Design and Development of Scales in Primary Care: Practical Steps and Statistical Methods (Postprint)

Authors: Wang Fei, Tang Jingqi, Sun Xiaonan, Sun Xinying, Li Jun, Meng Xingxing, Wu Yibo, Meng Xingxing, Yibo Wu

Date: 2022-11-23T00:00:00+00:00

Abstract

This study outlines the statistical methods and practical steps for designing and developing valid and reliable questionnaires in the field of primary health care. It reviews a series of studies on questionnaire development and scale design, and establishes a standard procedure for scale design within the primary care domain. This procedure involves key practical steps and statistical methods in the scale design process, and is illustrated through relevant research cases from this field. We propose a seven-step development method for primary health care questionnaires as follows: (1) Define the construct to be measured; (2) Generate an item pool; (3) Select the scoring system and response format; (4) Pretest (assessing content validity and face validity, etc.); (5) Eliminate items through item analysis; (6) Initial evaluation of the scale, including assessment of the scale' s reliability and validity, as well as factor analysis or Rasch analysis; (7) Re-evaluation of the scale, re-examining the scale's properties, including test-retest reliability and construct validity. In summary, scale design research should strictly follow the standard procedures of scale development, and the integrated use of Rasch models and factor analysis methods will render measurement results more objective.

Full Text

Design and Development of Scales in Primary Care: Practical Steps and Statistical Methods

WANG Fei[1], TANG Jingqi[2], SUN Xiaonan[3], SUN Xinying[4], LI Jun[5], MENG Xingxing[2], *WU Yibo*[4]

[1]State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing 100875, China

[2]School of Philosophy, Anhui University, Hefei 230039, China
[3]School of Humanities and Social Sciences, Harbin Medical University, Harbin 150081, China
[4]School of Public Health, Peking University, Beijing 100191, China
[5]Peking University Third Hospital, Beijing 100191, China

Corresponding authors: MENG Xingxing, Lecturer; E-mail: 614997175@qq.com; WU Yibo, PhD student; E-mail: bjmuwuyibo@outlook.com

Abstract

This study outlines statistical methods and practical steps for designing and developing valid and reliable questionnaires within the primary health care domain. We reviewed a series of studies on questionnaire development and scale design to establish a standardized process for scale design in primary care. This process addresses key practical steps and statistical methods in scale development, illustrated with examples from relevant previous research. We propose a seven-step approach for developing primary health care questionnaires: (1) define the construct to be measured; (2) generate an item pool; (3) select the scoring system and response format; (4) conduct pre-testing (assessing content validity, face validity, etc.); (5) eliminate items through item analysis; (6) conduct initial scale evaluation, including reliability and validity assessment, as well as factor analysis or Rasch analysis; and (7) re-evaluate the scale to re-examine its properties, including test-retest reliability and construct validity. Overall, scale design studies should strictly follow standard development procedures, and the integrated use of Rasch models and factor analysis will yield more objective measurement results.

Keywords: primary care; scale development; factor analysis; Rasch model

1 Introduction

The World Health Organization (WHO) proposed the ambitious goal of "Health for All" at the 30th World Health Assembly in 1977, identifying primary health care as the fundamental pathway and key to achieving this objective [**?**]. As primary providers of primary care services, general practitioners must make accurate assessments of patients' characteristics to offer appropriate recommendations. Scales, as tools for measuring specific traits, have been widely used in social sciences and medicine. Designing and developing scales within the primary care domain can help researchers and general practitioners measure the degree of specific traits in subjects.

However, scale design and development involve multiple complex and time-consuming steps that can be daunting and are often partially neglected [**?**]. This has led to problems in scale design, such as a study assessing athletes' and coaches' nutritional attitudes and knowledge finding that approximately 70% of included studies used tools with unknown validity and reliability, and 67% used unvalidated instruments [**?**]. In Chen Wenxiong' s autism screening scale,

some items with poor reliability and validity were retained in the final version [?]. Such scales without proper validation or with poor psychometric properties severely limit the conclusions that can be drawn and may even have negative effects.

Therefore, there is an urgent need for standardized procedures to guide scale design research in primary care. Additionally, we observe that most scale design studies in primary care are conducted within the framework of Classical Test Theory (CTT). While this approach is crucial for validating psychometric properties, its inherent limitations often fail to ensure measurement objectivity. The emergence of Rasch models provides an excellent solution to this problem. Rasch models use objective measurement in natural sciences as a benchmark to establish objective standards for measurement in social sciences, ensuring more objective and reliable information [?].

Based on this, our study summarizes current domestic and international questionnaire development and scale design methods in primary care from both CTT and Rasch model perspectives, helping researchers in this field conduct better studies through detailed explanations of specific steps and statistical methods.

2.1 Defining the Construct to be Measured

The most critical step in scale development within primary care is providing an accurate and comprehensive definition of the construct to be measured. The definition should explain both the connotation and extension of the construct, as well as its structure. Such definitions are typically derived from classic textbooks, guidelines, or authoritative experts in the field, or can be summarized based on extensive literature and surveys. The former is commonly used in clinical practice. To further expand the application of relevant methodologies, we use an example based on extensive surveys and expert interviews to establish a definition.

In Wang et al.' s study [?], the researchers used the definition established by Weiss-Laxer et al. based on extensive surveys and expert interviews: (1) researchers first contacted renowned researchers in family health to form an expert panel, together with an executive leadership team, to clarify the final objectives of expert interviews; (2) through the first round of expert consultation, the panel proposed and jointly revised the concept of "family health," which the leadership team divided into six distinct domains; (3) experts further confirmed the content and concepts included in each domain, ranking them by importance and feasibility. The final definition of family health was: "a resource at the family unit level, developed from the intersection of each family member' s health, their interactions and capabilities, and the family' s physical, social, emotional, economic, and medical resources" [?]. During scale development, four important factors were selected: family/social/emotional health processes, family healthy lifestyle, family health resources, and external family social support.

Weiss-Laxer et al. defined the construct before the study began, including the

exact topics of family health to be measured and its relevant dimensions, laying a foundation for smooth research progress. Their methodology is worthy of emulation. Researchers can also determine the initial dimensions and intended purpose of the questionnaire based on the definition, making the initial test as comprehensive as possible.

2.2 Generating the Item Pool

After completing the definition of the measured construct, researchers begin developing the initial dimensional item pool. The item pool representing the same dimension should be as redundant as possible to ensure it meets expected requirements and avoids insufficient items after later data processing. Generally, the number of items in the developed scale should be at least twice that of the final retained version.

Item pool generation typically uses classic textbooks, guidelines, literature, and theory as guidance, combined with previous research on clinical problems or existing questionnaires. Through evaluation of existing materials, questions measuring each dimensional characteristic are developed. Therefore, before developing scale items, it is essential to clearly define each dimension and develop questions that align with their meanings based on these definitions. For example, when Gao Zhiqiang et al. developed the Fear of Success Questionnaire, they reviewed existing research and summarized six structural dimensions of fear of success: quality of life, family happiness, physical health, mental health, interpersonal relationships, and romantic partner selection. They then developed the initial item pool around these six dimensions and conducted initial structured interviews and semi-open questionnaires for the target population [**?**].

Scale design language should also follow certain principles. The language used in item development should be as simple and clear as possible, avoiding professional jargon and double negatives, as these often confuse respondents. Item language should avoid social taboos and personal privacy to prevent respondent resistance that could interfere with the study. Additionally, language use must conform to the cultural norms of the respondents'region, with adjustments made when necessary. In the initial development of the Fear of Success Questionnaire, after completing content development, Chinese language experts were invited to evaluate the scale language, eliminating items with semantic repetition and ambiguity to obtain the initial scale.

2.3.1 Response Format

Response format selection typically proceeds concurrently with item pool generation. Researchers need to choose appropriate scoring systems and response formats based on actual circumstances and specific research purposes.

First, researchers must determine the response format for each question in the item pool—whether to use open-ended or closed-ended questions. Open-ended

questions require respondents to provide answers for each question, which is more difficult for both respondents and researchers, and the answers are often diverse, making coding and scoring challenging. The advantage of open-ended questions is that they can provide more ideas for researchers and are generally more suitable for initial surveys, but they are rarely used in finalized scales. Therefore, closed-ended questions are more commonly used in primary care research. Closed-ended questions provide specific options that are easier for respondents to answer, but this also creates other issues, such as whether answers should be single-choice or multiple-choice, and whether different answer options affect measurement results—these are all non-negligible issues in scale design research.

Single-choice questions are used in most scale design studies, but multiple-choice questions remain valuable because many questions do not have only one answer, and multiple-choice questions often provide more information about the issue. Sun Xinying et al. (2022) used item response theory to develop a diabetes functional health literacy scale containing 30 questions, three of which were multiple-choice, providing more information related to diabetes functional health literacy [**?**]. For scoring, Sun et al. assigned corresponding points for each correct option in multiple-choice questions, with "don' t know" scored as 0 [**?**]. However, this scoring method is relatively complex and can be affected by option settings. Generally, "select all correct options" questions may be difficult to "code" and score and should be avoided when possible [**?**]. Additionally, researchers should be cautious when setting options for closed-ended questions. For example, whether to include an "uncertain" option in scale options—Alsaffar used this option when translating a nutrition knowledge questionnaire [**?**]—but Folasire et al. questioned this practice [**?**], arguing that the "uncertain" option could lead those who understand the options well to avoid answering or choose to escape due to low confidence or laziness. Furthermore, researchers should avoid using "other" as an option category, though the decision to not provide an "other" option should only be made after carefully identifying almost all potential categories.

2.3.2 Scoring System

In a scale, the choice of scoring system often needs to be set in conjunction with specific items. Generally, when questions have clear right or wrong answers, correct choices are simply scored as 1 and incorrect choices as 0. However, in most cases, respondents cannot achieve absolute dichotomy, so the most commonly used scoring system in actual research is the Likert-type scoring system, such as 5-point, 7-point, or 9-point Likert scales. For example, Hu Haili et al. used a five-point scoring method when developing the Middle School Students' Psychological Resilience Scale, with five levels of "never," "occasionally," "sometimes,""often,"and "always"scored as 1, 2, 3, 4, and 5 respectively [**?**]. In attitude research, researchers prefer to use five levels of "strongly disagree," "somewhat disagree,""neutral,""somewhat agree,"and "strongly agree,"with scores still rang-

ing from 1 to 5. Both belong to the five-point Likert scale, while seven-point and nine-point scales further subdivide the options based on the five-point scale. So how should researchers choose the number of Likert scale points (e.g., 5-point, 7-point, 9-point) in research?

Berdie (1986) argued that when survey subjects have more knowledge and higher interest, the scale requires more attitude scale points, making seven-point or nine-point scales more appropriate than five-point scales because fewer attitude scale points result in greater skewness [**?**]. Additionally, even after data collection, different Likert scale point systems can still be converted. This conversion is achieved through Rasch models, which can systematically analyze the measurement characteristics of each option. By plotting Category Probability Curves (CPC), one can determine whether there is abuse or absence of option levels [**?**]. Using the French Tobacco Dependence Assessment Scale (FTND) from the 2021 China Family Health Index as an example, FTND item 1 reads: "How long after waking do you smoke your first cigarette? 60 minutes (Category 0), 31-60 minutes (Category 1), 6-30 minutes (Category 2), $ $5 minutes (Category 3)." Figure 1 [Figure 1: see original paper] shows the category probability curve for item 1, where each curve corresponds to an option, the horizontal axis represents the degree of tobacco dependence (increasing from left to right), and the vertical axis represents the probability of selection. For a subject with tobacco dependence of -4, the probability of selecting "Category 0" is about 95%, "Category 1" about 5%, and other options close to 0. Therefore, this subject is most likely to select "Category 0." Similarly, to the left of the intersection of Category 0 and Category 2, "Category 0" has the highest probability; between the intersection of Category 0 and Category 2 and the intersection of Category 2 and Category 3, "Category 2" has the highest probability; to the right of the intersection of Category 2 and Category 3, "Category 3" has the highest probability. We find that during measurement, the "Category 1" option has low usage, indicating Likert scale level abuse. According to Linacre' s recommendation, when Likert scale level abuse occurs, corresponding options should be considered for merging with adjacent options [**?**]. Therefore, Category 1 and Category 2 could be merged into 6-60 minutes. However, the scale after merging options still requires re-testing.

2.4 Pre-testing

Qualitative pre-testing is a critical stage in the development, translation, or revision of any questionnaire or psychometric tool. A small sample of respondents is selected for small-scale pre-testing to verify whether the target audience understands the items and options, and to evaluate from the respondents' perspective whether item wording is ambiguous. If issues such as semantic comprehension difficulties or unclear frameworks arise, items are modified and a new round of pre-testing is conducted until all respondents understand the item meanings and find the content acceptable [**?**].

Pre-testing mainly uses convenience sampling, selecting 30 or more samples

whenever possible to ensure stability and reliability of data analysis [**?**], and surveys the target population on questionnaire completion experience and comprehension. For example, Cheng Yanru et al. used convenience sampling to select 102 caregivers of disabled elderly individuals from three communities as pre-testing subjects when developing the Home Care Behavior Scale for Disabled Elderly Caregivers.

In the pre-testing phase, face validity assessment of the scale is required. Face validity examines whether the content of the assessment tool is consistent with its purpose from the respondents' perspective, though it is not a true validity indicator. In practical application, if the questionnaire' s measurement intent is obvious upon direct reading of the items, the questionnaire has high face validity. For example, a questionnaire measuring hand-washing conditions among nursing staff involving frequency, duration, and methods has face validity [**?**]. In primary care, researchers examining patient behaviors in health care domains or inquiring about specific conditions should increase scale face validity to ensure "what is answered is what is asked." However, in issues involving personal privacy or social image, excessively high face validity may lead to deception and concealment, so face validity settings should be determined based on specific research purposes.

2.5 Eliminating Items Through Item Analysis

In scale development within primary care, item analysis should be conducted after pre-testing. This step provides a basis for further scale revision and is a prerequisite for subsequent proper scale evaluation. The essence of item analysis is to explore differences in each item, test its quality, and revise or eliminate items based on certain criteria to ensure inter-item homogeneity and scale reliability. Researchers can examine items from three main aspects: item difficulty, item discrimination, and differential item functioning.

2.5.1 Item Difficulty

Item difficulty refers to the degree of difficulty encountered when completing test items, an indicator for evaluating respondents' performance. Higher correct response rates indicate lower difficulty. The purpose of setting test difficulty levels is to differentiate respondents as much as possible through the developed scale, maximizing the manifestation of respondent differences and demonstrating the scale' s discriminative power. As mentioned in Step 3, different scale types require different scoring systems. For non-dichotomous scoring items, difficulty can be calculated as the ratio of the average score of all respondents on an item to the item' s full score. For example, in a study on college students' health literacy, researchers recoded responses to multiple-choice questions into another proportion, re-evaluating items with correct values less than 0.2 or greater than 0.8 and considering deletion [**?**]. Overly high or low difficulty values affect score distribution and score dispersion. In practice, researchers should consider the scale' s nature and purpose to scientifically set reasonable difficulty thresholds.

Rasch models differ from CTT methods, emphasizing measurement objectivity and comparability. Therefore, for the difficulty indicator, the model states that item difficulty must be independent of the sample distribution—that is, the sampled population' s option selection is not influenced by item difficulty, while individual ability should also be independent of the difficulty distribution of test items. In other words, item difficulty does not change with sample variation and is not affected by respondents' ability levels. Consequently, Rasch measurement can provide interval-level scores for individual ability and item difficulty, placing individual ability levels and item difficulty levels on the same Logit scale for comparison, depicting a Person-Item Map (see Figure 2 [Figure 2: see original paper]). Figure 2 shows the Person-Item Map for the Life Satisfaction Scale. The black dots are mainly located between 0-2, meaning that the Life Satisfaction Scale items provide the most information for subjects with medium to high life satisfaction levels but are not suitable for assessing subjects with low life satisfaction levels. Different respondents and items are distributed on such a chart, providing researchers with more information. If calculated difficulty thresholds and means cluster around 0, this indicates moderate item difficulty. For example, in Hui Jianrong et al.' s quality analysis of a quality of life scale for stroke patients, statistical results showed all items' difficulty thresholds ranged from -0.32 to 0.67 (M=0.00, SD=0.34) [**?**], meaning all items had moderate acceptance levels and were relatively good. If item difficulty levels are too high or too low during scale development, this indicates that the behavior or dimension represented by the item does not occur frequently or is too difficult for respondents. Such scales often have higher accuracy only when targeting specific populations (with very high or very low levels).

2.5.2 Item Discrimination

The purpose of examining discrimination is to test whether the designed scale can truly distinguish between two different types of people as intended by the researcher. Methods mainly include the discrimination index method, correlation method, and CITC method.

The discrimination index calculation method is not complex. After calculating all respondents' total scores and ranking them by score, measurement theory generally divides the top and bottom 27% into high and low groups. Independent samples t-tests are conducted on each item' s scores between the two groups. Items showing no significant differences are considered separately and can be eliminated if necessary to ensure scale accuracy. Alternatively, the correlation coefficient between item scores and total test scores (PT-measure) can be used as a discrimination index. Larger correlation coefficients indicate higher discrimination, and items with poor correlation are considered for elimination. Corrected Item-Total Correlation (CITC) can also examine correlations between items within scale dimensions. Values greater than 0.5 indicate high correlation between the item and other items, while values below 0.5 may lead to consideration of deletion or revision after observing changes in Cronbach' s α coefficient.

For example, Hua Jing et al. used the discrimination index method to measure score differences between high and low groups across items in a study on children' s motor development, finding significant differences on 71 items and retaining all items at this stage [**?**]. Yang Zhen et al., when testing the reliability and validity of an elderly health promotion scale, found correlation coefficients between items and total scale scores ranged from 0.406 to 0.752 [**?**], showing moderate correlation (critical value 0.3), followed by further examination of each item combined with reliability coefficients.

In the Rasch model based on Item Response Theory, difficulty and discrimination are inseparable. At moderate difficulty levels, item discrimination is often highest. Therefore, item difficulty can also be seen in the Person-Item Map. In Figure 2, the bottom shows the Rasch scale, with measurement values increasing from left to right. For each respondent, the more rightward the position, the higher the life satisfaction. The bar height represents the number of respondents at that position. More concentrated respondent distribution indicates smaller scale discrimination, while more dispersed distribution indicates greater discrimination. In the figure, we can see that on five items, respondents' mastery levels are basically skewed and concentrated between 0 logit and 2 logit. This indicates poor scale discrimination on these five items, making it difficult to distinguish respondents with low life satisfaction. For example, Zhao Fuguo et al., when developing the Olweus Bullying Scale using Rasch models, found that difficulty distribution was very concentrated, resulting in poor differentiation of subjects with different bullying/victimization levels, especially difficulty in distinguishing high bullying/victimization groups [**?**].

2.5.3 Differential Item Functioning

Differential Item Functioning (DIF) refers to performance differences between two groups of respondents on an item, representing that the item has different statistical properties for different respondents. If the probability of correct response differs on the same item and reaches a certain threshold, the item is biased and requires further investigation of the source of difference [**?**]. Rasch models based on Item Response Theory tend to use statistical testing methods to calculate DIF. As the theoretical model' s influence has expanded, different scholars have proposed different calculation methods. The Mantel-Haenszel (M-H) test can be used to examine DIF caused by respondent characteristic variables, with differences greater than 0.5 and $p<0.05$ considered indicative of DIF [**?**]. For example, Du Haiyan and Li Fupeng found items 9, 39, and 58 showed moderate or severe DIF when applying the M-H method [**?**]. Lord' s chi-square test can also be used, employing R software for DIF testing, where X^2 is the DIF indicator, and $X^2 > 13$ with $p>0.05$ indicates DIF [**?**]. For example, Gao Shuang and Zhang Xiangkui used Lord' s chi-square test when applying Rasch models to analyze the Rosenberg Self-Esteem Scale, finding items 1 and 5 showed functional differences—that is, on these two items, gender differences led to different self-esteem levels [**?**]. For multi-level scoring items, ANOVA can

also be used for testing. For instance, in the development of the WHO Disability Assessment Schedule, researchers found different difficulty levels between gender groups and used ANOVA to compare gender and other potential DIF-producing items to identify inappropriate items for modification [**?**].

It is worth noting that the three major aspects of item analysis are not required to be used entirely in scale development but should be selected based on scale characteristics—whether the scale is single-choice or multiple-choice, dichotomous or multi-level scoring, and what the nature of the developed scale is. Whether problematic items found during item analysis should be eliminated cannot be generalized. Simply deleting items with excessive difficulty, poor discrimination, or poor fit is not advisable, as overly perfect models are difficult to exist in reality. They are only ideal assumptions and guidance, and decisions should be made based on comprehensive consideration of multiple indicators.

2.6.1 Initial Evaluation Based on Classical Test Theory

Classical Test Theory (CTT), also known as true score theory, became well-established in the 1950s. This theory holds that test scores X consist of true scores T and random error E, i.e., X=T+E, where the mean of error E is zero and the correlation between T and E is zero. Based on this, measurement indicators for test items were established, such as reliability, validity, difficulty, and discrimination, to screen test items, build item banks, and construct tests [**?**]. Previous sections have detailed how to use difficulty and discrimination to screen test items. This section introduces how to use CTT to complete initial test evaluation, namely conducting exploratory factor analysis and reliability and validity analysis.

2.6.1.1 Exploratory Factor Analysis

Exploratory Factor Analysis (EFA), as a CTT technique, has been widely used in scale design and development within primary care. EFA mainly uses mathematical methods to explore variables or factors in scales to determine specific dimensions and which items belong to which dimension. Next, we detail the EFA process. We believe EFA should include the following four key steps (see Figure 3 [Figure 3: see original paper]).

(1) Determine Variables and Sample
Determining variables and sample is preparatory work before data analysis, crucial for the entire study. This stage requires researchers to compile or collect as many items related to their research topic as possible based on previous research and theory, sometimes even including some items unrelated to the topic, because after EFA screening, remaining items are often much fewer than original items. How to decide which items to retain is also a concern for researchers. Common criteria include factor loading, item communality, and cross-factor loading. Generally, factor loading >0.71 in the component matrix is considered excellent, >0.63 very good, >0.55 good, >0.45 fair, and >0.32 poor [**?**]. Item

communality should not be too low, generally considered not less than 0.30 [?]. The same item should not have high loadings on two factors. For example, Chen Gui et al. eliminated items with similar loadings on different factors that were difficult to interpret [?]. Before factor analysis, sample size must also be considered. Factor analysis sample size should not be too low, otherwise results lack persuasiveness. Corsuch recommends a sample-to-variable ratio of 5:1, with a minimum sample size of 100. Nunnally recommends a sample-to-variable ratio of 10:1 [?].

(2) Determine Whether EFA Can Be Conducted

The purpose of EFA is to simplify data or identify the basic data structure of the scale. Researchers currently widely use principal component analysis for EFA. Before conducting EFA, it is necessary to ensure that theoretical and statistical assumptions for factor analysis are met. Factor analysis theoretical assumptions hold that a latent structure indeed exists in the variable set, while statistical assumptions require strong correlations between observed variables. Therefore, before EFA, the following conditions must be met: inter-item correlation >0.3, significant Bartlett's test of sphericity ($p<0.05$), and Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (MSA) of at least 0.6 [?]. Inter-item correlation >0.3 requires researchers to calculate correlations for all items; if all or most correlations are <0.3, EFA is not suitable. The same applies to sphericity and sampling adequacy tests. For example, Guo Jing conducted KMO and Bartlett' s tests when revising the Chinese version of the Psychological Vulnerability Questionnaire, showing KMO=0.89 and Bartlett's test ^{2}/df=25.31, $p<0.001$ [?]. Note that these parameters only indicate that factor analysis can be conducted, not that results are good.

(3) Determine Number of Factors

Determining the factor structure of selected variables and how many factors to retain is a critical step in EFA. Retaining too few or too many factors causes problems, but empirical research tends to retain more factors because over-extraction yields more accurate factor loading estimates than under-extraction. Researchers have proposed various testing methods to aid decision-making, mainly including: Eigenvalue >1, also known as the K1 rule, one of the most commonly used standards; Total variance explained, also developed based on principal component analysis. There is no unified standard for how much total variance factors should explain, with some researchers arguing it should not be less than 50% [?]. For example, Table 1 shows factor analysis results for the 8-item General Self-Efficacy Scale, where only one eigenvalue >1 (5.753>1), leading researchers to conclude the scale is unidimensional with only one factor. The table also shows the factor' s variance explanation (71.91%>50%), meaning this factor can explain 71.91% of variance in general self-efficacy and well reflects the construct; Scree plot, which provides a graphical representation of factor number and eigenvalue magnitude. Researchers simply select the factor number corresponding to the inflection point in the scree plot provided by EFA. This method is simple, convenient, and more intuitive. Figure 4 [Figure 4: see original paper] shows the scree plot for general self-efficacy. The figure shows

a huge turning point starting from the first component, so the first component can be considered the inflection point, indicating the scale contains only one factor.

(4) Factor Rotation
After determining the number of factors, the next step is to determine the factor rotation method. Factor rotation methods can be divided into two categories: oblique rotation and orthogonal rotation. Unlike oblique rotation, orthogonal rotation assumes no correlation between factors, while oblique rotation does not have this assumption. In empirical research within primary care, factors often have greater or lesser correlations, so oblique rotation is more objective. However, most published studies use orthogonal rotation, whose results are more conducive to researchers' interpretation of factor structure but can also mislead research conclusions. Therefore, we believe future researchers should first use oblique rotation, only considering orthogonal rotation if correlations between factors are found to be small or nonexistent. Table 2 shows results using the promax oblique rotation method, indicating the scale contains two factors, with J1, J2, J3, J4, J5, J7, J8 belonging to factor 1, and J6, J9, J10 belonging to factor 2.

2.6.1.2 Reliability Analysis

After eliminating items through EFA, the formal scale is finalized. At this point, the data should be used to test the formal scale' s reliability. Reliability refers to the stability of measurement results. If a person' s same trait can be measured repeatedly with the same measurement tool, the degree of agreement among various measurements is called reliability, sometimes also referred to as measurement reliability. In CTT, reliability methods typically include alternate-forms reliability, test-retest reliability, homogeneity reliability, split-half reliability, and inter-rater reliability. In clinical research, alternate-forms reliability is difficult to obtain and thus rarely used. Researchers prefer test-retest reliability, split-half reliability, and homogeneity reliability.

(1) Test-Retest Reliability
In scale design research, cross-time consistency of the scale is an important indicator for measuring tool reliability. Therefore, when developing and designing scales in primary care, it is necessary to report the consistency of results obtained from administering the scale twice to the same group of subjects, which can be expressed using the Pearson product-moment correlation coefficient between the two tests. For example, Liu Lei et al. reported a test-retest reliability of 0.883 for the Chinese version of the Exercise Psychological Needs Satisfaction Scale for the Elderly they developed, with three-dimensional test-retest reliability coefficients between 0.829 and 0.876 [**?**]. For test-retest reliability, generally accepted evaluation standards are: 0.65-0.70, minimum acceptable value; 0.70-0.80, quite good; 0.80-0.90, very good [**?**]. Therefore, the test-retest reliability of the scale developed by Liu Lei et al. is good. However, Liu Lei et al. did not report the interval between the two administrations, which is also an important

factor affecting test-retest reliability and should be noted in future research, as test-retest reliability can vary with different time intervals for the second measurement.

(2) Alternate-Forms Reliability
Designing two parallel tests to measure the same group of subjects yields alternate-forms reliability, which can be expressed using the Pearson product-moment correlation coefficient between the two parallel tests for the same group. Alternate-forms reliability is also an indicator of scale reliability, but due to the time and effort required to design parallel tests and the difficulty ensuring consistency in content and structure between the two tests, it is not widely used in measurement. Liu Aimei and Liu Yuanbin used this reliability when developing a health knowledge, attitude, and practice questionnaire for sudden deafness patients, using questionnaires with similar content and response format as parallel tests, finding alternate-forms reliability of 0.88 for the health knowledge section [**?**]. Evaluation standards for alternate-forms reliability are basically consistent with test-retest reliability [**?**], so this scale has good alternate-forms reliability.

(3) Split-Half Reliability
Also called internal consistency coefficient, researchers need to divide a complete test into equivalent halves and compare the consistency of test scores for subjects on the two newly obtained halves. Split-half reliability is one of the most commonly used reliabilities in current research, requiring only simple operations in SPSS statistical software to calculate.

(4) Homogeneity Reliability
Researchers can obtain homogeneity reliability by measuring the consistency among all items within the test, i.e., the internal consistency coefficient. Researchers generally use Cronbach' s alpha coefficient to measure a test' s internal consistency. Alpha coefficient is currently the most used reliability in research. Like split-half reliability, researchers only need simple operations in SPSS to calculate the alpha coefficient. Wu Minglong points out that alpha coefficient should preferably be above 0.80, with 0.70-0.80 being acceptable; subscales should preferably be above 0.70, with 0.60-0.70 being acceptable [**?**].

(5) Inter-Rater Reliability
When multiple raters score the same batch of answer sheets, the consistency of scores yields inter-rater reliability. Its magnitude equals Kendall' s coefficient of concordance between one rater' s set of scores and another rater' s set. Kendall' s coefficient of concordance is a measure of correlation for multi-column rank data, commonly used to evaluate consistency among multiple raters.

2.6.1.3 Validity Analysis

When conducting scale design research in primary care, the validity of the developed test should also be examined. Validity is the degree to which a test or scale can measure what it intends to measure. The theoretical definition of

validity is the ratio of true variation (effective variation caused by measured changes) to total variation (true variation) in a series of measurements related to the measurement purpose. Test validity can be divided into content validity, construct validity, and empirical validity.

(1) Content Validity

Content validity involves detailed, systematic judgment by relevant experts on the 吻合度 between the assessment tool' s items and content scope. The qualifications and professional scope of participating experts are basic guarantees for content validity assessment quality. For example, Cui Chuyun et al. selected six nursing experts (nursing professors from schools and hospitals, nursing department directors, and clinical nursing experts) to evaluate scale content validity, as selecting professors or clinical experts in the research field is the most common choice for content validity evaluation [**?**]. Additionally, quantitative assessment of content validity in item screening includes various indicator calculations, among which the Content Validity Index (CVI) is widely used due to its simple calculation, ease of understanding and communication, and ability to correct for random consistency: Item-level CVI (I-CVI) can evaluate each item' s content validity; Scale-level CVI (S-CVI) measures the entire scale' s content validity. For example, after preliminary development of a coronary heart disease patient secondary prevention medication adherence questionnaire, researchers prepared an expert rating form based on a Likert 4-point scale, with options set as irrelevant, relevant only if modified, very relevant but needs modification, and highly relevant, scored 1-4 respectively, distributed to experts for completion. After collection, I-CVI and S-CVI were both calculated as 1.00 [**?**], indicating good content validity.

(2) Construct Validity

The degree to which a test actually measures the intended theory and trait is the scale' s construct validity, representing how well a scale can explain some structure or trait of test theory. In empirical research, researchers can generally measure a scale' s construct validity through item analysis, exploratory factor analysis, and Confirmatory Factor Analysis (CFA). Item analysis examines correlations between scale items and their dimensions and between dimensions to test associations and independence among scale dimensions. For example, Yang Li et al. used item analysis to measure construct validity in a cognitive style questionnaire, showing correlations between items and their dimensions above 0.55, mainly distributed between 0.56 and 0.75, indicating good item discrimination. The four dimensions of the cognitive style questionnaire had moderate correlations, showing they were related yet relatively independent [**?**]. Exploratory factor analysis is basically as described in the previous section, except no items are deleted this time. Generally, questionnaires formed through EFA should collect new data to measure construct validity using EFA or CFA. For example, Wu Yibo et al. used AMOS software for CFA to test model fit when examining reliability and validity of the Chinese version of the Duke Anticoagulation Satisfaction Scale (DASS), finding all indicators showed good fit for the four-factor DASS model (CMIN/DF = 1.825 < 5, GFI = 0.854 > 0.85, CFI = 0.938 > 0.9,

RMSEA = 0.066 < 0.08, NFI = 0.875 < 0.9, TLI = 0.921 > 0.9), indicating good construct validity [**?**].

(3) Empirical Validity

If a test can effectively estimate subjects' behavior in specific situations, the test has good empirical validity or criterion-related validity. Criterion validity can be measured through correlation methods, differentiation methods, and hit rate methods, with correlation methods still being most commonly used in primary care scale design research. The correlation method measures the correlation between test scores and validity variables. The calculated correlation coefficient is the validity coefficient, whose square is validity. For example, You Yongheng et al. selected the General Well-Being Scale (GWB) as a criterion to test concurrent validity of the Beck Depression Inventory, requiring completion of the criterion scale when distributing the depression scale. Results showed significant correlations between all dimensions and total scores of general well-being and depression total scores (P<0.001), indicating good criterion validity for the BDI scale [**?**].

2.6.2 Initial Evaluation Based on Rasch Model

The Rasch model is a fundamental measurement model that measures latent traits through individual performance on items. The basic principle of the Rasch model is that an individual' s specific performance on a specific item is measured by the person' s ability and the item' s difficulty, so response quality depends entirely on individual ability and item difficulty. The Rasch model is an idealized mathematical model that proposes two requirements for objective measurement: (1) for any item, individuals with higher ability should have a greater probability of correct response than those with lower ability; (2) any individual should perform better on easy items and worse on difficult items [**?**]. Although Rasch models have been developed for decades, they still have not received sufficient attention, especially in primary care. Searching "Rasch" as a theme in CNKI (1915-2022) found only 160 core journal articles, with research from the past five years (2017-2021) accounting for as high as 46.25%. This means Rasch models have gradually attracted more researchers' attention in recent years, but these studies remain mainly concentrated in psychology and education, with only a few articles related to primary care. Therefore, conducting Rasch model research in primary care is very necessary.

2.6.2.1 Unidimensionality Test

Item Response Theory (IRT) is a mathematical formulation about the relationship between individuals' probability of answering questions and latent traits, another classic theory in measurement distinct from CTT. Common IRT models include one-parameter, two-parameter, and three-parameter models [**?**]. As a special case of the IRT one-parameter model, Rasch model usage has a prerequisite: the scale must be unidimensional. Unidimensionality means that only one latent trait influences respondents' answers during measurement. It is important

to note that one latent trait does not mean the scale can only have one dimension, as long as all dimensions in the scale point to the same trait. For example, Chen Yuanyuan et al., when translating a nutrition literacy assessment tool, found the tool contained six subscales, but items in the subscales all pointed to the nutrition literacy trait, so they conducted Rasch analysis on both subscales and the full scale [**?**]. Rasch model residual principal component analysis (PCA) is generally used to test scale unidimensionality. According to Raiche' s recommendation, if the first factor' s residual standardized eigenvalue is between 1.4 and 2.1, the data can be considered to meet unidimensionality requirements and be suitable for Rasch models [**?**]. For example, Chen Yuanyuan et al. found during unidimensionality testing that the first component residual eigenvalues for subscales 1-6 ranged between 1.6-1.8, and the total scale' s first component residual eigenvalue was 3.1, meaning the scale was suitable for Rasch analysis [**?**].

2.6.2.2 Model Fit

From the Wright map, we learn that Rasch models can estimate item difficulty and respondent ability levels. By comparing actual observed scores with the theoretical probability of each respondent answering correctly on each item, Rasch model fit can be evaluated. Rasch models typically require calculating two fit statistics: Weighted Mean Square Fit Statistic (Infit MNSQ) and Unweighted Mean Square Fit Statistic (Outfit MNSQ). Infit MNSQ and Outfit MNSQ close to 1 indicate good model fit. Generally, when data fit well, Outfit and Infit MNSQ are between 0.5 and 1.5 [**?**]. Using the Life Satisfaction Scale as an example, we collected 569 data points and used R for model fit testing, with results shown in Table 3 . The table shows all item parameters are basically within acceptable ranges, indicating good data-model fit. Item 5 ("If I could live my life over, there is almost nothing I would want to change," 1=disagree, 2=somewhat disagree, 3=neutral, 4=somewhat agree, 5=agree) has Outfit MNSQ and Infit MNSQ values of 1.52 and 1.40 respectively, both greater than 1.0. This means respondents with high life satisfaction chose low scores (disagree and somewhat disagree), while those with low life satisfaction chose high scores (agree and somewhat agree). Therefore, item 5 has large error in distinguishing respondents' life satisfaction and requires further consideration of whether to retain it.

Additionally, good items or scales should provide more information for testing and reduce errors in estimating subjects' trait levels. IRT holds that scales provide the most precise measurement results when testing subjects whose trait levels match the scale. In research, test information curves are generally used for measurement, reflecting the degree of accurate evaluation the entire scale can provide when subjects with different characteristic levels complete all items. Item difficulty can be seen on the horizontal axis, representing subjects' trait levels, with each scale representing one logit unit, and the vertical axis representing information amount, i.e., the Fisher information function [**?**]. Figure 5 [Figure

5: see original paper] shows the test information curve for the Life Satisfaction Scale, with the upper half showing each item' s test information curve and the lower half showing the total scale's test information curve. Overall, the scale has highest accuracy when life satisfaction estimates are between 0 and 2, providing maximum information for subjects with medium and high life satisfaction. For example, Gao Shuang and Zhang Xiangkui found after calculating the Fisher information function that self-esteem estimates between 0 and -2 can provide the highest measurement precision, offering the most information for medium and low self-esteem subjects [**?**].

2.6.2.3 Reliability

Rasch models use Person Separation Reliability (PSR) to measure scale reliability. Separation reliability can be obtained by calculating the ratio of "true" variation produced by individuals to total variation, typically used to examine the reliability of subjects' ratings on items [**?**]. Rasch model overall reliability is obtained by calculating explanatory power at the individual level, with values from 0 to 1. Generally, reliability indicators above 0.7 are acceptable, and above 0.8 are good [**?**]. The calculated reliability value for the Life Satisfaction Scale is 0.80, indicating good reliability.

2.7 Re-evaluation of the Scale

From Step 1 to Step 6, a scale is basically finalized. However, since scale item screening and reliability and validity testing all use the same sample, whether the scale has cross-sample and cross-time consistency remains unknown. Therefore, researchers should use the formal scale to collect a new sample and test the scale' s reliability and validity on the new sample. Note that if researchers need to test the scale' s test-retest reliability, the second batch of scale subjects should include some of the first batch. Since reliability and validity analysis content has been elaborated in the previous section, researchers only need to use the same methods for re-testing, so we will not elaborate further. Here, we only elaborate on using Confirmatory Factor Analysis (CFA) to test scale construct validity under CTT.

CFA refers to hypothesis testing conducted under the premise of clear 隶属 relationships between observed indicators and latent factors, and is theory-driven analysis. After EFA, we have clarified the formal scale's factor structure. Therefore, new data can be used to construct a CFA model to test scale construct validity. Based on the output results' fit, consideration can be given to whether model modification is needed. Main fit indices include chi-square/degrees of freedom ratio (^{2}/df), Goodness-of-Fit Index (GFI) and Adjusted Goodness-of-Fit Index (AGFI), Root Mean Square Error of Approximation (RMSEA), Normed Fit Index (NFI), Incremental Fit Index (IFI), Relative Fit Index (RFI), Comparative Fit Index (CFI), and Tucker-Lewis Index (TLI). Parameter fit standards are: ^{2}/df<2 (some researchers consider ^{2}/df<3) indicates good model fit [**?**]; RMSEA<0.08 means the model is acceptable [**?**]; AGFI and GFI should both

be >0.90, indicating good match between model and data [**?**]; NFI, RFI, IFI, TLI, CFI should all be >0.90 [**?**]. If these fit values do not meet good standards, researchers should consider model modification, specifically using MI values presented in AMOS reports to release relationships between two test error variables, i.e., establishing covariance relationships between them [**?**], thereby optimizing the model.

3 Discussion

Scale design methods have been fully applied in primary care, mainly reflected in the breadth of scale design research usage. Currently, most studies involve scale usage, so whether a scale' s design and development are reasonable determines whether the research is reliable. However, many non-standard aspects remain in scale design research, such as poor reliability and validity, lack of key steps, and statistical errors. Overall, conducting scale design research in primary care requires strict adherence to the above standard procedures, which can to some extent solve non-standard usage of steps and statistical methods during research. Of course, certain essential skills are also needed to better master this method.

Essential skills for scale design research mainly include theoretical guidance and statistical testing. Theoretical guidance is a top-down process driven by theory. It requires researchers to read extensive relevant literature before and during scale development to understand the structure of the trait to be measured and existing theories and scales. Only on the basis of these mature previous experiences can the validity of the developed scale be ensured as much as possible. Statistical testing is a bottom-up process driven by data. It can help researchers better discover problems in item development and is also an important reference for screening poor items. Researchers use statistics to test scale reliability and validity to ensure the objectivity and effectiveness of the measurement tool. In summary, theoretical guidance and statistical testing are two essential skills in scale design research. Only by combining these two well and considering them from both bottom-up and top-down perspectives can the reliability of the developed measurement tool be maximally ensured.

This study systematically elaborates on how to conduct scale design in primary care, but due to space and professionalism limitations, some clinicians may find it difficult to understand professional terminology in the text. Moreover, for most general practitioners, selecting an appropriate scale may be more directly effective than designing one. To this end, we provide explanations for some professional vocabulary in the text and relevant suggestions on how general practitioners should select scales in the appendix. Additionally, this study provides researchers with a list of learning resources for further in-depth study of scale design methods, see Table 4 . Overall, researchers should strictly follow standard procedures when conducting scale design, referring to relevant materials in the list for specific steps to ensure the objectivity and effectiveness of the designed scale.

4 Conclusion

In summary, we have outlined practical steps and statistical methods for researchers interested in developing or designing scales in primary care. We recommend that all scale design in primary care should consider the methods described in this review. Researchers should strictly follow standard scale development procedures and comprehensively use Rasch models and factor analysis methods to make measurement results more objective.

Author Contributions: WANG Fei proposed the research direction, was responsible for data processing, and wrote the initial draft; TANG Jingqi participated in writing the initial draft and conducted data management; SUN Xiaonan was responsible for manuscript revision; SUN Xinying provided critical suggestions; LI Jun revised and improved the article from a general practitioner' s perspective; MENG Xingxing and WU Yibo guided the research throughout, were responsible for quality control and review, and took overall responsibility for the article; all authors confirmed the final manuscript.

Conflict of Interest: This article has no conflict of interest.

Acknowledgments: We thank Associate Professor GAO Zhiqiang from the School of Philosophy at Anhui University for guidance in psychometrics, as it was his psychometrics course that introduced us to this field early on. We also thank all surveyors who participated in the 2021 Family Health Index survey, as their participation provided the substantial data supporting relevant charts in this article.

References

[1] WANG Ronghua, WANG Suping. Research on the Role of General Practitioners in China' s Medical and Health Service System [J]. Chinese General Practice, 2020, 23(04): 388-394+402.

[2] TRAKMAN G L, FORSYTH A, HOYE R, et al. Developing and validating a nutrition knowledge questionnaire: key methods and considerations [J]. Public Health Nutrition, 2017, 20(15): 2670-2679.

[3] KOUVELIOTI R, VAGENAS G. Methodological and statistical quality in research evaluating nutritional attitudes in sports [J]. International journal of sport nutrition and exercise metabolism, 2015, 25: 624-635.

[4] JIN Yingtong, CHEN Suqin, BAO Yinxin, et al. Research Progress on Assessment Tools for Autism Spectrum Disorder in Children [J]. Nursing Practice and Research, 2021, 18(09): 1325-1329.

[5] BOND T G, FOX C M. Applying the Rasch model. Fundamental measurement in the human sciences (3rd ed.) [M]. New York: NY. Routledge, 2015.

[6] WANG F, WU Y, SUN X, et al. Reliability and validity of the Chinese version of a short form of the family health scale [J]. Bmc Primary Care, 2022, 23(1).

[7] GAO Zhiqiang, ZHANG Tengxiao. Development and Application of the Fear of Success Questionnaire [J]. Chinese Journal of Clinical Psychology, 2011, 19(05): 602-605+86.

[8] SUN Xinying, ZHU Xiaorou, GONG Litong. Evaluation of Diabetes Functional Health Literacy Scale Based on Item Response Theory [J]. Chinese Journal of Health Education, 2022, 38(01): 18-22.

[9] ALSAFFAR A A. Validation of a general nutrition knowledge questionnaire in a Turkish student sample [J]. Public Health Nutrition, 2012, 15(11): 2074-2085.

[10] FOLASIRE O F, AKOMOLAFE A A, SANUSI R A. Does Nutrition Knowledge and Practice of Athletes Translate to Enhanced Athletic Performance? Cross-Sectional Study Amongst Nigerian Undergraduate Athletes [J]. Global journal of health science, 2015, 7(5): 215-225.

[11] HU Haili, ZHANG Hongbo, WANG Jun, et al. Development and Preliminary Evaluation of Middle School Students' Psychological Resilience Scale [J]. Chinese School Health, 2009, 30(12): 1097-1099.

[12] BERDIE D R. The optimum number of survey research scale points: what respondents say. the meeting of the American Educational Research Association [C]. San Francisco: CA.F, 1986.

[13] ZHAO Fuguo, HE Zhuang, YUAN Shuli, et al. Rasch Model Analysis of the Olweus Bullying Scale [J]. Journal of Southwest University (Social Sciences Edition), 2020, 46(05): 115-121.

[14] LINACRE J M. Optimizing rating scale category effectiveness [J]. Journal of applied measurement, 2002, 3(1): 85-106.

[15] PERNEGER T V, COURVOISIER D S, HUDELSON P M, et al. Sample size for pre-tests of questionnaires [J]. Quality of Life Research, 2015, 24(1): 147-151.

[16] Surface Validity [J]. Chinese Nursing Management, 2016, 16(07): 905.

[17] RABIN L A, MILES R T, 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) [J]. Psychiatry Research, 2021, 300.

[18] HUI Jianrong, PEI Jian, WANG Yuanchun, et al. Rasch Analysis of Acupuncture Intervention Quality of Life Scale for Stroke Patients [J]. Chinese Acupuncture, 2013, 33(4): 363-366.

[19] HUA Jing, ZHANG Lijun, GU Guixiong, et al. Preliminary Development of a Home Environment Scale for Motor Development of Urban Preschool Children [J]. Chinese School Health, 2011, 32(2): 161-163.

[20] YANG Zhen, ZHANG Huijun. Cross-Cultural Adaptation and Reliability and Validity Testing of the Elderly Health Promotion Scale [J]. Journal of Nursing, 2021, 36(19): 91-94.

[21] LAI J S, CELLA D, CHANG C H, 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 [J]. Quality of Life Research, 2003, 12(5): 485-501.

[22] ZWICK R, THAYER D T, LEWIS C. An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis [J]. Journal of Educational Measurement, 1999, 36(1): 1-28.

[23] DU Haiyan, LI Fupeng. Analysis of Sample Size Impact on Mantel-Haenszel Method for DIF Effect Testing [J]. Examination Research, 2016, (05): 55-62.

[24] CHOI S W, GIBBONS L E, CRANE P K. lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations [J]. Journal of Statistical Software, 2011, 39(8): 1-30.

[25] GAO Shuang, ZHANG Xiangkui. Application of Rasch Model in Analyzing Rosenberg Self-Esteem Scale [J]. Psychological Exploration, 2018, 38(05): 445-450.

[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) [J]. Quality of Life Research, 2021, 30(10): 2929-2938.

[27] YANG Zhiliang, HAO Xingchang. Psychology Dictionary [M]. Shanghai: Shanghai Dictionary Publishing House, 2016.

[28] COMREY A L. A first course in factor analysis. [M]. New York: Academic Press, 1973.

[29] CAO Chengxu, QI Shisan, JIN Tonglin. Development of College Students' Impression Management Efficacy Scale [J]. Modern Preventive Medicine, 2021, 48(17): 3199-3201+225.

[30] CHEN Gui, CAI Taisheng, HU Fengjiao, et al. Revision of the Emotional Eating Scale in Chinese Adolescents [J]. Chinese Journal of Clinical Psychology, 2013, 21(4): 572-575+88.

[31] WANG Mengcheng. Latent Variable Modeling and Mplus Application · Basic Volume [M]. 1 ed. Chongqing: Chongqing University Press, 2013.

[32] GUO Jing, WANG Ying, SONG Yuping, et al. Preliminary Revision of Chinese Version of Psychological Vulnerability Questionnaire and Analysis of Its Reliability and Validity in Community Residents [J]. Chinese Public Health, 2019, 35(02): 129-133.

[33] FOLYD F J, WIDAMAN K F. Factor analysis in the development and refinement of clinical assessment instruments [J]. Psychological Assessment, 1995, 7: 286-299.

[34] LIU Lei, LIU Huaping, GUO Hong, et al. Reliability and Validity Testing and Applicability Analysis of Chinese Version of Exercise Psychological Needs Satisfaction Scale for the Elderly [J]. Chinese General Practice, 2021, 24(05): 619-624.

[35] JIAN Xiaozhu, DAI Buyun. SPSS23.0 Statistical Analysis in Psychology and Education [M]. Beijing: Beijing Normal University Publishing House, 2017.

[36] LIU Aimei, LIU Yuanbin. Development and Reliability and Validity Testing of Health Knowledge, Attitude and Practice Questionnaire for Sudden Deafness Patients [J]. Journal of Audiology and Speech Pathology, 2012, 20(5): 444-448.

[37] CUI Chuyun, YUE Meng, LI Yufeng, et al. Study on Reliability and Validity of Chinese Version of Health Behavior Scale [J]. Journal of Nursing, 2017, 32(12): 62-65.

[38] ZHANG Wanyu, ZHOU Yi, CUI Shanshan. Development and Reliability and Validity Testing of Coronary Heart Disease Patient Secondary Prevention Medication Adherence Questionnaire [J]. Nursing Research, 2022, 36(6): 1004-1007.

[39] YANG Li, ZHAI Ruilong, QI Zhenya, et al. Development of Cognitive Style Questionnaire for Chinese Adults in Mental Health Quality Assessment System [J]. Psychological and Behavioral Research, 2012, 10(05): 332-339.

[40] WU Y, DONG S, LI X, et al. The Transcultural Adaptation and Validation of the Chinese Version of the Duke Anticoagulation Satisfaction Scale [J]. Frontiers in Pharmacology, 2022, 13.

[41] YOU Yongheng, YU Shaoping, LIANG Bin. Trial Use and Evaluation of Beck Depression Inventory Among Teachers in Disaster Areas [J]. Journal of Sichuan Normal University (Natural Science Edition), 2011, 34(03): 439-442.

[42] YAN Zi. Objective Measurement in Psychological Science—Characteristics and Development Trends of Rasch Model [J]. Advances in Psychological Science, 2010, 18(8): 1298-1305.

[43] CHEN Yuanyuan, YANG Chunjun, WANG Dongmei, et al. Translation and Reliability and Validity Study of Nutrition Literacy Assessment Tool in Diabetic Patients—Analysis Based on CTT and Rasch Model [J]. Chinese General Practice, 2020, 23(26).

[44] ZHAO Shouying, HE Feixia, LIU Yan. Application of Rasch Model in Academic Test Quality Analysis [J]. Educational Research and Experiment, 2013, (01): 87-91.

[45] BAGOZZI R P, YI Y. On the evaluation of structural equation models [J]. Journal of the Academy of Marketing Science, 1988, 16(1): 74-94.

[46] ZHANG Lihua, GU Ying, HUANG Miao, et al. Confirmatory Factor Analysis of Evidence-Based Practice Readiness Assessment Scale [J]. Nursing Journal of Chinese People' s Liberation Army, 2019, 36(02): 6-10+25.

Note: Figure translations are in progress. See original paper for figures.