A 13-item Health of the Nation Outcome Scale (HoNOS 13): Validation by Item Response Theory (IRT) in people with Substance Use Disorder

Anne Chatton
Geneva University Hospitals
Yasser Khazaal
University of Lausanne
Louise Penzenstadler (Louise.E.Penzenstadler@hcuge.ch)
Geneva University Hospitals

Research Article

Keywords: substance use disorders, symptom severity, HoNOS, Health of the Nation Outcome Scale, Item Response Theory

Posted Date: December 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1140799/v1

Abstract

Background

The Health of the Nation Outcome Scale (HoNOS) is a widely used 12-item tool to assess mental health and social functioning. The French version has an added 13th item measuring adherence to psychotropic medication. The aim of the current study is to uncover the unknown pattern of item 13 and to compare the unidimensional and multi-dimensional fit of both the original HoNOS 12 and the new HoNOS 13 using Item Response Theory (IRT) modelling. This research question was studied among inpatients with substance use disorder (SUD).

Methods

Six hundred and nine valid questionnaires of the HoNOS were analyzed using the multidimensional extension of the IRT graded-response modelling. For HoNOS 13, we fitted respectively a one-factor and a two-factor model.

Results

The two-factor model suggesting a first factor capturing psychiatric/impairment-related issues and a second factor reflecting social-related issues yielded better goodness-of-fit values compared to the one-factor solution.

Conclusions

We were able to validate the 13-item questionnaire including medication compliance and suggest that the HoNOS 13 can be recommended as a clinical evaluation tool to assess the problems and treatment needs for inpatients with SUD. In IRT analyses, the items related to substance use and item 13 showed moderate discriminative abilities to catch the severity of the latent construct whereas the items related to the second factor (social problems) showed higher discriminative abilities.

Trial registration

ClinicalTrials.gov, Identifier: NCT03551301, registered: 11.06.2018.

Background

The Health of the Nation Outcome Scale (HoNOS) was developed by Wing et al. (1) as a brief general assessment of mental health and social functioning designed to measure a large range of problems of psychiatric patients and their evolution.

This first version was validated by classical theory and gave rise to a 12-item scale evaluating 4 dimensions. Behavioral problems cover 3 items (1–3): overactive, aggressive, disruptive or agitated behavior, non-accidental self-injury and problem drinking or drug taking. Impairment covers 2 items (4–5): cognitive problems and physical illness or disability problems. Symptomatic problems include 3 items (6–8): problems with hallucinations and delusions, problems with depressed mood and other mental and behavioral problems. Social problems cover 4 items (9–12): problems with relationships, problems with activities of daily living, problems with living conditions and problems with occupation and activities. Each item is scored 0 (no problems during the reporting period) to 4 (severe to very severe problem), higher scores reflecting greater severity (2).

The reproducibility of the 4 dimensions listed above could not be demonstrated. Analyses involving individual HoNOS items have been undertaken in many studies (3–5). There is no universal agreement regarding the operationalization of HoNOS. Alternative models were suggested by Trauer (6) and Lauzon et al. (7). The study by Trauer (6) identified a five-scale solution using an exploratory factor analysis in a sample of 2137 patients from five different psychiatric services in Australia, with over 50% schizophrenia as a main diagnosis. The structure was replicated in a study by Eagar et al. (3) using confirmatory factor analysis. The five-scale solution of the HoNOS contains, a ‘Depression’ scale (items 2, 7–9), an ‘Impairment’ scale (items 4 and 5) and a ‘Hallucinations/delusions’ scale (item 6). Lauzon et al. (7) found that the scale was reliable only when the total score was used.

In the same vein, several other factor structures have been proposed but none of these have acceptable fit. Rasch analyses (8) demonstrate the absence of an underlying construct in the composite scale (9,10).

Moreover, perceptions of the value of the outcome measurement system seem to be mixed (11). For instance, Bebbington et al. (12) perceive HoNOS at best as a measure of social functioning and Bender (13) reported lack of studies related to the ability of HoNOS to serve for the improvement of mental health services. Despite these limitations, the HoNOS continues to be widely used to evaluate mental health patients in inpatient and ambulatory settings (14,15).

Medication non-adherence is known to be an important factor influencing clinical outcomes (16). In order to take this factor into account a 13th item concerning "Problems with psychotropic medication compliance" was added to the HoNOS (HoNOS-13) in its French version (17).

So far, the psychometric properties of HoNOS were measured for patients with general psychiatric disorders. Only few studies (18) have specifically measured these in patients with a main diagnosis of substance use disorders (SUD). In spite of several controversies related to HoNOS factorial structure, it was suggested that the items could help to identify sub-specific groups of patients with particular needs (19).

A well-established statistical model used to represent both item and test taker characteristics is Item Response Theory (IRT) modelling. IRT can help to understand the impact of each item on the latent construct. Such a statistical approach clearly contrasts with methods used in previous HoNOS-related studies, based on reliability scores (focus on the way each item relates to the total score) such as Exploratory Factor Analysis (EFA) used in Classical Test Theory context (CTT) or Confirmatory Factor Analysis (CFA).

An important feature of IRT compared to CTT is that item properties do not depend on a representative sample (20). IRT is based on the idea that the probability of a correct response to an item is a mathematical function of person and item parameters. Although its origins date back to the mid-20th century, its application did not become widely implemented until the late 1970-80s following the work of pioneers like Rasch (21), Samejima (22) and Bock (23). IRT typically uses a logistic model to estimate the probability of various types of item responses and thus to describe item functioning along a continuum (24). Under IRT, the primary purpose for administering a psychometric test is to locate the person taking it on the latent trait scale. If such a latent trait measure can be obtained for each person taking the test, two goals can be achieved. First, the respondent can be evaluated for the severity of the characteristic of interest and second, comparisons among respondents can be made to assign severity grades (25) under the appropriate IRT model. Within the IRT family, the logistic graded response model (GRM) is a cumulative probability model developed by Samejima (22) and designed for Likert-type items.

The aim of the current study is to uncover the unknown pattern of item 13 and to compare the fit of unidimensional and multi-dimensional of both the original HoNOS 12 and the new HoNOS 13 using IRT modelling in a sample of patients with SUD.

Methods

The data of this study were collected by experienced data extractors from the hospital electronic medical record system from February 2015 to September 2019. They concerned patients with SUD admitted to a specialized addiction unit of a large university hospital. The population were mainly men (70.7%), with a mean age of 43.3 (SD 11.5) years. During the reported period, the number of hospitalizations ranged from 1 to 13 with a median length of stay of 15 days (2-690). The median HoNOS score was 16 (1-44) at admission and 11 (0-37) at discharge. The questionnaire was administered by the psychiatrists working in the hospital unit who had received a training session for the use of this tool. The Geneva ethics comity approved this study (ClinicalTrials.gov, Identifier: NCT03551301). Six hundred nine (609) valid questionnaires of the

HoNOS were analyzed. The validated French version of HoNOS (HoNOS-F) (7) was used with an added 13th item (HoNOS 13) concerning medication adherence.

Statistical analysis

In this study, we used the multidimensional extension of the IRT (MIRT) graded-response modelling (GRM) because HoNOS is a polytomous-ordered categorical scale. The items are ranked on a 5-point Likert scale from 0 (no problem) to 4 (severe to very severe problem). In GRM, the following two types of parameters are estimated: the discrimination parameter and the difficulty parameter.

Because GRM is an ordered logistic model, parameters of each item were naturally estimated in increasing order.

For a K-category (*k = 0, …, K-1*) ordinal items, the multidimensional GRM model can be written as follows (26):

$$P(Y_{ij} = k \mid \theta_j) = \frac{1}{1 + e^{-(d_{j(k-1)} + a_i' \theta_j)}} - \frac{1}{1 + e^{-(d_{jk} + a_i' \theta_j)}}$$

In which k is the response category selected by individual j for item i. MIRT model parameters are estimates using the same procedures as for unidimensional models, the only difference being the number of dimensions included in the model.

These parameter estimates can be obtained using the Mirt package (27) of the free R program (28).

A high discrimination parameter suggests that an item has a high ability to differentiate subjects. In practice, a high discrimination parameter value means that the probability of endorsing an item response increases more rapidly as the latent trait or severity increases (29).

When discrimination is high (and the item response function is steep), the item provides more information on the latent trait and the information is concentrated around item difficulty. Items with low discrimination parameters, albeit less informative, may provide information over a wider range of the latent trait. With a logistic model for the item characteristic curve (ICC), Baker (25) proposed the following different ranges of values to better interpret the discrimination parameter: 0=non discriminative power; 0.01-0.34=very low; 0.35-0.64=low; 0.65-1.34=moderate; 1.35-1.69=high; >1.70=very high; and + infinity=perfect.

In GRM, the number of thresholds is equal to the outcome categories minus 1. In this study, we had five alternative responses yielding four thresholds. The item threshold in the GRM model refers to the level of the latent variable an individual needs to endorse the item with 50% probability (30).

Table 1 pictures the distribution of HoNOS and its four dimensions as found by Wing et al. (1).

Using the data at admission, we first fitted a four-factor (latent trait) model of HoNOS 12 as found by Wing et al. (1). Next, a one-factor model as found by Lauzon et al. (7) was fitted. For HoNOS 13, we fitted respectively a one-factor and a two-factor model. This two-factor model was identified by expert consensus, suggesting to group items 1 to 8 and 13 on one side and items 9 to 12 on the other side. The first factor would capture psychiatric/impairment-related issues while the second factor would reflect social-related issues.

Nested models were compared with the anova function of the Mirt package of R.

Good fit of the models was assessed by the root mean square error of approximation (RMSEA) of <0.08 and <0.06, respectively, and the comparative fit index (CFI) values of >0.90 and >0.95, respectively (31,32). Other information criteria, specifically the Akaike information criterion (AIC), AIC corrected (AICc), Bayesian information criterion (BIC), and the sample size adjusted BIC (SABIC) were also used, knowing that AIC and BIC are specifically designed to penalize for model complexity. Robust values for RMSEA and CFI were reported.

A cross-validation study was performed on HoNOS data recorded at discharge.

All analyses and plots were obtained using R program. More specifically, the multidimensional item response theory package using the full-information maximum likelihood (FIML) estimator was used. The FIML estimator is recommended when estimating IRT models with relatively small sample sizes (33). We obtained the CFA analysis with the Lavaan package.

Sample size requirements

Forero and Maydeu-Olivares (33) cited by Depaoli et al. (34) have found that sample sizes as small as 200 were sufficient for the parameter estimation of a graded response model. On the other hand, Jiang and al. also cited by Depaoli et al. (34) showed that a sample size of 500 provided accurate parameter estimates in the case a three-dimensional GRM composed from 30 to 90 items each with four response categories (35). The sample size at hand (609) is adequate for two-dimensional GRM with 13 items between and fulfilled the necessary requirements.

Results

The results of the CFA performed in HoNOS 12 and HoNOS 13 respectively are presented in Table 2.

Model fit

We note that while the one-factor model of HoNOS 12 yielded mediocre fits RMSEA (0.093) and CFI (0.899), the 4-Factor model as advocated by Wing et al. (1) was not identified. There are several reasons for this phenomenon, the most likely being that the model was too complex.

As for HoNOS 13, the expert consensus two-factor model yielded better goodness-of-fit values compared to the one-factor solution. AIC, AICc, BIC and SABIC were lower than in the one-factor model. Moreover, the two-factor model fulfilled the criteria of satisfactory RMSEA and CFI statistics (0.075 and 0.929 compared to 0.088 and 0.901). We conclude with these empirical findings that the 13-item scale can be conceptualized as a two-factor model because it produces a significantly improved fit over the unidimensional model. Moreover, the significant p-value yielded by an anova test comparing the two nested Mirt objects, suggests that the two-factor model is superior to the competing one-Factor model ($p<0.001$). The consensus across indices therefore made the two-factor model the best candidate for further evaluation.

Table 3 presents the results of item loading in a one-factor solution compared to a two-factor solution. It can be seen that most of the items have a higher loading onto their respective factor in the two-factor model than in the one-factor model. The standardized loadings of all items in their respective factor exceed 0.30, except for item 5 that presents a value loading of 0.20. The high loading of item 13 onto its component, the highest compared to the others, strongly legitimates its presence in the conceptualization of the scale.

In Table 4 we present the GRM estimates. In terms of the ranges proposed by Baker (25), we observed that items 9, 10, 11 and 12 had very high discriminative power with a range of 1.75-2.73, items 1, 2, 3, 4, 7, 8 and 13 had moderate discriminative power (range: 0.70 to 1.17) and items 5 and 6 showed very low to low discriminative power (range: 0.33 and 0.57). Items with negative difficulty parameters are considered to be more frequently endorsed (this is the case of items 3, item 7 to item 13), and items with positive difficulties are linked to the less frequently endorsed ones (items 1 and 2, 4 to 6). Two other links can be established between the most and the less endorsed items: 1) the discrimination parameters are higher in the former than in the latter and 2), the most endorsed items tend to be clustered along Factor 2 while the less endorsed ones are clustered along Factor 1.

The GRM being defined in terms of cumulative probabilities allows cumulative comparisons. The difficulties represent a point at which a person with $\theta = b_{ik}$ has a 50% chance of responding in category k or higher (36). For example, looking at the estimated parameters for Item 13, we see that a person with $\theta=-0.14$ has a 50% chance of answering 1 versus greater than or equal to 2, a person with $\theta=0.47$ has a 50% chance of answering 1 or 2 versus greater than or equal to 3. Similarly, a person

with θ=1.48 has a 50% chance of answering 1, 2, or 3 versus greater than or equal to 4, and a person with θ=2.64 has a 50% chance of answering 1, 2, 3, or 4 versus 5. We note that the ratings for Item 13 span a broad range of the latent trait and that its discrimination parameter is relatively high.

Unlike the one-dimensional item characteristic curve (ICC) in IRT, for a general Mirt with latent variable of dimension p, the expected total score is a response surface in a p-dimensional surface. In our case, Figure 1 (Expected Total Score) represents the total expected score in a two-dimension space. It can be seen that the first component (items 1 to 8 and 13) accounts less in the expected total than the second one after standardization.

The cross-validation study conducted on HoNOS data at discharge confirmed the Expert-model as superior to the one-component model (RMSEA = 0.062 and CFI statistics =0.955).

Discussion

The present study investigated the psychometric properties of the HoNOS 13 compared to HoNOS 12 in a large sample of in-patients with SUD. The present study is the first to our knowledge to cover such differences using an IRT model. We found that a one-dimensional instrument was not reliable to use as a primary outcome. The two-factor model of HoNOS 13 which resulted by expert consensus, seemed to reflect the data best. This model groups psychiatric/impairment-related issues (symptoms) while the second factor reflects social-related issues (problems). The lower loading observed for factor 1 (especially for items 5 and 6) is likely due to the heterogeneity of the psychiatric symptoms (19,37) assessed by the HoNOS. This is probably the main reason that leads to the heterogeneity of results of CTT-based studies (38,39), giving more importance to specific item scores than to a global score. The higher loadings observed for the social-related issues may reflect a form of commonality of such problems among individuals with SUD and/or psychiatric disorders. Similar figures for the social-related items were observed in another study using a sample with psychiatric disorders (19).

We also found that the discrimination estimates for the items ranged from 0.33 to 2.73, indicating that some items of HoNOS 13, show rather low discrimination ability whereas others have high levels (Figure 2 (Item Characteristic Curves (ICC)) and Table 4). However, the strength of the factor loadings of items 5 and 6 in the two-component model is a matter of concern which their item characteristic curves (Cf. Figure 2) reflect. Item 5 measures physical impairment and item 6 hallucinations. These items seem to be less important in our specific group of patients with SUD. As the sample was taken from a specialized addiction unit, patients were typically treated for substance withdrawal and were less commonly admitted for acute psychiatric disorders. This may explain fewer problems with hallucinations (item 6) as found in a study by Andreas et al. (18). Even though comorbid substance use is common among patients with psychotic disorders (40) these are more likely to be treated in psychiatric units. In the present sample, 22.9% of the subjects scored higher than zero in this item showing some kinds of symptoms, however not enough linked to overall severity of the latent trait (Table 4). A similar comment could be made for the items 5 (physical illness or disabilities problems) where 37.4% of the participants (scored from 1 to 4) on this item showing that such issues are common among patients with SUD (41,42) however without having a strong contribution to catch the severity of the latent trait. Patients presenting important physical impairments are perhaps more often admitted to general hospital units for withdrawal and treatment of comorbid physical disorders. The removal of items 5 and 6 could yield stronger goodness-of-fit measures. But recalling that the development of a scale is not solely a question of statistical matter, model modification based on modification indices may result in models that lack validity, highly susceptible to capitalization on chance. Therefore, the modifications should be defensible from a theoretical point of view (43). For these reasons, a safe approach is to consider the scale in its integrality, that is, using all 13 items. Particularly removing such items could be problematic when considering other populations such as the ones admitted in acute psychiatric wards. However, the present data lead to expect loadings and IRT results variation according to the specific population (specially for the Factor 1, symptoms related items). For instance, in the present study, item 3 (problem drinking or drug taking), show the highest discriminative ability among the Factor 1 (moderate discriminative power) related items, as expected for patients admitted in a specialized addictive disorders unit. A similar figure is observed for item 13 (Figure 2, Table 4). This item may contribute in

a more transdiagnostic way to the latent construct. Further studies using IRT on other populations are needed to assess the role of this item.

By contrast, the issues assessed by the Factor 2-related items were found to have very high discriminative power. These problems are common among patients with SUD as well as patients with other mental disorders (44,45) and were also observed in studies using HoNOS in inpatients admitted for psychiatric disorders (19). Importance of social problems among people with addictive disorders (46,47), and their influence in the rate of service use (48) were repeatedly observed especially for more severe forms and longer duration of substance use. Social problems-related symptoms seem to play an important role in the overall severity. This highlights the importance of community and recovery-oriented interventions (49,50) as well as for approaches focusing on transdiagnostic factors involved in such difficulties such as theory of mind (51) or self-stigma (52).

HoNOS 13 can be recommended as a clinical evaluation tool to assess the problems and treatment needs for inpatients with SUD. It is necessary to assess the two-factor model suggested in this study in other patient groups. It could be hypothesized that loadings and discriminative power may change across items depending on the clinical characteristics of a given population. For people with psychiatric and addictive disorders, the items related to the second factor and probably item 13 may show more constant characteristics.

Nevertheless, IRT seems to be of particular interest when analyzing symptoms using the HoNOS scale.

This analysis presents one main limitation as it used routinely collected administrative and clinical data. It was therefore not possible to have more detailed information about individual patients such as specific measures on addiction severity, duration of treatment, and marital or family status.

Conclusions

The 13-item questionnaire including medication compliance was validated in this analysis. In spite of the above limitation, the HoNOS-13 scale including a question “Problems with psychotropic medication compliance” can be recommended as a valid clinical evaluation tool to assess the problems and treatment needs for inpatients with SUD. In IRT analyses, the items related to substance use and item 13 showed moderate discriminative abilities to catch the severity of the latent construct whereas the items related to the second factor (social problems) showed higher discriminative abilities.

Abbreviations

AIC	Akaike information criterion
AICc	AIC corrected
BIC	Bayesian information criterion
CFA	Confirmatory Factor Analysis
CFI	Comparative fit index
CTT	Classical Test Theory context
EFA	Exploratory Factor Analysis
FIML	Full-information maximum likelihood estimator
GRM	Graded-response modelling

HoNOS	Health of the Nation Outcome Scale
ICC	Item characteristic curve
IRT	Item Response Theory
MIRT	Multidimensional extension of the IRT
RMSEA	Root mean square error of approximation
SABIC	Sample size adjusted BIC
SUD	Substance use disorder

Declarations

Funding:

No funding was obtained for this study.

Disclosure statement:

The authors declare that they have no conflict of interest.

Competing interests:

Not applicable.

Data availability:

Data can be made available by the corresponding author upon request.

Authors' contribution:

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by AC. The first draft of the manuscript was written by AC and LP and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval:

Approval was obtained from the Geneva Ethics Committee (ID 2017-00733). The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Consent to participate:

This study used routinely collected anonymized hospital data. Therefore, the Geneva Ethics Committee did not request individual consent from each participant.

Consent to publish:

Not applicable.

Acknowledgements:

Not applicable.

References

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

2. Wing J, Curtis RH, Beevor A. Health of the Nation Outcome Scales (HoNOS). Glossary for HoNOS score sheet. Br J Psychiatry J Ment Sci. 1999 May;174:432–4.

3. Eagar K, Gaines P, Burgess P, Green J, Bower A, Buckingham B, et al. Developing a New Zealand casemix classification for mental health services. World Psychiatry. 2004 Oct;3(3):172–7.

4. Tulloch AD, Khondoker MR, Thornicroft G, David AS. Home treatment teams and facilitated discharge from psychiatric hospital. Epidemiol Psychiatr Sci. 2015 Oct;24(5):402–14.

5. Tulloch AD, David AS, Thornicroft G. Exploring the predictors of early readmission to psychiatric hospital. Epidemiol Psychiatr Sci. 2016 Apr;25(2):181–93.

6. Trauer T. The subscale structure of the Health of the Nation Outcome Scales (HoNOS). J Ment Health. 1999;8(5):499–509.

7. Lauzon S, Corbière M, Bonin JP, Bonsack C, Lesage AD, Ricard N. [Validation of the French version of the Health of the Nation Outcome Scales (HoNOS-F)]. Can J Psychiatry Rev Can Psychiatr. 2001 Nov;46(9):841–6.

8. Speak B, Muncer S. The structure and reliability of the Health of the Nation Outcome Scales. Australas Psychiatry. 2015 Feb 1;23(1):66–8.

9. Lovaglio PG, Monzani E. Validation aspects of the health of the nation outcome scales. Int J Ment Health Syst. 2011 Sep 6;5(1):20.

10. Lovaglio PG, Monzani E. Health of the nation outcome scales evaluation in a community setting population. Qual Life Res. 2012 Nov 1;21(9):1643–53.

11. Jacobs R. Investigating Patient Outcome Measures in Mental Health [Internet]. 2009 [cited 2020 Feb 26]. Available from: http://eprints.whiterose.ac.uk/139380/

12. Bebbington P, Brugha T, Hill T, Marsden L, Window S. Validation of the Health of the Nation Outcome Scales. Br J Psychiatry J Ment Sci. 1999 May;174:389–94.

13. Bender KG. The meagre outcomes of HoNOS. Australas Psychiatry. 2020;28(2):206–9.

14. 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):63–8.

15. 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 Nov 28;3(1):76.

16. Semahegn A, Torpey K, Manu A, Assefa N, Tesfaye G, Ankomah A. Psychotropic medication non-adherence and associated factors among adult patients with major psychiatric disorders: a protocol for a systematic review. Syst Rev. 2018 Jan 22;7(1):10.

17. Bonsack C, Borgeat F, Lesage A. Mesurer la sévérité des problèmes des patients et leur évolution dans un secteur psychiatrique : une étude sur le terrain du Health of Nation Outcome Scales en français (HoNOS-F). Ann Méd-Psychol Rev Psychiatr. 2002 Sep 1;160(7):483–8.

18. Andreas S, Harries-Hedder K, Schwenk W, Hausberg M, Koch U, Schulz H. Is the Health of the Nation Outcome Scales appropriate for the assessment of symptom severity in patients with substance-related disorders? J Subst Abuse Treat. 2010 Jul 1;39(1):32–40.

19. Golay P, Basterrechea L, Conus P, Bonsack C. Internal and Predictive Validity of the French Health of the Nation Outcome Scales: Need for Future Directions. PLoS ONE [Internet]. 2016 Aug 2 [cited 2019 Oct 21];11(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4970811/

20. De Ayala RJ. Methodology in the social sciences. The theory and practice of item response theory. Guilford Press. https://doi. org/10.3102/10769986030003295; 2009.

21. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. MESA Press, 5835 S; 1993.

22. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychom Monogr Suppl 17. 1969;

23. Bock RD. A Brief History of Item Theory Response. Educ Meas Issues Pract. 1997;16(4):21–33.

24. Reeve BB, Fayers P. Applying item response theory modeling for evaluating questionnaire item and scale properties. Assess Qual Life Clin Trials Methods Pract. 2005;2:55–73.

25. Baker F. The Basics of Item Response Theory [Internet]. ERIC Clearinghouse on Assessment and Evaluation; 2001. Available from: http://www.edres.org/irt/baker/final.pdf

26. Immekus JC, Snyder KE, Ralston PA. Multidimensional item response theory for factor structure assessment in educational psychology research. In: Frontiers in Education. Frontiers; 2019. p. 45.

27. Chalmers RP. mirt: A multidimensional item response theory package for the R environment. J Stat Softw. 2012;48(6):1–29.

28. R Core Team. R Foundation for Statistical Computing: R: A language and environment for statistical computing. Vienna, Austria; 2018.

29. An X, Yung Y-F. Item response theory: What it is and how you can use the IRT procedure to apply it. SAS Inst Inc SAS364-2014. 2014;10(4).

30. Lipscomb J, Gotay C, Snyder C, editors. Outcomes assessment in cancer: measures, methods and applications. [Internet]. Cambridge University Press; 2004 [cited 2020 Feb 26]. Available from: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Lipscomb+J%2C+Gotay+C%2C+Snyder+C.+Outcomes+Asjsessment+in+Cancer%3A+Measures%2C

+Methods+and+Applications.+Cambridge%3A+Cambridge+University+Press%3B+2005.&btnG=

31. Hooper D, Coughlan J, Mullen MR. Structural equation modelling: Guidelines for determining model fit. Electron J Bus Res Methods. 2008;6(1):53–60.

32. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Model Multidiscip J. 1999;6(1):1–55.

33. Forero CG, Maydeu-Olivares A. Estimation of IRT graded response models: Limited versus full information methods. Psychol Methods. 2009;14(3):275.

34. Depaoli S, Tiemensma J, Felt JM. Assessment of health surveys: fitting a multidimensional graded response model. Psychol Health Med. 2018;23(sup1):1299–317.

35. Jiang S, Wang C, Weiss DJ. Sample size requirements for estimation of item parameters in the multidimensional graded response model. Front Psychol. 2016;7:109.

36. StataCorp. StataCorp. Stata Statistical Software: Release 141. College Station, TX: StataCorp LP; 2016.

37. Kraus L, Baumeister SE, Pabst A, Orth B. Association of Average Daily Alcohol Consumption, Binge Drinking and Alcohol-Related Social Problems: Results from the German Epidemiological Surveys of Substance Abuse. Alcohol Alcohol. 2009 May 1;44(3):314–20.

38. Andreas S, Harfst T, Dirmaier J, Kawski S, Koch U, Schulz H. A psychometric evaluation of the German version of the 'health of the nation outcome scales, HoNOS-D': on the feasibility and reliability of clinician-performed measurements of severity in patients with mental disorders. Psychopathology. 2007;40(2):116–25.

39. Turton R. An exploratory factor analysis of HONOS-LD scales. Adv Ment Health Intellect Disabil. 2020;

40. Pennou A, Lecomte T, Potvin S, Khazaal Y. Mobile Intervention for Individuals With Psychosis, Dual Disorders, and Their Common Comorbidities: A Literature Review. Front Psychiatry. 2019;10:302.

41. Han BH, Termine DJ, Moore AA, Sherman SE, Palamar JJ. Medical multimorbidity and drug use among adults in the United States. Prev Med Rep. 2018 Dec;12:214–9.

42. Wu L-T, Zhu H, Ghitza UE. Multicomorbidity of chronic diseases and substance use disorders and their association with hospitalization: Results from electronic health records data. Drug Alcohol Depend. 2018 01;192:316–23.

43. Schermelleh-Engel K, Moosbrugger H, Müller H. Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. Methods Psychol Res Online. 2003;8(2):23–74.

44. Khan S. Concurrent mental and substance use disorders in Canada. Health Rep. 2017 Aug 16;28(8):3–8.

45. Moulin A, Evans E, Xing G, Melnikow J. Substance Use, Homelessness, Mental Illness and Medicaid Coverage: A Set-up for High Emergency Department Utilization. West J Emerg Med. 2018 Oct 18;19(6):902–6.

46. Cerdá M, Moffitt TE, Meier MH, Harrington H, Houts R, Ramrakha S, et al. Persistent cannabis dependence and alcohol dependence represent risks for midlife economic and social problems: A longitudinal cohort study. Clin Psychol Sci. 2016;4(6):1028–46.

47. Rhemtulla M, Fried EI, Aggen SH, Tuerlinckx F, Kendler KS, Borsboom D. Network analysis of substance abuse and dependence symptoms. Drug Alcohol Depend. 2016;161:230–7.

48. Penzenstadler L, Gentil L, Huỳnh C, Grenier G, Fleury M-J. Variables associated with low, moderate and high emergency department use among patients with substance-related disorders. Drug Alcohol Depend. 2020;207:107817.

49. Penzenstadler L, Machado A, Thorens G, Zullino D, Khazaal Y. Effect of Case Management Interventions for Patients with Substance Use Disorders: A Systematic Review. Front Psychiatry. 2017;8:51.

50. Penzenstadler L, Soares C, Anci E, Molodynski A, Khazaal Y. Effect of Assertive Community Treatment for Patients with Substance Use Disorder: A Systematic Review. Eur Addict Res. 2019;25(2):56–67.

51. Pennou A, Lecomte T, Khazaal Y, Potvin S, Vézina C, Bouchard M. Does theory of mind predict specific domains of social functioning in individuals following a first episode psychosis? Psychiatry Res. 2021;113933.

52. Oexle N, Müller M, Kawohl W, Xu Z, Viering S, Wyss C, et al. Self-stigma as a barrier to recovery: a longitudinal study. Eur Arch Psychiatry Clin Neurosci. 2018;268(2):209–12.

Tables

Table 1: Description of HoNOS items by the four dimensions as found by Wing and colleagues

Section	Range of scores for section	Item name	Item score	Response rate
Behaviour	0-12	1 Overactive, aggressive, disruptive or agitated behaviour	0	68.9
			1	15.2
			2	10.6
			3	3.8
			4	1.6
		2 Non-accidental self injury	0	82.1
			1	9.2
			2	5.3
			3	2.8
			4	0.6
		3 Problem drinking or drug taking	0	12.0
			1	12.0
			2	19.7
			3	33.5
			4	22.8
Impairment	0-8	4 Cognitive problems	0	72.9
			1	14.7
			2	8.5
			3	3.4
			4	0.6
		5 Physical illness or disability problems	0	62.7
			1	16.5
			2	13.8
			3	5.9
			4	1.2
Symptoms	0-12	6 Problems with hallucinations and delusions	0	77.1
			1	8.9
			2	6.9
			3	4.6
			4	2.5
		7 Problems with depressed mood	0	18.5
			1	23.8
			2	37.6
			3	15.6

			4	4.5
		8 Other mental and behavioural problems	0	27.8
			1	18.3
			2	37.2
			3	13.5
			4	3.3
Social	0-16	9 Problems with relationships	0	31.3
			1	31.5
			2	26.5
			3	8.5
			4	2.2
		10 Problems with activities of daily living	0	38.7
			1	24.9
			2	24.4
			3	9.3
			4	2.7
		11 Problems with living conditions	0	37.6
			1	22.8
			2	22.4
			3	11.5
			4	5.6
		12 Problems with occupation and activities	0	20.8
			1	24.6
			2	35.3
			3	15.7
			4	3.7
(additional item)		**13 Problems with psychotropic medication compliance**	**0**	**60.3**
			1	**12.9**
			2	**14.6**
			3	**7.5**
			4	**4.7**

Table 2: Comparison of model fit results of the confirmatory factor analysis of HoNOS

		AIC	AICc	BIC	SABIC	RMSEA	CFI
HoNOS 12	4-Factor (Wing et al.)	Model not identified					
	1- Factor	17787.1	17800.5	18051.9	17861.4	0.093	0.899
HoNOS 13	1- Factor	19389.4	19405.2	19676.2	19469.8	0.088	0.901
	2- Factor (expert consensus)	19327.0	19343.3	19618.1	19408.59	0.075	0.929

Table 3: Item loadings of HoNOS 13: one-factor versus two-factor solution by confirmatory factor analysis

Items and factor loadings of HoNOS13			
		two-factor solution	
	one-factor solution	Factor 1	Factor 2
Item 1	0.40	0.46	
Item 2	0.35	0.42	
Item 3	0.50	0.57	
Item 4	0.32	0.38	
Item 5	0.18	0.20	
Item 6	0.29	0.33	
Item 7	0.35	0.40	
Item 8	0.42	0.48	
Item 9	0.71		0.72
Item 10	0.80		0.82
Item 11	0.68		0.70
Item 12	0.74		0.76
Item 13	0.51	0.58	
RMSEA	0.088	0.075	
CFI	0.901	0.929	

RMSEA: root mean square error of approximation, CFI: comparative fit index

Table 4: GRM IRT estimates

IRT parameter estimates for the graded response models						
	Discrimination (Slope)		Difficulty (Threshold)			
Item N°	a1	a2	b1	b2	b3	b4
Item 1	0.85		0.54	1.65	3.38	5.22
Item 2	0.77		1.51	2.59	3.54	4.76
Item 3	1.16		-3.02	-2.57	-1.86	-0.02
Item 4	0.70		0.77	2.35	4.71	7.89
Item 5	0.33		0.55	2.53	6.10	10.97
Item 6	0.57		2.11	3.33	4.61	6.33
Item 7	0.78		-2.80	-1.46	0.74	3.15
Item 8	0.99		-1.65	-0.87	1.01	2.99
Item 9		1.90	-1.07	-0.12	0.95	2.24
Item 10		2.73	-0.88	-0.33	0.66	1.74
Item 11		1.75	-0.91	-0.19	0.78	1.86
Item 12		2.16	-1.54	-0.76	0.41	1.59
Item 13	1.17		-0.14	0.47	1.48	2.64

GRM: graded-response modelling, IRT: Item Response Theory

Figures

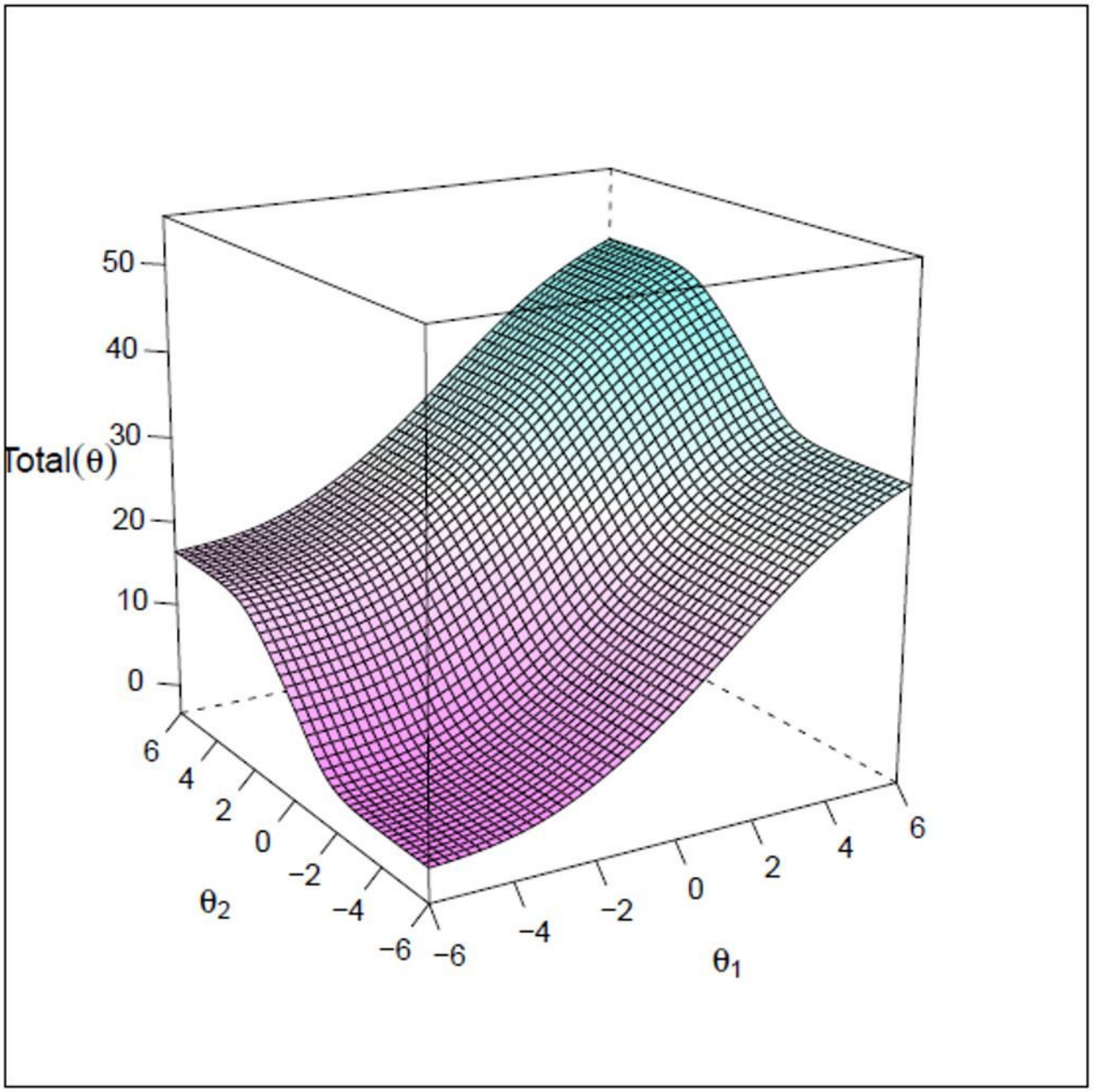

Figure 1

Expected Total Score

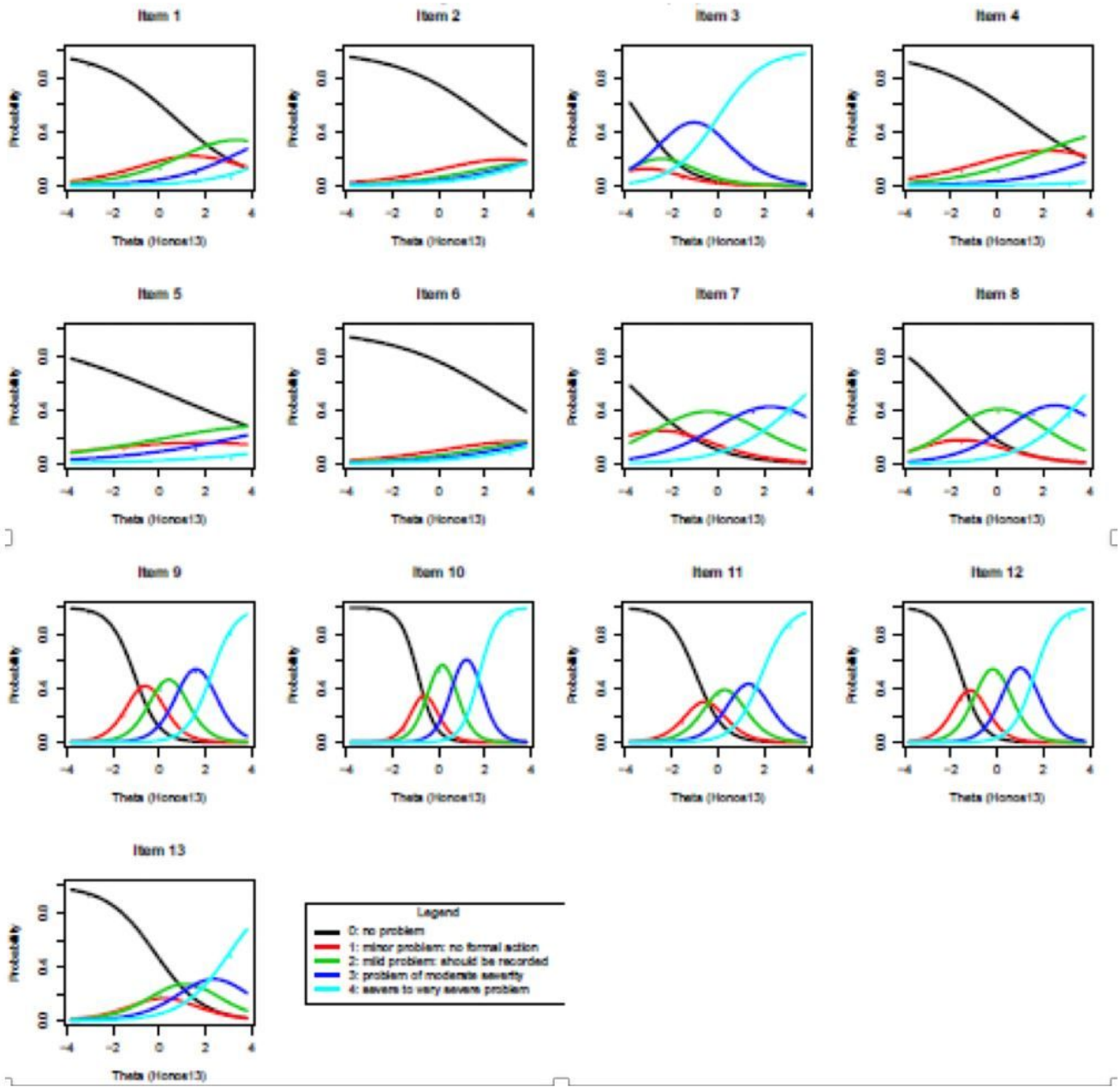

Figure 2

Item Characteristic Curves (ICC)