Rasch analysis of a patient-reported outcome measure for self-perceived health among psychiatric patients in Denmark

Valentin, Jan Brink; Mainz, Jan; Johnsen, Søren Paaske; Kristensen, Klaudia; Kristensen, Solvejg

Published in:
BMJ Open Quality

DOI (link to publication from Publisher):
10.1136/bmjoq-2023-002519

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Valentin, J. B., Mainz, J., Johnsen, S. P., Kristensen, K., & Kristensen, S. (2024). Rasch analysis of a patient-reported outcome measure for self-perceived health among psychiatric patients in Denmark. *BMJ Open Quality*, *13*(1), Article e002519. https://doi.org/10.1136/bmjoq-2023-002519

BMJ Open Quality

Rasch analysis of a patient-reported outcome measure for self-perceived health among psychiatric patients in Denmark

Jan Brink Valentin ,[1] Jan Mainz,[1,2,3,4] Søren Paaske Johnsen,[1] Klaudia Kristensen,[2] Solvejg Kristensen[2]

To cite: Valentin JB, Mainz J, Johnsen SP, *et al*. Rasch analysis of a patient-reported outcome measure for self-perceived health among psychiatric patients in Denmark. *BMJ Open Quality* 2024;**13**:e002519. doi:10.1136/bmjoq-2023-002519

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjoq-2023-002519).

Received 26 July 2023
Accepted 25 January 2024

[1]Danish Center for Health Services Research, Department of Clinical Medicine, Aalborg University, Aalborg, Denmark
[2]Psychiatry, Aalborg University Hospital, Aalborg, Denmark
[3]Department for Community Mental Health, University of Haifa, Haifa, Israel
[4]Department of Health Economics, University of Southern Denmark, Odense, Denmark

Correspondence to
Dr Jan Brink Valentin; jvalentin@dcm.aau.dk

ABSTRACT

Background Patient-reported outcome measures (PROMs) are valuable and necessary tools for establishing and maintaining patient-centred healthcare. The PRO-Psychiatry initiative was primarily initiated to support the patient's voice in treatment decision-making and secondarily to monitor patient-perceived quality of care. The result of the initiative is a patient-reported instrument developed in collaboration between patients and clinicians. We aimed to validate the PROM developed for measuring self-perceived health among psychiatric patients in North Denmark Region, in terms of internal consistency, criterion validity and responsiveness.

Method Patients in contact with a psychiatric hospital in the North Denmark Region from September 2018 to March 2021 were included in the study. The PROM constitutes a scale of 17 items covering various aspects of self-perceived health including well-being (7 items), lack of well-being (5 items) and social functioning (5 items), where the former domain entails the WHO-5 Well-Being Index. The potential range of the total scale score is 0–85. We applied McDonald's omega, average inter-item correlation (AIIC) and differential item functioning (DIF). In addition, we used mixed effects analyses to estimate temporal correlations. The instrument was compared with self-rated overall mental and psychiatric health.

Results The patient population consisted of 1132 unique patients and a total of 2476 responses corresponding to one response per patient pathway. McDonald's omega was found to be 0.92 (95% CI 0.92 to 0.93), while the AIIC was found to be 0.42 (95% CI 0.39 to 0.44). For DIF, the largest systematic variation resulted in a maximum difference of 2.3 points on the total score when adjusting for the latent trait and was found when comparing initial measurements with follow-up measurements. The correlation between the total score and the outcomes regarding overall physical and mental health was 0.52 (95% CI 0.48 to 0.56) and 0.74 (95% CI 0.72 to 0.76). Similar correlations were found for the corresponding changes over time.

Conclusion The scale showed high consistency and little systematic variation between the comparison groups. The concurrent correlations and analyses of responsiveness coincided with the prespecified hypotheses. Overall, we deem the Danish PRO-Psychiatry instrument to possess suitable psychometric properties for measuring self-perceived health among a psychiatric population.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Patient-reported outcome measures are key elements in patient-centred psychiatric healthcare.

WHAT THIS STUDY ADDS

⇒ Validation of a patient-reported outcome measure of self-perceived psychiatric health developed in cocreation with patients in terms of inter-item consistency, differential item functioning, concurrent validity and responsiveness.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The instrument may aid the development in quality of psychiatric healthcare towards a more patient-centred orientation.

INTRODUCTION

The essence of patient-centred healthcare is a healthcare system that accommodates the patient's needs and invites shared decision-making on treatment to the extent possible.[1 2] Cocreation with patients is a key factor in patient-centred care but is not limited to active engagement during treatment consultations.[3 4] The patient-reported outcome psychiatry (PRO-Psychiatry) project is a joint collaboration between regional hospital-based psychiatry in the North Denmark Region and two national clinical quality registries: the Danish Depression Database and the Danish Schizophrenia Registry.[5 6] The initiative supports and encourages patient involvement in many aspects of healthcare. The product of this novel collaboration is patient-reported outcome measures (PROMs), which have been developed through a series of sessions involving both care providers and receivers.[7]

PROMs in psychiatric healthcare serve as an important tool for enhancing patient-centred care, as these instruments intend to measure traits such as needs, agony and well-being as experienced by the patient.[1,8–10] A psychiatric PROM may even to a larger extent reflect the important aspects of the individual patient's condition other than those inferred from the clinician-administered scales compared with non-psychiatric PROMs.

The purpose of the PROM developed by the PRO-Psychiatry collaboration is twofold, where the primary aim is to enhance patient involvement during treatment consultations, in that the questions that constitute the PROM are constructed in order to guide the consultation towards subjects that are important to the patient in their current state of illness.[7] At the same time, the instrument ensures that the patient is perceived, even in situations where the patient otherwise has difficulties stating their concerns about treatment or disease progression. Moreover, the PROM allows patients to monitor their own health state over time. The secondary purpose is to guide quality of care improvement to a more patient-centred care setting by monitoring and optimising according to indicators of what matters to the patient.[2,7] To accommodate the second purpose, the scale is implemented for registration in the national Danish clinical quality registers for patients with unipolar depression respectively schizophrenia, which did not previously contain information on PROMs.[5,6]

The current study constitutes a paper in a series of consecutive articles, which report the main findings of the work produced by the PRO-Psychiatry collaboration. More specifically, the four papers describe the process of tool development, implementation of the tool, evaluation of the implementation and finally validation of the tool.[7–9,11,12] The subjective validation of the PROM has already been finalised as the instrument has been repeatedly assessed by healthcare professionals and patients during the development phase.[7] However, it is also important to conduct an objective validation of the scale, when the purpose includes quality of care assessment, to ensure that the instrument does not measure some other patient characteristic than the trait intended.[13] Item response theory (IRT) provides a valuable methodology for such assessment. One of the tools of IRT is differential item functioning (DIF), for which the purpose is to investigate whether any patient subgroups defined by the available baseline characteristics have certain prerequisites for giving a particular response to each item.[14] The purpose of IRT is to ensure that the scale items under investigation possess unidimensionality, monotonicity, local independence and absence of DIF.[15] In fact, according to Rosenbaum, these requirements are sufficient criteria to display criterion-related construct validity.[16] However, the COnsensus-based Standards for the selection of health Measurement Instruments checklist suggests assessment of internal reliability, concurrent validity and responsiveness in addition to IRT.[17]

The objective assessment of the instrument has partly been explored as the PROM constitutes various aspects of well-being, which entails the WHO-5 Well-Being Index (WHO-5 WBI) instrument.[18,19] However, a complete objective validation of the complete scale is yet to be conducted, thus, we aim to assess the internal reliability, criterion validity, concurrent validity and responsiveness of the PROM developed for patients admitted to psychiatric hospitals in the North Denmark Region.

METHOD

Design

Data were collected as part of a local quality improvement project in the North Denmark Region, Denmark. The patient characteristics were collected at the first visit during the project period. Only patients who were recurrent in this period were given follow-up questions, however, all measurements including follow-up were used for cross-sectional analysis. Temporal analyses were conducted on the subgroup of patients who were measured at multiple occurrences.

Population

Inpatients and outpatients in contact with a psychiatric hospital in the North Denmark Region in the period from September 2018 to March 2021, older than 18 years of age at the time of contact, and who were willing to answer the questionnaire during this period were included in the study. Healthcare in Denmark is free of charge for all citizens. This includes psychiatric healthcare, which implies that the patient sample to a large extent represents the general population of referable patients with psychiatric disorders in Denmark.

Data

Besides the 17-item PROM, the subjects were asked about overall physical and mental health as well as the effect of medication on quality of life. In addition, the subjects were asked about their education, work status, civil status, alcohol consumption and drug abuse, however, these questions were only administered at baseline. The date of measurement was also recorded.

The questionnaire was linked to data on psychiatric hospital contacts using a unique personal identification number.[20] These data were collected from local patient records ranging from November 2016 to October 2021 and comprised information on diagnoses and date of related and previous contacts as well as sex and date of birth.

Instrument

The scale under investigation consists of the first 17 items of the PROM questionnaire (see online supplemental table S1). These 17 items can be subdivided into 3 domains; well-being (7 items), lack of well-being (5 items) and social functioning (5 items), where well-being is positively formulated while the other two domains are negatively formulated. There are six response categories for each item and the scoring is performed such that a positive answer triggers a high score, and a negative response triggers a low score. Thus, the items of the well-being domain are scored from 0 to 5, where 0 means at no time and 5 means all the time, while scoring for the responses of the remaining domains are reversed such that 0 means all the time and 5 means at no time.

Outcomes

The three outcomes were patient-reported single-item measures and included overall mental health, overall physical health and impact of medication on quality of life. Specifically, these outcomes were defined as the following:

- In general, would you say your physical health is: Poor, fair, good, very good or excellent.
- In general, would you say your mental health is: Poor, fair, good, very good or excellent.
- During the past 2 weeks, I have experienced side effects of my medication which have influenced my quality of life: All of the time, most of the time, more than half of the time, less than half of the time, some of the time or at no time.

The latter question was only given to subjects confirming they were on medication for their psychiatric disorder. All three outcomes were measured at baseline as well as follow-up.

The outcomes are scored such that a positive answer triggers a large value, with the largest value being 5, and a negative answer triggers a low value, with the lowest value being 0 or 1 depending on the number of response categories.

Covariates

The following variables were used for dichotomous groupings considered in the DIF analysis:

- Sex, male versus female.
- Education, graduated high school or other profession versus no education or primary school.
- Work status, currently working versus unemployed.
- Civil status, living together versus living alone.
- Alcohol consumption, drinking alcohol more than three times a week or more than nine units at the time versus less.
- Drug abuse, abused drugs more than nine times in the last 12 months vs less.
- The number of previous contacts with the psychiatry, recurrent versus incident patient.
- Age at inclusion, above versus below median age.
- Disease severity, severe versus non-severe psychiatric disorder.
- Disease severity (broad definition), severe versus non-severe psychiatric disorder.
- Initial measurement versus follow-up measurement.

We used two definitions of disease severity. In the first definition, severe psychiatric disorder was defined as an International Classification of Diseases, Tenth Revision (ICD-10) diagnosis of F20, F22, F25, F30 or F31. In the second and broader version, severe psychiatric disorder was defined as an ICD-10 diagnosis of F20–F29, F30, F31, F323 or F333.[21] For both definitions, the diagnosis was acquired at any timepoint in the period from November 2016 to inclusion.

Statistical analysis

Initially, we conducted a descriptive analysis of baseline characteristics as well as a number of times subjects were measured in total and stratified by disease severity. Variables of continuous nature were be presented as medians and IQR and categorical variables as frequencies and percentages.

For all analyses, we excluded measurements where all items in the instrument under investigation were missing. Any remaining missingness in the data was managed by single value imputation using chained equations.[22] Results are presented with 95% CIs where appropriate. All analyses were conducted in Stata V.16 (StataCorp. 2019. Stata Statistical Software: Release V.16, StataCorp).

Internal reliability

We estimated the McDonald's omega to assess scale consistency.[23] In addition, we estimated the average inter-item correlations (AIIC) as well as minimum and maximum inter-item correlations. The AIIC should reflect a trade-off, such that the scale items are consistent without being isomorphic, where the latter implies that two items may not be identical, or one item may not set limits of response for other items. Thus, the AIIC is preferable between 0.2 and 0.4.[24] Finally, we calculated categorical percentages of each item as well as the outcomes to assess ceiling and floor effects. We estimated 95% CIs using clustered bootstrap with subjects constituting a cluster.

Differential item functioning

For the Rasch analysis, we used a graded response model, which implies item specific discrimination and difficulty parameters.[25 26] For each of the covariates and items, we estimated the difference in average difficulty displacement (ADD), average difficulty separation (ADS) and discrimination between patient groups, defined by the covariates while constraining the parameters of the remaining items to be equal between patient groups.

For the covariates in which the instrument displayed significant DIF, we compared the expected outcomes of the patient groups as a function of the latent trait. Again, we used the graded response model where the parameters were allowed to vary between patient groups among the items which were considered differential, while constraining the parameters for the remaining items.

We only estimated the expected outcomes for covariates which induced the largest amount of DIF based on the number of items with a statistically significant difference in ADD, ADS or discrimination as well as the absolute value for a single item using the same measures of differential functioning.

Concurrent validity

For concurrent validity, we estimated the correlation between the total score of the 17 items with each of the outcomes. Given that the population consists of psychiatric patients, we hypothesised that the total score will to some extend correlate positively with impact of medication on quality of life, show moderate to high positive correlation with overall physical health and correlate the most with overall mental health. We estimated 95% CIs using clustered bootstrap with subjects constituting a cluster.

Responsiveness

Since responsiveness refers to the ability of measuring change over time, we excluded patients with only one measurement during the inclusion period. Moreover, the subjects were measured at different follow-up times with a various number of follow-ups. Thus, we applied repeated measures mixed effects analysis on the total score with a subject specific random slope. Afterward we estimated each of these slopes, which denoted the change in total score per time unit. We estimated similar patient-specific slopes for each of the outcomes and calculated the correlation between the slopes of the instrument and the slopes of each of the outcomes. As for concurrent validity, we hypothesised that the slope of the total score will to some extend correlate positively with the slopes of the impact of medication on quality of life, show moderate to high positive correlation with the slopes of the overall physical health and correlate the most with the slopes of the overall mental health. We estimated 95% CIs using bootstrap.

RESULTS

The patient population consisted of 1132 unique patients and a total of 2476 responses corresponding to one response per patient pathway. Of the total number of subjects, 557 (49.2%) patients were only measured at baseline. The median number of measurements per subject was 2 (IQR 1–2). Almost 50% of the population had a severe psychiatric disorder at baseline using the broad definition, while less than 14% of both severe and non-severe patients recorded that they had abused drugs more than nine times in the last 12 months prior to baseline. Baseline characteristics are presented in table 1.

Internal reliability

The consistency of the complete 17 items of the instrument represented by McDonald's omega was found to be satisfactory with a value of 0.92 (95% CI 0.92 to 0.93). The AIIC was, however, found to be in the high end with a value of 0.42 (95% CI 0.39 to 0.44). This is explained by the rather high maximum inter-item correlation of 0.80 found between item 11 (I have had thoughts indicating it would be better if I was dead) and 12 (I have had thoughts about harming myself). The minimum inter-item correlation was found between item 10 (I have experienced changes in my normal eating habits) and 13 (because of my health problems, my ability to work/take an education is impaired) with a value of 0.20.

The categorical percentages of each item are presented in online supplemental table S2. There was no strong indication of either ceiling or floor effect of either items nor the outcomes, however, the subjects tended to answer at no time or some of the time to the outcome regarding experiencing side effects of their medication. Moreover, less than 4% of the responders answered all of the time to items 1–5 of the instrument. The same was displayed for the two outcomes regarding overall health where less than 4% of the responders answered excellent.

Differential item functioning

The results of the DIF analysis represented by difference in ADD, ADS and discrimination between patient groups are presented in online supplemental table S3. These results indicated that the differences in item responses between groups were relatively small when adjusted for the latent trait. The results for item 1 displayed the most systematic variation especially for recurrent versus incident patient for which the difference in discrimination, ADD and ADS was −0.99 (95% CI −1.57 to −0.42), −0.69 (95% CI −1.07 to −0.32) and −0.74 (95% CI −1.22 to −0.26). The corresponding boundary characteristics curve is shown in figure 1. Likewise, the results for currently working versus unemployed displayed a large variation in DIF measures between groups for item 1. Specifically, the difference in discrimination, ADD and ADS was 0.91 (95% CI 0.31 to 1.50), 0.35 (95% CI −0.02 to 0.71) and 0.70 (95% CI 0.25 to 1.14) and the corresponding boundary characteristics curve is shown in figure 2.

The differences in expected outcome between recurrent and incident patients are shown in figure 3, where the left figure is for an analysis for which only the parameters of item 1 is allowed to differ between groups, while the right figure is for an analysis where the parameters of all items which displayed statistically significant differences on any of the DIF measures were allowed to differ between groups. These items included items 1, 5 and 10–13. The figure shows that the difference in expected outcome between a recurrent and an incident patient with the same latent trait differs by up to 1.1, thus, a difference in the total score of less than 1.1 cannot for certain be attributed to a general difference when comparing groups with unequal distributions of recurrent and incident patients.

The differences in expected outcome between groups displaying the highest number of statistically significant differences in discrimination, ADD and ADS, respectively, are shown in figure 4 and online supplemental figure S1. For discrimination the covariate with maximum number of items displaying statistically significant differences was initial measurements versus follow-up measurements. The number of statistically significant differences in discrimination for this covariate was 13 and included items 1–3, 5–9, 11–12 and 15–17. When allowing the parameters for these items to vary between groups, while constraining the parameters of the remaining items the expected outcome varied between groups by as much as 2.3 on the total scale

Table 1 Baseline characteristics

				No of missing		
	Total	Non-severe psychiatric disorder*	Severe psychiatric disorder*	Total	Non-severe psychiatric disorder	Severe psychiatric disorder
	N=1132	N=664	N=468			
Male	516 (45.6%)	285 (42.9%)	231 (49.4%)	0	0	0
Graduated high school or other profession	537 (50.7%)	323 (51.1%)	214 (50.1%)	73	32	41
Currently working	194 (18.3%)	150 (23.8%)	44 (10.3%)	74	33	41
Living together	332 (33.3%)	234 (39.3%)	98 (24.3%)	134	69	65
High level alcohol consumption	178 (16.8%)	117 (18.6%)	61 (14.3%)	75	34	41
High level substance abuse	140 (13.3%)	85 (13.5%)	55 (12.9%)	77	36	41
More than 0 previous contacts	861 (76.1%)	453 (68.2%)	408 (87.2%)	0	0	0
Above 37 years of age	584 (51.6%)	337 (50.8%)	247 (52.8%)	0	0	0
Severe psych. disorder (broad def.)*	554 (48.9%)	86 (13.0%)	468 (100.0%)	0	0	0
DF0*	43 (3.8%)	27 (4.1%)	16 (3.4%)	0	0	0
DF1*	339 (29.9%)	207 (31.2%)	132 (28.2%)	0	0	0
DF2*	409 (36.1%)	53 (8.0%)	356 (76.1%)	0	0	0
DF3*	475 (42.0%)	305 (45.9%)	170 (36.3%)	0	0	0
DF4*	487 (43.0%)	400 (60.2%)	87 (18.6%)	0	0	0
DF5*	33 (2.9%)	23 (3.5%)	10 (2.1%)	0	0	0
DF6*	162 (14.3%)	124 (18.7%)	38 (8.1%)	0	0	0
DF7*	45 (4.0%)	26 (3.9%)	19 (4.1%)	0	0	0
DF8*	96 (8.5%)	69 (10.4%)	27 (5.8%)	0	0	0
DF9*	129 (11.4%)	92 (13.9%)	37 (7.9%)	0	0	0
No of previous contacts, median (IQR)	3 (2–6)	3 (1–5)	5 (3–8)	0	0	0
Age, median (IQR)	38.0 (26.6–54.2)	37.7 (25.9–54.2)	38.7 (27.9–54.2)	0	0	0

Total and stratified by severity of psychiatric diagnosis.
*Diagnoses groups account primary and secondary diagnoses given at any time between November 2016 to index date.

(see figure 4). For ADD, the covariate with a maximum number of items displaying statistically significant differences was education. Here, the number of statistically significant differences in ADD was 12 and included items 1–3, 8 and 10–17, and when constraining the remaining items, we found that the expected outcome varied between groups with as little as 0.1 on the total scale (see online supplemental figure S1 left). Finally, for ADS the covariate with a maximum number of items displaying statistically significant differences was work status. Here, the number of statistically significant differences in ADD was 15 and included items 1–7, 9–15 and 17, and when constraining the remaining items, we found that the expected outcome varied between groups with 0.5 on the total scale (see online supplemental figure S1 right).

Concurrent validity

As hypothesised the correlation between the total score and the outcome regarding experiencing side effects of their medication was small but positive as well as statistically significant with a coefficient of 0.17 (95% CI 0.11 to 0.23). Likewise, the correlation between the total score and the outcomes regarding overall physical and mental health was both high with coefficients of 0.52 (95% CI 0.48 to 0.56) and 0.74 (95% CI 0.72 to 0.76), where the latter was larger than the former in accordance with the hypothesised correlations.

Responsiveness

The level of responsiveness was only investigated in 575 responders with more than one baseline measurement. The median delay between consecutive measurements within each patient was 22 days (IQR 14–49), while the median number of follow-up measurements was 3 (IQR 1–5). The correlation between the slope of the total score and the slope of the outcome regarding experiencing side effects of their medication was slightly statistically insignificant with a value of 0.14 (95% CI −0.01 to 0.29). However, the slopes of the outcomes regarding overall physical and mental health correlated highly with the slopes of the total score with a statistically significant value of 0.56 (95% CI 0.46 to 0.63) and 0.83 (95% CI 0.79

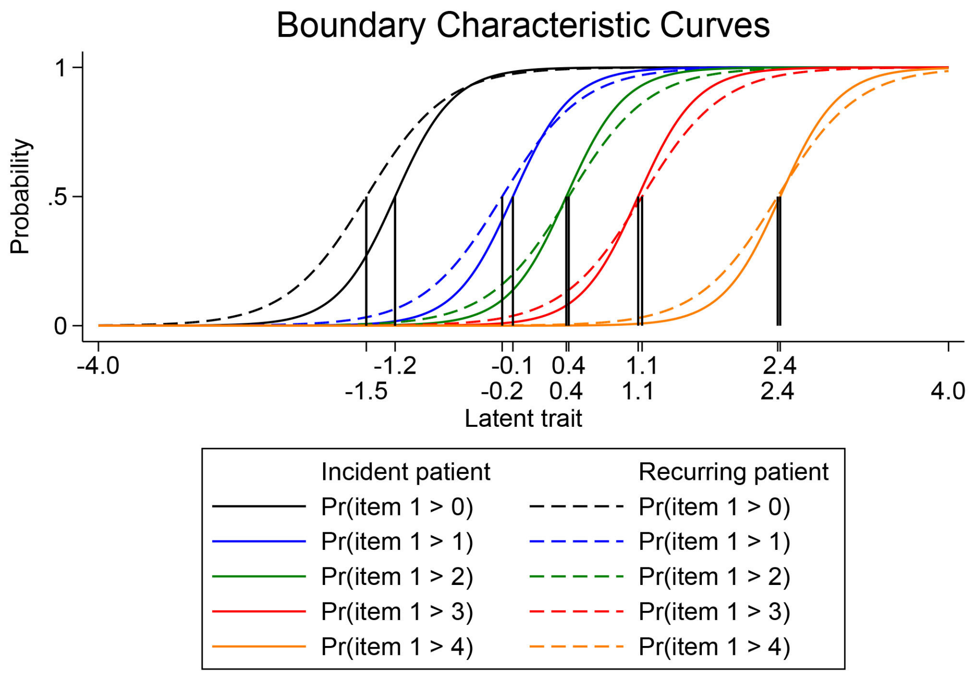

Figure 1 Boundary characteristics curve of item 1 with differential parameters for recurrent versus incident patient while constraining to equal parameters between groups for the remaining items.

to 0.87). Except for the slight insignificance of the one correlation coefficient, the results were in accordance with the hypothesised correlations of the slopes.

DISCUSSION

In general, the scale showed high consistency and little systematic variation for the included covariates. Only a difference in the total score of less than 2.3 appeared to be attributed to systematic variance. This should be viewed relative to a score that ranges from 0 to 85. Moreover, this systematic variance was found when lifting constraints on 13 items out of 17, thus, it could be argued that this difference in total score was a result of the change in the latent trait rather than a systematic error.

The WHO-5 WBI is a validated instrument, which has often been shown to possess adequate psychometric abilities.[18 19] As such, it is surprising that some of the items from the WHO-5 WBI instrument repeatedly display

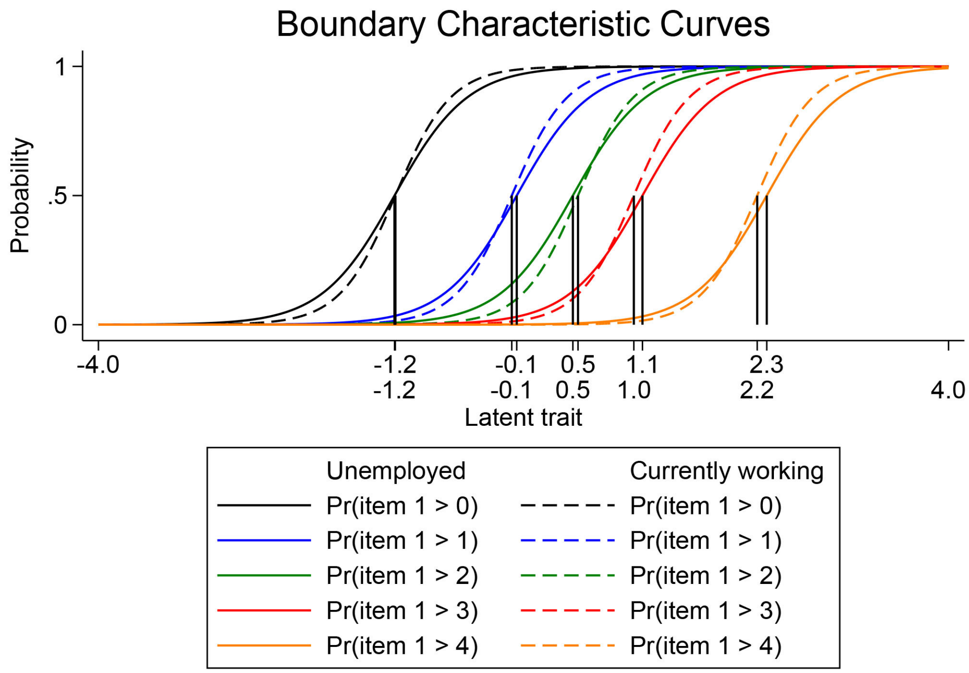

Figure 2 Boundary characteristics curve of item 1 with differential parameters for currently working versus unemployed while constraining to equal parameters between groups for the remaining items.

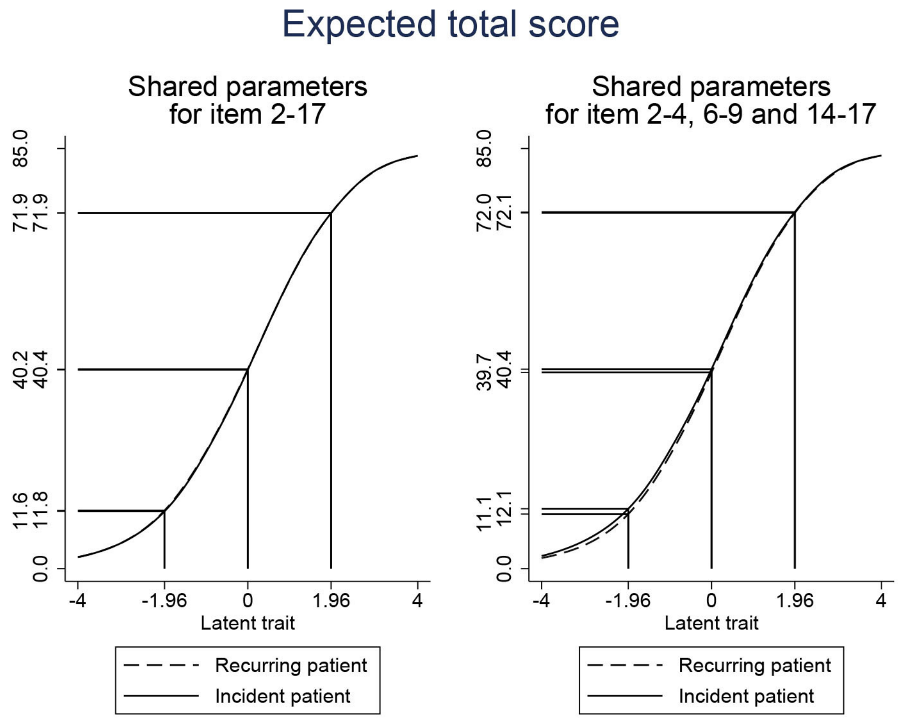

Figure 3 Left: expected outcome for the total score with differential parameters for recurrent versus incident patient of item 1 while constraining to equal parameters between groups for the remaining items. Right; same as left but with differential parameters of items 1, 5 and 10–13.

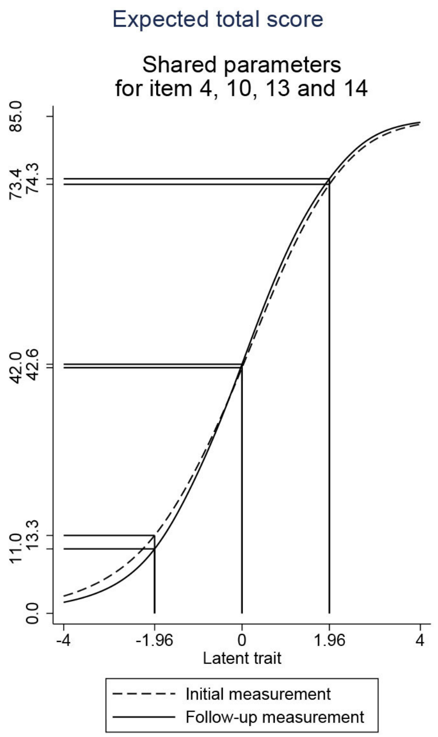

Figure 4 Expected outcome for the total score with differential parameters for initial measurement versus follow-up measurements of items 1–3, 5–9, 11–12 and 15–17 while constraining to equal parameters between groups for the remaining items.

systematic variation, especially item 1 (I have felt cheerful and in good spirits), which has been the subject of scrutiny in all analyses on the expected outcome (figures 3 and 4 and online supplemental figure S1). However, since the measures of DIF were generally low, the fact that item 1 was considered the source of most systematic error reflects confidently on the remaining items.

Finally, while unidimensionality was considered adequate according to McDonald's omega, the inter-item correlations were somewhat high indicating that local independence might be impaired for items 11 and 12, which displayed the largest inter-item correlation among all item pairs. However, it is not surprising that these two items are highly correlated, since they both reflect self-harm, although the response of one item does not constrain the response of the other.

The PROM under investigation in the current study has not previously been validated, however, the WHO-5 WBI scale has previously been subject to scrutiny using Rasch analysis, in which the DIF was investigated across several countries and displayed much higher difference in discrimination.[18] The main explanation for this discrepancy may be that cultural differences weigh higher than any of the groupings under investigation in the current study.

The results regarding concurrent validity and responsiveness were not surprising. The total score was correlated with patient reported general mental and physical health, and the questions on general health were part of the same questionnaire as the scale under the loop. Moreover, the questions on general health were stated in continuation of the scale items, thus, the results of the analyses on concurrent validity and responsiveness may simply be viewed as quality control rather than measures of validity. The WHO-5 WBI scale has previously been assessed using concurrent validity and has shown to be positively correlated with symptom burden among patients with depression and anxiety.[27] It is likely that the PRO-Psychiatry tool possesses the same properties.

Although the primary aim of the instrument regarded shared decision-making, the results of the current study show that the scale displays adequate psychometric properties for measuring self-perceived health across a heterogeneous psychiatric population. Thus, the instrument is a valid and reliable tool for quality improvement in psychiatric healthcare. The PRO-Psychiatry measure may even impact the development in healthcare quality more than the commonly used clinician-rated scales such as the Hamilton depression severity scale, as the PROM reflects the state of the patient's health as perceived by the patient, resulting in increased patient satisfaction and consequently life quality.[28 29] In addition, the ability to discriminate changes in outcome following an intervention compared with treatment as usual is an important property for applying PROMs in clinical research.[30]

Validation of psychiatry-related PROMs in the current scientific literature is subject to a diverse methodology with little attention given to IRT. A systematic review from 2022, which mapped PROMs for life engagement in mental health, identified 49 distinct and allegedly validated PROMs.[13] However, the reported results of this review regarding assessment of validity were reduced to Cronbach's alpha and correlation analyses with other outcome measures. Cronbach's alpha is a measure of internal reliability generalising to a fixed number of items, and thus, does not compare well across measures comprising different numbers of items.[31] Moreover, subjective assessment such as content validity was not given any attention in the systematic review from 2022 in the dissemination of results. While objective validation exhibits important aspects of the psychometric properties of a PROM, a subjective assessment of validity is at minimum just as important. A systematic review from 2023 investigating content validity of PROMs assessing health-related quality of life in children with cancer, likewise, revealed a gap in the evidence of content validity among PROMs.[32]

Nationwide implementation of PROMs is a process that requires considerable resources.[33] The objective and subjective validation plays an important but small part of this process. The current study finalises the development, implementation and validation of a nationwide PROM of self-perceived health for psychiatry in Denmark.[7 11]

Strength and limitations

The strength of the study is a sizeable and diverse study population of psychiatric patients with a broad spectrum of disorders. Moreover, the patients' 5-year register-based clinical history allows for a more reliable investigation of systematic differences in item response, beyond self-reported socioeconomic and civil status as well as alcohol and drug consumption. However, the methods for investigating DIF consider the different probabilities of picking a certain item on a multi-item scale between particular population groupings after adjusting for the latent trait of that scale. Thus, any general displacements in item discriminations or item difficulties cannot be captured by this method. In addition, the analysis of concurrent validity and responsiveness is limited by the lack of clinical outcomes and other rater-based scales.

Conclusion

The scale developed under the PRO-Psychiatry initiative showed high consistency and little systematic error between the comparison groups. The concurrent correlations and analyses of responsiveness coincided with the prespecified hypotheses. Overall, we deem the Danish PRO-Psychiatry instrument to possess suitable psychometric properties for measuring self-perceived health among a heterogeneous psychiatric population.

Patient and public involvement

Patients were directly involved in the development of the instrument but were not involved in any parts of the assessment of quantitative validity and reliability.

Contributors JBV has conducted the analyses and drafted the manuscript, KK and SK have collected the data, JBV, JM, SPJ and SK have contributed to the conception and design of the study and all authors have edited, read and approved the final manuscript. JBV is the guarantor and responsible for the overall content.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Ethics approval The study was approved by the management of Aalborg University Hospital, the involved departments, and the local Danish Data Protection Authority (ID-number: 2016-135). According to Danish law, approval from a National ethical committee is not required for quality improvement project, which implies patient consent was waived. All participating patients were informed about the project and assured anonymity in the communication of study results.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Data can be shared on reasonable requests and in accordance with Danish legislation.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Jan Brink Valentin http://orcid.org/0000-0002-8205-7179

REFERENCES

1 Basch E, Spertus J, Dudley RA, *et al*. Methods for Developing Patient-Reported Outcome-Based Performance Measures (PRO-PMs). *Value Health* 2015;18:493–504.

2 Coulter A. Measuring what matters to patients. *BMJ* 2017;356:j816. 10.1136/bmj.j816 Available: https://doi.org/10.1136/bmj.j816

3 Kötter T, Schaefer FA, Scherer M, *et al*. Involving patients in quality indicator development - A systematic review. *Patient Prefer Adherence* 2013;7:259–68.

4 Wiering B, de Boer D, Delnoij D. Patient involvement in the development of patient-reported outcome measures: The developers' perspective. *BMC Health Serv Res* 2017;17:635. 10.1186/s12913-017-2582-8 Available: https://doi.org/10.1186/s12913-017-2582-8

5 Baandrup L, Cerqueira C, Haller L, *et al*. The Danish Schizophrenia Registry. *Clin Epidemiol* 2016;8:691–5. 10.2147/CLEP.S99488 Available: https://doi.org/10.2147/CLEP.S99488

6 Videbech P, Deleuran A. The Danish depression database. *Clin Epidemiol* 2016;8:475–8. 10.2147/CLEP.S100298 Available: https://doi.org/10.2147/CLEP.S100298

7 Kristensen S, Mainz J, Baandrup L, *et al*. Conceptualizing patient-reported outcome measures for use within two Danish psychiatric clinical registries: description of an iterative co-creation process between patients and healthcare professionals. *Nord J Psychiatry* 2018;72:409–19.

8 de Bienassis K, Kristensen S, Hewlett E, *et al*. Measuring patient voice matters: setting the scene for patient-reported indicators. *Int J Qual Health Care* 2021;34:ii3–6.

9 de Bienassis K, Kristensen S, Hewlett E, *et al*. Patient-reported indicators in mental health care: towards international standards among members of the OECD. *Int J Qual Health Care* 2021;34:ii7–12.

10 Mainz J, Kristensen S, Roe D. The power of the patient's voice in the modern health care system. *Int J Qual Health Care* 2022;34:ii1–2.

11 Kristensen S, Holmskov J, Pølund K, *et al*. Using patient-reported outcome measures in psychiatric hospital care: an observational study describing an iterative implementation process in Denmark. *Int J Qual Health Care* 2022;34:ii40–8.

12 Kristensen S, Holmskov J, Baandrup L, *et al*. Evaluating the implementation and use of patient-reported outcome measures in a mental health hospital in Denmark: a qualitative study. *Int J Qual Health Care* 2022;34:ii49–58.

13 McIntyre RS, Ismail Z, Watling CP, *et al*. Patient-reported outcome measures for life engagement in mental health: a systematic review. *J Patient Rep Outcomes* 2022;6:62.

14 Zumbo B. *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: National Defense Headquarters, 1999.

15 Christensen KB, Kreiner S, Mesbah M. *Rasch models in health*. 2012.

16 Rosenbaum PR. Criterion-related construct validity. *Psychometrika* 1989;54:625–33.

17 Gagnier JJ, Lai J, Mokkink LB, *et al*. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Qual Life Res* 2021;30:2197–218.

18 Sischka PE, Costa AP, Steffgen G, *et al*. The WHO-5 well-being index – validation based on item response theory and the analysis of measurement invariance across 35 countries. *J Affect Dis Report* 2020;1:100020.

19 Topp CW, Østergaard SD, Søndergaard S, *et al*. The WHO-5 Well-Being Index: A systematic review of the literature. *Psychother Psychosom* 2015;84:167–76.

20 Mainz J, Hess MH, Johnsen SP. The Danish unique personal identifier and the Danish Civil Registration System as a tool for research and quality improvement. *Int J Qual Health Care* 2019;31:717–20. 10.1093/intqhc/mzz008 Available: https://doi.org/10.1093/intqhc/mzz008

21 Nesvåg R, Jönsson EG, Bakken IJ, *et al*. The quality of severe mental disorder diagnoses in a national health registry as compared to research diagnoses based on structured interview. *BMC Psychiatry* 2017;17:93.

22 White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011;30:377–99.

23 Dunn TJ, Baguley T, Brunsden V. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *Br J Psychol* 2014;105:399–412.

24 Piedmont RL. Inter-item Correlations. *Encyclopedia of Quality of Life and Well-Being Research* 2014. 10.1007/978-94-007-0753-5 Available: https://doi.org/10.1007/978-94-007-0753-5_1493

25 Thissen D, Steinberg L. A taxonomy of item response models. *Psychometrika* 1986;51:567–77.

26 Zheng X, Rabe-Hesketh S. Estimating Parameters of Dichotomous and Ordinal Item Response Models with Gllamm. *Stata Journal* 2007;7:313–33.

27 Bech P, Austin SF, Lau ME. Patient reported outcome measures (PROMs): examination of the psychometric properties of two measures for burden of symptoms and quality of life in patients with depression or anxiety. *Nord J Psychiatry* 2018;72:251–8.

28 Bagby RM, Ryder AG, Schuller DR, *et al*. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? *Am J Psychiatry* 2004;161:2163–77. 10.1176/appi.ajp.161.12.2163 Available: https://doi.org/10.1176/appi.ajp.161.12.2163

29 Østergaard SD. Do not blame the SSRIs: blame the Hamilton Depression Rating Scale. *Acta Neuropsychiatr* 2018;30:241–3. 10.1017/neu.2017.6 Available: https://doi.org/10.1017/neu.2017.6

30 Pape LM, Adriaanse MC, Kol J, *et al*. Patient-reported outcomes of lifestyle interventions in patients with severe mental illness: a systematic review and meta-analysis. *BMC Psychiatry* 2022;22:261.

31 Bland JM, Altman DG. Statistics notes: Cronbach's alpha. *BMJ* 1997;314:572.

32 Rothmund M, Meryk A, Rumpold G, *et al*. A critical evaluation of the content validity of patient-reported outcome measures assessing health-related quality of life in children with cancer: A systematic review. *J Patient Rep Outcomes* 2023;7:2. 10.1186/s41687-023-00540-8 Available: https://doi.org/10.1186/s41687-023-00540-8

33 Roe D, Mazor Y, Gelkopf M. Patient-reported outcome measurements (PROMs) and provider assessment in mental health: A systematic review of the context of implementation. *Int J Qual Health Care* 2021;34:ii28–39.