ORIGINAL RESEARCH

Psychometric Evaluation of the Chinese Version of the Mental Health System Responsiveness Questionnaire for Psychiatric Outpatients: Classical Test Theory and Item Response Theory Approaches

Zhenzhu Qian [1–3], Yanyan Yang[3], Jianfeng Tan[4], You Li[4], Jin Zhou[4], Jiewen Huang[1], Sawitri Assanangkornchai [2], Jialong Chen [1]

[1]The First Dongguan Affiliated Hospital, Guangdong Medical University, Dongguan, Guangdong Province, People's Republic of China; [2]Department of Epidemiology, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla, Thailand; [3]Department of Hospital Management, Guangdong Medical University, Dongguan, Guangdong Province, People's Republic of China; [4]School of Humanities and Management, Guangdong Medical University, Dongguan, Guangdong Province, People's Republic of China

Correspondence: Sawitri Assanangkornchai, Department of Epidemiology, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla, 90110, Thailand, Email savitree.a@psu.ac.th; Jialong Chen, The First Dongguan Affiliated Hospital, Guangdong Medical University, Dongguan, Guangdong, 523710, People's Republic of China, Email jialongc@gdmu.edu.cn

Background: The Health System Responsiveness Questionnaire has been an important tool for evaluating the quality of health services from non-clinical perspectives. However, there is a lack of reliable and valid tools to assess the responsiveness of the mental health system in China, despite the increasing attention to the quality of mental health services.

Purpose: This study aimed to adapt a Chinese version of the Mental Health System Responsiveness Questionnaire based on the WHO Health System Responsiveness Questionnaire and assess its psychometric properties among psychiatric outpatients.

Patients and Methods: We adapted the Health System Responsiveness Questionnaire into Chinese and used it on psychiatric outpatients through face-to-face interviews, resulting in 1067 valid responses. We conducted a reliability and validity analysis based on classical test theory (CTT) and applied the graded response model based on item response theory (IRT) to examine the performance of each item.

Results: From the CTT analysis, values of Cronbach's alpha, test-retest reliability, and split-half reliability of the questionnaire were all greater than 0.8. Confirmatory factor analysis exhibited a good fit; convergent and discriminant validity were achieved with average variance extracted > 0.5 and heterotrait-monotrait ratio < 0.9. Results from the IRT analysis showed that all items demonstrated acceptable discrimination (discrimination parameters $\geq$ 0.65) and appropriate difficulty parameters (ranging from −4 to 4). Additionally, all items provided sufficient information, except for two items in the prompt attention domain.

Conclusion: Our adapted Mental Health System Responsiveness Questionnaire possessed strong psychometric properties and could serve as an effective tool for assessing the responsiveness of mental health systems in China. However, improving two items in the domain of prompt attention may contribute to more optimal psychometric properties of the questionnaire.

Keywords: health system responsiveness, questionnaire, psychiatric patients, psychometric properties

Introduction

According to the World Mental Health Report in 2022, approximately one in every eight individuals globally suffers from a mental disorder.[1] A national survey in China revealed that the lifetime prevalence of any mental disorder was 16.6%.[2] As the enormous demand for mental health services and the insufficiency of their supply have become well-recognized issues,[1,3,4] the quality of mental health services also warrants serious attention.[5–7]

Health system responsiveness, introduced by the WHO in its 2000 World Health Report, refers to the ability of health systems to address patients' non-clinical needs and meet their universally legitimate expectations.[8] The Health System Responsiveness Questionnaire (HSRQ) is recommended as a key tool for evaluating the quality of health services from these perspectives.[8–10] Along with health outcomes and equity in health financing, health system responsiveness serves as one of the three main indicators for evaluating health system performance.[9]

The Health Systems Responsiveness Questionnaire was developed as part of the World Health Survey to evaluate the responsiveness of health systems worldwide,[8] Its main goal was to support surveys in various countries. However, due to differences in sociocultural factors, language subtleties, healthcare practices, and target populations, it is essential to conduct cultural adaptation before applying the questionnaire in different countries and among diverse populations.

Most studies to date on health system responsiveness have used the HSRQ to survey general patients,[11–13] with only a few studies focusing specifically on evaluating mental health system responsiveness among psychiatric patients.[14–16] Accurate and comparable data on evaluations of mental health system responsiveness are essential for monitoring and improving the performance of the mental health system. However, there is a notable lack of such data, as well as a shortage of tools with high reliability and validity that are tailored to assess mental health system responsiveness in China, although there have been attempts to do this.[17] Only two studies have been found in China to evaluate the HSRQ among psychiatric patients. One study focused on applying the questionnaire to community management services for patients with severe mental disorders.[18] The other study collected data from both inpatient and outpatient psychiatric patients to assess the psychometric properties of the questionnaire. However, the sample sizes in the two studies were relatively small, and community management services for psychiatric patients differ significantly from hospital outpatient services. Additionally, the study conducted among inpatients and outpatients did not assess test-retest reliability, and certain items contributed minimally to the questionnaire's internal consistency.

In the few existing studies, the validation of the HSRQ has predominantly utilized methods based on classical test theory (CTT). While CTT remains widely used and allows for a macro-level assessment of a scale's reliability and validity, it has certain limitations. These include sensitivity to sample size, simplistic assumptions about measurement error, difficulty in providing precise individual-level measurements, and inability to assess item functioning differences.[19–21] Item response theory (IRT), developed to overcome the shortcomings of CTT, is gaining popularity for offering several advantages, such as precise individual item analysis, sample independence of item parameters, and independence of test results from specific items.[20,22,23] By utilizing the methods of both CTT and IRT, we can combine the generalizability and simplicity of CTT with the item-level precision and analytical depth provided by IRT to enable a more comprehensive and detailed evaluation of the questionnaire.[21]

Therefore, in this study, we aimed to translate and validate the HSRQ to develop a Chinese version of the Mental Health System Responsiveness Questionnaire (MHSRQ) using both CTT and IRT methods, providing a tool with strong psychometric properties for assessing mental health system responsiveness.

Materials and Methods

Assessment Questionnaire

The MHSRQ assesses health services for outpatients from 7 domains: prompt attention, dignity, communication, autonomy, confidentiality, choice of healthcare, and quality of basic amenities. It consists of 12 items, each rated on a 5-point Likert scale ranging from "1-very bad" to "5-very good", with higher scores indicating better responsiveness. The composition of each domain in the MHSRQ is shown in Supplementary File 1.

Study Phase

We conducted the study in two phases. The first phase involved translating and culturally adapting the MHSRQ, which followed the established guidelines for the process of cross-cultural adaptation,[24] while the second phase involved assessing its psychometric properties based on CTT and IRT.

Phase I: Cross-Cultural Adaptation

Forward Translation

Two native Chinese speakers fluent in Chinese and English worked independently: one a professional in health services management and the other an experienced translator with expertise in English but unfamiliar with health system responsiveness concepts. The two individuals each provided a translated version, called A1 and A2.

Synthesis of the Translations

The two translators and the primary researcher synthesized the two translated versions. All three discussed the differences between the two versions until they reached a consensus, resulting in the combined version, called A12.

Back-Translation

The back-translation was independently completed by two native English speakers who are proficient in Chinese but completely unaware of the HSRQ. They translated the Chinese A12 version back into English, resulting in two back-translated versions, called B1 and B2. This process ensured that the original meaning was preserved and allowed for an assessment of equivalence between the original and adapted versions.

Expert Review

The expert panel consisted of the primary researcher, a health management professional, a psychiatrist, an English language professional, and the translators involved in the process (both forward and back translators). They meticulously reviewed the original HSRQ along with all translated versions and engaged in repeated discussions and revisions to achieve semantic, idiomatic, experiential, and conceptual equivalence. Particular attention was paid to theoretical constructs, assessing whether the definitions and domains of health system responsiveness were universally applicable (etic) or required culturally specific adaptations (emic). This collaborative process continued until all discrepancies were resolved, ultimately resulting in the prefinal Chinese version of the MHSRQ.

Pilot Study

Subsequently, we conducted face-to-face interviews with 35 psychiatric outpatients using the prefinal MHSRQ to gather their feedback. We encouraged them to report any wording, layout, or instructions issues. The expert panel then evaluated these results and revised the MHSRQ.

Phase II: Psychometric Evaluation

Study Setting and Participants

We conducted a cross-sectional survey from May to September 2023 at three psychiatric hospitals in Dongguan, Foshan, and Zhongshan cities in Guangdong Province, China. We adopted a consecutive sampling method to select outpatients from these hospitals as study subjects. The inclusion criteria were: (1) age between 18 and 65 years, (2) having visited hospitals at least once for mental health problems in the past 12 months, (3) being in stable medical condition, and (4) having sufficient cognitive ability to engage in the survey. Patients with organic mental disorders, such as dementia or delirium, which might impair cognitive function, were excluded. To ensure that participants were in a good mental state, mental health specialists conducted an initial cognitive screening. The Mini-Mental State Examination (MMSE) was used exclusively for cases where cognitive status was unclear after this screening. Initially, we were unaware that the scale was copyright-protected; the study team used an unauthorized version of the Chinese MMSE without permission; however, this has now been rectified with PAR. The MMSE is a copyrighted instrument and may not be used or reproduced in whole or in part, in any form or language, or by any means without written permission of PAR (www.parinc.com).

Sample Size

For CTT analysis, it is generally recommended that the sample size be at least 5 to 10 times the number of items in the scale.[25] Given that our questionnaire consists of 12 items, a sample size of 60 to 120 participants would be necessary under CTT guidelines.

Compared to CTT, IRT models typically require larger sample sizes, ranging from 500 to 1000 participants, due to the need to estimate item parameters (eg, difficulty and discrimination) and underlying traits (eg, ability level) of the

subjects.[26] Therefore, a sample size of 1000 would be adequate to ensure a robust and reliable assessment of the questionnaire's measured attributes. In addition, for assessing test-retest reliability, a minimum sample size of 50 to 100 participants is often recommended to achieve reliable results.[27]

Data Collection

All stages of the study were implemented in accordance with the Declaration of Helsinki. After obtaining ethical approval from the Human Research Ethics Committee, Faculty of Medicine, Prince of Songkla University and the Ethics Committee of Dongguan Seventh People's Hospital, we recruited and systematically trained 11 investigators who were undergraduate or graduate students majoring in psychology and health management. We conducted face-to-face interviews with eligible patients using a paper-based MHSRQ. Given that psychiatric patients were a vulnerable population, and to enhance privacy protection and allow for more objective and impartial responses, the ethics committees waived the requirement for written informed consent. Therefore, we assured participants of data anonymity and obtained verbal consent after thoroughly explaining the study procedures. Additionally, to assess the test-retest reliability of the questionnaire, we invited about 10% of the total survey respondents to participate in the survey again when they came to the hospital to retrieve their medication or visit the hospital within 2–4 weeks of completing the initial survey.

Statistical Analysis

We used EpiData 3.1 for double data entry, verification, and documentation. For data analysis, we initially used descriptive statistics to present the characteristics of the participants. Subsequently, we assessed the MHSRQ using both CTT and IRT approaches. All statistical analyses were performed in R version 4.3.2.[28]

Analysis Based on CTT

Reliability Analysis

We measured the reliability of the MHSRQ using internal consistency, test-retest reliability, and split-half reliability. The evaluation indicators and criteria are presented in Table 1.

Validity Analysis

We carried out validity analysis from three aspects: construct validity by performing confirmatory factor analysis, convergent validity using average variance extracted (AVE) in structural equation modeling, and discriminant validity

Table 1 Evaluation Indicators and Criteria for the Questionnaire

Theory	Evaluation Aspects	Evaluation Indicators	Evaluation Criterion
CTT	Reliability	Internal consistency	• Cronbach's alpha: > 0.7 is acceptable, > 0.8 is good;[27,29] • McDonald's ω coefficient: > 0.7 is acceptable, > 0.8 is good;[27,29] • Item-total correlation: > 0.4 is good;[30] • item-rest correlation: > 0.4 is good.[20]
		Test-retest reliability	Pearson correlation > 0.7 is good.[29]
		Split-half reliability	Spearman-Brown coefficient > 0.7 is good.[29]
	Validity	Convergent validity	AVE > 0.5 is good.[31]
		Discriminant validity	HTMT: < 0.9 is acceptable,< 0.85 is good.[32]
		Construct validity	CFI ≥ 0.90, TLI ≥ 0.90, RMSEA ≤ 0.08.[33]

(*Continued*)

Table 1 (Continued).

Theory	Evaluation Aspects	Evaluation Indicators	Evaluation Criterion
IRT	Item response	Item discrimination	Discrimination parameter: values within [0.65–1.34] are considered acceptable, ≥ 1.35 are good.[22]
		Item difficulty	Difficulty parameters should fall between [−4,4], and with a monotonically increasing trend (b1<b2<b3<b4).[34]
		Item information	Average information amount > 0.42 (calculated by dividing 5 by the total number of items, 12).[34]

Abbreviations: CTT, classic test theory; IRT, item response theory; AVE, average variance extracted; HTMT, heterotrait-monotrait ratio of correlations; CFI, Comparative Fit Index; TLI, Tucker-Lewis Index; RMSEA, Root Mean Square Error of Approximation.

with the heterotrait-monotrait ratio of correlations (HTMT). We chose HTMT because it is considered more robust than the other methods.[31,32]

Analysis Based on IRT

Unidimensionality Test

An essential prerequisite for applying IRT is assessing unidimensionality. Thus, we first tested this to assess the suitability of the unidimensional IRT analysis by performing a principal component analysis. We followed Hambleton's criterion, which states that if the ratio of the first to the second eigenvalue exceeds 3 and the first factor explains over 50% of the variance, the data supports unidimensionality.[35,36]

Considering that the MHSRQ utilizes a five-point Likert scoring method, we employed Samejima's Graded Response Model,[37] which is recommended for data with ordered polytomous categories, to calculate item discrimination (a), item difficulty (b), and the item average information amount.

Item Discrimination

The values of the discrimination parameters represent a measure of how well an item differentiates respondents with different levels of the latent trait (in this study, referring to perceived health system responsiveness). Larger values indicate better item discrimination.

Item Difficulty

Item difficulty parameters refer to the points on the latent trait scale where a respondent is equally likely to choose between two adjacent categories. Since our questionnaire had five categories for each item, there were four difficulty parameters b1, b2, b3, and b4. The difficulty parameters for well-performing items should typically range from −4 to 4 and follow a monotonically increasing pattern.[34]

Category Probability Curves

To illustrate each item's performance, we plotted probability curves for each item. For a good item, the curves should be well-separated, with each category having a distinct range where its selection probability is the highest.

Item Information Amount and Item Information Curves

In IRT, a test is considered to provide adequate measurement if the total information amount of the questionnaire exceeds 5.[34] Therefore, for our 12-item questionnaire, the average amount of information provided by a well-functioning item should exceed 0.42 (5/12). Additionally, we plotted item information curves to visualize the amount of information each item contributes at different levels of the latent trait. Larger areas under the curve indicate greater information and higher measurement accuracy.

Results

Phase I: Cross-Cultural Adaptation

There was a high degree of consistency in the translations across the various versions. The adapted MHSRQ preserved the structure and core content of the original version. However, considering the linguistic and cultural differences, we made adjustments to some item descriptions to enhance their relevance and alignment with theoretical constructs. For example, in English, the term "healthcare providers" is very broad and refers to a wide range of professionals and organizations that deliver medical services. In contrast, such a general term is rarely used in Chinese, where specific terms like "medical institutions" and "medical staff" are used to clearly differentiate between these entities. As a result, in the original HSRQ, the phrase "healthcare provider" in "traveling time to the [healthcare provider]" (item r1) was modified to "mental health center" to explicitly refer to a mental health facility. Similarly, the terms "healthcare providers" in "talk privately to [healthcare providers]" (item r9) and "choose your [healthcare provider]" (item r11) were translated as "medical staff", specifically referring to individuals. Additionally, considering that patient involvement in discussions occurs more frequently than direct decision-making, and patients in China have a strong awareness of their involvement in healthcare activities and seek not only to participate in decision-making but also in discussions, we revised item r8 from "being involved in making decisions about your health care or treatment" to "being involved in discussions and decision-making regarding your health care or treatment". The revised MHSRQ is shown in Supplementary File 1.

Phase II: Psychometric Evaluation

Participants Characteristics

Valid questionnaires were obtained from 1067 participants (Supplementary File 2), with a mean (standard deviation) age of 34 (13.8) years; 62.4% were female, 80.5% reported no religious affiliation, and 42.8% had achieved a bachelor's degree or higher. Additionally, 55.9% were unmarried, 51.3% were unemployed, and 45.5% had been diagnosed with mood disorders. Of the 1067 respondents, 151 participated in the retest.

Results of the CTT Approach

Reliability of the Questionnaire

Both the Cronbach's alpha coefficient and McDonald's ω coefficient for the overall questionnaire exceeded 0.9. Additionally, the item-total correlation and item-rest correlation were both above 0.4, indicating excellent internal consistency. The test-retest reliability and split-half reliability of the overall questionnaire were 0.898 and 0.941, respectively (Table 2).

Validity of the Questionnaire

The results of confirmatory factor analysis confirmed construct validity with a good model fit (χ^2/df = 4.68, CFI = 0.982, TLI=0.966, RMSEA = 0.058). Furthermore, the AVE values for all domains exceeded the threshold of 0.5, and the HTMT values were below the recommended threshold of 0.9 (Table 3), indicating reasonable discriminant and convergent validity.

Results of the IRT Approach

Unidimensionality Test

Principal component analysis revealed an eigenvalue ratio of 5.23 (6.18/1.18), with the first factor explaining 51.5% of the model's variance, supporting the assumption of unidimensionality.

Item Discrimination and Difficulty

Table 4 presents the discrimination and difficulty parameters of each item. From the perspective of discrimination parameters, aside from items r1 and r2, whose difficulty parameters fell within the acceptable range of [0.65–1.34], all other items had difficulty parameters greater than 1.35, which are considered excellent. The difficulty parameters (b) for the 12 items ranged from −3.97 to 2.27, which fell within the acceptable range of [−4, 4], and exhibited a monotonically increasing trend ($b_1 < b_2 < b_3 < b_4$). These results indicate that the response options for the items were ordered, with higher response categories becoming more likely to be endorsed as the latent trait increased.

Table 2 Reliability of the Questionnaire

Domains	Items	Mean	S.D.	Item-Total Correlation	Item-Rest Correlation	Cronbach's Alpha[a]	McDonald's ωCoefficient	Test-Retest Reliability	Split-Half Reliability
PA	r1	3.390	0.930	0.403	0.404	0.911	0.916	0.851	0.941
	r2	3.460	0.918	0.472	0.468	0.909	0.914		
DI	r3	4.170	0.733	0.696	0.692	0.897	0.904	0.814	
	r4	4.210	0.730	0.690	0.690	0.897	0.904		
CM	r5	4.080	0.811	0.744	0.742	0.894	0.901	0.792	
	r6	3.980	0.842	0.748	0.745	0.894	0.901		
AU	r7	3.870	0.870	0.726	0.725	0.895	0.903	0.799	
	r8	3.900	0.850	0.705	0.707	0.896	0.903		
CF	r9	4.250	0.736	0.679	0.679	0.898	0.904	0.794	
	r10	4.280	0.722	0.671	0.674	0.898	0.904		
CH	r11	4.190	0.740	0.675	0.669	0.898	0.905	0.800	
QU	r12	4.090	0.776	0.537	0.530	0.904	0.911	0.788	
Overall		3.988	0.568	-	-	0.907	0.913	0.898	

Notes: [a]The "Cronbach's Alpha" column displays internal consistency estimates under two conditions: the rows r1 through r12 correspond to the Cronbach's alpha of the scale if each respective item is deleted; the last row, labeled "Overall", shows a value of 0.907, which represents the overall Cronbach's alpha for the questionnaire when all items remain.
Abbreviations: PA, prompt attention; DI, Dignity; CM, communication; AU, autonomy; CF, confidentiality; CH, choice of healthcare; QU, quality of basic amenities.

Table 3 Convergent Validity and Discriminant Validity

Domains	AVE	Heterotrait-Monotrait Ratio of Correlations				
		PA	DI	CM	AU	CF
PA	0.504					
DI	0.618	0.628				
CM	0.703	0.566	0.894			
AU	0.743	0.519	0.742	0.881		
CF	0.782	0.406	0.746	0.668	0.645	

Notes: The domains "choice of healthcare" and "quality of basic amenities" each contained only a single item, making it impossible to calculate the AVE and HTMT ratio.
Abbreviations: PA, prompt attention; DI, Dignity; CM, communication; AU, autonomy; CF, confidentiality.

Table 4 Discrimination, Difficulty Parameters, and Information Amount of the Items

Items	Discrimination (a)	Difficulty (Threshold)				Average Information Amount[a]
		b1	b2	b3	b4	
r1	0.902	−3.974	−2.562	0.453	2.268	**0.228**[b]
r2	1.155	−3.557	−2.235	0.109	1.895	**0.371**[b]
r3	2.629	−2.867	−2.347	−1.208	0.411	1.203
r4	2.698	−3.109	−2.395	−1.209	0.292	1.226

(*Continued*)

Table 4 (Continued).

Items	Discrimination (a)	Difficulty (Threshold)				Average Information Amount[a]
		b1	b2	b3	b4	
r5	3.350	−2.497	−1.922	−0.904	0.420	1.922
r6	3.334	−2.473	−1.772	−0.767	0.549	1.888
r7	2.814	−2.648	−1.827	−0.580	0.694	1.521
r8	2.696	−2.593	−1.958	−0.662	0.710	1.420
r9	2.708	−3.119	−2.360	−1.255	0.201	1.258
r10	2.771	−3.083	−2.396	−1.298	0.149	1.253
r11	2.581	−2.896	−2.353	−1.203	0.344	1.174
r12	1.680	−3.688	−2.762	−1.211	0.606	0.618

Notes: [a]The overall information amount for the questionnaire is 14.082 (the sum of the average information of all items). [b]Bold numbers indicate values that do not meet the criterion.

Category Probability Curves

Figure 1 displays the category probability curves for each of the 12 items (r1 to r12). Overall, the curves suggested that most items on the questionnaire exhibit good discrimination, with clear peaks at different θ levels for different response categories, indicating that the items were effective at distinguishing between respondents with varying levels of the
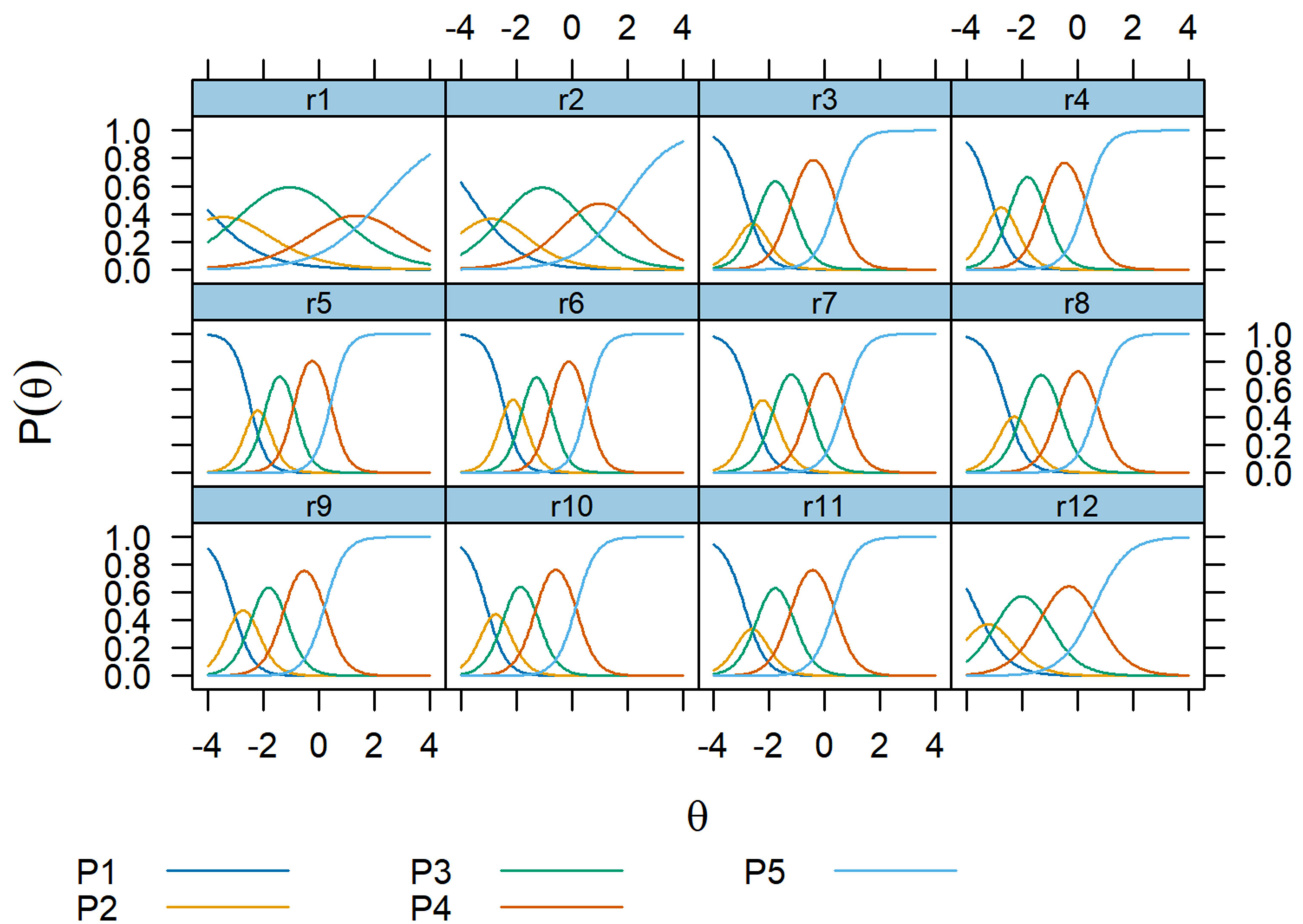
Figure 1 Category probability curves (CPCs) for all items.
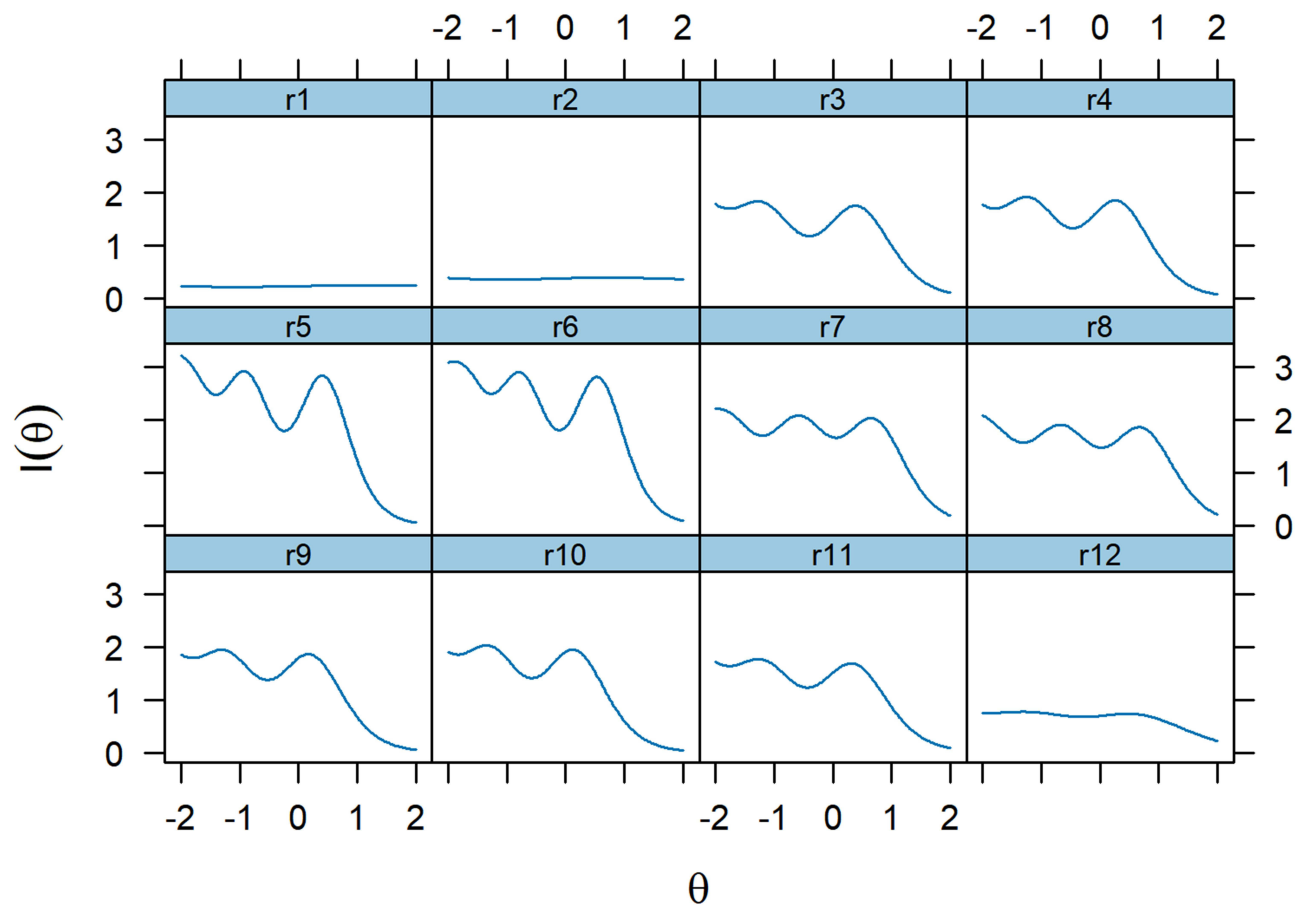
Figure 2 Item information curves (IICs) for all items.

underlying trait. However, items r1 and r2 showed less distinct separation between response options, which may suggest overlap or redundancy in these categories.

Item Information Amount and Item Information Curves

As shown in Table 4, the total amount of information provided by the 12 items was 14.08, surpassing the required total amount of 5. Of the 12 items, 9 provided adequate average information amount, while Items r1 and r2 fell below the required value of 0.42. The item information curves (Figure 2) also confirm these findings, showing that the area under the curve for items r1 and r2 was smaller compared to the other items.

Discussion

In this study, we utilized both CTT and IRT methods to evaluate the psychometric properties of the questionnaire, whereas previous evaluations of the HSRQ have primarily relied on CTT. CTT assesses the overall reliability and validity of a questionnaire from a macroscopic point of view, while IRT focuses on the performance of individual items in a microscopic way. Together, the two methods complemented each other, enhancing the reliability of our evaluation results.

According to the results of CTT, our MHSRQ showed high reliability and validity with excellent internal consistency, good test-retest reliability, excellent split-half reliability, desirable construct validity, good convergent validity, and acceptable discrimination validity. Compared to the 2017 Farsi version,[38] and the 2022 Chinese version of the HSRQ,[17] our version had significantly higher internal consistency, and we achieved good test-retest reliability, which was not assessed in the other studies. We believe that test-retest reliability is crucial, even considering the unique

characteristics of psychiatric patients, and our results demonstrated that the test-retest reliability of the MHSRQ was strong among psychiatric outpatients.

According to the IRT results, all items performed well in terms of item differentiation and difficulty parameters. However, the average information for items r1 and r2 was relatively low compared to the other items, indicating that the measurement precision of these two items may be insufficient. Previous studies have also reported the poor performance of these two items,[38] which can mainly be attributed to the following reasons. First, item r1 assesses "travel time to a mental health center" and item r2 evaluates "waiting time before being attended to". Both these items are more appropriately measured as continuous variables, whereas the questionnaire uses categorical options, which may not accurately capture the real situation. Second, these two items fall, to some extent, outside the scope of services provided by healthcare providers (medical staff and medical facilities). For example, travel time to the hospital is primarily determined by geographical accessibility. Additionally, some studies have suggested classifying item r1 as part of a new domain, namely "Access to healthcare".[14]

Based on the combined results of CTT and IRT, although the average information amounts of items r1 and r2 was relatively low, the overall questionnaire passed the comprehensive assessment of internal consistency, test-retest reliability, split-half reliability, construct validity, convergent validity, discriminant validity, item difficulty, and item discrimination. Moreover, the results of the questionnaire's internal consistency (Table 2) showed that the item-total correlations for items r1 and r2 were lower than those of the other items, but remained above 0.4, which could still be considered good indicators. Therefore, we believe that these two items are also acceptable, and do not affect the excellent performance of the whole questionnaire.

Limitations

This study has certain limitations. First, we did not assess the criterion validity of the MHSRQ due to a lack of a gold standard suitable for measuring mental health system performance. Second, the predominance of patients with mood disorders (45.5%) and the limited representation of other psychiatric diagnoses may have influenced the measurement results, potentially affecting the generalizability of the findings to broader psychiatric populations.

Conclusions

In conclusion, the overall reliability and validity of the MHSRQ were quite high. Each item demonstrated a reasonable degree of discrimination, ordered and appropriate difficulty parameters, and most items provided sufficient average information amount. These results indicate that our adapted MHSRQ possessed strong psychometric properties and could serve as an effective tool for assessing the responsiveness of mental health systems in China. However, improving the two items in the domain of prompt attention may contribute to more optimal psychometric properties of the MHSRQ.

Data Sharing Statement

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Ethical Approval and Informed Consent

This study was approved by the Human Research Ethics Committee, Faculty of Medicine, Prince of Songkla University (REC. 65-366-19-9) and the Ethics Committee of Dongguan Seventh People's Hospital (Approval No. 2023-013) and conducted in accordance with the Declaration of Helsinki. Verbal consent was obtained from all participants after they were provided detailed information about the study's purpose and procedures. Written consent was waived for two reasons: First, as psychiatric patients are a vulnerable population, requiring a signature could cause undue concern. Second, the questionnaire assesses health services, waiving written consent enhanced privacy protection and allowed participants to evaluate the health system responsiveness more objectively.

Acknowledgments

at the Mental Health Centers in Dongguan, Shunde Foshan, and Zhongshan City, for their invaluable support throughout the data collection process. Finally, we thank Professor Edward McNeil for his insightful advice and meticulous proofreading of the manuscript.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

Disclosure

The authors report no conflicts of interest in this work.

References

1. World Health Organization. *World Mental Health Report: Transforming Mental Health for All*. 1st ed. Geneva, Switzerland: World Health Organization; 2022.
2. Huang Y, Wang Y, Wang H, et al. Prevalence of mental disorders in China: a cross-sectional epidemiological study. *Lancet Psychiatry.* 2019;6 (3):211–224. doi:10.1016/S2215-0366(18)30511-X
3. Munakampe MN. Strengthening mental health systems in Zambia. *Int J Ment Health Syst.* 2020;14:28. doi:10.1186/s13033-020-00360-z
4. Moitra M, Owens S, Hailemariam M, et al. Global mental health: where we are and where we are going. *Curr Psychiatry Rep.* 2023;25(7):301–311. doi:10.1007/s11920-023-01426-8
5. Kilbourne AM, Beck K, Spaeth-Rublee B, et al. Measuring and improving the quality of mental health care: a global perspective. *World Psychiatry off J World Psychiatr Assoc WPA.* 2018;17(1):30–38. doi:10.1002/wps.20482
6. Lundqvist L-O, Ahlström G, Wilde-larsson B, Schröder A. The patient's view of quality in psychiatric outpatient care. *J Psychiatr Ment Health Nurs.* 2012;19(7):629–637. doi:10.1111/j.1365-2850.2012.01899.x
7. D'Avanzo B, Barbato A, Monzio Compagnoni M, et al. The quality of mental health care for people with bipolar disorders in the Italian mental health system: the QUADIM project. *BMC Psychiatry.* 2023;23(1):424. doi:10.1186/s12888-023-04921-7
8. Darby C, Valentine N, Jl C, de Silva A. World health organization (WHO): strategy on measuring responsiveness. *GPE Discuss Pap Ser.* 2000;23. Available from: https://www.researchgate.net/publication/268295796/. Accessed October 20, 2024.
9. Valentine N, de Silva A, Kawabata K, Darby C, Murray C, Evans D. Health system responsiveness: concepts, domains and operationalization. *Health Syst Perform Debates Methods Empiricism.* 2003. Available from: https://www.researchgate.net/publication/242475744/. Accessed October 20, 2024.
10. Murray CJL, Frenk J. A framework for assessing the performance of health systems. *Bull World Health Organ.* 2000;78(6):717–731.
11. Chao J, Lu B, Zhang H, Zhu L, Jin H, Liu P. Healthcare system responsiveness in Jiangsu Province, China. *BMC Health Serv Res.* 2017;17(1):31. doi:10.1186/s12913-017-1980-2
12. Amani PJ, Tungu M, Hurtig AK, Kiwara AD, Frumence G, San Sebastián M. Responsiveness of health care services towards the elderly in Tanzania: does health insurance make a difference? A cross-sectional study. *Int J Equity Health.* 2020;19(1):179. doi:10.1186/s12939-020-01270-9
13. Negash WD, Tsehay CT, Yazachew L, Asmamaw DB, Desta DZ, Atnafu A. Health system responsiveness and associated factors among outpatients in primary health care facilities in Ethiopia. *BMC Health Serv Res.* 2022;22(1):249. doi:10.1186/s12913-022-07651-w
14. Forouzan S, Padyab M, Rafiey H, Ghazinour M, Dejman M, San Sebastian M. Measuring the mental health-care system responsiveness: results of an outpatient survey in Tehran. *Front Public Health.* 2015;3:285. doi:10.3389/fpubh.2015.00285
15. Bramesfeld A, Klippel U, Seidel G, Schwartz FW, Dierks ML. How do patients expect the mental health service system to act? Testing the WHO responsiveness concept for its appropriateness in mental health care. *Soc Sci Med 1982.* 2007;65(5):880–889. doi:10.1016/j.socscimed.2007.03.056
16. Guerriero C, Gillan MGC, Cairns J, Wallis MG, Gilbert FJ. Testing the WHO responsiveness concept in the Iranian mental healthcare system: a qualitative study of service users. *BMC Health Serv Res.* 2011;11:11. doi:10.1186/1472-6963-11-325
17. Zhou W, Xiao S, Feng C, et al. Measuring the quality of mental health services from the patient perspective in China: psychometric evaluation of the Chinese version of the World Health Organization responsiveness performance questionnaire. *Glob Health Action.* 2022;15(1):2035503. doi:10.1080/16549716.2022.2035503
18. Wei Z, Yu Y, Xiao X, Xiao S. Psychometric tests of WHO health system responsiveness questionnaire in evaluation on community management and services for psychosis. *Chin Ment Health J.* 2021;35(12):1007–1012.
19. Crocker, Algina. *Introduction To Classical And Modern Test Theory*. Wadsworth Publishing; 2008. Available from: http://archive.org/details/introductiontoclassicalandmoderntesttheory. Accessed September 23, 2024.
20. DeVellis RF. *Scale Development: Theory and Applications*. 4th ed. Sage Publications; 2016.
21. Wyse ARJDEAYALA. (2009) The theory and practice of item response theory. *Psychometrika.* 2010;75:778–779. doi:10.1007/s11336-010-9179-z
22. Bichi AA, Talib R. Item response theory: an introduction to latent trait models to test and item development. *Int J Eval Res Educ IJERE.* 2018;7 (2):142. doi:10.11591/ijere.v7i2.12900

739

23. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement | quality of Life Research. Available from: https://link.springer.com/article/10.1007/s11136-007-9198-0. Accessed October 26, 2024.

24. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*. 2000;25 (24):3186–3191. doi:10.1097/00007632-200012150-00014

25. Hair JF, L.D.S. Gabriel M, da Silva D, Braga Junior S. Development and validation of attitudes measurement scales: fundamental and practical aspects. *RAUSP Manag J*. 2019;54(4):490–507. doi:10.1108/RAUSP-05-2019-0098

26. Kutscher T, Eid M, Crayen C. Sample size requirements for applying mixed polytomous item response models: results of a monte carlo simulation study. *Front Psychol*. 2019;10:2494. doi:10.3389/fpsyg.2019.02494

27. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15 (2):155–163. doi:10.1016/j.jcm.2016.02.012

28. R Core Team. R: the R Project for Statistical Computing. Available from: https://www.r-project.org/. Accessed October 20, 2024.

29. Wang X, Ren J, Wang H. Psychometric characteristics of the Chinese version of the Pregnancy Exercise Attitudes Scale (C-PEAS). *BMC Pregnancy Childbirth*. 2024;24(1):679. doi:10.1186/s12884-024-06817-0

30. Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use (5th edition). *Aust N Z J Public Health*. 2016;40(3):294–295. doi:10.1111/1753-6405.12484

31. Henseler J, Ringle CM, Sarstedt M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J Acad Marking Sci*. 2015;43(1):115–135. doi:10.1007/s11747-014-0403-8

32. Uzir MUH, Al Halbusi H, Lim R, et al. Applied Artificial Intelligence and user satisfaction: smartwatch usage for healthcare in Bangladesh during COVID-19. *Technol Soc*. 2021;67:101780. doi:10.1016/j.techsoc.2021.101780

33. 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*. 2003;8(2):23–74.

34. Shi H, Ren Y, Xian J, Ding H, Liu Y, Wan C. Item analysis on the quality of life scale for anxiety disorders QLICD-AD(V2.0) based on classical test theory and item response theory. *Ann Gen Psychiatry*. 2024;23(1):19. doi:10.1186/s12991-024-00504-2

35. Culpepper SA. The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Appl Psychol Meas*. 2013;37(3):201–225. doi:10.1177/0146621612470210

36. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of Item Response. 1st*. Thousand Oaks, CA: Sage Publications, Inc; 1991. Vols. x, 174: x,174.

37. Samejima F. Estimation of latent ability using a response pattern of graded scores. *ETS Res Bull Ser*. 1968;1968(1):i–169. doi:10.1002/j.2333-8504.1968.tb00153.x

38. Forouzan AS, Rafiey H, Padyab M, Ghazinour M, Dejman M, Sebastian MS. Reliability and validity of a mental health system responsiveness questionnaire in Iran. *Glob Health Action*. 2014;7:24748. doi:10.3402/gha.v7.24748