JIDR Journal of Intellectual Disability Research
Published on behalf of mencap and in association with IASSID

An investigation into the factor structure of the Health of the Nation Outcome Scales for People with Learning Disabilities

R. Hunt,[1] D. Dagnan,[2] S. Muncer[3] & L. Copping[4]

1 *Department of Clinical Psychology, Tees, Esk & Wear Valley NHS Foundation Trust, Darlington, UK*
2 *Community Learning Disability Services, Cumbria, Northumberland Tyne & Wear NHS Foundation Trust and University of Cumbria, Workington, UK*
3 *Department of Psychology, Durham University, Durham, UK*
4 *Department of Psychology, Teesside University, Middlesbrough, UK*

Abstract

Background The Health of the Nation Outcome Scales for People with Learning Disabilities (HoNOS-LD) is one of the most used outcome measures in learning disability services in the United Kingdom. There is relatively little known of the psychometric properties of the scales.
Method A data set of HoNOS-LD scales from 571 people with learning disabilities was randomly split into two halves. Exploratory Mokken analysis was applied to the first dataset, and confirmatory scale factor analysis was applied to the second dataset to test the fit of scale structures.
Results Two-factor and three-factor solutions were explored in the Mokken analysis, with the three-factor option having somewhat better characteristics. One-factor, three-factor and seven-factor solutions were explored using confirmatory factor analysis; a three-factor solution with items 8, 16, 17 and 18 used separately offers the best characteristics.
Conclusions The HoNOS-LD is best conceptualised as consisting of three scales, accounting for 14 items that can be labelled as 'Cognitive and Physical Functioning', 'Behaviour and Mood Disturbances' and 'Functional Difficulties'.

Keywords Factor structure, HONOS-LD, Intellectual disability, Mokken analysis, Psychometric analysis

Introduction

Perhaps the most widely used, clinician-completed outcome measures in mental health-related services in the United Kingdom are the Health of the Nation Outcome Scales (HoNOS – Wing *et al.* 1998; Delaffon *et al.* 2012) family of measures. These tools originated as a means of supporting clinicians to record the presentation of service users whilst maintaining the brevity required for use in daily clinical practice. The popularity of the HoNOS has prompted further development of the measure for different client groups, including children and adolescents (HoNOS-CA; Gowers *et al.* 1999), older adults (HoNOS 65+; Burns *et al.* 1999), people with acquired brain injuries (HoNOS ABI; Fleminger and

Correspondence: Prof Dave Dagnan, Community Learning Disability Services, Cumbria, Northumberland Tyne & Wear NHS Foundation Trust and University of Cumbria, Unit 9, Lillyhall Business Centre, Jubilee Road, Lillyhall, Workington, CA14 4HA Workington, UK (e-mail: dave.dagnan@cntw.nhs.uk).

Powell 1999) and people with mental health problems who are in forensic settings (HoNOS-secure; Dickens *et al.* 2007). The focus for this paper is a version of the HoNOS that has been adapted for people with intellectual disabilities (HoNOS-LD; Roy *et al.* 2002).

The Health of the Nation Outcome Scales for People with Learning Disabilities (HoNOS-LD; Roy *et al.* 2002) is an 18-item clinician-rated measure for use with people with intellectual disability with an emphasis on additional mental health and behavioural challenges. Published investigations into the psychometric properties of the measure have demonstrated preliminary support for some elements of validity and reliability (Roy *et al.* 2002; Tenneij *et al.* 2009; Esteba-Castillo *et al.* 2018).

Extensive research into the latent variable structure of the original HoNOS (e.g. Williams *et al.* 2014) has highlighted complex and inconsistent latent variable structures in different clinical populations, which complicates the interpretation of the measure in healthcare practice. Items within outcome measures are commonly collated into composite scores and/or total scores and used for service evaluation and individual outcome monitoring; to do this coherently requires an understanding of the latent variable structure of the measure. A small number of studies have begun to address this issue for the HoNOS-LD. Esteba-Castillo *et al.* (2018) analysed data for 111 people with intellectual disabilities using a Spanish translation of the HoNOS-LD and reported a single latent variable containing all 18 items derived from a combined parallel analysis and theoretical inclusion criteria; however, several items had a poor fit to the unidimensional model. Skelly and D'Antonio (2008) analysed 155 clinical records using principal components analysis to identify four scales that include all but four items. The four components included items that loaded at or above an absolute value of 0.50 and the authors labelled the factors as 'cognitive and communicative competence', 'functional behaviour and attachment disturbance', 'loss of functioning and community presence' and 'internal dysregulation'. Turton (2020) analysed 2109 clinical records, used 'principal factor analysis' with an oblimin rotation and tested three-factor, four-factor, and five-factor solutions. A four-factor solution was preferred although only 12 of 22 items were identified as having 'strong' loadings of greater than 0.5. The four-factor solution included three 'strong' factors: 'cognitive skills', 'social competence' and 'mood' with a fourth behavioural factor with only two items that loaded strongly. A further latent variable structure has been suggested in a recent protocol for a systematic review (Harris *et al.* 2018), which suggests a seven-factor model, albeit with no reported statistical or theoretical rationale for this.

Traditional investigations into latent variable structures of psychometric tools employ methodologies from classical test theory (CTT), which, researchers argue, can fail to adequately account for data from populations where scores may fall in the extremes of a normal distribution (Van der Linden and Hambleton 1997). An alternative is to use methodologies drawn from item response theory (IRT), such as Mokken scale analysis (Mokken 1971). This approach provides an IRT-based methodology for assessing whether ordinal items measure the same underlying construct. It uses probabilistic modelling of the unidimensionality of a measure with non-parametric assumptions, providing a more appropriate analytic methodology for psychometrics than similar analyses such as Rasch and Guttman scales (Van 2003). Mokken scale analysis can be used in both exploratory and confirmatory approaches (Snijders, 2008) and has previously been utilised to investigate the latent variable structure of the original HoNOS (e.g. Muncer *et al.* 2016; Muncer and Speak 2016).

The current paper explores the structure of the HoNOS-LD to inform its use within clinical practice. We use Mokken analysis to understand the factor structure of the HoNOS-LD and subsequent confirmatory scale analysis to investigate the goodness of fit of the statistically derived model and of previously suggested structures for the HoNOS-LD.

Method

Ethics

All procedures contributing to this work comply with the ethical standards of the Helsinki Declaration of 1975, as revised in 2008. The study was approved by the Health Research Authority (REC 19/EE/0148) and the Teesside University School of Social Sciences, Humanities and Law ethics sub-committee. All data were collected for the purpose of evaluation and anonymised before this study was conceived and before being transferred between organisations.

HoNOS-LD

The HoNOS-LD (Roy *et al.* 2002) is an 18-item clinician-rated measure for use with people with any degree of intellectual disability with an emphasis on additional mental health and behavioural challenges. The measure is reported to have good inter-rater reliability and to be sensitive to change (Roy *et al.* 2002; Tenneij *et al.* 2009; Esteba-Castillo *et al.* 2018).

Each item is rated on a 5-point scale with anchoring statements for each item. Item 3 of the HoNOS-LD is 'Other mental and behavioural problems', which can include a rating of five different behaviours 'A, behaviour destructive to property; B, problems with personal behaviours e.g. spitting, eating rubbish, self-induced vomiting, continuous eating or drinking, hoarding rubbish, inappropriate sexual behaviour; C, rocking, stereotyped and ritualistic behaviour; D, Anxiety, phobias, obsessive, compulsive behaviour & E, others', with an instruction to 'rate the most prominent behaviours present'. This item has been treated differently in previous papers. Authors have treated item 3 as either a single rating of the most prominent behaviour (e.g. Tenneij *et al.* 2009; Esteba-Castillo *et al.* 2018) or as five separate ratings, resulting in the measure being reported as having 22 items (e.g. Skelly and D'Antonio 2008; Turton, 2020). Harris *et al.* (2018) described a systematic review of the HoNOS family of measure and proposes a seven-factor structure that incorporates a single rating for item 3. In the data available for this study clinicians identified the most prominent behaviour, and it was this rating that was used in the analysis; thus, this paper also uses a single item rating for item 3.

Participants

The data for this study were a complete sample of individuals seen by a community learning disabilities team over a 2-year period (covering January 2010 to December 2011 inclusive). Ratings were completed by 87 qualified members of the community teams consisting of 85.0% community nurses, 6.5% clinical psychologists, 4.5% allied health professionals (occupational therapists, physiotherapists and speech and language therapists) and 4.0% psychiatrists. The HoNOS-LD was introduced into the service with 1 day's training for all staff and regular updates to its use during regular team days.

All people with intellectual disabilities were over the age of 18 years old and had a diagnosed learning disability based on criteria from the International Classification of Diseases (ICD-10; World Health Organisation 1992); level of disability was coded by the clinician based upon clinical judgement. HoNOS-LD measures were scored by clinicians within the service for each service user as part of routine clinical assessment at initial assessment, follow-up appointments and at discharge. The dataset was initially collected for service evaluation purposes and subsequently anonymised. The total dataset consists of 1703 HoNOS-LD ratings of 650 service users.

The dataset was cleaned and all measures where scoring was incomplete or contained items coded as 'missing' or 'unknown' were removed from analysis (55 measures); 17 participants had only a single discharge assessment available; these were typically measures collected early in the project and did not have full demographic data available; these measures were also excluded to ensure a more complete dataset and a more similar assessment context. The chronologically first measure meeting the above criteria for each participant was selected for analysis (43% initial and 57% review assessment). This resulted in a sample of 571 participants (42% male, 57% female). There were no significant differences between the 571 participants included in the final analysis and the 71 who were not by age, sex, level of disability or place of accommodation.

Full demographic information for this final sample can be seen in Table 1.

For the purposes of exploratory and subsequent scale analysis, this database was randomly split into two halves, creating an exploratory first database (n = 285) and a second confirmatory database (n = 286); these groups were of a size sufficient to evaluate the scalability of scales (Watson *et al.* 2018), but item-level scalability should be interpreted with caution. All data analysis was conducted using the statistics package *R* (available at https://cran.r-project.org/index.html).

Table 1 Demographic information for full sample (N = 571)

Factor	N	%
Biological sex		
Male	241	42.21
Female	329	57.62
Unknown	1	0.18
Age		
18–25	115	20.14
26–35	93	16.29
36–45	123	21.54
46–55	128	22.42
56–65	67	11.73
66+	41	7.18
Unknown	4	0.70
Assessment type		
Initial	245	42.91
Review	326	57.09
Level of disability		
Mild	173	30.4
Moderate	243	42.6
Severe	126	22.1
Profound	26	4.6
Unknown	3	0.4
Accommodation		
Lives independently	63	11.03
Family home	147	25.74
Acute hospital	10	1.75
Long-stay hospital	9	1.58
Group home (staffed)	262	45.88
Group home (unstaffed)	5	0.88
Other	67	11.73
Unknown	8	1.40

Exploratory analysis

Exploratory Mokken analysis was applied to the first dataset (n = 285) using the R 'Mokken' package (Van der Ark 2007). This package uses an automated item selection process (AISP) with a bottom-up approach to defining scales using the scalability coefficient Loevinger's H (Loevinger 1948). This represents the frequency to which items are similarly endorsed (with a score of one indicating the presence of a perfect Guttman scale and a score of zero indicating no scalability) and can be calculated for items within a scale, item pairs and the scales themselves. The AISP initially selects the pair of items with the highest item-pair scalability coefficient (H_{ij}) and compiles scales by ensuring that subsequent items have both positive correlations with the existing scale items and a scalability coefficient, with the items currently in the scale, of at least the lower bound criterion (Sijtsma and Molenaar 2002). For the purposes of this study, exploratory analysis was conducted with lower bounds between 0.00 and 0.30 with 0.01 increments in line with suggestions regarding setting appropriate lower bounds (Hemker *et al.* 1995) and previous research using HoNOS measures (Sijtsma and Molenaar 2002; Muncer and Speak 2016).

Confirmatory factor analysis

To clarify the goodness of fit of the derived model, confirmatory scale factor analysis was applied to the second dataset utilising the R 'lavaan' package (Rosseel 2012). The ordinal nature of the data warranted the use of the diagonally weighted least squares (DWLS) estimator, with robust standard errors and scaled test statistics as described by Li (2016). To ensure a robust confirmatory scale analysis, five different goodness-of-fit measurements were used. Chi-square (χ^2) tests the difference between the theorised model and the current sample, with a large χ^2 indicating greater differences between the models. However, χ^2 is sensitive to sample size and can provide misleading results for samples where $N > 200$ (Schumacker and Lomax 2016). Both the comparative fit index (CFI; Bentler 1990) and the Tucker–Lewis index (TLI; Tucker and Lewis 1973) compare the performance of the proposed fit model against a null model in which there are no correlations between items (Hooper *et al.* 2008), outputting values between 0 (no fit of the proposed model) and 1 (proposed model fits the data perfectly). The root-mean-square error of approximation (RMSEA) provides a measure of fit by considering the degrees of freedom (representing model complexity) and sample size. The standardised root-mean-square residual index (SRMR) measures goodness of fit by using the square root of the mean-squared differences between the model-implied and observed matrix elements. Commonly accepted criteria for goodness of fit suggest that for a model to be considered an acceptable fit, it should have a CFI and TLI of greater than 0.90 (ideally >0.95), and both an RMSEA and SRMR below 0.08 (Cangur and Ercan 2015).

The reliability of scales is reported throughout using Cronbach's alpha and Guttman's Lambda.2 (Bendermacher 2010).

Results

Exploratory analysis

At more liberal criteria, the AISP identifies two-scale models: the first containing items *4, 5, 6, 7, 12, 13, 14* and *15*; the second comprising the remaining items of *1, 2, 3, 8, 9, 10, 11, 16, 17* and *18*. As the criteria become stricter, items *16* and *8* are dropped from the second scale, whilst the first remains intact. When the lower bound criteria for clustering reached 0.22, the

Table 2 Mean scores and item scalability coefficients with standard errors for emergent scales from Mokken analysis

		Item scalability coefficients and standard errors (SEs in brackets) on emergent scales by lower bound criteria							
		Lower bound = 0.00		**Lower bound = 0.22**			**Lower bound = 0.30**		
HoNOS-LD item	**Mean (standard deviation)**	**S1**	**S2**	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
1 Behaviour towards others	0.49 (0.90)		0.23 (0.035)		0.45			0.39 (0.047)	
1 Self-injurious behaviour	0.96 (1.09)		0.29 (0.031)		0.53			0.46 (0.038)	
1 Other behaviour problems	1.34 (1.18)		0.33 (0.034)		0.53			0.51 (0.039)	
1 Attention and concentration	1.49 (1.16)	0.59 (0.034)		0.59 (0.034)			0.59 (0.034)		
1 Memory and orientation	0.74 (1.14)	0.56 (0.038)		0.56 (0.038)			0.56 (0.038)		
1 Communication (understanding)	0.76 (1.03)	0.60 (0.033)		0.60 (0.033)			0.60 (0.033)		
1 Communication (expression)	0.95 (1.03)	0.62 (0.031)		0.62 (0.031)			0.62 (0.031)		
1 Hallucinations and delusions	0.16 (0.54)		0.18 (0.032)			0.23	ns	ns	ns
1 Mood disturbance	0.94 (1.04)		0.34 (0.032)		0.49			0.41 (0.046)	
1 Problems with sleeping	0.61 (0.91)		0.25 (0.032)		0.34				0.32 (0.061)
1 Problems with eating and drinking	0.28 (0.69)		0.24 (0.048)		0.32				0.34 (0.071)
1 Physical problems	0.50 (0.97)	0.31 (0.060)		0.31 (0.060)			0.31 (0.060)		
1 Seizures	0.23 (0.67)	0.29 (0.078)		0.29 (0.078)			0.29 (0.078)		
1 Activities of daily living at home	1.29 (1.07)	0.54 (0.045)		0.54 (0.045)			0.54 (0.045)		
1 Activities of daily living outside home	1.20 (1.15)	0.48 (0.044)		0.48 (0.044)			0.48 (0.044)		
1 Self-care	0.50 (0.78)		0.13 (0.038)	ns	ns	ns	ns	ns	Ns
1 Relationship problems	0.85 (0.91)		0.22 (0.039)		0.38		ns	ns	Ns
1 Occupation and activities	1.21 (1.07)		0.22 (0.036)			0.28			0.31 (0.062)

Ns, item unselected for clust.

solution shifted to a three-scale model, which remains consistent until the lower bound criterion of 0.30, only dropping item *17* from the second scale as it reaches this criterion. The Loevinger's *H* coefficients for each item within the emergent structures at lower bounds on 0.00, 0.22 and 0.39, and the item mean scores are outlined in Table 2.

Mokken (1971) suggested that scales with an *H* coefficient below 0.30 should be considered as demonstrating poor or no scalability; 0.30–0.40 denoting a useful but weak scalability; 0.40–0.50 as indicating medium scalability; and an *H* of greater than 0.50 as indicating good scalability. Throughout items 2, 3, 4, 5, 6, 7, 12, 13 and 14 demonstrate strong *H* values, with items 1 and 9 consistently demonstrating moderate scalability. Whilst there is some fluctuation in items within clusters, several items are consistent; items 1, 2, 3 and 9 form a consistent cluster across all possible solutions, as do items 4, 5, 6, 7, 12, 13, 14 and 15.

Further Mokken analysis replicated structures across both the initial and second datasets. Analysis of scale coefficients for the two-scale model can be seen in Table 3. The two-scale model demonstrates acceptable consistency across both datasets. However, whilst it depicts reliability over the accepted cut-off of $\alpha = 0.70$ for both scales (Cronbach 1951), the scalability co-efficient for the second scale is particularly low, falling within the range identified by Mokken (1971) as 'no scalability' ($H < 0.3$). Across both halves of the dataset, the two-scale structure demonstrated no significant violations in monotonicity for either scale. Within the second dataset, Scale 1 demonstrated violations to invariant item ordering in items 5, 6 and 12. Whilst Scale 2 demonstrated no significant violations to monotonicity across both datasets, there were numerous violations to invariant item ordering although, given that the items on the HoNOS-LD appear to measure quite different domains theoretically it would not be expected that the order of items would be invariant.

The scalability and reliability of the three-scale model are summarised in Table 4. The three-scale model shows greater discrepancy across datasets than the two-scale model. Scales 1 and 2 both report meaningful scalability ranging from medium to high scalability based on the criteria by Mokken (1971) and an acceptable reliability in both datasets. Scale 3 shows weak scalability in the initial dataset and no evidence of scalability in the second. In both datasets, Scale 3 fails to meet the criteria given by Cronbach (1951) for acceptable internal consistency, although this may be a result of the small number of items within this scale (e.g. Eisinga *et al.* 2013). Given the particularly poor performance of the second scale in the two-scale model, further analysis will focus on the applicability and utility of the three-scale model alone.

Considering the content of the items within each cluster of the three-item scale, Scale 1 may be best conceptualised as a 'Cognitive and Physical Functioning' cluster (containing items investigating *Attention and Concentration*, *Memory and Orientation*, *Receptive Communication*, *Expressive Communication*, *Physical Health*, *Seizures*, *Activities of Daily Living at Home* and *Activities of Daily Living outside the home*). Scale 2 may be best described as 'Behaviour and

Table 3 Scalability and reliability for extracted two-scale model in each dataset

	First dataset (n = 285)		Second dataset (n = 286)	
Statistic	**S1**	**S2**	**S1**	**S2**
Scalability (H)	0.52	0.25	0.43	0.25
Standard error (H)	0.033	0.023	0.034	0.027
Reliability (α)	0.84	0.72	0.82	0.71
Reliability (λ)	0.87	0.74	0.84	0.71

Table 4 Scalability and reliability for extracted three-scale model in each dataset

	First dataset (n = 285)			**Second dataset (n = 286)**		
Statistic	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
Scalability H	0.52	0.45	0.32	0.43	0.46	0.21
Standard error H	0.033	0.035	0.057	0.034	0.039	0.052
Reliability (α)	0.86	0.73	0.50	0.82	0.73	0.39
Reliability (λ)	0.87	0.73	0.51	0.84	0.74	0.40

Mood Disturbances' and contains measures of *Behaviour towards others*, *Self-destructive behaviour*, *Other behavioural problems* and *Mood*. Scale 3 contains items measuring *Problems with Eating and Drinking*, *Problems with Sleep* and *Occupation and Activities* and as such may be best described as 'Functional Difficulties'.

Confirmatory factor analysis

Confirmatory factor analysis of the three-factor model indicated a significant difference between the observed data and the proposed model, χ^2(87, 286) = 313.26, $p < 0.01$, although this is not unexpected given the large sample used for this analysis. The model demonstrated an acceptable goodness of fit based on the CFI and TLI criteria; however, it failed to meet the criteria of the RMSEA and SRMR. On further examination, the model demonstrated significant co-variances between Scales 1 and 3 (covariance estimate 0.456, SE = 0.095, $P < 0.001$), and Scales 2 and 3 (covariance estimate 0.545, SE = 0.107, $P < 0.001$), and a small non-significant co-variance between Scales 1 and 2 (covariance estimate 0.052, SE = 0.073, $P = 0.47$). Theoretically, a *prima facie* case can be made for the correlations between Scales 1 and 3 and Scales 2 and 3 (considering the natures of these scales); however, given the more tenuous theoretical links between Scales 1 and 2 alongside the non-significant covariance, a decision was made to remove this covariance. Investigation of modification indices between scales and items also revealed factorial complexities around Item 18 (*Occupation and Activities*), which was loading poorly onto multiple scales. A clear similarity exists between the other two items in Scale 3 (*Difficulties eating and drinking*, and *Difficulties sleeping*), and item 18 is an item concerned with general activity and engagement, which is likely to be associated with challenging behaviour, mental ill-health and physical difficulties and severity of intellectual disability, and it is not surprising that multiple items would interact with it. Thus, a decision was made to remove *item 18* from the model. The results of the initial confirmatory factor analysis can be seen in Table 5.

When the covariance between Scale 1 and Scale 2 is controlled and item 18 is removed, the scale shows an improvement in goodness of fit and more theoretical connection between items, providing a more appropriate scale for clinical usage. This revised solution provided a significantly better fit to the data than the originally extracted three-scale model indicating that the removal of item 18 provides a significantly more acceptable model, χ^2_{diff} (32, 286) = 82.01, $P < 0.01$. Thus, confirmatory analysis of 286 HoNOS-LD ratings has yielded a three-scale model containing 14 items, which best describes the data. The final model is shown in Figure 1.

To further confirm the goodness of fit, the three-scale model (excluding item 18) has been applied to the first dataset and the combined dataset using confirmatory factor analysis. Goodness-of-fit statistics are reported in Table 5. Comparison of goodness of fit of the three-scale model between samples demonstrate a similar goodness of fit for all samples.

Table 5 Robust confirmatory scale analysis test statistics

Three-scale model and subsequent amendments on second dataset	χ^2	df	CFI	TLI	RMSEA (95% CI)	SRMR
Three-scale model	313.26*	87	0.92	0.91	0.10 (0.08–0.11)	0.10
Three-scale model (controlling for covariance and *item 18* removed)	231.25*	75	0.94	0.93	0.09 (0.08–0.10)	0.11
Final three-scale model on initial, second and combined dataset						
First dataset (*N* = 285)	249.56*	75	0.97	0.96	0.09 (0.08–0.10)	0.12
Second dataset (*N* = 286)	231.25*	75	0.94	0.93	0.09 (0.08–0.10)	0.11
Combined dataset (N = 571)	372.06*	75	0.96	0.95	0.08 (0.08–0.09)	0.10
Model for all published structures for the HoNOS-LD						
Unidimensional model	1978.96*	135	0.77	0.74	0.16 (0.15–0.16)	0.16
Seven-scale model (Harris *et al.* 2018)	508.46*	114	0.95	0.93	0.08 (0.07–0.09)	0.08
All published structures for the HoNOS-LD, minus items excluded in exploratory analysis						
Three-factor model	372.06*	75	0.96	0.95	0.08 (0.08–0.09)	0.10
Unidimensional model	1632.95*	77	0.80	0.76	0.19 (0.18–0.20)	0.17
Seven-scale model	298.74*	70	0.97	0.95	0.08 (0.07–0.09)	0.08

*Statistically significant to $p < 0.01$.

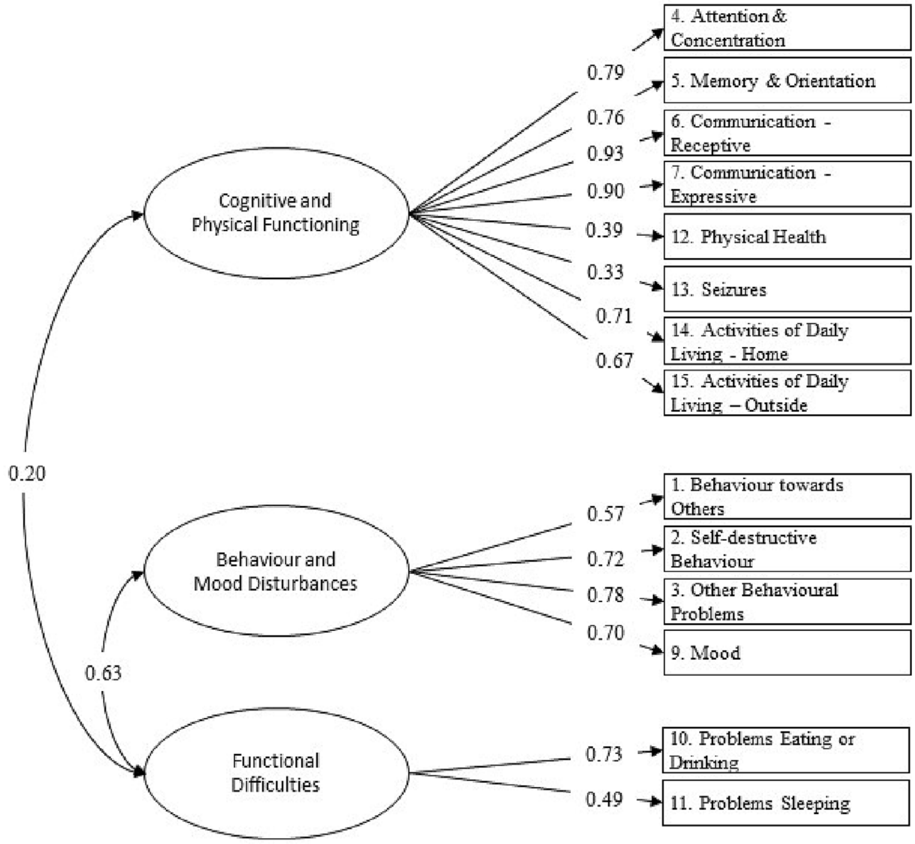
4. Attention & Concentration
5. Memory & Orientation
6. Communication - Receptive
7. Communication - Expressive
12. Physical Health
13. Seizures
14. Activities of Daily Living - Home
15. Activities of Daily Living – Outside
0.79
0.76
0.93
0.90
0.39
0.33
0.71
0.67
Cognitive and Physical Functioning
0.20
1. Behaviour towards Others
2. Self-destructive Behaviour
3. Other Behavioural Problems
9. Mood
0.57
0.72
0.78
0.70
Behaviour and Mood Disturbances
0.63
Functional Difficulties
0.73
0.49
10. Problems Eating or Drinking
11. Problems Sleeping

Figure 1. Scale structure diagram of the three-scale model of the HoNOS-LD with scale loadings and covariances.

Comparative analysis

Two additional confirmatory analyses were conducted to compare the final three factor model to a unidimensional model using a single total score (e.g. Esteba-Castillo *et al.*, 2018) and the 7-scale model suggested by Harris *et al.* (2018). Table 5 depicts the results of these analyses on the full dataset. The unidimensional model demonstrates a poor fit to the data on all goodness-of-fit criterion. Whilst the seven-scale model demonstrated acceptable goodness of fit on all indicators, further analysis resulted in negative variance, indicating that regardless of the goodness of fit, the model does not demonstrate real-world application. Both the unidimensional and seven-factor analyses were repeated, excluding the items that had shown non-significant scaling in the exploratory analysis; the results are presented in Table 5. The goodness of fit of the unidimensional model following the removal of items 8, 16, 17 and 18 shows an improved fit over the 18-item model; however, fit is significantly weaker than for the final three-factor model, χ^2_{diff} (2, 571) = 1260.89, $P < 0.01$. The Harris *et al.* (2018) model showed an improved goodness-of-fit with items 8, 16, 17 and 18 removed; however, the analysis again returned negative variance, implying that the model does not provide a meaningful summation of the data.

Conclusion

Exploratory and confirmatory Mokken scale analysis combined with confirmatory factor analysis suggests that the HoNOS-LD is best conceptualised as consisting of three scales, accounting for 14 items. The analysis indicates two robust scales with acceptable internal consistency, measuring *Cognitive and Physical Functioning* (items 4, 5, 6, 7, 12, 13, 14 and 15) and *Behaviour and Mood Disturbance* (items 1, 2, 3 and 9), and a third, less reliable, scale, which is labelled *Functional Difficulties* (items 10 and 11). Items assessing *Hallucinations and Delusions, Impairments in Self-Care* and *Relationship Problems* (items 8, 16 and 17) were excluded from analysis, as these items demonstrate a poor relationship with the underlying scales identified within the HoNOS-LD and may be best considered as independent items. Item 18 (*Occupation and Activities*) shows a complex relationship with the latent variables underlying the HoNOS-LD as it appears to load onto multiple variables and as such cannot be considered as independent from other items on the scale.

The analyses presented in this paper are comprehensive and use exploratory IRT approaches and both IRT and CTT confirmatory approaches. There are potential limitations to these analyses. The decision to split the sample into two for the purpose of exploratory and confirmatory analysis results in a degree of caution in the interpretation of item-level scalability data. The data were collected routinely, which results in less detailed control of the assessment process; however, all raters were trained in the use of the HoNOS-LD and were qualified clinicians with experience of using routine measures, and the number of raters contributing data is high, and therefore, data are unlikely to be affected by a particular rater's bias. The data were initially used in a local evaluation and subsequently offered for research use, a process that took several years and was subsequently delayed by COVID-19 factors; however, there is no reason to suggest that the structure of the data would be affected by having been collected 10 years ago.

The HoNOS-LD is one of several scales recommended for use in English intellectual disability services (e.g. NHS England 2017) and has been used in several clinical studies, as an outcome measure for hospital wards (Hillier *et al.* 2010), for therapy evaluations (Skelly *et al.* 2018) and as a means of managing caseload complexity (Clifford and Kemp 2020). Published studies have either used the total score for the HoNOS-LD (e.g. Skelly *et al.* 2018; Clifford and Kemp 2020) or studies have used clusters of scores based on clinical opinion or face validity as suggested by Harris *et al.* 2018 (e.g. Hillier *et al.* 2010). The current paper suggests that a 'global problem severity' score of all 18 items does not provide a good fit to the data and should be used cautiously, although omitting items 8, 16, 17 and 18 provides a somewhat more robust scale. Similarly, more complex structures such as suggested by Harris *et al.* (2018) are not a good representation of the structure of the HoNOS-LD. From the current analysis, the three-factor model reported here would be justified, with items 8, 16, 17 and 18 being considered as individual items. Further study of the performance of the HoNOS-LD using the structure suggested here is recommended, and it would be important to demonstrate that the scales are sensitive to change and to understand patterns of change in clinical work across areas such as mental health, challenging behaviour and physical health.

Acknowledgements

None required.

Source of funding

There is no funding associated with this paper.

Conflict of interest

There are no conflicts of interest to declare.

Ethics statement

The authors assert that all procedures contributing to this work comply with the ethical standards of the Helsinki Declaration of 1975, as revised in 2008. The study was approved by the Health Research Authority (REC 19/EE/0148) and the Teesside University School of Social Sciences Humanities and law ethics sub-committee. All data were anonymised before this study was conceived and before being transferred between organisations.

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.

References

Bendermacher N. (2010) Beyond alpha: lower bounds for the reliability of tests. *Journal of Modern Applied Statistical Methods* **9**, 11–02.

Bentler P. M. (1990) Comparative fit indexes in structural models. *Psychological Bulletin* **107**, 238–46.

Burns A., Beevor A., Lelliott P., Wing J., Blakey A., Orrell M. *et al.* (1999) Health of the nation outcome scales for elderly people (HoNOS 65+). *The British Journal of Psychiatry* **174**, 424–7.

Cangur S. & Ercan I. (2015) Comparison of model fit indices used in structural equation modeling under multivariate normality. *Journal of Modern Applied Statistical Methods* **14**, 14–67.

Clifford A. & Kemp F. G. (2020) A pragmatic mixed-methods review of changing "case-complexity" of referrals to an intensive support service. *Advances in Mental Health and Intellectual Disabilities* **14**, 111–24.

Cronbach L. J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* **16**, 297–334.

Delaffon V., Anwar Z., Noushad F., Ahmed A. & Brugha T. (2012) Use of health of the nation outcome scales in psychiatry. *Advances in Psychiatric Treatment* **18**, 173–9.

Dickens G., Sugarman P. & Walker L. (2007) HoNOS-secure: a reliable outcome measure for users of secure and forensic mental health services. *The Journal of Forensic Psychiatry and Psychology* **18**, 507–14.

Eisinga R., Grotenhuis M. T. & Pelzer B. (2013) The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health* **58**, 637–42.

Esteba-Castillo S., Torrents-Rodas D., García-Alba J., Ribas-Vidal N. & Novell-Alsina R. (2018) Translation and validation of the Spanish version of the health of the nation outcome scales for people with learning disabilities (HoNOS-LD). *Revista de Psiquiatría y Salud Mental* **11**, 141–50.

Fleminger S. & Powell J. (1999) Evaluation of outcomes in brain injury rehabilitation. *Neuropsychological Rehabilitation* **9**, 225–30.

Gowers S. G., Harrington R. C., Whitton A., Lelliot P., Beevor A., Wing J. *et al.* (1999) Health of the nation outcome scales for children and adolescents (HoNOSCA): glossary for HoNOSCA score sheet. *The British Journal of Psychiatry* **174**, 428–31.

Harris M. G., Sparti C., Scheurer R., Coombs T., Pirkis J., Ruud T. *et al.* (2018) Measurement properties of the health of the nation outcome scales (HoNOS) family of measures: protocol for a systematic review. *BMJ Open* **8**, e021177.

Hemker B. T., Sijtsma K. & Molenaar I. W. (1995) Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement* **19**, 337–52.

Hillier B., Wright L., Strydom A. & Hassiotis A. (2010) Use of the HoNOS–LD in identifying domains of change. *The Psychiatrist* **34**, 322–6.

Hooper D., Coughlan J. & Mullen M. R. (2008) Model fit. *Electronic Journal of Business Research Methods* **6**, 53–60.

Li C. H. (2016) Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods* **48**, 936–49.

Loevinger J. (1948) The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin* **45**, 507–29.

Mokken R. J. (1971) *A Theory and Procedure of Scale Analysis*. De Gruyter, Berlin, Germany.

Muncer S., Bass M. & Dawkin M. (2016) Mokken analysis of the health of the nation outcome scales in acute inpatient and community samples. *Australasian Psychiatry* **24**, 459–61.

Muncer S. J. & Speak B. (2016) Mokken scale analysis and confirmatory factor analysis of the health of the nation outcome scales. *Personality and Individual Differences* **94**, 272–6.

NHS England (2017). Transforming care model service specifications: supporting implementation of the service models. Available at: https://www.england.nhs.uk/publication/transforming-care-service-model-specification-january-2017/ (retrieved 26 August 2022).

Rosseel Y. (2012) lavaan: an R package for structural equation modelling. *Journal of Statistical Software* **48**, 1–36.

Roy A., Matthews H., Clifford P., Fowler V. & Martin D. M. (2002) Health of the nation outcome scales for people with learning disabilities (HoNOS–LD). *The British Journal of Psychiatry* **180**, 61–6.

Schumacker E. & Lomax G. (2016) *A Beginner's Guide to Structural Equation Modelling*, 4th edn. Routledge.

Sijtsma K. & Molenaar I. W. (2002) *Introduction to nonparametric item response theory*, vol. **5**. Sage.

Skelly A. & D'Antonio M. L. (2008) Factor structure of the HoNOS-LD: Further evidence of its validity and use as a generic outcome measure. *The British Psychological Society* **6**, 3–7.

Skelly A., McGeehan C. & Usher R. (2018) An open trial of psychodynamic psychotherapy for people with mild-moderate intellectual disabilities with waiting list and follow up control. *Advances in Mental Health and Intellectual Disabilities* **12**, 153–62.

Snijders T. (2008) Intermediate Social Statistics Lecture 6: Scale Construction. Available at: http://www.stats.ox.ac.uk/~snijders/ (retrieved 26 August 2022).

Tenneij N., Didden R., Veltkamp E. & Koot H. M. (2009) Reliability and validity of the HoNOS-LD and HoNOS in a sample of individuals with mild to borderline intellectual disability and severe emotional and behavior disorders. *Journal of Mental Health Research in Intellectual Disabilities* **2**, 188–200.

Tucker L. R. & Lewis C. (1973) A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* **38**, 1–10.

Turton R. (2020) An exploratory factor analysis of HONOS-LD scales. *Advances in Mental Health and Intellectual Disabilities* **14**, 33–44.

Van der Ark L. A. (2007) Mokken scale analysis in R. *Journal of Statistical Software* **20**, 1–19.

Van der Linden W. J. & Hambleton R. K. (1997) *Handbook of Item Response Theory*. Taylor & Francis Group.

Van W. H. (2003) Mokken scale analysis: between the Guttman scale and parametric item response theory. *Political Analysis* **11**, 139–63.

Watson R., Egberink I. J., Kirke L., Tendeiro J. N. & Doyle F. (2018) What are the minimal sample size requirements for Mokken scaling? An empirical example with the Warwick-Edinburgh mental well-being scale. *Health Psychology and Behavioral Medicine* **6**, 203–13.

Williams B., Speak B., Hay P. & Muncer S. J. (2014) An evaluation of the independence of the health of the nation outcome scales. *Australasian Psychiatry* **22**, 473–5.

Wing J. K., Beevor A. S., Curtis R. H., Park S. G. B., Hadden J. & Burns A. (1998) Health of the nation outcome scales (HoNOS): research and development. *The British Journal of Psychiatry* **172**, 11–8.

World Health Organization (1992) *ICD-10: Alphabetical Index*, vol. **3**. World Health Organization.

Accepted 11 July 2023