-CVI and AD index values derived from experts' responses to the core questions about the relevance, comprehensiveness and comprehensibility of the HoNOS 2018 scales.

The I-CVI values show that 'positive' ratings were made by at least 50% (i.e., I-CVI ≥0.5) of experts on all but one of the total of 72 core questions, and by 70% of experts (i.e., I-CVI ≥0.7) on nearly 70% of core questions.

The number of scales that met the *a priori* criterion for excellent content validity (I-CVI ≥0.75) varied according to the question asked. On the question assessing *importance for determining overall clinical significance* (relevance), 11 of the 12 scales met the criterion. On the question assessing *the helpfulness of the glossary for determining what to rate and/or include* (comprehensibility), 9 scales met the criterion. On the questions assessing *likelihood of capturing change during a period of mental health care* (relevance), and *correspondence between descriptors and levels of severity* (comprehensibility) and *consistency of wording with contemporary mental health practice* (comprehensibility), 6 scales met the criterion. On the question assessing *coverage of problems typically seen among adult mental health service consumers/patients* (comprehensiveness), 5 scales met the criterion.

Several scales met the criterion for excellent content validity on all questions; these were Scale 6 (Problems associated with hallucinations and /or delusions), Scale 7 (Problems with depressed mood), and Scale 9 (Problems with relationships). Scale 4 (Cognitive problems) and Scale 10 (Problems with activities of daily living) met the criterion on all questions except the *likelihood of capturing change during a period of mental health care* question. At the other end of the spectrum, Scale 5 (Physical illness or disability problems) only met the criterion on the question assessing *importance for determining overall clinical significance*.

Almost all AD index values were equal to or below the critical 0.68 threshold, indicating acceptable and statistically significant agreement between experts, with 2 exceptions. For the question about correspondence between the descriptors and the severity of scale-specific problems, Scale 2 (Non-accidental self-injury) and Scale 5 (Physical illness or disability problems) had an AD index values of 0.75 and 0.69, respectively. Inspection of the distribution of ratings showed this was due to equal numbers of experts holding 'positive' and 'negative' views. This pattern of responses also contributed to the relatively low I-CVI values of 0.50 for these two scales.

**Table 3.** Experts' ratings of the content validity of the HoNOS 2018 scales: relevance and comprehensiveness

| | Relevance | | | | | | | | Comprehensiveness | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | How important is this scale for determining overall clinical severity for adult mental health service consumers/patients? | | | | How likely are repeat ratings on this scale to capture change in [scale-specific problems] during a period of mental health care? | | | | How well do the descriptors for each rating of 0-4 cover the range of [scale-specific problems] typically seen among adult mental health service consumers/patients? [a] | | | |
| **HoNOS 2018 scale** | **n** | **Range** | **I-CVI** | **AD** | **n** | **Range** | **I-CVI** | **AD** | **n** | **Range** | **I-CVI** | **AD** |
| Scale 1. Overactive or aggressive or disruptive or agitated behaviour | 31 | 2-4 | **0.81** | 0.48 | 30 | 1-4 | 0.67 | 0.53 | 32 | 1-4 | 0.72 | 0.50 |
| Scale 2. Non-accidental self-injury | 31 | 2-4 | **0.90** | 0.55 | 32 | 2-4 | 0.66 | 0.47 | 31 | 1-4 | 0.65 | 0.48 |
| Scale 3. Problem drinking or drug-taking | 32 | 2-4 | **0.94** | 0.38 | 31 | 1-4 | 0.55 | 0.61 | 31 | 1-4 | 0.65 | 0.61 |
| Scale 4. Cognitive problems | 32 | 2-4 | **0.91** | 0.41 | 32 | 1-4 | 0.66 | 0.63 | 32 | 1-4 | **0.88** | 0.31 |
| Scale 5. Physical illness or disability problems | 30 | 1-4 | **0.77** | 0.40 | 32 | 1-4 | 0.56 | 0.66 | 31 | 1-4 | 0.71 | 0.55 |
| Scale 6. Problems associated with hallucinations and /or delusions | 31 | 2-4 | **0.97** | 0.52 | 32 | 2-4 | **0.88** | 0.50 | 32 | 1-4 | **0.81** | 0.56 |
| Scale 7. Problems with depressed mood | 32 | 2-4 | **0.97** | 0.50 | 32 | 2-4 | **0.81** | 0.50 | 32 | 1-4 | **0.88** | 0.41 |
| Scale 8. Other mental and behavioural problems | 32 | 2-4 | **0.88** | 0.44 | 32 | 1-4 | 0.69 | 0.44 | 32 | 1-4 | 0.69 | 0.50 |
| Scale 9. Problems with relationships | 31 | 2-4 | **0.87** | 0.39 | 32 | 1-4 | **0.81** | 0.31 | 32 | 1-4 | **0.78** | 0.34 |
| Scale 10. Problems with activities of daily living | 32 | 2-4 | **0.91** | 0.25 | 31 | 1-4 | 0.74 | 0.39 | 31 | 1-4 | **0.81** | 0.39 |
| Scale 11. Problems with housing and living conditions | 31 | 1-4 | 0.71 | 0.45 | 29 | 1-4 | **0.79** | 0.45 | 32 | 1-4 | 0.66 | 0.50 |
| Scale 12. Problems with occupation and activities | 30 | 1-4 | **0.77** | 0.47 | 32 | 1-4 | **0.75** | 0.34 | 32 | 1-4 | 0.66 | 0.56 |

AD, average deviation. I-CVI, item-level content validity index. n, number. Bold I-CVI values meet the criterion for excellent content validity (i.e., I-CVI ≥ 0.75). [a] To fit the wording of Scale 8, the equivalent question for Scale 8 was: How well do problems A-O cover the range of other mental and behavioural problems typically seen among adult mental health service consumers/patients?

**Table 4.** Experts' ratings of the content validity of the HoNOS 2018 scales: comprehensibility

| | Comprehensibility | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | How helpful is the glossary for determining what to include when rating [scale-specific problems]? [a, b] | | | | How well do the descriptors for each rating of 0-4 correspond to the different levels of severity of [scale-specific problems]? | | | | How consistent is the wording of the glossary with language used in contemporary mental health practice? | | | |
| **HoNOS 2018 scale** | **n** | **Range** | **I-CVI** | **AD** | **n** | **Range** | **I-CVI** | **AD** | **n** | **Range** | **I-CVI** | **AD** |
| Scale 1. Overactive or aggressive or disruptive or agitated behaviour | 32 | 2-4 | **0.78** | 0.44 | 32 | 1-4 | 0.59 | 0.50 | 30 | 1-4 | **0.80** | 0.33 |
| Scale 2. Non-accidental self-injury | 31 | 2-4 | **0.81** | 0.35 | 32 | 1-4 | 0.50 | 0.75 | 32 | 1-4 | **0.75** | 0.38 |
| Scale 3. Problem drinking or drug-taking | 32 | 1-4 | **0.75** | 0.50 | 31 | 1-4 | 0.65 | 0.58 | 32 | 1-4 | 0.69 | 0.53 |
| Scale 4. Cognitive problems | 31 | 2-4 | **0.84** | 0.42 | 30 | 1-4 | **0.87** | 0.27 | 30 | 1-4 | **0.83** | 0.27 |
| Scale 5. Physical illness or disability problems | 22 [c] | 1-4 | 0.45 | 0.64 | 32 | 1-4 | 0.50 | 0.69 | 30 | 1-4 | 0.67 | 0.50 |
| Scale 6. Problems associated with hallucinations and /or delusions | 32 | 2-4 | **0.88** | 0.44 | 32 | 2-4 | **0.88** | 0.41 | 32 | 2-4 | **0.88** | 0.28 |
| Scale 7. Problems with depressed mood | 32 | 1-4 | **0.78** | 0.66 | 31 | 2-4 | **0.81** | 0.45 | 32 | 1-4 | **0.81** | 0.34 |
| Scale 8. Other mental and behavioural problems | 32 | 2-4 | **0.78** | 0.47 | 31 | 2-4 | 0.68 | 0.42 | 32 | 1-3 | **0.81** | 0.22 |
| Scale 9. Problems with relationships | 32 | 2-4 | **0.75** | 0.31 | 32 | 2-4 | **0.78** | 0.38 | 31 | 1-4 | **0.77** | 0.35 |
| Scale 10. Problems with activities of daily living | 32 | 2-4 | **0.88** | 0.28 | 32 | 1-4 | **0.91** | 0.22 | 32 | 2-4 | **0.88** | 0.22 |
| Scale 11. Problems with housing and living conditions | 32 | 1-4 | 0.69 | 0.50 | 31 | 1-4 | 0.74 | 0.42 | 31 | 1-4 | **0.77** | 0.29 |
| Scale 12. Problems with occupation and activities | 32 | 1-4 | 0.59 | 0.53 | 32 | 1-4 | **0.75** | 0.38 | 32 | 1-4 | 0.69 | 0.50 |

AD, average deviation. I-CVI, item-level content validity index. n, number. Bold I-CVI values meet the criterion for excellent content validity (i.e., I-CVI ≥ 0.75). [a] Question text differed across scales; depending on the glossary, "what to rate and include" or "what to rate and consider" was substituted for the phrase "what to include". [b] To fit the wording of Scale 8, the equivalent question for Scale 8 was: How helpful is the glossary for determining which other mental and behavioural problem to rate on this scale? [c] This question was inadvertently omitted from the England survey.

## 3.3 Experts' concerns

For each scale, experts were invited to elaborate on their reasons for giving a 'negative' rating (i.e., a rating of 1 or 2) on any question. Analysis of these open-ended comments revealed nine themes that corresponded to one of the three aspects of content validity defined in the COSMIN framework. Most (six) of these themes related to comprehensibility, with two themes relating to relevance, and one theme relating to comprehensiveness. An additional theme highlighted the important role of HoNOS training. The themes are summarised in this order below, with illustrative quotations.

### 3.3.1 Themes related to comprehensibility

#### 3.3.1.1 Too many phenomena

A recurring concern from the experts was that some scales combine too many different phenomena together.

> "The item confuses and conflates a number of different clinical symptoms. What is it trying to capture? Overactive behaviour is not the same as aggressive behaviour and both cannot be sensibly combined into a single rating" (*Scale 1. Overactive or aggressive or disruptive or agitated behaviour*).

> "The item conflates two independent behaviours and assumes an equivalence" (*Scale 2. Non-accidental self-injury*).

> "The wording asks scorer to rate excessive, harmful, craving, dependence and adverse consequences all on one scale too many variables" (*Scale 3. Problem drinking or drug-taking*).

> "Conflating iatrogenic or highly transitory states with long term and enduring disability is problematic" (*Scale 5. Physical illness or disability problems*).

This merging of a variety of phenomena into one scale has several consequences, as illustrated in the following themes of ambiguity, need for more description or examples, assessment challenges, and incomplete coverage.

#### 3.3.1.2 Ambiguity

The HoNOS review project aimed to increase the clarity and reduce the ambiguity of the measure. However, the experts continued to identify sources of ambiguity in terminology and instructions for some scales.

> "More serious overactivity is open to interpretation" (*Scale 1. Overactive or aggressive or disruptive or agitated behaviour*).

> "'D' is labelled 'Reactions to stressful events and trauma.' - however, the descriptor seems much more specific in linking this to an acute stress reaction and/or response to traumatic events. It is not clear whether only acute stressors and traumas are to be coded (and if so how recent the event [might] have been). This is ambiguous" (*Scale 8. Other mental and behavioural problems*).

Some comments identified ambiguity relating to the problem of 'too many phenomena'.

> "Craving, dependence, level of use and subsequent harm are all potentially important but they

> don't co-vary in a linear fashion. Suppose I have no craving or dependence but get very drunk, fall and sustain a serious head injury? Or develop psychosis from a one off use of amphetamines? The descriptors don't always correlate." (*Scale 3. Problem drinking or drug taking*).

Other comments identified that ambiguity can arise when a rating requires a comparison to cultural or contextual norms.

> "It is not clear, and is open to subjectivity, what is meant by cultural and contextual factors or how these may alter ratings" (*Summary rating instructions*).

> "What does "Excessive" mean. More than the rater? This needs better anchors. Would any Ice use be excessive?" (*Scale 3. Problem drinking or drug-taking*).

#### 3.3.1.3 Need for more description or examples

Given concerns about rating too many phenomena and ambiguity it is not surprising that there were calls for more descriptions or examples to be added to the glossary to guide ratings.

> "Descriptors would be far more useful if they [simply] gave examples of the types of acts one would expect at each rating level. Examples are only given for a minority of the descriptors" (*Scale 1. Overactive or aggressive or disruptive or agitated behaviour*).

> ""Adverse consequences" and "severe adverse consequences" would benefit from definition/examples" (*Scale 3. Problem drinking or drug-taking*).

> "Whatever descriptor is used [for non-accidental self-injury], it would be better to elaborate what is intended in this rating (to exclude clearly accidental self-injury, by giving a few more examples of these)" (*Scale 2. Non-accidental self-injury*).

#### 3.3.1.4 Assessment challenges

Given the variety of phenomena to be considered, discriminating between these phenomena across and within scales for the purposes of rating can be an assessment challenge.

> "Making a distinction between behavioural aspects of drug/alcohol use (rated here) and aggressive/destructive behaviour rated in scale 1 can be problematic" (*Scale 3. Problem drinking or drug-taking*).

> "The anchors for [rating level] two are challenging... excessive drinking but no craving... this distinction will be hard to judge" (*Scale 3. Problem drinking or drug-taking*).

This is particularly the case at Scale 8:

> "It can be difficult to decide what to capture on this scale when a service user has multiple things they find problematic" (*Scale 8. Other mental and behavioural problems*).

These discrimination challenges are not limited to the disentanglement of multiple phenomena but also the context within which the assessment may be taking place. Not having access to certain information can make rating a challenge.

> "The glossary seems entirely focussed on community patients and does not describe how to approach this scale if a patient is being treated in a residential setting e.g. inpatient ward in hospital" (*Scale 11. Problems with housing and living conditions*).

> "Knowledge of patients usual occupation / activities is required - may be difficult for inpatient staff" (*Scale 12. Problems with occupation and activities*).

#### 3.3.1.5 Lack of fit with clinical thinking

Although the glossary's instructions regarding what to rate and/or include were generally well-understood, experts identified a lack of fit with their thinking about certain clinical problems.

> "Self-harm often occurs independently of suicidal intent and has quite a different clinical meaning and significance. A confused and incoherent item" (*Scale 2. Non-accidental self-injury*).

> "Considering cognitive issues from wide variety of causes which clinically does not fit well & is confusing to the rater" (*Scale 4. Cognitive problems*).

> "People with serious mental disorders are at risk of (or may have actually been diagnosed with) metabolic disorders have a reduced life expectancy of 15 to 20 years - they can be under-rated here as the illness may not yet effect their mobility or activity - I think the risk of metabolic disorders causing early and avoidable death is not well identified by HoNOS, but it is a massive risk for people" (*Scale 5. Physical illness or disability problems*).

> "The difficulty I have with this catch-all item is that it contains the most common presentations […] in one question. In an ideal world, there would be an optional drop-box that permits these to be rated separately" (*Scale 8. Other mental and behavioural problems*).

The primacy of "clinical thinking" is exemplified by this comment:

> "Staff may continue to think in terms of depression rather than depressed mood irrespective of how it is worded" (*Scale 7. Problems with depressed mood*).

#### 3.3.1.6 Problems with language

There was feedback that some of the wording used in the glossary does not align with clinical language or constructs:

> "I don't like the 'ending it all' phrase - it is inconsistent with the other ratings wordings and I don't think it is precise or related to clinical practice wording" (*Scale 2. Non-accidental self-injury*).

> "Language again does not align with clinical work that well" (*Scale 12. Problems with occupation and activities*).

> "Occupation is a bit narrow (both the language as well as the construct)" (*Scale 12. Problems with occupation and activities*).

or could be viewed as pejorative:

> "The language used is quite accusatory and not client-focused" (*Scale 1. Overactive or aggressive or disruptive or agitated behaviour*).

> ""Passive" is not an ideal term - requires a judgement which is not easily made and is potentially pejorative" (*Scale 2. Non-accidental self-injury*).

> "Patient could be replaced with consumer or person" (*Scale 8. Other mental and behavioural problems*).

### 3.3.2 Themes related to relevance

#### 3.3.2.1 Importance

There were few concerns about the importance of the scales in determining clinical severity. However, some noted that Scale 11 (Problems with housing and living conditions) does not directly involve rating patient need and questioned its relevance to some assessment contexts.

> "Housing is not part of the clinical formulation but part of the contextual background" (S*cale 11. Problems with housing and living conditions*).

> "Not sure of the value of rating inpatient setting" (S*cale 11. Problems with housing and living conditions*).

#### 3.3.2.2 Challenges to capturing change

The experts had various concerns about the idea of change and its measurement in clinical practice. The most common concern was perceived lack of sensitivity to detect frequent, delayed or subtle changes during an episode of care.

> "Difficult to capture patients with emotionally unstable personality disorder who can have daily ideas suicide & frequent self-harm attempts" (*Scale 2. Non-accidental self-injury*).

> "These difficulties can be longstanding issues so small changes unlikely to be captured within scale" (*Scale 9. Problems with relationships*).

> "May be less likely to pick up change in capacity in an episode of care compared with most other scales, as there is often a lag in these resuming as clinical state improves" (*Scale 10. Problems with activities of daily living*).

Some commented that the cause of the behaviour is an important consideration:

> "Depending on the cause of the problem, change may be slow/absent/minor" (*Scale 4. Cognitive problems*).

Another concern was whether the HoNOS should be measuring disability and distress compared to a norm for the individual or a societal norm:

> "If we were to rate the severity of the condition based on 'change' from what is 'normal' for a person it may provide a more valid picture of the condition and its impact" (*Summary rating instructions*).

### 3.3.3 Themes related to comprehensiveness

#### 3.3.3.1 Incomplete coverage

For a few scales, experts suggested specific behaviours or problems that should be included in the descriptors:

> "Self-harming behaviour, e.g. cutting, skin picking / hair pulling/ head banging/ burning (cigarette burns) without suicidal thoughts especially when these present as longer term chronic mal-adaptive behaviour aimed at self-management of emotions are not included in descriptors. e.g. 2 or 3" (*Scale 2. Non-accidental self-injury*).

or changes to the scope of the descriptors:

> "Scale appears to be useful for the most seriously impaired, but not fine grained enough, doesn't include wider range of roles - parenting, caregiving, training, cultural" (*Scale 12. Problems with occupation and activities*).

### 3.3.4 Need for training

Some comments pointed to areas of clarification that could be a focus for training.

> "I am somewhat uneasy about the incorporation of cultural aspects into the HoNOS tools. I do not consider myself qualified to judge other cultures other than the one I was born into." (*Summary rating instructions*).

> "Instructions regarding the need to incorporate cultural and contextual factors into ratings? - minimal guidance is provided as to how such factors may need to be considered, hopefully this would be addressed in any training package." (*Summary rating instructions*).

This highlights that cultural competence is a broader framework guiding clinical practice. HoNOS training needs to fit within this framework.

> "Whether we are asking for a rating of existence of symptoms (such as hallucinations, physical illness) and the effects on behaviour of those symptoms ('Problems associated with ...') causes a great deal of confusion" (*Scale 6. Problems associated with hallucinations and /or delusions*).

Training reinforces that rating involves the need to consider degree of distress and impact on behaviour.

> "Again, it would be good to clarify if this is to be rated from the clinician's perspective or, more consistent with a recovery approach, the patient's perspective?" (*Scale 12. Problems with occupation and activities*).

Training reinforces that the HoNOS is a measure of the clinician's perspective, taking the consumer's/patient's cultural context into account.

Although described in the glossary, the subtlety that motivation is measured on Scale 10, while Scales 11 and 12 are availability and suitability of living conditions and occupational activities is often missed. These scales are important for determining severity, but less directly so than others.

> "Is this item attempting to capture the availability of occupation/activity or the patient's ability and or motivation to engage in activity?" (*Scale 12. Problems with occupation and activities*).

## 3.4 Experts' summary comments

At the end of the survey, experts were provided with the opportunity to make final comments about the HoNOS 2018. The survey tasks did not involve comparing the original HoNOS to the HoNOS 2018. Nonetheless, several experts favoured the HoNOS 2018 over the original HoNOS.

> "Prefer 2018 version."

> "Useful clarifications in the glossaries compared to HoNOS (4) and in changes to specific items including inclusion of thought disorder. Item 8 now more relevant to today's presenting difficulties and clarification of stress/ trauma is a significant improvement, as is separating anxiety and phobia."

Some went on to say that they expect the revisions will improve validity, reliability and the ability to detect change.

> "The revisions in HoNOS 2018 brings more clarity to the scales within HoNOS which is likely to improve the overall validity and reliability of the scale. The revisions are well thought through as they maintain the integrity of the original measure."

> "Overall I think this is an improvement and will lead to more accurate information and detection of change. It is "somewhat useful" as a broad "brush" stroke instrument""

Some experts endorsed the HoNOS as a measure but did not indicate a specific preference for one version over the other.

> "I have always believed that HoNOS is a credible baseline assessment that captures necessary aspects of mental health."

There were some negative comments regarding practical issues that limit utility of the measure. A frequent comment was that the value of the HoNOS is limited, as it is not used to guide clinical decision making and care.

> "It remains a flawed tool that has little relevance to clinicians and patients, and consequently is not widely used in clinical practice. It is too broad in scope to enable it to be sensitive enough to change to be clinically useful, it takes too long to complete relative to its value; the purpose of the scale has never been clearly articulated, and many of the items are confused and unclear."

> "Key issue is clinicians using these rating scales to guide care provision. This will drive up accuracy & consistency. Unfortunately scales are seen as performance measure to be completed not one of range of tools to help with assessment of patient's needs."

> "It's not greatly different. Convenient and quick but I am ambivalent about it being useful clinically. More for research."

Other experts noted implementation issues that need to be addressed, including gaps in completion rates and rating consistency.

> "The key practical issue is of course the variable of 'raters' ...time, habit vs reading the rules, brain space, stereotyping, lack of information required..."

“Mental health services need to be informed by science and outcome measures. Response rates poor and need to be addressed.”

# 4. Discussion

## 4.1 Summary of findings

To our knowledge, this is the first empirical study of the revised HoNOS glossary. A key finding was the strong consensus between experts that the HoNOS 2018 scales are important for determining overall clinical severity among adults in contact with mental health services. This is consistent with a previous study of the original HoNOS[28] and provides some reassurance that the glossary revisions have not altered this core aspect of content validity. The exception was Scale 11 (Problems with housing and living conditions) which had an I-CVI (0.71) slightly below the 0.75 threshold for acceptability, likely reflecting some concerns about its relevance to clinical severity and inpatient settings.

Evaluations of each scales' ability to capture change, comprehensiveness and comprehensibility were more variable, although the majority of experts rated most scales positively on these aspects of content validity. Thematic analysis revealed possible explanations for this variability. For example, one concern was that some scales combine multiple phenomena, which may result in ambiguity in item wording and discrimination challenges for raters. Indeed, several scales consistently met the criterion for excellent content validity. These scales were: Scale 4 (Cognitive problems), Scale 6 (Problems associated with hallucinations and /or delusions), Scale 7 (Problems with depressed mood), Scale 9 (Problems with relationships) and Scale 10 (Problems with activities of daily living). These scales tend to focus on a single phenomenon, or a relatively narrower range of phenomena.

Conversely, although the HoNOS review project aimed to increase the clarity and reduce the ambiguity of the measure, this may not have been completely achieved for some scales. For example, the scales that describe behavioural problems - Scale 1 (Overactive or aggressive or disruptive or agitated behaviour), Scale 2 (Non-accidental self-injury) and Scale 3 (Problem drinking or drug-taking) - were frequently noted by experts as entailing multiple phenomena and as being insufficiently illustrated with examples, which would make it challenging to determine a severity rating. This corresponded to lower I-CVIs for these scales on the survey questions about correspondence between the descriptors and severity levels (comprehensibility) and coverage of the descriptors (comprehensiveness).

Another theme was a perceived lack of fit between the intention of the ratings and usual clinical thinking for certain problems. For example, for Scale 2 (Non-accidental self-injury) several experts were concerned that the descriptors for the rating levels do not align with commonly used suicide risk paradigm of ideation, plan and attempt. For Scale 5 (Physical Illness or disability problems) some experts perceived the focus on activity restrictions to be too narrow and wanted an opportunity to include issues relating to chronic physical health problems (e.g., risk of future adverse consequences). For Scale 8 (Other mental and behavioural problems), several experts expressed a desire to rate multiple problems on whereas the intention of the scale is to rate only the most severe problem. This view has been reported previously,[29, 30] but was out of scope for the HoNOS revisions as it would conflict with the 'rate the worst problem' rule.[12] The views expressed regarding Scale 5 and Scale 8 may reflect the growing recognition of the prevalence and outcomes of multi-morbidity among people with severe mental illness.[31, 32]

This study is a first step in examining the revised HoNOS 2018. In services where the HoNOS 2018 is already in use, the information obtained in this study could be used to refine training and support materials. For example, although the HoNOS 2018 includes additional guidance about incorporating cultural and contextual factors into ratings,[12] some experts called for further explanation and examples in the glossary or via training. These comments underscore the importance of cultural competence as a broader framework to guide clinical practice, including HoNOS ratings. Training provides an opportunity to address identified assessment challenges – for example, distinguishing when to rate patient motivation

versus opportunities in their environment, which is a difficult task that may require additional support materials.[33] Training also provides an opportunity to reinforce that, although the HoNOS 2018 permits a summary of assessments across a broad range of important constructs, it does not replace clinical judgement or preclude other clinical issues being documented.

This study provides evidence that may help inform decisions about HoNOS 2018 implementation in services where this is being considered. However, other information is also likely to be needed to guide such decisions, including evidence regarding inter-rater reliability and other measurement properties, evidence regarding clinical utility, as well as consideration of infrastructure costs and training implications. Findings may also assist in interpreting results from future studies of other measurement properties of the HoNOS 2018.[15]

## 4.2 Strengths and limitations

A key strength of our study is the multi-site international design which incorporates 3 countries that have invested heavily in implementing the HoNOS in their national mental health outcome measurement efforts but have different service systems as well as HoNOS training materials and delivery. This increases the likelihood that our findings are applicable across a range of real-world mental health service contexts. Other strengths include the sample size; according to COSMIN, a sample size of 30 or greater is considered adequate for quantitative studies of content validity. We calculated inter-rater agreement statistics, which is an appropriate statistical approach according to COSMIN.[14]

Our study also included a qualitative component, from which we were able to identify several themes that facilitated a deeper understanding of experts' perspectives. The inclusion of experts from clinical, research and evaluation, and service development domains is consistent with the different uses of the HoNOS. It is also consistent with the range of opinions sought in the HoNOS revision project.[12]

Some limitations should be noted. First, it is possible that our findings reflect the specific mix of experts we recruited, and another study using a different set of experts might yield different results. In addition, approximately one-quarter of invited experts did not complete the survey. Non-completers may have held different views to completers; however, survey responses revealed a mix of positive and negative views among participating experts. The use of multiple strategies to identify experts may have mitigated potential selection biases and making the survey anonymous may have limited potential response bias. Second, while every effort was made to select appropriate experts, we did not conduct preselection interviews to verify expertise against stringent recruitment criteria.[17, 34] However, experts reported that they had worked with the HoNOS for many years, with the vast majority having expertise with HoNOS ratings, usually coupled with research or service-related expertise. Third, to reduce respondent burden, we only asked experts to elaborate on their 'negative' responses. This means the qualitative results emphasise concerns. Therefore, it is important that interpretation of the results of this study considers the quantitative results, which represent both positive and negative views, as well as the qualitative results.

## 4.3 Comparison to previous studies and future directions

The present study was not designed to compare the content validity of the original HoNOS and the HoNOS 2018. This was considered too onerous for an online survey but could be explored in future research using focus groups or other qualitative methods. However, some limited comparison with studies of the content validity of the original HoNOS can be made. For example, we found strong consensus between experts about the importance of the HoNOS 2018 scales for determining overall clinical severity, consistent with a previous study of the original HoNOS by our group.[28] This provides

some reassurance that the glossary revisions have not altered this core aspect of content validity. In the current study, experts identified some potential rating challenges that have also been identified in previous studies. One of these is that the availability of knowledge about a consumer's/patient's usual environment can be a challenge when rating Scale 11 (Problems with housing and living conditions).[28] As already noted, this may require additional support materials.[33] Another is the desire to rate multiple problems on Scale 8 (Other mental and behavioural problems).[12, 29, 30] This finding is unsurprising, as structural changes were outside the scope of revisions to the glossary.[12] Another is the need for additional guidance about incorporating cultural and contextual factors into ratings.[12] As noted earlier, this may be an area of focus for HoNOS 2018 training but, more broadly, reinforces that cultural competence is essential skill for good quality clinical assessment.[12]

Although outside the scope of the current study, several experts expressed concerns about the clinical utility of the HoNOS 2018, regardless of the revisions. It is important to acknowledge that these concerns may, at least in part, reflect broader views about the value of routine outcome measurement,[35-38] as much as specific limitations of the HoNOS or HoNOS 2018. That said, information about the utility of the HoNOS 2018 has been identified as one of several pieces of information necessary for informing decisions about whether the HoNOS 2018 should be implemented in Australia and New Zealand. This information could be gathered in several ways. Experienced raters could be asked about their views of the utility of the HoNOS 2018 for different purposes (e.g., monitoring consumer/patient outcomes, evaluating service effectiveness), either on its own or in comparison to the original HoNOS. Another avenue could be to explore barriers and facilitators to using the HoNOS 2018 in clinical decision-making in places where it has been implemented. This could be done using surveys, case studies or service audits. This research would help fill a gap in knowledge about how the HoNOS is used by clinicians.[39] Previous research has shown that clinician perspectives about the utility of the original HoNOS are mixed,[40] but that some approaches (e.g., focusing case review meetings on all clinically significant HoNOS scores, integrating HoNOS scores into referral forms and care planning documents, and information technology systems) can improve the utility of the HoNOS in clinical settings.[41 42]

## 4.4 Conclusions

After 20 years of use in clinical practice, the HoNOS glossary was revised resulting in an updated measure known as the HoNOS 2018. In this study, there was strong consensus among experts that the HoNOS 2018 scales remain important for determining clinical severity of adults in contact with specialised mental health services. Although evidence on other aspects of content validity was more variable, the majority of experts who participated in this study viewed the scales' ability to capture change, comprehensiveness and comprehensibility of the HoNOS 2018 scales positively. Given the measure's breadth of content, findings reinforce the important role of training and support materials to address residual areas of ambiguity and encourage rating fidelity. Overall, findings are sufficiently encouraging to warrant further exploration of the inter-rater reliability and other measurement properties of the HoNOS 2018.

# References


1. Wing JK, Beevor AS, Curtis RH, Park SB, Hadden S, Burns A. Health of the Nation Outcome Scales (HoNOS). Research and development. Br J Psychiatry. 1998;172:11-8.

2. Burgess P, Pirkis J, Coombs T. Routine outcome measurement in Australia. Int Rev Psychiatry. 2015;27(4):264-75.

3. Holloway F. Outcome measurement in mental health--welcome to the revolution. Br J Psychiatry. 2002;181:1-2.

4. Macdonald AJ, Fugard AJ. Routine mental health outcome measurement in the UK. Int Rev Psychiatry. 2015;27(4):306-19.

5. Smith M, Baxendine S. Outcome measurement in New Zealand. Int Rev Psychiatry. 2015;27(4):276-85.

6. Delaffon V, Anwar Z, Noushad F, Ahmed AS, Brugha TS. Use of Health of the Nation Outcome Scales in psychiatry. Adv Psychiatr Treat. 2012;18(3):173-9.

7. Almanasreh E, Moles R, Chen TF. Evaluation of methods used for estimating content validity. Res Social Adm Pharm. 2019;15(2):214-21.

8. Yeomans D. Clustering in mental health payment by results: a critical summary for the clinician. Adv Psychiatr Treat. 2018;20(4):227-34.

9. Te Pou. The HoNOS Family of Measures: A technical review of their psychometric properties. Auckland: Te Pou o Te Whakaaro Nui, 2012.

10. Pirkis JE, Burgess PM, Kirk PK, Dodson S, Coombs TJ, Williamson MK. A review of the psychometric properties of the Health of the Nation Outcome Scales (HoNOS) family of measures. Health Qual Life Outcomes. 2005;3:76.

11. Burgess PM, Harris MG, Coombs T, Pirkis JE. A systematic review of clinician-rated instruments to assess adults' levels of functioning in specialised public sector mental health services. Aust N Z J Psychiatry. 2017;51(4):338-54.

12. James M, Painter J, Buckingham B, Stewart MW. A review and update of the Health of the Nation Outcome Scales (HoNOS). BJPsych Bull. 2018;42(2):63-68.

13. Royal College of Psychiatrists. Health of the Nation Outcome Scales (HoNOS) [Internet]. 2020 [cited 2020 24/07]. Available from: https://www.rcpsych.ac.uk/events/in-house-training/health-of-nation-outcome-scales.

14. Terwee C, Prinsen C, Chiarotto A, de Vet H, Bouter L, Alonso J, et al. COSMIN methodology for assessing the content validity of PROMS. User manual version 1.0. February 2018 2018. Available from: https://cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf.

15. Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res. 2018;27(5):1159-70.

16. Grant J, Kinney M, Guzzetta C. A methodology for validating nursing diagnoses. Adv Nurs Sci. 1990;12(3):65-74.

17. Grant JS, Davis LL. Selection and use of content experts for instrument development. Res Nurs Health. 1997;20(3):269-74.

18. Grant JS, Kinney MR. Using the Delphi technique to examine the content validity of nursing diagnoses. Nurs Diagn. 1992;3(1):12-22.

19. Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, et al. Content validity--establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1--eliciting concepts for a new PRO instrument. Value Health. 2011;14(8):967-77.

20. Lynn MR. Determination and quantification of content validity. Nurs Res. 1986;35(6):382-5.

21. Wynd CA, Schmidt B, Schaefer MA. Two quantitative approaches for estimating content validity. West J Nurs Res. 2003;25(5):508-18.

22. Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, et al. Content validity--establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2--assessing respondent understanding. Value Health. 2011;14(8):978-88.

23. Powers III JH, Patrick DL, Walton MK, Marquis P, Cano S, Hobart J, et al. Clinician-reported outcome assessments of treatment benefit: report of the ISPOR Clinical Outcome Assessment Emerging Good Practices Task Force. Value Health. 2017;20(1):2-14.

24. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health. 2007;30(4):459-67.

25. Polit DF, Beck CT. The content validity index: are you sure you know what's being reported? Critique and recommendations. Res Nurs Health. 2006;29(5):489-97.

26. Burke MJ, Dunlap WP. Estimating interrater agreement with the average deviation index: A user's guide. Organ Res Methods. 2002;5(2):159-72.

27. Smith-Crowe K, Burke MJ. Interpreting the Statistical Significance of Observed AD Interrater Agreement Values: Correction to Burke and Dunlap (2002). Organ Res Methods. 2016;6(1):129-31.

28. Burgess P, Trauer T, Coombs T, McKay R, Pirkis J. What does 'clinical significance' mean in the context of the Health of the Nation Outcome Scales? Australas Psychiatry. 2009;17:141-48.

29. McClelland R, Trimble P, Fox ML, Stevenson MR, Bell B. Validation of an outcome scale for use in adult psychiatric practice. Qual Health Care. 2000;9(2):98-105.

30. National Mental Health Information Development Expert Advisory Panel. Mental Health National Outcomes and Casemix Collection: NOCC Strategic Directions 2014 – 2024. Canberra: 2013.

31. Germack HD, Noor EAM, Wang X, Hanrahan N. Association of Comorbid Serious Mental Illness Diagnosis With 30-Day Medical and Surgical Readmissions. JAMA Psychiatry. 2019;76(1):96-98.

32. Firth J, Siddiqi N, Koyanagi A, Siskind D, Rosenbaum S, Galletly C, et al. The Lancet Psychiatry Commission: a blueprint for protecting physical health in people with mental illness. The Lancet Psychiatry. 2019;6(8):675-712.

33. Australian Mental Health Outcomes and Classification Network (AMHOCN). Frequently Asked Questions [Internet]. AMHOCN; 2020. Available from: https://www.amhocn.org/resources/frequently-asked-questions.

34. Leyden KN, Hanneman SK. Validity of the Modified Richmond Agitation-Sedation Scale for use in sedated, mechanically ventilated swine. J Am Assoc Lab Anim Sci. 2012;51(1):63-68.

35. Gilbody SM, House AO, Sheldon TA. Psychiatrists in the UK do not use outcomes measures. National survey. Br J Psychiatry. 2002;180:101-3.

36. Duncan EA, Murray J. The barriers and facilitators to routine outcome measurement by allied health professionals in practice: a systematic review. BMC Health Serv Res. 2012;12:96.

37. Davison S, Hauck Y, Martyr P, Rock D. How mental health clinicians want to evaluate the care they give: a Western Australian study. Aust Health Rev. 2013;37(3):375-80.

38. Garland AF, Kruse M, Aarons GA. Clinicians and outcome measurement: what's the use? J Behav Health Serv Res. 2003;30(4):393-405.

39. Bender KG. The meagre outcomes of HoNOS. Australas Psychiatry. 2020;28(2):206-09.

40. Callaly T, Hyland M, Coombs T, Trauer T. Routine outcome measurement in public mental health: results of a clinician survey. Aust Health Rev. 2006;30(2):164-73.

41. Coombs T, Stapley K, Pirkis J. The multiple uses of routine mental health outcome measures in Australia and New Zealand: experiences from the field. Australas Psychiatry. 2011;19(3):247-53.

42. McKay R, Coombs T, Burgess P, Pirkis J, Christo J, delle-Vergini A. Development of clinical prompts to enhance decision support tools related to the National Outcomes and Casemix Collection. Version 1.0. Sydney, NSW: 2008.

# Appendix

**Appendix Table A.1.** COSMIN criteria for good content validity addressed in the survey

| Criteria | Addressed in survey? | Reason |
|---|---|---|
| Relevance | | |
| 1. Are the included items relevant for the construct of interest? | Yes[a] | Changes to content may have affected scales' relevance to determining overall clinical severity |
| 2. Are the included items relevant for the target population of interest? | No | No change to the number or title of the scales. |
| 3. Are the included items relevant for the context of use of interest? | Yes[a] | Changes to content may have affected their relevance to context of use (e.g., assessing change over time). |
| 4. Are the response options appropriate? | Yes[a] | The summary rating instructions were modified to improve clarity about the meaning of the severity ratings in relation to clinical significance. |
| 5. Is the recall period appropriate? | Yes[a] | The summary rating instructions were modified to improve clarity about the rating period. Minor modifications to Scales 11 and 12. |
| Comprehensiveness | | |
| 6. Are all key concepts included? | Yes[a] | Moderate changes were made to the rating instructions and descriptions to some Scales (e.g., Scales 2, 3, 4, 8 and 9) to improve the coverage of behaviours/problems to be rated. |
| Comprehensibility | | |
| 7. Are the instructions understood by the population of interest as intended? | Yes[a, b] | The overarching rating guidelines were modified to improve clarity about (a) what is to be taken into account when making ratings and (b) the meaning of the severity ratings.<br>The instructions and descriptions for several Scales (e.g., Scales 8, 11 and 12) were modified to improve clarity about what is to be taken into account when making ratings. |
| 8. Are the items and response options understood by the population of interest as intended? | Yes[a] | Scale descriptions were modified to improve clarity regarding the meaning of severity ratings across the severity ratings and across Scales (e.g., Scales 3, 7 and 9-12). |
| 9. Are the items appropriately worded? | Yes[a] | Some rewording of the rating descriptions was done to: (1) remove any subjective aspects of the wording; and (2) make language more contemporary and broadly applicable (e.g. Scale 12 - "public baths and library" became "public facilities", reference to "giro cheques" removed; Scale 6 "odd" changed to "unusual"). |
| 10. Do the response options match the question? | No | Format of response options (i.e., rating levels 0-4) was not changed. |

[a] Covered in 'core' questions asked about every scale; [b] Covered in section-specific questions.

**Appendix Table A.2.** Characteristics of experts who completed the survey, by country

| | Australia (N = 13) | | England (N = 10) | | New Zealand (N = 9) | |
|---|---|---|---|---|---|---|
| | **n** | **%** | **n** | **%** | **n** | **%** |
| **Main professional background** [a] | | | | | | |
| Nurse | 2 | 15 | 1 | 11 | 4 | 44 |
| Psychologist | 2 | 15 | 1 | 11 | 0 | 0 |
| Clinical psychologist | 0 | 0 | 1 | 11 | 1 | 11 |
| Social worker | 0 | 0 | 1 | 11 | 0 | 0 |
| Psychiatrist | 8 | 62 | 5 | 56 | 3 | 33 |
| Psychiatric registrar | 0 | 0 | 0 | 0 | 0 | 0 |
| Occupational therapist | 0 | 0 | 0 | 0 | 0 | 0 |
| Other | 1 [b] | 8 | 0 | 0 | 1 [c] | 11 |
| **Expertise with HoNOS** [d] | | | | | | |
| Rating HoNOS or reviewing HoNOS ratings made by others | 11 | 85 | 8 | 80 | 8 | 89 |
| Research in the measurement properties of the HoNOS and/or measuring clinical effectiveness | 11 | 85 | 4 | 40 | 5 | 56 |
| HoNOS staff training and/or using HoNOS results at a macro level | 6 | 46 | 9 | 90 | 5 | 56 |
| **Mental health settings worked with HoNOS** [d] | | | | | | |
| Inpatient | 9 | 69 | 7 | 70 | 6 | 67 |
| Residential [e] | 7 | 54 | - | - | - | - |
| Community | 12 | 92 | 10 | 100 | 8 | 89 |
| Other, non-clinical setting | 2 | 15 | 1 | 10 | 1 | 11 |
| **Aware of HoNOS 2018 prior to survey** | | | | | | |
| No, I was not aware of the HoNOS 2018 at all | 7 | 54 | 3 | 30 | 2 | 22 |
| Yes, I was aware of the HoNOS 2018, but have not used it in my work | 6 | 46 | 5 | 50 | 5 | 55 |
| Yes, I have used the HoNOS 2018 in my work | 0 | 0 | 2 | 20 | 1 | 11 |
| Not sure | 0 | 0 | 0 | 0 | 0 | 0 |
| Other | 0 | 0 | 0 | 0 | 1 [f] | 11 |
| | **M (SD)** | **Range** | **M (SD)** | **Range** | **M (SD)** | **Range** |
| **Years worked in mental health** [g] | 31 (8) | 12-43 | 24 (7) | 10-35 | 27 (9) | 16-42 |
| **Years worked with the HoNOS** | 14 (5) | 5-25 | 14 (4) | 10-21 | 16 (5) | 7-28 |

HoNOS, Health of the Nation Outcome Scales. M, mean. N, number. SD, standard deviation.

[a] Missing data for one respondent (England, n = 9). [b] "Clinical epidemiologist". [c] "Consumer and Family Leader". [d] Categories not mutually exclusive. [e] 'Residential' category included only in the Australian version of the survey. [f] "I have not seen the revised HoNOS 2018". [g] Due to missing data, n=8 for New Zealand and n = 9 for England.

**Appendix Table A.3.** Australian experts' ratings of the content validity of the HoNOS 2018 scales: relevance and comprehensiveness

| | Relevance | | | | | | | | Comprehensiveness | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | How important is this scale for determining overall clinical severity for adult mental health service consumers/patients | | | | How likely are repeat ratings on this scale to capture change in [scale-specific problems] during a period of mental health care? | | | | How well do the descriptors for each rating of 0-4 cover the range of [scale-specific problems] typically seen among adult mental health service consumers/patients? [a] | | | |
| **HoNOS 2018 scale** | **n** | **Range** | **Median** | **% positive** | **n** | **Range** | **Median** | **% positive** | **n** | **Range** | **Median** | **% positive** |
| Scale 1. Overactive or aggressive or disruptive or agitated behaviour | 12 | 2-4 | 3 | 75 | 12 | 1-4 | 3 | 58 | 13 | 1-4 | 3 | 77 |
| Scale 2. Non-accidental self-injury | 13 | 2-4 | 3 | 85 | 13 | 2-4 | 3 | 62 | 12 | 1-4 | 3 | 75 |
| Scale 3. Problem drinking or drug-taking | 13 | 2-4 | 3 | 85 | 12 | 2-3 | 2 | 33 | 12 | 1-4 | 2.5 | 50 |
| Scale 4. Cognitive problems | 13 | 2-4 | 3 | 85 | 13 | 1-4 | 2 | 46 | 13 | 2-4 | 3 | 92 |
| Scale 5. Physical illness or disability problems | 11 | 2-4 | 3 | 55 | 13 | 1-3 | 2 | 23 | 13 | 1-4 | 3 | 54 |
| Scale 6. Problems associated with hallucinations and /or delusions | 12 | 2-4 | 3.5 | 92 | 13 | 2-4 | 3 | 92 | 13 | 1-4 | 3 | 77 |
| Scale 7. Problems with depressed mood | 13 | 2-4 | 4 | 92 | 13 | 2-4 | 3 | 77 | 13 | 2-4 | 3 | 92 |
| Scale 8. Other mental and behavioural problems | 13 | 2-3 | 3 | 69 | 13 | 1-3 | 3 | 62 | 13 | 2-4 | 3 | 69 |
| Scale 9. Problems with relationships | 13 | 2-4 | 3 | 85 | 13 | 1-3 | 3 | 77 | 13 | 2-4 | 3 | 92 |
| Scale 10. Problems with activities of daily living | 13 | 2-4 | 3 | 85 | 12 | 2-3 | 3 | 75 | 12 | 1-4 | 3 | 83 |
| Scale 11. Problems with housing and living conditions | 12 | 1-4 | 3 | 58 | 10 | 1-3 | 3 | 80 | 13 | 1-3 | 3 | 54 |
| Scale 12. Problems with occupation and activities | 11 | 2-4 | 3 | 73 | 13 | 2-3 | 3 | 77 | 13 | 1-3 | 3 | 62 |

% positive, percentage of ratings of 3 or 4. HoNOS, Health of the Nation Outcome Scales. n, number. [a] To fit the wording of Scale 8, the equivalent question for Scale 8 was: How well do problems A-O cover the range of other mental and behavioural problems typically seen among adult mental health service patients?

**Appendix Table A.4.** Australian experts' ratings of the content validity of the HoNOS 2018 scales: comprehensibility

| | Comprehensibility | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | How helpful is the glossary for determining what to include when rating [scale-specific problems]? [a, b] | | | | How well do the descriptors for each rating of 0-4 correspond to the different levels of severity of [scale-specific problems]? | | | | How consistent is the wording of the glossary with language used in contemporary mental health practice? | | | |
| **HoNOS 2018 scale** | **n** | **Range** | **Median** | **% positive** | **n** | **Range** | **Median** | **% positive** | **n** | **Range** | **Median** | **% positive** |
| Scale 1. Overactive or aggressive or disruptive or agitated behaviour | 13 | 2-4 | 3 | 92 | 13 | 2-4 | 3 | 62 | 12 | 2-4 | 3 | 83 |
| Scale 2. Non-accidental self-injury | 13 | 2-4 | 3 | 85 | 13 | 1-4 | 3 | 54 | 13 | 2-4 | 3 | 69 |
| Scale 3. Problem drinking or drug-taking | 13 | 2-4 | 3 | 69 | 12 | 1-4 | 2.5 | 50 | 13 | 1-4 | 3 | 62 |
| Scale 4. Cognitive problems | 13 | 2-4 | 3 | 85 | 12 | 2-4 | 3 | 83 | 12 | 3-4 | 3 | 100 |
| Scale 5. Physical illness or disability problems | 13 | 1-4 | 2 | 23 | 13 | 1-4 | 2 | 31 | 11 | 1-3 | 2 | 45 |
| Scale 6. Problems associated with hallucinations and /or delusions | 13 | 2-4 | 3 | 92 | 13 | 2-4 | 3 | 92 | 13 | 2-4 | 3 | 92 |
| Scale 7. Problems with depressed mood | 13 | 2-4 | 3 | 77 | 12 | 2-4 | 3 | 75 | 13 | 3-4 | 3 | 100 |
| Scale 8. Other mental and behavioural problems | 13 | 2-4 | 3 | 69 | 12 | 2-3 | 3 | 58 | 13 | 2-3 | 3 | 69 |
| Scale 9. Problems with relationships | 13 | 2-4 | 3 | 77 | 13 | 2-4 | 3 | 85 | 12 | 2-3 | 3 | 83 |
| Scale 10. Problems with activities of daily living | 13 | 2-4 | 3 | 85 | 13 | 2-4 | 3 | 92 | 13 | 2-4 | 3 | 92 |
| Scale 11. Problems with housing and living conditions | 13 | 2-3 | 3 | 62 | 13 | 2-3 | 3 | 77 | 12 | 2-3 | 3 | 67 |
| Scale 12. Problems with occupation and activities | 13 | 1-4 | 3 | 54 | 13 | 1-3 | 3 | 77 | 13 | 2-4 | 3 | 69 |

% positive, percentage of ratings of 3 or 4. HoNOS, Health of the Nation Outcome Scales. n, number. [a] Question text differed across scales; depending on the glossary, "what to rate and include" or "what to rate and consider" was substituted for the phrase "what to include". [b] To fit the wording of Scale 8, the equivalent question for Scale 8 was: How helpful is the glossary for determining which other mental and behavioural problem to rate on this scale?

**Appendix Table A.5.** England experts' ratings of the content validity of the HoNOS 2018 scales: relevance and comprehensiveness

| | Relevance | | | | | | | | Comprehensiveness | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | How important is this scale for determining overall clinical severity for adult mental health service consumers/patients? | | | | How likely are repeat ratings on this scale to capture change in [scale-specific problems] during a period of mental health care? | | | | How well do the descriptors for each rating of 0-4 cover the range of [scale-specific problems] typically seen among adult mental health service consumers/patients? [a] | | | |
| **HoNOS 2018 scale** | **n** | **Range** | **Median** | **% positive** | **n** | **Range** | **Median** | **% positive** | **n** | **Range** | **Median** | **% positive** |
| Scale 1. Overactive or aggressive or disruptive or agitated behaviour | 10 | 2-4 | 3 | 80 | 9 | 2-4 | 3 | 56 | 10 | 2-4 | 3 | 60 |
| Scale 2. Non-accidental self-injury | 9 | 2-4 | 3 | 89 | 10 | 2-4 | 3 | 60 | 10 | 2-4 | 2 | 30 |
| Scale 3. Problem drinking or drug-taking | 10 | 3-4 | 3 | 100 | 10 | 1-3 | 2.5 | 50 | 10 | 1-4 | 3 | 60 |
| Scale 4. Cognitive problems | 10 | 3-4 | 3 | 100 | 10 | 1-3 | 3 | 70 | 10 | 2-3 | 3 | 80 |
| Scale 5. Physical illness or disability problems | 10 | 1-4 | 3 | 80 | 10 | 1-4 | 3 | 70 | 10 | 1-4 | 3 | 80 |
| Scale 6. Problems associated with hallucinations and /or delusions | 10 | 3-4 | 3.5 | 100 | 10 | 2-4 | 3 | 70 | 10 | 2-4 | 3 | 70 |
| Scale 7. Problems with depressed mood | 10 | 3-4 | 3.5 | 100 | 10 | 2-4 | 3 | 70 | 10 | 1-4 | 3 | 90 |
| Scale 8. Other mental and behavioural problems | 10 | 3-4 | 4 | 100 | 10 | 1-4 | 3 | 70 | 10 | 1-4 | 3 | 60 |
| Scale 9. Problems with relationships | 9 | 2-4 | 3 | 78 | 10 | 1-3 | 3 | 80 | 10 | 2-3 | 3 | 70 |
| Scale 10. Problems with activities of daily living | 10 | 2-4 | 3 | 90 | 10 | 1-4 | 3 | 60 | 10 | 1-4 | 3 | 80 |
| Scale 11. Problems with housing and living conditions | 10 | 2-3 | 3 | 70 | 10 | 1-3 | 3 | 70 | 10 | 1-3 | 3 | 60 |
| Scale 12. Problems with occupation and activities | 10 | 1-4 | 3 | 70 | 10 | 1-3 | 3 | 60 | 10 | 1-4 | 2.5 | 50 |

% positive, percentage of ratings of 3 or 4. HoNOS, Health of the Nation Outcome Scales. n, number. [a] To fit the wording of Scale 8, the equivalent question for Scale 8 was: How well do problems A-O cover the range of other mental and behavioural problems typically seen among adult mental health service consumers/patients?

**Appendix Table A.6.** England experts' ratings of the content validity of the HoNOS 2018 scales: comprehensibility

| | Comprehensibility | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | How helpful is the glossary for determining what to include when rating [scale-specific problems]? [a, b] | | | | How well do the descriptors for each rating of 0-4 correspond to the different levels of severity of [scale-specific problems]? | | | | How consistent is the wording of the glossary with language used in contemporary mental health practice? | | | |
| HoNOS 2018 scale | n | Range | Median | % positive | n | Range | Median | % positive | n | Range | Median | % positive |
| Scale 1. Overactive or aggressive or disruptive or agitated behaviour | 10 | 2-4 | 3 | 70 | 10 | 1-3 | 2.5 | 50 | 10 | 1-4 | 3 | 70 |
| Scale 2. Non-accidental self-injury | 10 | 2-4 | 3 | 70 | 10 | 1-3 | 2 | 30 | 10 | 1-3 | 3 | 70 |
| Scale 3. Problem drinking or drug-taking | 10 | 1-4 | 3.5 | 70 | 10 | 1-4 | 3 | 60 | 10 | 1-3 | 3 | 60 |
| Scale 4. Cognitive problems | 9 | 2-4 | 3 | 89 | 9 | 2-4 | 3 | 89 | 10 | 2-3 | 3 | 80 |
| Scale 5. Physical illness or disability problems | - [c] | - | - | - | 10 | 2-4 | 2.5 | 50 | 10 | 1-3 | 3 | 70 |
| Scale 6. Problems associated with hallucinations and /or delusions | 10 | 2-4 | 3 | 80 | 10 | 2-4 | 3 | 80 | 10 | 2-4 | 3 | 80 |
| Scale 7. Problems with depressed mood | 10 | 1-4 | 3 | 70 | 10 | 2-4 | 3 | 90 | 10 | 1-3 | 3 | 80 |
| Scale 8. Other mental and behavioural problems | 10 | 2-4 | 3 | 80 | 10 | 2-4 | 3 | 80 | 10 | 1-3 | 3 | 90 |
| Scale 9. Problems with relationships | 10 | 2-3 | 3 | 80 | 10 | 2-3 | 3 | 80 | 10 | 1-3 | 3 | 80 |
| Scale 10. Problems with activities of daily living | 10 | 2-4 | 3 | 80 | 10 | 1-4 | 3 | 90 | 10 | 2-4 | 3 | 90 |
| Scale 11. Problems with housing and living conditions | 10 | 1-4 | 2.5 | 50 | 9 | 1-3 | 3 | 67 | 10 | 1-3 | 3 | 80 |
| Scale 12. Problems with occupation and activities | 10 | 1-3 | 3 | 60 | 10 | 1-4 | 3 | 60 | 10 | 1-3 | 3 | 60 |

% positive, percentage of ratings of 3 or 4. HoNOS, Health of the Nation Outcome Scales. n, number. [a] Question text differed across scales; depending on the glossary, "what to rate and include" or "what to rate and consider" was substituted for the phrase "what to include". [b] To fit the wording of Scale 8, the equivalent question for Scale 8 was: How helpful is the glossary for determining which other mental and behavioural problem to rate on this scale? [c] Question was inadvertently omitted from the England survey.

**Appendix Table A.7.** New Zealand experts' ratings of the content validity of the HoNOS 2018 scales: relevance and comprehensiveness

| | Relevance | | | | | | | | Comprehensiveness | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | How important is this scale for determining overall clinical severity for adult mental health service consumers/patients? | | | | How likely are repeat ratings on this scale to capture change in [scale-specific problems] during a period of mental health care? | | | | How well do the descriptors for each rating of 0-4 cover the range of [scale-specific problems] typically seen among adult mental health service consumers/patients? [a] | | | |
| **HoNOS 2018 scale** | **n** | **Range** | **Median** | **% positive** | **n** | **Range** | **Median** | **% positive** | **n** | **Range** | **Median** | **% positive** |
| Scale 1. Overactive or aggressive or disruptive or agitated behaviour | 9 | 2-4 | 3 | 89 | 9 | 2-4 | 3 | 89 | 9 | 2-4 | 3 | 78 |
| Scale 2. Non-accidental self-injury | 9 | 3-4 | 3 | 100 | 9 | 2-4 | 3 | 78 | 9 | 2-4 | 3 | 89 |
| Scale 3. Problem drinking or drug-taking | 9 | 3-4 | 4 | 100 | 9 | 2-4 | 3 | 89 | 9 | 2-4 | 3 | 89 |
| Scale 4. Cognitive problems | 9 | 2-4 | 3 | 89 | 9 | 1-4 | 3 | 89 | 9 | 1-4 | 3 | 89 |
| Scale 5. Physical illness or disability problems | 9 | 3-4 | 3 | 100 | 9 | 2-4 | 3 | 89 | 8 | 1-3 | 3 | 88 |
| Scale 6. Problems associated with hallucinations and /or delusions | 9 | 3-4 | 4 | 100 | 9 | 3-4 | 4 | 100 | 9 | 3-4 | 3 | 100 |
| Scale 7. Problems with depressed mood | 9 | 3-4 | 3 | 100 | 9 | 3-4 | 3 | 100 | 9 | 2-4 | 3 | 78 |
| Scale 8. Other mental and behavioural problems | 9 | 3-4 | 3 | 100 | 9 | 1-3 | 3 | 78 | 9 | 2-3 | 3 | 78 |
| Scale 9. Problems with relationships | 9 | 3-4 | 3 | 100 | 9 | 1-4 | 3 | 89 | 9 | 1-4 | 3 | 67 |
| Scale 10. Problems with activities of daily living | 9 | 3-4 | 3 | 100 | 9 | 1-4 | 3 | 89 | 9 | 1-3 | 3 | 78 |
| Scale 11. Problems with housing and living conditions | 9 | 2-4 | 3 | 89 | 9 | 1-4 | 3 | 89 | 9 | 1-4 | 3 | 89 |
| Scale 12. Problems with occupation and activities | 9 | 2-4 | 3 | 89 | 9 | 2-4 | 3 | 89 | 9 | 1-4 | 3 | 89 |

% positive, percentage of ratings of 3 or 4. HoNOS, Health of the Nation Outcome Scales. n, number. [a] To fit the wording of Scale 8, the equivalent question for Scale 8 was: How well do problems A-O cover the range of other mental and behavioural problems typically seen among adult mental health service patients?

**Appendix Table A.8.** New Zealand experts' ratings of the content validity of the HoNOS 2018 scales: comprehensibility

| | Comprehensibility | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | How helpful is the glossary for determining what to include when rating [scale-specific problems]? [a, b] | | | | How well do the descriptors for each rating of 0-4 correspond to the different levels of severity of [scale-specific problems]? | | | | How consistent is the wording of the glossary with language used in contemporary mental health practice? | | | |
| **HoNOS 2018 scale** | **n** | **Range** | **Median** | **% positive** | **n** | **Range** | **Median** | **% positive** | **n** | **Range** | **Median** | **% positive** |
| Scale 1. Overactive or aggressive or disruptive or agitated behaviour | 9 | 2-4 | 3 | 67 | 9 | 2-4 | 3 | 67 | 8 | 2-4 | 3 | 88 |
| Scale 2. Non-accidental self-injury | 8 | 2-4 | 3 | 88 | 9 | 2-4 | 3 | 67 | 9 | 2-4 | 3 | 89 |
| Scale 3. Problem drinking or drug-taking | 9 | 2-3 | 3 | 89 | 9 | 1-4 | 3 | 89 | 9 | 1-4 | 3 | 89 |
| Scale 4. Cognitive problems | 9 | 2-4 | 3 | 78 | 9 | 1-3 | 3 | 89 | 8 | 1-4 | 3 | 63 |
| Scale 5. Physical illness or disability problems | 9 | 1-4 | 3 | 78 | 9 | 1-4 | 3 | 78 | 9 | 1-4 | 3 | 89 |
| Scale 6. Problems associated with hallucinations and /or delusions | 9 | 2-4 | 3 | 89 | 9 | 2-4 | 3 | 89 | 9 | 2-4 | 3 | 89 |
| Scale 7. Problems with depressed mood | 9 | 2-4 | 3 | 89 | 9 | 2-4 | 3 | 78 | 9 | 2-4 | 3 | 56 |
| Scale 8. Other mental and behavioural problems | 9 | 2-4 | 3 | 89 | 9 | 2-3 | 3 | 67 | 9 | 2-3 | 3 | 89 |
| Scale 9. Problems with relationships | 9 | 2-4 | 3 | 67 | 9 | 2-4 | 3 | 67 | 9 | 1-4 | 3 | 67 |
| Scale 10. Problems with activities of daily living | 9 | 3-4 | 3 | 100 | 9 | 2-4 | 3 | 89 | 9 | 2-4 | 3 | 78 |
| Scale 11. Problems with housing and living conditions | 9 | 3-4 | 3 | 100 | 9 | 1-4 | 3 | 78 | 9 | 2-4 | 3 | 89 |
| Scale 12. Problems with occupation and activities | 9 | 1-3 | 3 | 67 | 9 | 1-3 | 3 | 89 | 9 | 1-4 | 3 | 78 |

% positive, percentage of ratings of 3 or 4. HoNOS, Health of the Nation Outcome Scales. n, number. [a] Question text differed across scales; depending on the glossary, "what to rate and include" or "what to rate and consider" was substituted for the phrase "what to include". [b] To fit the wording of Scale 8, the equivalent question for Scale 8 was: How helpful is the glossary for determining which other mental and behavioural problem to rate on this scale?

**Appendix Table A.9.** Summary of themes identified through the qualitative assessment, by country

| Themes | Australia | England | New Zealand |
|---|---|---|---|
| **Experts' concerns** | | | |
| Too many phenomena | ✓ | ✓ | ✓ |
| Ambiguity | ✓ | ✓ | ✓ |
| Need for more description or examples | ✓ | ✓ | ✓ |
| Assessment challenges | ✓ | ✓ | ✓ |
| Lack of fit with clinical thinking | ✓ | ✓ | ✓ |
| Problems with language | ✓ | ✓ | ✓ |
| Importance | ✓ | ✓ | |
| Challenges to capturing change | ✓ | ✓ | ✓ |
| Incomplete coverage | ✓ | ✓ | ✓ |
| Need for training | ✓ | ✓ | ✓ |
| **Experts' summary comments** | | | |
| HoNOS 2018 is preferred/an improvement | ✓ | ✓ | ✓ |
| Endorse HoNOS but no preference | | ✓ | |
| Lacks clinical utility | ✓ | ✓ | |
| Need to address completion rates/rating consistency | ✓ | ✓ | ✓ |