Received: 25 October 2023 | Revised: 29 February 2024 | Accepted: 11 March 2024

DOI: 10.1111/jpm.13047

ORIGINAL ARTICLE

WILEY

Reliability testing of the Health of the Nation Outcome Scales 2018

Jon Painter[1] | Mick James[2]

[1]Department of Nursing and Midwifery, Sheffield Hallam University, Sheffield, UK

[2]Royal College of Psychiatrists, London, UK

Correspondence
Jon Painter, Department of Nursing and Midwifery, Sheffield Hallam University, Sheffield, UK.
Email: j.painter@shu.ac.uk

Accessible Summary

What is known on the subject?

- The Health of the Nation Outcome Scales (HoNOS) is a widely used clinical measure designed to rate and monitor the outcomes of service users accessing specialist mental healthcare.
- Since its development (in 1996), numerous research studies have confirmed the HoNOS captures the aspects of care that it purports to (validity), and that clinicians' ratings are consistent both over time, and between different raters (reliability).

What the paper adds to existing knowledge?

- In 2018, the HoNOS was reviewed with updates made to some terminology and other revisions intended to remove ambiguity in the guidance for raters. However, although the new version (HoNOS 2018) was accompanied by a recommendation that its validity and reliability be re-tested this was not undertaken.
- To our knowledge, this is the first study to re-assess the updated tool's reliability by measuring the level of agreement between different raters. Our findings confirm that there is an acceptable level of consistency between student mental health nurses that have been trained to use the (new) HoNOS 2018.

What are the implications for practice?

- The HoNOS is nationally mandated for use by all specialist mental healthcare providers in the UK.
- Our findings provide some assurance that, with appropriate update training and monitoring of organisational-level data sets, the original HoNOS glossary can safely be replaced with the HoNOS 2018 to ensure more contemporary routine outcome measurement can occur.

Abstract

Introduction: The Health of the Nation Outcome Scales (HoNOS) is a well-established clinician rated outcome measure for use in mental health services. Following an international review, an updated version (HoNOS 2018) was published with a recommendation that its psychometric properties be re-tested prior to widespread implementation. To date, only one such study has been published.

Aims: To test the inter-rater agreement levels for HoNOS 2018.

Method: Third-year student mental health nurses received training to complete the HoNOS 2018. Following this timetabled session, they were each invited to independently rate two, randomly selected, videos of (simulated) patient interviews. The resulting data were then analysed to calculate the tool's internal consistency and inter-rater agreement levels.

Results: The 55 participants provided 106 ratings from four vignettes. Cronbach's alphas and McDonalds omegas confirmed the revised tool's internal consistency was acceptable. Average measure intraclass correlation coefficients for the four patient vignettes indicated excellent reliability.

Implications for practice: This study provides initial assurance that the HoNOS 2018 is a reliable clinician rated outcome measure suitable for use in routine clinical practice by relatively inexperienced mental health practitioners with limited training.

KEYWORDS
HoNOS, mental health, outcome measurement, psychiatry, psychometrics, reliability

1 | INTRODUCTION

The Health of the Nation Outcomes Scales (HoNOS) (Wing et al., 1996) is a 12-item clinical measure originally designed to assist practitioners in secondary mental health care to quantify service users' treatment outcomes. It is arguably the most widely used needs assessment tool used in UK mental health services (Gilbody et al., 2002) with reports of up to 85% of secondary care service users nationally having been rated (Childs, 2015). Since its development, numerous studies have ascertained and/or confirmed different aspects of its validity, reliability and utility (e.g., Adams et al., 2000; McClelland et al., 2000; Trauer & Buckingham, 2006; Wing et al., 1998). However, after more than two decades of use, the Royal College of Psychiatrists orchestrated an international review in order to update some of the tool's anachronistic language, remove ambiguity and improve consistency of ratings. This resulted in an updated version, entitled HoNOS 2018 that was published with a clear recommendation that it should be subject to re-testing before widespread implementation and use (James et al., 2018).

Since then, one of the only published studies into the psychometric properties of the new version was undertaken by Harris et al. (2022). Their validity study comprised an online survey of clinicians, with expertise in using HoNOS, to ascertain the content validity of each HoNOS 2018 item in terms of relevance, comprehensiveness and comprehensibility. The I-CVI values (which represent the proportion of positive responses) were analysed for 72 core survey questions. The majority of ratings were positive, with 50% or more experts giving positive ratings for all but one question and around 70% of experts giving positive ratings for nearly 70% of the core questions.

When asked to consider each item's clinical significance, 11 out of 12 HoNOS items achieved excellent content validity based on the a priori criterion (I-CVI ≥0.75). Nine items met this criterion for the helpfulness of its glossary, and six items met this criterion for their ability to capture change, accurately depict differing severity levels, and their relevance to contemporary mental health practice. Overall, three items met the criterion on all questions, two items met the criterion on all questions except one (related to capturing change), whilst three items met the criterion on only one or two questions. For the survey responses regarding each HoNOS item, the average deviation (AD) index values, which measure agreement between experts, were all below the critical value of 0.68, indicating acceptable and significant agreement (Harris et al., 2022). However, against this generally positive picture, the expert clinicians did express some concerns including items measuring multiple phenomena; residual ambiguity of wording and other linguistic issues; a lack of illustrative examples; a disconnect with 'clinical thinking'; relevance; coverage, and ability of some items to capture change.

In light of the dearth of evidence regarding the impact of the recent updates made to HoNOS (Wing et al., 1996), the aim of this study was to investigate the reliability of the HoNOS 2018, using rating data obtained from HoNOS 2018 training sessions, before any formal recommendations to use it in routine clinical practice be made.

2 | MATERIALS AND METHODS

2.1 | Participants and recruitment

As part of their routine pre-registration mental health nurse training syllabus, 95 of the cohort of 105 third-year students from Sheffield Hallam University attended one of two timetabled sessions on outcome measurement during which they were trained in the use of the HoNOS 2018. As the session was being offered to all

available students (i.e., the entire cohort), a formal a priori power calculation was not performed. This particular 3-h training session was delivered by the RCPsych National HoNOS Adviser (MJ) and a senior lecturer in mental health nursing (JP) both of whom were experienced HoNOS trainers and had been members of the International Expert Advisory Board that revised the HoNOS and published the HoNOS 2018. The training sessions mirrored the Royal College of Psychiatrists (RCPsych) routine HoNOS training package in that they used standardised materials and involved a detailed 'walk through' of the tool's glossary. To reinforce the key learning points for each item, the trainers used clinical examples and encouraged participants to ask questions to clarify points as required.

All trained students were then offered the opportunity to obtain a Royal College of Psychiatrists (RCPsych) Continuing Professional Development (CPD) Certificate, by participating in this research study. This created a self-selecting, convenience sample of 55 third-year student mental health nurses.

2.2 | Measure

The HoNOS 2018 is an updated version of the original HoNOS (Wing et al., 1996). A copy of the HoNOS 2018's glossary has been published previously (see James et al., 2018); however, in brief, it maintains the same core rules, structures and 5-point scale structure (from no problem to severe/very severe problem) across 12 items. In this regard, reductions in total HoNOS scores over time represent a reduction in symptom severity (recovery); however, it is important to note that a reduced rating in one item can be compensated by an increase in another meaning the total score may hide clinically significant changes in presentations. As a consequence, numerous studies have sought to understand the latent structure of the HoNOS with a view to creating more clinically meaningful and statistically robust sub-totals, for example, Wing et al. (1996), Preston (2000), Newnham et al. (2009), McClelland et al. (2000), Trauer (1999), Speak and Muncer (2015) and Lovaglio and Monzani (2012). It is also still intended to be a short, and simple enough, measure for use in routine practice to capture the breadth of issues experienced by service users of secondary care mental health services, and to be sensitive to changes in their presentation over time.

2.3 | Data collection

In order to obtain a set of HoNOS 2018 ratings for analysis, four short (approximately 15-minute) videos of simulated clinical interviews were created with the RCPsch HoNOS advisor playing the role of a clinician in each and four actors each portraying a different service user. These were as follows: a female with symptoms frequently seen in individuals diagnosed with an emotionally unstable personality disorder; a male with memory problems indicative of mild dementia; a female exhibiting florid psychotic phenomena and a female with moderate depressive features. With 38% of the UK's psychiatric bed days used by people diagnosed with psychosis, 14% with mood disorders, 10% with personality disorder and 5% with dementia (RCPsych, 2019), these different clinical presentations were selected to ensure the ratings were collectively representative of a range of service users for whom the tool was designed.

Each of the students were allocated two of the vignettes by printing equal numbers of the video weblinks onto cards for them to draw from a hat. This simple randomisation method (Suresh, 2011) ensured students did not, for example, select clinical presentations they were more confident in assessing. Once they had selected their vignettes, students were asked to rate them independently, with the importance of doing so stressed to them during training. They were able to do this at their own pace and to re-watch all, or part of each video, as many times as they needed to prior to submitting their ratings. Provided they had rated their allocated videos, they were also permitted to rate one or both of the remaining two if they wished (with the choice of vignette left to the student). They were required to submit their ratings online within 3 days of their training to minimise the effects of memory fade.

2.4 | Data analysis

Data were cleansed and formatted in Microsoft Excel prior to being exported to Jamovi version 2.3.21 (The jamovi project, 2023) and SPSS version 26 (IBM Corp, 2019) for analysis. According to Boateng et al. (2018), two key assessments of a measure's reliability are its internal consistency and its test–retest reliability.

Internal consistency is the degree to which the items in a measure correlate, and hence measure the same underlying construct. One of the most commonly used reliability indexes is the Cronbach's alpha; however, concerns have been raised that the data in question often violate the assumptions necessary for a meaningful result (McNeish, 2018). These include unidimensional measures consisting of continuous scales that adhere to tau-equivalence (i.e., that all items contribute equally to the total scale score), that errors are uncorrelated, and that ratings that are normally distributed. In contrast, the assumptions for McDonald's omegas are less restrictive; hence, it is being increasingly advocated (Hayes & Jacob, 2020). In light of this, both tests of consistency were performed (hence requiring Jamovi to calculate omegas).

The data did not contain repeat assessments of the same service user (vignette) by the same rater, from which test–retest reliability can be deduced. Instead, multiple ratings of each vignette were made by different raters which still allowed assessment of inter-rater agreement. It has been suggested that, despite their theoretical pros and cons (O'Neil, 2017), in practice most agreement indices tend to produce very similar results and that the choice of index is therefore largely a matter of personal preference (LeBreton & Senter, 2008). Given its popularity and ease of interpretation, the intraclass

correlation coefficient (ICC) was calculated for each vignette. More specifically, the study design and nature of the data dictated a two-way, consistency, average measures ICC was performed (McGraw & Wong, 1996). Results were then classified according to Koo and Li (2016) thresholds of poor (<0.5); moderate (0.5–0.74); good (0.75–0.9) and excellent.

2.5 | Ethics

The project was ethically approved by Sheffield Hallam University's Ethics Committee (ER40282792). In line with good research practice, a verbal explanation of the project was given to students as well as a written Participant Information Sheet, and the opportunity to ask questions. All students that agreed to participate, and for their data to be used, were then provided with weblinks to access the videos and to record their HoNOS 2018 ratings. This online data collection included a written consent form prior to the HoNOS 2018 rating fields. The only identifiable data requested were the student's university email address, which was required when sending their CPD certificate and to identify their data should they have subsequently asked to withdraw from the study. All data were held on the university's dedicated research server in a folder only accessible to the principal investigator.

3 | RESULTS

In total, 106 ratings were returned by the 55 participants with the mean number of ratings per student = 1.93 (SD 0.57). Vignette one was rated by 30 students; vignette two by 23; vignette three by 28 and vignette four by 25 students. With reference to Nunally's rule of thumb (Nunnally & Bernstein, 1994), internal consistency for the whole dataset was found to be acceptable—Cronbach's Alpha = 0.713 95% CI [0.624, 0.788]; McDonalds Omega = 0.777.

For each of the four sets of ratings (one set per vignette), a high degree of reliability was found. For vignette one, the average measure ICC was 0.988 with a 95% confidence interval from 0.975 to 0.996 ($F = 82.378$, $p < .001$). For vignette two, the average measure ICC was 0.971 with a 95% confidence interval from 0.940 to 0.991 ($F = 34.916$, $p < .001$). For vignette three, the average measure ICC was .983 with a 95% confidence interval from 0.963 to 0.995 ($F = 58.725$, $p < .001$). Finally, for vignette four, the average measure ICC was 0.900 with a 95% confidence interval from 0.764 to 0.976 ($F = 10.009$, $p < .001$).

4 | DISCUSSION

In previous studies of the internal consistency of the HoNOS, Cronbach's alphas have ranged from 0.59 to 0.76 (Pirkis et al., 2005). Our results suggest that the updates made to the HoNOS have not adversely affected the new tool's (HoNOS 2018) internal consistency (as assessed with both Cronbach's Alpha and McDonalds Omega). Additionally, ratings for each of four, clinically diverse service user vignettes, were found to have excellent reliability (ICCs ranging from 0.94 to 0.99). This compares favourably to Wing et al.'s (1998) original development study where total score ICCs were 0.86 for one site and 0.77 for the second, with individual item ICCs extending between 0.49 and 0.99.

Therefore, whilst far from definitive, given the lack of published evidence for the reliability of the updated HoNOS 2018, the results obtained are encouraging. When combined with the original HoNOS's well-established validity (Wing et al., 1998), the clinical expertise harnessed for the update (James et al., 2018), and the generally positive results from Harris et al.'s (2022) content validity study it seems reasonable to suggest the HoNOS 2018 could now be introduced into routine clinical practice. The nature of the updates made was limited to linguistic changes rather than more fundamental changes to the number of items etc. As a result, this would only require relatively minor changes to electronic patient record systems as well as a relatively brief training update (which is likely to improve inter-rater agreement in its own right). Following its introduction, HoNOS data guardians at all levels (from ward to board and beyond) should monitor ratings and investigate any significant shifts in mean severity ratings to assure organisations that these are legitimate.

4.1 | Limitations

As with all research, this study had a number of limitations. For example, an a priori power calculation to determine sample adequacy was not performed; however, this was deemed unnecessary as the entire study population were invited to participate. Convenience sampling, sample size and a lack of information regarding the participants' demographics for example, limit generalisability (Acharya et al., 2013) however, once recruited, participants were randomly allocated vignettes which increases confidence in results. Likewise, this study lacked some of the controls commonly found in true experimental designs (e.g., we did not capture how many times students watched each video before rating it and they were not supervised whilst independently rating their vignettes); however, its pragmatic nature can also be seen as more reflective of routine clinical training/practice and therefore an inherent strength. Finally, the fact that participants were not (yet) registered nurses could be seen as a challenge to the study's ecological validity. Conversely however, in line with the findings of Spengler et al.'s (2009) meta-analysis, we would argue that using students, with relatively little clinical experience to draw upon, is potentially a more stringent test of a psychometric measure's rating instructions than using more experienced nurses as raters as they are less experienced at using such clinical measures and at assessing service users generally.

In conclusion, this study has provided initial assurance that, like its predecessor, the HoNOS 2018 is a reliable clinician rated

outcome measure which is suitable for use in routine clinical practice by relatively inexperienced/junior mental health practitioners with limited training.

5 | RELEVANCE STATEMENT

The Health of the Nation Outcome Scales (HoNOS) (Wing et al., 1996) is a clinician rated outcome measure used internationally. Its routine use in specialist mental health settings is mandated in the NHS, Australia, and New Zealand. An updated version (HoNOS 218) was developed by James et al. (2018) and published with a recommendation that its validity and reliability be confirmed before widespread implementation. The only such study published to date (Harris et al., 2022), confirmed its clinical face validity. Therefore, our study provides important evidence of the reliability of this new version prior to its implementation across the NHS and beyond.

CONFLICT OF INTEREST STATEMENT

MJ is the National HoNOS adviser for the RCPSych who own the copyright for the HoNOS and HoNOS 2018. JP and MJ were members of the RCPsych's expert advisory group that updated the HoNOS and published the HoNOS 2018.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ETHICAL APPROVAL STATEMENT

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures involving human subjects/patients were ethically approved by Sheffield Hallam University (ER40282792).

ORCID

Jon Painter https://orcid.org/0000-0003-1589-4054

REFERENCES

Acharya, A., Prakash, A., & Nigam, A. (2013). Sampling: Why and how of it? *Indian Journal of Medical Specialities*, *4*(2), 330–333. https://doi.org/10.7713/ijms.2013.0032

Adams, M., Palmer, A., O'Brien, J. T., & Crook, W. (2000). Health of the nation outcome scales for psychiatry: Are they valid? *J Ment Health UK*, *9*(2), 193–198. https://doi.org/10.1080/09638230050009186

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Front Public Health*, *6*(49), 1–18. https://doi.org/10.3389/fpubh.2018.00149

Childs. (2015). *Mental Health and Learning Disability Statistics*. Monthly Report, V1.0 https://files.digital.nhs.uk/publicationimport/pub19xxx/pub19578/mhlds-monthly-exec-sep-2015.pdf

Gilbody, S., House, A., & Sheldon, T. (2002). Psychiatrists in the UK do not use outcomes measures: National survey. *Br J Psychiatry*, *180*(2), 101–103. https://doi.org/10.1192/bjp.180.2.101

Harris, M., Tapp, C., Arnautovska, U., Coombs, T., Dickinson, R., James, M., Painter, J., Smith, M., Jury, A., Lai, J., & Burgess, P. M. (2022). Assessing the content validity of the revised health of the nation outcome scales (HoNOS 2018). *Int J Environ Res Public Health*, *19*(16), 9895. https://doi.org/10.3390/ijerph19169895

Hayes, A., & Jacob, C. (2020). Use omega rather than Cronbach's alpha for estimating reliability but…. *Commun Methods Meas*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

IBM Corp. (2019). *IBM SPSS statistics for windows, version 26.0*. IBM Corp.

James, M., Painter, J., Buckingham, B., & Stewart, M. (2018). A review and update of the health of the nation outcome scales (HoNOS). *BJPsych Bulletin*, *42*(2), 63–68. https://doi.org/10.1192/bjb.2017.17

Koo, T., & Li, M. (2016). A guideline of selecting and reporting Intraclass correlation coefficients for reliability research. *J Chiropr Med*, *15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

LeBreton, J., & Senter, J. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organ Res Methods*, *11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Lovaglio, P. G., & Monzani, E. (2012). Health of the nation outcome scales evaluation in a community setting population. *Qual Life Res*, *21*, 1643–1653. https://doi.org/10.1007/s11136-011-0071-9

McClelland, R., Trimble, P., Fox, M. L., Stevenson, M. R., & Bell, B. (2000). Validation of an outcome scale for use in adult psychiatric practice. *Qual Health Care*, *9*(2), 98–105. https://doi.org/10.1136/qhc.9.2.98

McGraw, K., & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychol Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychol Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Newnham, E. A., Harwood, K. E., & Page, A. C. (2009). The subscale structure and clinical utility of the health of the nation outcome scale. *J Ment Health*, *18*, 326–334. https://doi.org/10.1080/09638230802522486

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill. ISBN: 978–0070478497.

O'Neil, T. (2017). An overview of interrater agreement on Likert scales for researchers and practitioners. *Front Psychol*, *8*, 777. https://doi.org/10.3389/fpsyg.2017.00777

Pirkis, J. E., Burgess, P. M., Kirk, P. K., Dodson, S., Coombs, T. J., & Williamson, M. K. (2005). A review of the psychometric properties of the health of the nation outcome scales (HoNOS) family of measures. *Health Qual Life Outcomes*, *3*, 76. https://doi.org/10.1186/1477-7525-3-76

Preston, N. J. (2000). The health of the nation outcome scales: Validating factorial structure and invariance across two health services. *Aust N Z J Psychiatry*, *34*, 512–519. https://doi.org/10.1046/j.1440-1614.2000.00726.x

Royal College of Psychiatrists. (2019). *Exploring Mental Health Inpatient Capacity*. https://www.strategyunitwm.nhs.uk/sites/default/files/2019-11/Exploring%20Mental%20Health%20Inpatient%20Capacity%20accross%20Sustainability%20and%20Transformation%20Partnerships%20in%20England%20-%20191030_1.pdf

Speak, B., & Muncer, S. J. (2015). The structure and reliability of the health of the nation outcome scales. *Australas Psychiatry*, *23*, 66–69. https://doi.org/10.1177/1039856214563851

Spengler, P., White, M., Ægisdóttir, S., Maugherman, A., Anderson, L., Cook, R., Nichols, C., Lampropoulos, G., Walker, B., Cohen, G., & Rush, J. (2009). The meta-analysis of clinical judgement project. *Couns Psychol*, *37*(3), 350–399. https://doi.org/10.1177/0011000006295149

Suresh, K. (2011). An overview of randomization techniques: An unbiased assessment of outcome in clinical research. *Journal of Human*

Reproductive Sciences, *4*(1), 8–11. https://doi.org/10.4103/0974-1208.82352

The jamovi project. (2023). jamovi (*Version 2.3.21*). [Computer Software]. https://www.jamovi.org

Trauer, T. (1999). The structure of the health of the nation outcome scales (HoNOS). *Journal of Mental Health*, *8*(5), 499–509. https://doi.org/10.1080/09638239917193

Trauer, T., & Buckingham, B. (2006). The health of the nation outcomes scales (HoNOS), general adult version: Towards an agenda for future development. *Mental health Information Development* http://www.amhocn.org/publications/health-nation-outcomes-scales-honos-general-adult-version-towards-agenda-future

Wing, J., Beevor, A., Curtis, R., Park, S., Hadden, S., & Burns, A. (1998). Health of the nation outcomes scale (HoNOS): Research and Development. *Br J Psychiatry*, *172*(1), 11–18. https://doi.org/10.1192/bjp.172.1.11

Wing, J., Curtis, R., & Beevor, A. (1996). *Health of the nation outcome scales: Glossary for HoNOS: 1–14*. Royal College of Psychiatrists. https://www.rcpsych.ac.uk/docs/default-source/events/2021/mhct-iapt-23-march-2021/honos-glossary.pdf?sfvrsn=d4b6638b_2

How to cite this article: Painter, J., & James, M. (2024). Reliability testing of the Health of the Nation Outcome Scales 2018. *Journal of Psychiatric and Mental Health Nursing*, *00*, 1–6. https://doi.org/10.1111/jpm.13047