Chinese General Practice Journal 2 (2025) 100091

Contents lists available at ScienceDirect

Chinese General Practice Journal

journal homepage: https://www.keaipublishing.com/en/journals/chinese-general-practice-journal/

Design and development of scales in primary care: Practical steps and statistical methods ☆,☆☆

Wang Fei [a], Tang Jingqi [b], Sun Xiaonan [c], Sun Xinying [d], Li Jun [e], Meng Xingxing [f,*], Wu Yibo [d,*]

[a] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing 100875, China
[b] School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China
[c] School of Public Health, Harbin Medical University, Harbin 150081, China
[d] School of Public Health, Peking University, Beijing 100191, China
[e] Department of General Practice, Peking University Third Hospital, Beijing 100191, China
[f] School of Philosophy, Anhui University, Hefei 230039, China

ARTICLE INFO

Keywords:
Primary health care
Scale development;Research design;Factor analysis;Rasch model

ABSTRACT

This article outlines the statistical methods and practical steps involved in designing and developing valid and reliable questionnaires in primary care. Based on literature review on questionnaire development and scale design, we proposed a standardized protocol for scale development in primary care. This process encompasses key practical steps and statistical methods, illustrated by cases from prior research.

The recommended seven-step approach includes: (1) defining the construction of measurement; (2) generating the pool of items; (3) selecting the scoring system and response format; (4) pre-testing (assessing content validity and face validity, etc.); (5)eliminating items by item analysis; (6)evaluating the scale initially, including evaluating the reliability and validity of the scale, and factor analysis or Rasch analysis; (7)re-evaluating the scale to re-examine the nature of the scale, including retesting reliability and constructing validity. In conclusion, scale development studies should adhere to standardized procedures, and the integrated use of Rasch model and factor analysis can make the measurements more objective.

In 1977, at the 30th World Health Assembly, the World Health Organization (WHO) proposed "Health for All," identifying primary health care (PHC) as the key to achieving this vision.[1] As the main providers of PHC services, general practitioners (GPs) face the dual pressure of practice and research. While GPs rarely use scales for diagnosis in daily clinical practice, these tools become a popular research paradigm due to their convenience and efficiency. However, scale development involves complex and time-consuming procedures. These rigorous steps can be daunting, leading researchers to frequently overlook critical aspects of the process.[2] Consequently, methodological flaws are common in current scale development research. For instance, a systematic review of nutritional assessment tools for athletes found that approximately 70% of studies used instruments with unknown validity and reliability, and 67% employed unvalidated tools.[3] Similarly, the autism screening scale developed by Chen et al.[4] retained items with poor psychometric properties in its final version. The use of unvalidated or psychometrically poor scales can severely compromise the accuracy of findings. Therefore, it is essential to establish a standardized protocol to guide scale development in primary care. Furthermore, most scale development in primary care currently relies on Classical Test Theory (CTT). While CTT is fundamental for evaluating psychometric properties, it has limitations of measurement error, which can compromise measurement objectivity. The Rasch model offers a robust solution to this issue. By establishing objective measurement standards comparable to those in the natural sciences, the Rasch model ensures more reliable and objective data in social science research.[5] In this context, this study summarizes common questionnaire development and scale design methods in primary care, integrating both CTT and Rasch perspectives. By elucidating specific steps and statistical methods, we aim to assist researchers in conducting more rigorous scale development.

☆ Peer review under the responsibility of Editorial Office of Chinese General Practice Journal.

☆☆ The Chinese version of this paper was published in Chinese General Practice on [2024-03-06] (DOI: 10.12114/j.issn.1007-9572.2022.0819). The current English paper is a compliant secondary publication by Chinese General Practice Journal after obtaining copyright permission from both the authors and Chinese General Practice.

* Corresponding authors.
E-mail addresses: 614997175@qq.com (M. Xingxing), bjmuwuyibo@outlook.com (W. Yibo).

https://doi.org/10.1016/j.cgpj.2025.100091
Received 5 November 2025; Received in revised form 1 December 2025; Accepted 20 December 2025

Practical steps and statistical methods

Defining the construction of measurement

The foundational step in developing a scale for primary care is to establish an accurate and comprehensive definition of the target construct. This definition must delineate both the conceptual boundaries and the internal structure of the construct. Operational definitions are typically derived from authoritative guidelines, expert consensus, clinical experience, or literature and empirical data. To illustrate the methodological rigor required for scale development, this section focuses on the latter approach.

For instance, Wang et al.[6] utilized the definition originally established by Weiss-Laxer et al. based on surveys and expert interviews. The procedure was as follows.

(1) Preparation: A steering committee was formed to recruit a panel of recognized experts in family health and to clarify the objectives of the consultation.
(2) Conceptualization: During the first round of consultation, the expert panel proposed and refined the concept of "family health," which the steering committee then categorized into six domains.
(3) Validation: Experts further validated the content of each domain and prioritized the constituent concepts based on importance and feasibility.

This process ultimately defined family health as "a resource at the family unit level that emerges from the intersection of the health, interactions, and capabilities of family members, alongside the family's physical, social, emotional, economic, and medical resources." Based on this definition, four key dimensions were selected for scale development: family/social/emotional health processes, family healthy lifestyle, family health resources, and external social support. By clearly operationalizing the construct in advance, the researchers were able to specify the exact themes and dimensions of family health. This rigorous preparatory phase provided a solid foundation for the study and serves as an exemplary model for future research. Furthermore, a robust definition guides the determination of initial domains and item generation, ensuring the content validity and comprehensiveness of the initial item pool.

Generating the pool of items

Following the definition of the construct, the next phase is to generate an initial item pool. It is standard practice to make the initial pool intentionally over-inclusive for each dimension. This redundancy ensures sufficient content coverage and buffers against the inevitable item attrition in statistical analysis. The initial number of items should be at least twice the number intended for the final validated version. Item generation is typically informed by authoritative textbooks, clinical guidelines, existing literature, and theoretical frameworks. Researchers should also review prior studies on similar clinical issues to formulate items that accurately capture the characteristics of each dimension. Crucially, operational definitions for each dimension must be established prior to item writing to ensure alignment. For instance, in developing the Fear of Success Questionnaire, Gao et al.[7] synthesized existing research to identify six structural dimensions: quality of life, family happiness, physical health, mental health, interpersonal relationships, and mate selection. Based on these dimensions, an initial item pool was created, followed by preliminary structured interviews and semi-open questionnaires with the target population.

Item formulation must also adhere to strict principles: (1)Items should be concise and simple, avoiding professional terms and double negatives to minimize the cognitive load on respondents. (2)Questions involving social taboos or privacy should be carefully phrased to avoid inducing respondent resistance or non-response bias. (3)Language must align with the cultural norms of the target population. For example, in the study by Gao et al., linguistic experts were invited to review the draft scale to eliminate redundancy and ambiguity, ensuring the precision of the instrument.[7]

Selecting the response format and scoring system

Response format

Determining the response format is typically a concurrent process with item generation. Researchers must select a format and scoring system that align with the specific research objectives and the practical context of the study. The primary distinction to be made is between open-ended and closed-ended formats for each item. Open-ended questions require respondents to formulate their own answers, it imposes a higher burden on respondents and researchers and presents challenges in coding and quantification for statistical analysis. While open-ended questions can provide researchers with more ideas and are generally more suitable for initial survey stage, but are not frequently employed in a well-established scale. Consequently, open-ended questions are rarely used in finalized psychometric scales. Closed-ended questions are the predominant choice in PHC research. By providing predefined options, they standardize responses and facilitate ease of administration and data processing. However, this approach requires careful methodological consideration regarding the specific configuration of options (e.g., single-response vs. multiple-response selection). Researchers must critically evaluate how different response sets might influence measurement outcomes, as these decisions significantly impact the validity of the scale.

While single-response items are predominant in scale development, multiple-response formats remain valuable when a construct cannot be adequately captured by a single answer. For instance, Sun et al. utilized Item Response Theory to develop the Diabetes Functional Health Literacy Scale. By including three multiple-response items within the 30-item instrument, they were able to capture multidimensional information relevant to the patients' health literacy. However, scoring such items introduces complexity. Although a common method involves awarding one point per correct selection, this approach is sensitive to the specific configuration of options. Generally, "select all correct answers" formats are difficult to code and analyze, and are often discouraged in standardized scaling.[2]

Furthermore, the design of response options requires careful deliberation. Such as the "uncertain" option, while Alsaffar[8] included an "uncertain" option in a nutrition questionnaire, Folasire et al.[9] argued against this practice, they cautioned that such options can encourage "satisficing", where respondents driven by low confidence or fatigue select the neutral option to avoid the cognitive effort required to answer truthfully. In addition, researchers should avoid including an "other" category in quantitative scales to facilitate standardized scoring. The decision to omit this option is scientifically justifiable only after rigorous pilot testing has ensured that the provided response categories are exclusive and collectively exhaustive.

Scoring system

The scoring system must be tailored to the items. For objective items with definitive correct answers, a simple dichotomous scoring method (e.g., 1 = correct, 0 = incorrect) is appropriate

However, most constructs exist on a continuum rather than as binary states. Consequently, Likert-type scales are the standard in applied research, typically appearing in 5-, 7-, or 9-point formats. For example, in developing a Psychological Resilience Scale for adolescents, Hu et al.[10] employed a 5-point frequency scale ranging from "never" to "always" (scored 1–5). Conversely, attitudinal research typically utilizes agreement anchors ranging from "strongly disagree" to "strongly agree." While both represent 5-point Likert formats, 7-point and 9-point variations offer finer granularity by further subdividing the response continuum. The selection of the optimal number of scale points requires careful consideration. Podsakoff et al.[11] argue that when respondents

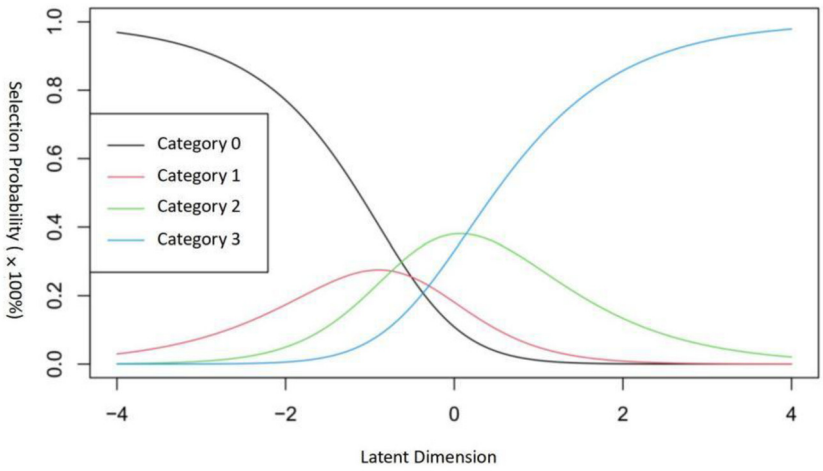

Fig. 1. Probability curve of item 1 options in the FTND.

possess high issue-relevant knowledge or interest, a greater number of response categories is warranted. In such contexts, 7- or 9-point scales are preferable to 5-point versions, as restricted scale points can exacerbate distributional skewness and limit the variance of the data.

Furthermore, the structure of a Likert scale is not immutable; post-hoc optimization of response categories is possible even after data collection. This process is facilitated by Rasch analysis, which systematically evaluates the functioning of rating scales. By inspecting Category Probability Curves (CPCs), researchers can identify issues such as category disordering or redundancy.[12] To illustrate, we examine Item 1 of the Fagerström Test for Nicotine Dependence(FTND), administered as part of the 2021 Psychological and Behavioral Investigation of Chinese Residents (PBICR).[13] Item 1 asks: "How soon after you wake up do you smoke your first cigarette?" The response categories are: >60 min (Category 0), 31–60 min (Category 1), 6–30 min (Category 2), and ≤5 min (Category 3). Fig. 1 displays the CPCs for this item. Each curve represents a specific response category. The horizontal axis denotes nicotine dependence, increasing from left to right, while the vertical axis represents the probability of endorsing. For a respondent with a dependence measure of -4, the probability of selecting Category 0 is approximately 95%, while the probability for Category 1 is only about 5%, with other categories being negligible. Thus, Category 0 is the most probable response. Crucially, to the left of the intersection between Categories 0 and 2, Category 0 is dominant; between the intersections of 0/2 and 2/3, Category 2 becomes dominant.

The research team observed low probability of Category 1, indicating Likert level abuse. Following Linacre's guidelines,[14] when such category disordering is detected, researchers should consider collapsing the problematic category with an adjacent one. Consequently, Categories 1 and 2 could be combined into a single category (6–60 min). However, it is imperative that the psychometric properties of the rescaled instrument be re-validated. Note that as FTND Item 1 utilizes polytomous response options, the Partial Credit Model (PCM) was employed for the analysis.

Pre-testing

Qualitative pilot testing is a pivotal stage in the development, translation, or revision of psychometric instruments. The primary objective is to verify that the target population interprets the items and response options as intended. Researchers evaluate whether the wording is ambiguous or the structural framework is unclear from the respondents' perspective. If semantic discrepancies or cognitive hurdles are identified, the items must undergo revision and re-test to achieve universal comprehension and content acceptability.[15]. The pre-testing mainly adopts the convenience sampling method to select samples, and tries to choose 30 or more samples to ensure the stability and reliability of data analysis.[15] For instance, in developing a Home-Care Behavior Scale for caregivers of functionally impaired older adults, Cheng et al.[16] conducted a pre-test with 102 caregivers recruited from three communities. Face validity in pre-testing is a subjective judgment of whether the instrument appears to measure what it claims to measure. High face validity implies that the measurement intent is transparent to the respondent. For example, a hand hygiene questionnaire asking about washing frequency and duration possesses high face validity.[17] In primary care research, high face validity is generally desirable for behavioral assessments or clinical inquiries to ensure construct alignment. However, for personal privacy or social stigma, high face validity may trigger deceptive and concealing behaviors. Consequently, the level of face validity must be calibrated according to the research objectives.

Eliminating items by item analysis

In the context of scale development for primary care, item analysis is an indispensable step following pilot testing. It serves as the foundation for scale refinement and prerequisite for subsequent validation. Item analysis aims to identify the psychometric performance of individual items. By applying specific statistical criteria, researchers determine whether items should be retained, modified, or eliminated, thereby ensuring item homogeneity and bolstering the scale's overall reliability. Generally, items are evaluated across three primary dimensions: item difficulty, item discrimination, and differential item functioning (DIF).

Item difficulty

Item difficulty quantifies the level of challenge an item presents and serves as a metric for evaluating respondent performance. A higher proportion of correct responses corresponds to a lower difficulty level. The objective of calibrating item difficulty is to maximize the instrument's ability to distinguish between respondents with varying levels of the trait, thereby enhancing the scale's discriminant power. As discussed in Section Selecting the Scoring System and Response Format, scoring architectures vary by scale type. For polytomous items, difficulty is typically calculated as the ratio of the item's mean score across all respondents to its maximum possible score. For instance, in a study assessing health literacy among college students, researchers recoded multiple-choice responses to calculate difficulty indices. Items with indices falling outside the range of 0.2 to 0.8 were flagged for re-evaluation or potential removal.[18] Extreme difficulty values can skew score distributions and restrict variance. Consequently, researchers must establish scientifically grounded difficulty thresholds tailored to the specific nature and objectives of the scale during the optimization process.

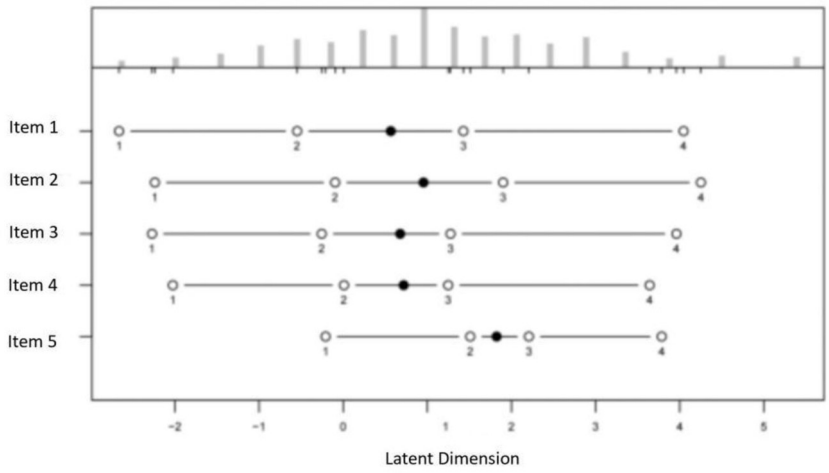

Fig. 2. Person-item map of the life satisfaction scale.

Distinct from CTT, the Rasch model operates on a different paradigm that prioritizes measurement objectivity and comparability.

Regarding item difficulty, the Rasch model posits that measurement parameters must exhibit specific objectivity. This implies that the estimation of item difficulty should remain stable regardless of the sample's ability distribution, and the estimation of an individual's ability should not depend on the specific difficulty of the items administered. Consequently, Rasch analysis generates interval-level estimates for both person ability and item difficulty, calibrating them on a common logit scale to enable direct comparison. These relationships are visually represented in a Person-Item Map(or Wright Map). Fig. 2 displays the Person-Item Map for the Life Satisfaction Scale. The distribution of respondents(black dots) clusters primarily between 0 and 2 logits. This indicates that the scale provides the maximum information(measurement precision) for individuals with moderate-to-high life satisfaction but lacks targeting for those with lower levels. By plotting persons and items on the same continuum, the map offers rich diagnostic insights. Ideally, for a well-targeted scale, item difficulty measures (thresholds) should center around 0 logits. For instance, in Hui et al.'s[19] psychometric evaluation of a stroke-specific quality-of-life scale, item difficulty thresholds ranged from–0.32 to 0.67 logits (M=0.00, SD=0.34). This centering suggests that the items were moderately difficult and well-endorsed by the sample. Conversely, if item difficulty levels are extremely high or low, it implies the behaviors or attributes represented are either too rare or too common (or challenging/easy). Such a scale would suffer from poor targeting, yielding accurate measurements only for specific subgroups at the extremes of the trait continuum.

Item discrimination

The assessment of item discrimination aims to verify whether the designed instrument can effectively differentiate between respondents possessing varying levels of the latent trait, thereby fulfilling the researcher's measurement objectives. The primary analytical approaches include the discrimination index method, item-scale correlation, and the corrected item-total correlation (CITC).

(1) The calculation of discrimination index is computationally straightforward. Respondents are ranked by total score in descending order. Following psychometric convention, the top and bottom 27% of the sample are designated as the high-scoring and low-scoring groups, respectively. Independent samples t-tests are then performed to compare item scores between these two groups. Items failing to demonstrate statistically significant differences are flagged for potential removal to ensure the instrument's precision.
(2) Discrimination can also be assessed via PT-mesure. Higher coefficients denote stronger discriminatory power. Items exhibiting suboptimal correlations require comprehensive review to determine suitability for retention.
(3) CITC evaluates the correlation between an item and the sum of the remaining items in the dimension. A CITC value≥0.5 indicates strong internal consistency. Conversely, if the value falls below 0.5, researchers should consider revising the item or removing it, contingent upon the resulting change in the "Cronbach's alpha if item deleted." For instance, Hua et al.[20] employed the extreme groups method to assess item discrimination in their Home Environment Scale for Children's Motor Development. The results demonstrated statistically significant differences between high- and low-scoring groups across all 71 items, justifying their retention at that stage. Similarly, in validating the Health Promotion Scale for older adults, Yang et al.[21] reported item-total correlations ranging from 0.406 to 0.752. These values exceeded the critical threshold of 0.300, indicating acceptable correlation. Consequently, items were further scrutinized in conjunction with reliability coefficients.

In Item Response Theory, item difficulty and discrimination are inextricably linked; items typically yield maximal discriminatory power at moderate difficulty levels relative to the target population. Consequently, the Person-Item Map serves as a vital visual tool for evaluating these parameters. In Fig. 2, the horizontal axis represents the logit scale, with values increasing from left to right. Respondents positioned further to the right possess higher levels of life satisfaction. The height of the bars corresponds to the frequency of respondents at specific ability levels. A highly concentrated distribution of respondents suggests limited discriminatory power, whereas a dispersed distribution indicates superior discrimination. For the five items depicted, the distribution of respondent ability is skewed, clustering predominantly between 0 and 2 logits. This clustering implies that the scale possesses suboptimal discriminatory power, particularly regarding its ability to differentiate respondents with lower levels of life satisfaction. Similarly, in their development of the Olweus Bully/Victim Questionnaire, Zhao et al.[12] observed that item difficulty estimates were highly concentrated. This resulted in poor differentiation across varying degrees of bullying/victimization, specifically failing to distinguish individuals with high involvement. Notably, the standard Rasch model estimates only item difficulty (assuming equal discrimination across items). To explicitly model varying item discrimination slopes, a Two-Parameter Logistic model would be required.

Differential item functioning(DIF)

DIF occurs when respondents from different subgroups who possess the same level of the underlying latent trait have differing probabilities of endorsing or correctly answering a specific item. Essentially,

Table 1
Factor analysis results of the general self-efficacy scale using principal component analysis.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	Variance (%)	Cumulative (%)	Total	Variance (%)	Cumulative (%)
1	5.753	71.910	71.910	5.753	71.910	71.910
2	0.515	6.441	78.351			
3	0.388	4.845	83.196			
4	0.306	3.829	87.024			
5	0.295	3.683	90.708			
6	0.276	3.444	94.151			
7	0.244	3.055	97.206			
8	0.224	2.794	100.000			

DIF indicates that an item exhibits biased statistical properties across groups. When this discrepancy exceeds a critical threshold, the item is flagged as potentially biased, necessitating further investigation into the source of the variance.[22] Within the Rasch framework, DIF is typically assessed using statistical inference. As the theory has evolved, various detection methods have emerged. Mantel-Haenszel (M-H) Method is a widely used approach for detecting uniform DIF, it evaluates performance differences after matching respondents on total score. Typically, an item is considered to exhibit significant DIF if the DIF contrast exceeds 0.5 logits combined with statistical significance ($P<0.05$).[23] For instance, Du et al.[24] utilized the M-H method and identified moderate to severe DIF in Items 9, 39, and 58. Lord's Chi-Square Test is implemented via R software in a Rasch analysis of the Rosenberg Self-Esteem Scale by Gao et al,[25] this method employed Lord's chi-square test and detected DIF in Items 1 and 5, implying that gender influenced the expression of self-esteem on these specific items. For items with multiple response categories, ANOVA can be utilized. In the development of the WHO Disability Assessment Schedule, researchers observed gender-based difficulties. By conducting ANOVA on gender and other potential grouping variables, they identified and revised problematic items.[26]

Crucially, the application of these item analysis techniques should not be mechanistic. Researchers must select methods appropriate for their specific characteristics (e.g., dichotomous vs. polytomous, unidimensional vs. multidimensional). Furthermore, decisions regarding item deletion should not be based solely on statistical flags. Blindly discarding items due to high difficulty, poor discrimination, or suboptimal fit is ill-advised. A theoretically "perfect" model is an idealized standard rather than a practical reality; thus, decisions should be made by synthesizing multiple psychometric indicators within the broader context of the construct being measured.

Evaluating the scale initially

Initial evaluation based on CTT

CTT also referred to as True Score Theory, became fully established in the 1950s.

The theory posits that an observed test score (X) is the additive sum of a true score (T) and a random error component (E), formulated as X=T+EX =T+EX=T+E. Fundamental assumptions include that the expected value(mean) of the error term is zero, and that the correlation between the true score and the error term is zero. Based on these axioms, CTT establishes core psychometric indices, such as reliability, validity, difficulty, and discrimination.[27] While preceding sections have detailed the use of difficulty and discrimination for item screening, this section focuses on the initial evaluation of the scale within the CTT framework. Specifically, it addresses exploratory factor analysis (EFA) alongside comprehensive reliability and validity assessments.

As a cornerstone technique within the framework of Classical Test Theory, EFA has been extensively applied to the design and development of scale in primary care. EFA employs mathematical algorithms to uncover the latent structure of variables, thereby delineating the specific dimensionality of the scale and determining the attribution of items to factors. The execution of EFA typically encompasses four critical steps: selection of variables and samples; assessment of data factorability (suitability); determination of the number of factors to extract; and factor rotation.

①Selection of variables and samples

This preparatory phase is pivotal to the integrity of the analysis.

Researchers must generate an initial item pool based on existing literature and theoretical frameworks, maximizing the inclusion of relevant items. Given that EFA typically results in a reduced final item set, a broad initial pool is necessary. The retention or exclusion of items is governed by specific psychometric criteria, primarily factor loadings, communalities, and cross-loadings. According to established guidelines, factor loadings in the component matrix are classified as follows: >0.71 (excellent), >0.63 (very good), >0.55 (good), >0.45 (fair), and >0.32 (poor).[28] Regarding communalities, values falling below 0.30 are generally considered unacceptable.[29] Furthermore, items exhibiting significant cross-loadings (i.e., high loadings on multiple factors) should be scrutinized. For instance, Chen et al.[30] eliminated items with similar loadings across factors that lacked interpretability. Sample size is another critical prerequisite; insufficient data can compromise the stability of the factor structure. Gorsuch recommended a subject-to-variable ratio of 5:1 with a minimum sample size of 100, while Nunnally advocated for a stricter ratio of 10:1.[31]

②Assessment of data factorability

The primary objective of EFA is data reduction or the identification of the scale's underlying structure. Given that Principal Component Analysis is a prevalent extraction method, researchers must ensure that both theoretical and statistical assumptions are satisfied. Theoretically, it postulates the existence of a latent structure among variables; statistically, it necessitates strong correlations between observed variables. Consequently, the following prerequisites must be met: inter-item correlations > 0.30, a significant Bartlett's Test of Sphericity ($P<0.05$), and a Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (MSA) of at least 0.60.[2] Specifically, the correlation matrix should be inspected; if the majority of coefficients are≤0.30, the data are deemed unsuitable for factor analysis. For instance, in revising the Chinese version of the Psychological Vulnerability Scale, Guo et al.[32] reported a KMO of 0.89 and a significant Bartlett's test (χ^2/df=25.31,$P<0.001$). Crucially, meeting these criteria indicates only that the data are factorable, it does not guarantee the quality or interpretability of the resulting factor solution.

③Determination of the number of factors

Determining the optimal number of factors to extract is a pivotal step in EFA.

While both under-extraction and over-extraction pose analytical risks, empirical evidence generally favors slight over-extraction, as it tends to yield more accurate parameter estimates than under-extraction.

To guide this decision, researchers typically employ three primary criteria. First, the Kaiser Criterion (Eigenvalue>1) is the most prevalent standard, dictating that only factors with eigenvalues exceeding 1.0 should be retained. Second, cumulative variance explained is originated from PCA principles, this criterion assesses the proportion of total variance accounted for by the extracted factors. While no universal con-

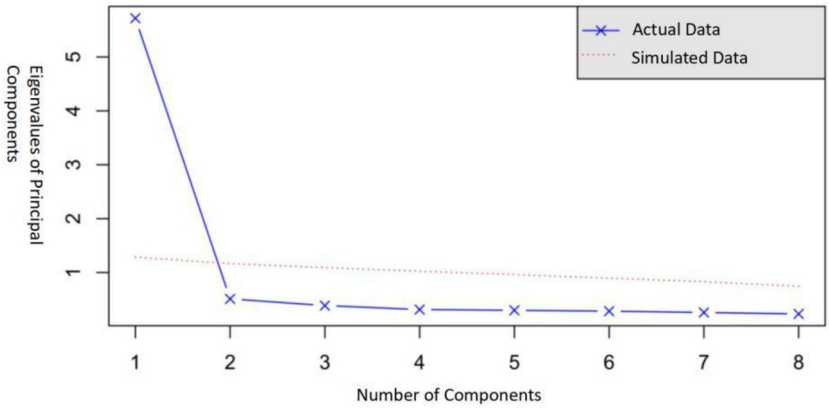

Fig. 3. Scree plot of the general self-efficacy scale.

sensus exists regarding the cutoff, it is often recommended that the retained factors explain at least 50% of the total variance.[33] For example, Table 1 presents the EFA results for the eight-item General Self-Efficacy Scale,[13] only one principal component exhibited an eigenvalue > 1, indicating a unidimensional structure. Furthermore, this single factor accounted for 71.91% of the total variance, demonstrating substantial explanatory power regarding the construct. Third, Scree Plot plots eigenvalues against the number of factors. Researchers identify the number of factors by locating the inflection point.

Fig. 3 depicts the scree plot for the General Self-Efficacy Scale. The plot shows a precipitous drop in eigenvalues after the first component, followed by a plateau. This distinct inflection point supports the conclusion that the scale comprises a single underlying factor.

④Factor rotation

Following the determination of the number of factors, the subsequent phase involves selecting an optimal rotation method. Rotation techniques are classified into oblique rotation and orthogonal rotation. Oblique rotation permits correlations among factors, whereas orthogonal rotation assumes factors are uncorrelated. In the context of empirical research in primary care, varying degrees of correlation leads to oblique rotation for objectivity. Despite this, the majority of published studies employ orthogonal rotation, primarily due to the interpretability of its results. However, imposing orthogonality on correlated data can distort the factor structure and lead to misleading conclusions. Therefore, it is recommended that researchers initially employ oblique rotation. If the resulting factor correlation matrix reveals negligible correlations among factors, one may then justifiably proceed with orthogonal rotation.

(2) Reliability analysis

Following the exclusion of items via EFA, the formal scale is established. At this stage, it is imperative to assess the reliability. Reliability refers to the stability of measurement results. Specifically, it reflects the degree of concordance between repeated measurements of the same trait in the same individual using the same scale. Under the framework of CTT, standard methods for assessing reliability include alternate form reliability, test-retest reliability, homogeneity reliability, split-half reliability, and scorer reliability. In clinical research, alternate form reliability is rarely utilized due to challenges in obtain. Consequently, researchers predominantly rely on test–retest reliability, split-half reliability, and homogeneity reliability as the primary indices of reliability.

(1) Test-Retest Reliability

In the context of scale development, temporal stability serves as a fundamental metric of reliability. Consequently, when developing scales for primary care, researchers should report the consistency between two administrations of the scale to the same cohort. This relationship is typically quantified using the Pearson product-moment correlation coefficient (rrr). For instance, in validating the Chinese version of the Psychological Need Satisfaction in Exercise Scale for older adults, Liu et al.[34] reported an overall test-retest reliability of 0.883, with coefficients for the three subdimensions ranging from 0.829 to 0.876. Conventional guidelines suggest that coefficients between 0.65 and 0.70 represent the minimum acceptable threshold, values between 0.70 and 0.80 indicate good reliability, while those between 0.80 and 0.90 are considered excellent. While the reliability reported by Liu et al.[34] is satisfactory, the authors failed to specify the time interval between administrations. This is a critical omission, as test-retest reliability is sensitive to the duration of the interval (i.e., reliability coefficients may fluctuate over time). Future studies must report this interval to facilitate an accurate assessment of temporal stability.

(2) Alternate form reliability

Alternate form reliability refers to the degree of consistency between two parallel tests administered to the same group. Its magnitude is typically quantified using the Pearson product-moment correlation coefficient between scores obtained from the two parallel forms. Alternate form reliability serves as an important indicator of measurement stability, however, it is not widely implemented in practice due to the challenges associated with designing two genuinely equivalent tests. For instance, Liu Aimei et al.[35] developed a Knowledge-Attitude-Practice (KAP) questionnaire tailored for patients with sudden sensorineural hearing loss, employing alternate form reliability by using a second questionnaire with similar content and response formats. They reported a alternate form reliability coefficient of 0.88 for the health-related knowledge section. The evaluation standards for alternate form reliability are consistent with those applied to test–retest reliability,[36] indicating that this measure demonstrates good alternate form reliability.

(3) Split-Half Reliability

Split-half reliability is a measure of internal consistency that involves partitioning a single test into two equivalent halves and comparing the consistency of scores across these two subscales. This method is frequently employed in research due to its primary advantage: it requires simple operations within the statistical software.

(4) Homogeneity Reliability

Homogeneity reliability assesses the extent to which items on a scale are interrelated and measure the same underlying construct. As widely used metric for this purpose, Cronbach's α is similar to split-half reliability, requires only a single test administration and is readily calculated using statistical software. Established benchmarks for interpreting Cronbach's α are as follows:

For an overall scale, a coefficient of 0.80 or higher is considered excellent, while values between 0.70 and 0.80 are deemed acceptable; for subscales, a coefficient of 0.70 or higher is considered good, while values between 0.60 and 0.70 represent the minimum acceptable threshold.

(5) Scorer Reliability

Scorer reliability quantifies the degree of agreement among multiple independent raters who are evaluating the same Scale. When the evaluation data are ordinal, Kendall's coefficient of concordance (W) is a

particularly suitable non-parametric statistic for this purpose. Kendall's W assesses the level of association among several sets of ranks provided by three or more raters.

(3) Assessment of validity

Validity evaluation is a crucial step in the development of any scale within the primary care context. Validity refers to the extent to which a test scale can measure the characteristics it intends to measure. The theoretical definition of validity is: in a series of measurements related to the purpose of the measurement, the ratio of the actual change (the change caused by the measured change) to the total change (the actual change).

Validity can be classified into content validity, construct validity and empirical validity.

Content Validity

Content validity refers to the degree to which an instrument's items are representative of the entire theoretical domain it is intended to measure. This form of validity is not established through statistical analysis but rather through a rigorous, systematic evaluation by a panel of subject matter experts. The credibility of a content validity assessment is therefore fundamentally dependent on the qualifications, expertise, and diversity of the selected expert panel. A common and recommended practice is to recruit established experts from both academic and clinical settings. For instance, Cui et al.[37] convened a panel of six nursing experts, including university professors, nursing department directors, and clinical specialists, to ensure a comprehensive evaluation of their scale's content. To quantify the judgments of these experts, the Content Validity Index (CVI) is the most widely used metric, it offers a straightforward method for assessing the consensus on item relevance; Item-Level CVI (I-CVI) is the proportion of experts who rate an individual item as "relevant" or "highly relevant"; Scale-Level CVI (S-CVI) represents the overall content validity of the entire instrument, typically calculated as the average of the I-CVI values for all items on the scale. The application of this method is illustrated in the development of a medication adherence questionnaire for coronary heart disease patients. Researchers designed an evaluation form asking experts to rate each item on a 4-point relevance scale (e.g., from 1 = "not relevant" to 4 = "highly relevant").

After collecting the expert ratings, the resulting I-CVI for every item and the overall S-CVI were both 1.00. This perfect score indicates unanimous expert agreement on the relevance of all items, providing strong evidence for the questionnaire's content validity.

(2) Construct Validity

Construct validity is the extent to which the scores of an instrument accurately reflect its underlying theoretical construct. Item analysis, EFA and confirmatory factor analysis(CFA) are employed to evaluate construct validity. Item analysis examines the relationship between individual items and the overall scale or their designated subscale. Strong item-total or item-subscale correlations provide initial evidence that the items are collectively measuring a unified construct. For example, Yang et al.[38] found that in their cognitive style questionnaire, all items correlated strongly with their respective dimensions (correlation coefficients > 0.55), while the dimensions were moderately intercorrelated. This pattern suggests that the items are well-differentiated and appropriately grouped within a cohesive theoretical structure. EFA is basically the same as described in the previous section. However, this time there is no need to delete items. Generally speaking, when testing the structural validity of a questionnaire formed through EFA, new data should be collected and EFA or CFA should be used to measure the new data. This confirmatory approach is exemplified in the validation of the Chinese version of the Duke Anticoagulation Satisfaction Scale (DASS) by Wu et al.[39] The researchers used CFA to test the fit of the established four-factor model. The analysis yielded a range of model fit indices that met or exceeded established benchmarks for a good fit: Chi-square/degrees of freedom (CMIN/DF): 1.825 (< 5.00); Comparative Fit Index (CFI): 0.938 (> 0.90); Tucker-Lewis Index (TLI): 0.921 (> 0.90); Root Mean Square Error of Approximation (RMSEA): 0.066 (< 0.08). The scale has good structural validity.

(3) Empirical validity

Empirical validity evaluates how well the scores on a scale correlate with a distinct, external outcome, known as the criterion. If a scale can effectively estimate the behaviors of the subjects in a specific context, then it is said to have good empirical validity or criterion-related validity. This form of validity is essential for demonstrating the practical utility of a scale. The relevant law refers to the degree of correlation between the test scores and the validity variables. The calculated correlation coefficient is the validity coefficient, and the square of the validity coefficient represents validity. The study by You et al.[40] provides a clear example of assessing concurrent validity. To validate the Beck Depression Inventory (BDI), they selected the General Well-Being Scale (GWB) as the external criterion. Participants completed both scales in the same session. The analysis revealed a statistically significant negative correlation($P<0.001$) between the BDI scores and the GWB scores. This finding provides strong evidence for the BDI's criterion-related validity.

Advanced psychometric assessment using the Rasch model

The Rasch model represents a foundational approach within the broader framework of Item Response Theory (IRT). It provides a sophisticated method for evaluating the psychometric properties of a scale by modeling the relationship between an individual's underlying latent trait and their responses to specific items.

The central tenet of the Rasch model is that the probability of a person endorsing a particular item is a logistic function of the difference between their individual trait level and the difficulty of that item. This elegant principle allows for objective measurement, which rests on two fundamental requirements. First, for any given item, an individual with a higher level of the latent trait must have a greater probability of a successful outcome than an individual with a lower level of the trait. Second, for any given individual, the probability of a successful outcome must be higher for an easier item than for a more difficult one.[41] Despite its development several decades ago, the Rasch model remains underutilized in primary care research. A literature search of the CNKI database for the term "Rasch" shows that from 1915 to 2022, only 160 articles were published in core journals. While there has been a notable increase in interest, with 46.25% of these articles published between 2017 and 2021, the research is predominantly concentrated in the fields of psychology and education. Given the scarcity of its application to primary care, there is a clear and compelling need to adopt the Rasch model in this domain.

(1) Unidimensionality test

Item Response Theory (IRT) is a modern measurement framework distinct from Classical Test Theory (CTT), refers to the mathematical modeling of the relationship between a person's latent trait and their probability of a specific response to an item. Common IRT models include single-parameter models, two-parameter models and three-parameter models.[41] The Rasch model, often considered the one-parameter model within IRT, is governed by a strict primary assumption of unidimensionality, which posits that a single latent trait is the primary driver of a respondent's answers. Crucially, this does not mean a scale must be simplistic or contain only one dimension. An instrument can be multidimensional, comprising several distinct subscales, so long as those subscales collectively measure a single, overarching higher-order construct. For example, in their adaptation of a Nutrition Literacy Assessment Tool, Chen et al.[42] performed a Rasch analysis on six distinct subscales as well as the total scale, because all items were hypothesized to reflect the single, unifying trait of "nutrition literacy." The standard method for testing this assumption is a Principal Component Analysis (PCA) of the standardized residuals. After the primary Rasch dimension is extracted from the data, the PCA examines the remaining variance to see if any other significant dimensions exist. A common guideline, based on the work of Raiche, suggests that the eigenvalue of the first residual component should ideally be less than 2.0. Values between 1.4 and 2.1 are often considered acceptable, indicating that no substantial

Table 2
Model fitting parameters of the life satisfaction scale.

Item	x^2 value	df	*P*-value	Outfit MNSQ	Infit MNSQ	Outfit t value	Infit t value
1	343.86	538	1	0.64	0.63	-6.99	-7.10
2	324.13	538	1	0.60	0.60	-7.89	-7.86
3	307.71	538	1	0.57	0.59	-8.59	-8.26
4	496.31	538	0.90	0.92	0.91	-1.36	-1.54
5	817.96	538	0	1.52	1.4	7.21	5.90

secondary dimension is present.[42] In the study by Chen et al.[42] the results of this test were twofold: The first residual eigenvalues for the six individual subscales ranged from 1.6 to 1.8. The eigenvalue for the total scale, however, was 3.1. An eigenvalue of this magnitude is typically interpreted as evidence of a meaningful secondary dimension.

(2) Evaluating model-data fit

Once unidimensionality is established, the next step is to evaluate how well the collected data fit the expectations of the Rasch model. This is achieved by comparing the observed response patterns to the patterns predicted by the model, which simultaneously estimates item difficulty and person ability. The primary diagnostic tools for this are the infit and outfit mean-square (MNSQ) statistics. Both are chi-square statistics divided by their degrees of freedom, with an expected value of 1.0 indicating perfect fit. In practice, values between 0.5 and 1.5 are generally considered productive for measurement.[43] For instance, in an analysis of the Life Satisfaction Scale,[13] most items demonstrated good fit (Table 2). However, Item 5 ("If I could live my life over, I would change almost nothing") showed elevated fit statistics: an outfit MNSQ of 1.52 and an infit MNSQ of 1.40. This means that those with higher life satisfaction chose the lower scores, that is, they disagreed/ somewhat disagreed; while those with lower life satisfaction chose the higher scores, that is, they agreed/ somewhat agreed. This suggests that the responses to Item 5 were erratic and did not consistently align with the respondents' overall life satisfaction levels, thereby introducing measurement error. Such an item may be a candidate for revision or removal.

Beyond item-level fit, it is crucial to assess the scale's overall precision. In Item Response Theory, this is visualized using a Test Information Function (TIF). The TIF illustrates the amount of measurement information the entire scale provides across the full spectrum of the latent trait. Information is the mathematical inverse of measurement error; therefore, where the information curve is highest, the scale's precision is greatest, and the standard error of measurement is lowest. The difficulty of item can be seen on the horizontal axis, which represents the trait level of the participants. Each scale represents 1 Logit unit. The vertical axis represents the amount of information, that is, the Fisher information function.[12] Fig. 4 shows that the TIF for the Life Satisfaction Scale.[13] This means the scale functions most precisely when measuring individuals with moderate to high levels of life satisfaction. Similarly, Gao et al.[25] found that their self-esteem scale's TIF peaked between -2 and 0 logits, indicating it was most effective for measuring individuals with low to moderate self-esteem. This analysis is critical for understanding a scale's practical utility, as it reveals for which specific populations the instrument is most and least reliable.

(3) Reliability

Within the Rasch framework, the primary index of reliability is the Person Separation Reliability (PSR). PSR represents the ratio of the true variance in person ability estimates to the total observed variance.[12] PSR quantifies how effectively a scale can differentiate or separate the sample of respondents into distinct strata based on their estimated trait level. The PSR value ranges from 0 to 1, the reliability indicators above 0.7 are acceptable, and those above 0.8 are considered excellent.

Re-evaluating the scale

The initial development process (steps 1-6) yields a preliminary version of the scale. However, because the item analysis and validation were conducted on a single dataset, it may be uniquely tailored to the specific characteristics of that sample. Therefore, the crucial final stage is to establish the scale's generalizability and robustness. Furthermore, to assess test-retest reliability, a subset of participants from the original sample should be included in the second wave of data collection. The analytical methods for this re-evaluation are identical to those previously described. This section will focus on CFA for rigorously testing the scale's structural validity. Unlike the data-driven Exploratory Factor Analysis(EFA), CFA is a theory-driven, hypothesis-testing technique.

Its purpose is to statistically test whether a new set of data fits a pre-specified, a priori theoretical structure. In this context, the factor structure derived from the initial EFA becomes the hypothesized model to be confirmed in the new sample. The evaluation of the CFA model is based on a battery of goodness-of-fit indices, which quantify the degree of correspondence between the theoretical model and the observed data. Key indices and their widely accepted thresholds include: (1)Chi-Square to Degrees of Freedom Ratio (χ^2/df): Should be < 3 (with < 2 indicating a more stringent fit)[44]; (2)Root Mean Square Error of Approximation(RMSEA): Should be < 0.08 for acceptable fit[6]; (3)Goodness-of-Fit Index (GFI) & Adjusted GFI (AGFI): Should be > 0.90, indicating good absolute fit[45]; (4)Incremental Fit Indices (NFI, IFI, TLI, CFI): Should be > 0.90.[45] Should the initial model demonstrate poor fit, model respecification may be considered. This is often guided by MI, which suggest potential paths to add to the model to improve fit. A common modification involves allowing the error terms of two items to covary.[43]

Discussion

The method of scale design has been widely applied in the field of primary care, which is mainly reflected in the extensive use of scale design research. Most research projects involve the use of scales, so the rationality of a scale's design and development determines the reliability of the entire research. Despite this, a review of the literature reveals that many scale development studies in this field suffer from significant methodological weaknesses. Common problems include insufficient evidence of reliability and validity, the omission of crucial procedural steps, and the incorrect application of statistical analyses. Following a rigorous, step-by-step process ensures that key psychometric principles are respected, thereby enhancing the quality and trustworthiness of the final version.

High-quality scale development is a process that relies on the synthesis of two complementary pillars: a top-down, theory-driven approach and a bottom-up, data-driven approach.

The first pillar, the theoretical framework, is a deductive process that must precede and guide all empirical work. It requires the researcher to conduct a comprehensive review of the existing literature to develop a deep understanding of the target construct's theoretical domain. This foundational step ensures that the scale is grounded in established knowledge, which is a prerequisite for achieving content and construct validity. The second pillar, statistical test, is an inductive, data-driven process that tests the theoretical framework against observed evidence. Statistical analyses provide the objective criteria necessary to evaluate item performance, identify and remove flawed or poorly functioning items, and formally assess the scale's overall reliability and validity. The researchers used statistics to test the reliability and validity of the scale, in order to ensure the objectivity and effectiveness of the scale. In con-

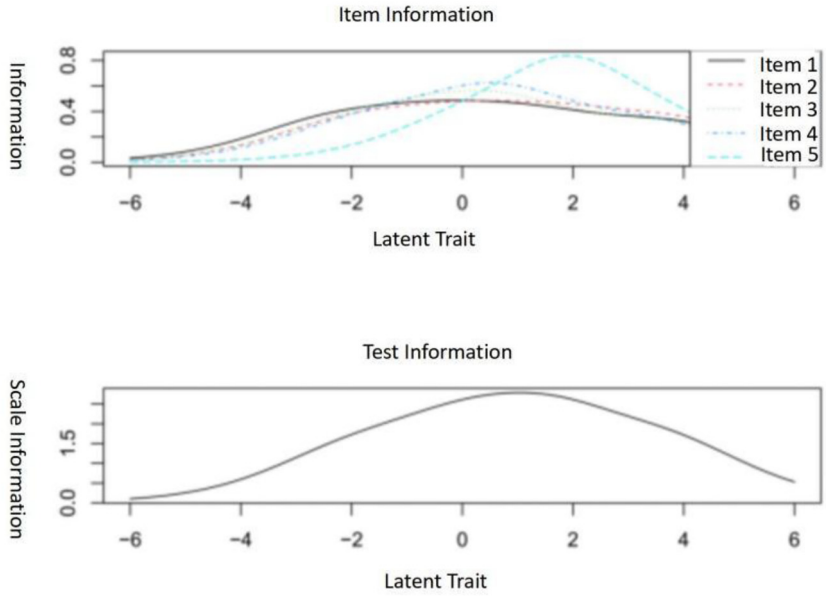

Fig. 4. Information curve of life satisfaction test.

clusion, neither approach is sufficient on its own. Only by integrating the deductive, top-down guidance of theory with the inductive, bottom-up evidence from statistical evaluation can researchers develop scales that are not only reliable and valid but also truly meaningful.

Furthermore, from a statistical perspective, traditional factor analysis and Rasch analysis (Item Response Theory) represent two distinct methods of data analysis. Factor analysis tends to interpret subject responses (i.e., the selection of a 0-to-4 score) as continuous variables, whereas Item Response Theory treats them as five distinct categories.[46] Therefore, during the process of scale development or adaptation, both methods can be used in conjunction to test the reliability and validity of the scale. However, it is crucial to avoid mixing them indiscriminately—for instance, using Classical Test Theory to reduce items and subsequently using Item Response Theory to construct the statistical model.

This study systematically elaborates on how to conduct scale design research within the field of primary care. However, due to space limitations and the specialized nature of the content, some clinicians may find the terminology used in the text difficult to understand. Additionally, for the majority of GPs, selecting an appropriate existing scale may be more effective than designing a new one. To this end, explanations of certain technical terms appearing in the text, along with recommendations for GPs on selecting scales, are provided in the appendix (please scan the QR code in the article to access). In addition, this study provides researchers with references for further in-depth study of scale design methods, such as Latent Variable Modeling and Mplus Application: Basic Edition, Principles and Practice of Health Questionnaire Design, R Language: Scale Development, Statistical Analysis, and Item Response Theory, and Handbook of Quantitative Research in Psychology and Behavior. In summary, researchers must adhere to standardized procedures when designing scales. Specific steps can be executed by referring to the relevant resources in the checklist to ensure the objectivity and validity of the designed scale.

Declarations

Not applicable.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

Funding

Not applicable.

Declaration of competing interest

All authors declare that there are no competing interests

CRediT authorship contribution statement

Wang Fei: Conceptualization, Writing – original draft. **Tang Jingqi:** Methodology, Data curation, Formal analysis, Funding acquisition, Project administration, Resources, Supervision, Validation. **Sun Xiaonan:** Writing – review & editing. **Sun Xinying:** Writing – review & editing. **Li Jun:** Writing – review & editing. **Meng Xingxing:** Methodology, Data curation, Formal analysis, Funding acquisition, Project administration, Resources, Supervision, Validation. **Wu Yibo:** Methodology, Data curation, Formal analysis, Funding acquisition, Project administration, Resources, Supervision, Validation.

Acknowledgements

We are deeply grateful to Associate Professor Gao Zhiqiang of the School of Philosophy, Anhui University, for his valuable guidance in the field of psychometrics. His course on psychometrics provided the authors with an early and foundational understanding of this discipline. We also acknowledge the contributions of all investigators involved in the Psychology and Behavior Investigation of Chinese Residents(PBICR) in 2021. Their dedicated participation enabled the collection of extensive data that underpins the figures and analyses presented in this study.

References

1. Wang R, Wang S. Research on the role of general practitioners in China's medical and health service system. *Chin Gen Pract*. 2020;23(4):388–394 402. Chinese. doi:10.12114/j.issn.1007-9572.2020.00.084.
2. Trakman GL, Forsyth A, Hoye R, et al. Developing and validating a nutrition knowledge questionnaire: key methods and considerations. *Public Health Nutr*. 2017;20(15):2670–2679. doi:10.1017/S1368980017001471.
3. Kouvelioti R, Vagenas G. Methodological and statistical quality in research evaluating nutritional attitudes in sports. *Int J Sport Nutr Exerc Metab*. 2015;25(6):624–635. doi:10.1123/ijsnem.2014-0010.
4. Jin Y, Chen S, Bao Y, et al. Research progress on assessment tools for children with autism spectrum disorder. *Nurs Pract Res*. 2021;18(9):1325–1329 Chinese. doi:10.3969/j.issn.1672-9676.2021.09.016.
5. Bond TG, Yan Z, Heene M. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. New York: Routledge; 2020:1–348.
6. Wang F, Wu YC, Sun XN, et al. Reliability and validity of the Chinese version of a short form of the family health scale. *BMC Prim Care*. 2022;23(1):108. doi:10.1186/s12875-022-01702-1.
7. Gao Z, Zhang T. Development and application of the success fear questionnaire. *Chin J Clin Psychol*. 2011;19(5):602–605 686. Chinese. doi:10.16128/j.cnki.1005-3611.2011.05.009.
8. Alsaffar AA. Validation of a general nutrition knowledge questionnaire in a Turkish student sample. *Public Health Nutr*. 2012;15(11):2074–2085. doi:10.1017/S1368980011003594.
9. Folasire OF, Akomolafe AA, Sanusi RA. Does nutrition knowledge and practice of athletes translate to enhanced athletic performance? Cross-sectional study amongst Nigerian undergraduate athletes. *Glob J Health Sci*. 2015;7(5):215–225. doi:10.5539/gjhs.v7n5p215.
10. Hu H, Zhang H, Wang J, et al. Development and preliminary evaluation of the psychological resilience scale for middle school students. *Chin J Sch Health*. 2009;30(12):1097–1099 Chinese.
11. Podsakoff PM, MacKenzie SB, Lee JY, et al. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *J Appl Psychol*. 2003;88(5):879.
12. Zhao F, He Z, Yuan S, et al. Rasch model analysis of the Olweus bullying scale. *J Southwest Univ (Soc Sci Ed)*. 2020;46(5):115–121 Chinese. doi:10.13718/j.cnki.xdsk.2020.05.012.
13. Wang YJ, Kaierdebieke A, SY F, et al. Study protocol: a cross-sectional study on psychology and behavior investigation of Chinese residents, PBICR. *Psychosom Med Res*. 2022;4(3):19.
14. Linacre JM. Optimizing rating scale category effectiveness. *J Appl Meas*. 2002;3(1):85–106.
15. Perneger TV, Courvoisier DS, Hudelson PM, et al. Sample size for pre-tests of questionnaires. *Qual Life Res*. 2015;24(1):147–151. doi:10.1007/s11136-014-0752-2.
16. Cheng Y, Wang Y, Li Y, et al. Development and reliability and validity test of the home care behavior scale for caregivers of disabled elderly. *Chin J Gerontol*. 2018;38(21):5314–5316 Chinese.
17. Gao Y, Pan B. Reliability and validity test of the stamps nurse job satisfaction scale. *Chin Nurs Res*. 2010;24(7):645–646 Chinese.
18. Rabin LA, Miles RT, Kamata A, et al. Development, item analysis, and initial reliability and validity of three forms of a multiple-choice mental health literacy assessment for college students (MHLA-c). *Psychiatry Res*. 2021;300:113897. doi:10.1016/j.psychres.2021.113897.
19. Hui J, Pei J, Wang Y, et al. Rasch analysis of the stroke specific quality of life scale for acupuncture intervention. *Chin Acupunct Moxibustion*. 2013;33(4):363–366 Chinese. doi:10.13703/j.0255-2930.2013.04.029.
20. Hua J, Zhang L, Gu G, et al. Preliminary development of the home environment scale for motor development of urban preschool children. *Chin J Sch Health*. 2011;32(2):161–163 Chinese. doi:10.16835/j.cnki.1000-9817.2011.02.015.
21. Yang Z, Zhang H. Cross-cultural adaptation and reliability and validity test of the elderly health promotion scale. *J Nurs Sci*. 2021;36(19):91–94 Chinese. doi:10.3870/j.issn.1001-4152.2021.19.091.
22. Lai JS, Cella D, Chang CH, et al. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT Fatigue Scale. *Qual Life Res*. 2003;12(5):485–501. doi:10.1023/a:1025014509626.
23. Zwick R, Thayer DT, Lewis C. An empirical Bayes approach to Mantel-Haenszel DIF analysis. *J Educ Meas*. 1999;36(1):1–28. doi:10.1111/j.1745-3984.1999.tb00543.x.
24. Du H, Li F. Analysis of the influence of sample size on the Mantel-Haenszel method for testing DIF effects. *Exam Res*. 2016;12(5):55–62 Chinese.
25. Gao S, Zhang X. Applying the Rasch model to analyze the Rosenberg self-esteem scale. *Psychol Explor*. 2018;38(5):445–450 Chinese.
26. Vaganian L, Bussmann S, Boecker M, et al. An item analysis according to the Rasch model of the German 12-item WHO disability assessment schedule (WHODAS 2.0). *Qual Life Res*. 2021;30(10):2929–2938. doi:10.1007/s11136-021-02872-8.
27. Yang Z., House; HX.DP.S:SLP 2016. Chinese.
28. Comrey AL. *A First Course in Factor Analysis*. New York: Academic Press; 1973.
29. Cao C, Qi S, Jin T. Development of the impression management self-efficacy scale for college students. *Mod Prev Med*. 2021;48(17):3199–3201 3225. Chinese..
30. Chen G, Cai T, Hu F, et al. Revision of the emotional eating scale in Chinese adolescents. *Chin J Clin Psychol*. 2013;21(4):572–575 588. Chinese. doi:10.16128/j.cnki.1005-3611.2013.04.026.
31. Wang M. *Latent Variable Modeling and Mplus Application: Basic Edition*. Chongqing: Chongqing University Press; 2014 Chinese.
32. Guo J, Wang Y, Song Y, et al. Preliminary revision of the Chinese version of the psychological vulnerability questionnaire and reliability and validity analysis in community residents. *Chin J Public Health*. 2019;35(2):129–133 Chinese.. doi:10.11847/zgggws1119908.
33. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. *Psychol Assess*. 1995;7(3):286–299. doi:10.1037/1040-3590.7.3.286.
34. Liu L, Liu H, Guo H, et al. Reliability and validity test and applicability analysis of the Chinese version of the psychological need satisfaction in exercise scale for the elderly. *Chin Gen Pract*. 2021;24(5):619–624 Chinese. doi:10.12114/j.issn.1007-9572.2020.00.459.
35. Liu A, Liu Y. Development and reliability and validity test of the health knowledge, attitude and practice questionnaire for patients with sudden deafness. *J Audiol Speech Pathol*. 2012;20(5):444–448 Chinese. doi:10.3969/j.issn.1006-7299.2012.05.010.
36. Jian X, Dai B. *SPSS 23.0 Statistical Analysis: Applications in Psychology and Education*. Beijing: Beijing Normal University Publishing Group; 2017 Chinese.
37. Cui C, Yue M, Li Y, et al. Reliability and validity study of the Chinese version of the Health Behavior Scale. *J Nurs Sci*. 2017;32(12):62–65 Chinese. doi:10.3870/j.issn.1001-4152.2017.12.062.
38. Yang L, Zhai R, Qi Z, et al. Mental health quality assessment system: development of the cognitive style questionnaire for Chinese Adults. *Psychol Behav Res*. 2012;10(5):332–339 Chinese.
39. Wu YB, Dong SJ, Li XY, et al. The transcultural adaptation and validation of the Chinese version of the Duke anticoagulation satisfaction scale. *Front Pharmacol*. 2022;13:790293. doi:10.3389/fphar.2022.790293.
40. You Y, Yu S, Liang B. Trial and evaluation of the beck depression inventory among teachers in disaster areas. *J Sichuan Norm Univ (Nat Sci Ed)*. 2011;34(3):439–442 Chinese. doi:10.3969/j.issn.1001-8395.2011.03.033.
41. Yan Z. Objective measurement in the field of psychological science: characteristics and development trends of the Rasch model. *Adv Psychol Sci*. 2010;18(8):1298–1305 Chinese.
42. Chen Y, Yang C, Wang D, et al. Sinicization of the nutritional literacy assessment tool and reliability and validity study in patients with diabetes: analysis based on CTT and Rasch models. *Chin Gen Pract*. 2020;23(26):3342–3347 Chinese. doi:10.12114/j.issn.1007-9572.2020.00.009.
43. Zhao S, He F, Liu Y. Application of the Rasch model in the quality analysis of academic achievement tests. *Educ Res Exp*. 2013(1):87–91.
44. Bagozzi RP, Yi Y. On the evaluation of structural equation models. *J Acad Mark Sci*. 1988;16(1):74–94. doi:10.1007/BF02723327.
45. Zhang L, Gu Y, Huang M, et al. Confirmatory factor analysis of the evidence-based practice readiness assessment scale. *Nurs J Chin PLA*. 2019;36(2):6–10 25. Chinese. doi:10.3969/j.issn.1008-9993.2019.02.002.
46. Sun X. *Principles and Practice of Health Questionnaire Design*. Beijing: Peking University Medical Press; 2020 Chinese.