sustainability

MDPI

Article

Evaluating a National Traditional Chinese Medicine Examination via Cognitive Diagnostic Approaches

Lingling Xu [1], Zhehan Jiang [1,*] and Yuting Han [2]

[1] Institute of Medical Education, Peking University, Beijing 100191, China; linglingxu@bjmu.edu.cn
[2] School of Psychology, Beijing Language and Culture University, Beijing 100083, China; hanyuting716@gmail.com
* Correspondence: jiangzhehan@gmail.com

Abstract: The current research utilized diagnostic classification models (DCMs), an advanced psychometric theory, to evaluate the examination's quality using psychometric methods for a more precise and comprehensive understanding of health professionals' competence. Data was gathered from 16,310 fourth-year Traditional Chinese Medicine undergraduates who completed the Standardized Competence Test for Traditional Chinese Medicine Undergraduates (SCTTCMU) comprising 300 multiple-choice items. The study examined the fundamental assumptions, model-data fit, and cognitive diagnostic theory models' item and test properties. The generalized deterministic input, noisy, "and" gate model applied in this research demonstrated a strong alignment with the real response data, meeting all the necessary assumptions. Cognitive diagnostic analysis indicated that all items exhibited satisfactory psychometric characteristics, and the reported scores offered insights into candidates' proficiency in cognitive skills. It is expected that the advent of modern psychometric technology will contribute to the improvement of refined diagnostic information for health professional candidates. Furthermore, this research holds the potential to significantly enhance sustainability in healthcare practices, knowledge, economics, resource use, and community resilience.

Keywords: traditional Chinese medicine; health professions education; cognitive diagnostic analysis; assessment

Citation: Xu, L.; Jiang, Z.; Han, Y. Evaluating a National Traditional Chinese Medicine Examination via Cognitive Diagnostic Approaches. *Sustainability* **2024**, *16*, 5400. https://doi.org/10.3390/su16135400

Academic Editor: Grigorios L. Kyriakopoulos

Received: 21 April 2024
Revised: 5 June 2024
Accepted: 11 June 2024
Published: 25 June 2024

1. Introduction

Various professionals in health professions education offer a wide array of health services in various settings. Formative assessment plays a key role in facilitating learning, while summative assessment serves as the definitive evaluation of a candidate's skills [1]. In the realm of health professions education, summative assessments are utilized to gauge candidates' advancement through their courses and/or training programs [2]. Establishing competency and identifying qualifications for advanced training necessitates that summative assessment scores accurately forecast candidates' future performance.

In the last few decades, classical test theory (CTT), generalizability theory (G-theory), and item response theory (IRT) have become the dominant psychometric frameworks used in carrying out summative assessments [3,4]. CTT focuses on differentiating between the variances of "true score" and "measurement error". By independently estimating these variances, it acknowledges the indirect nature of educational and psychological assessments [5]. G-theory allows for a comprehensive view of measurement error and its various elements, allowing researchers to break down measurement error into distinct sources [6]. Furthermore, as the foundation of contemporary psychometric methods, IRT approaches can offer insights into individual latent traits and item features [7]. Nevertheless, while there have been significant developments in psychometric techniques, there remains a heated discussion regarding the effectiveness and suitability of these methods in health professions education. Particularly, it is argued that the initial purpose and

structure of these evaluations do not offer detailed qualitative diagnostic insights, nor do they allow for personalized performance adjustments. Consequently, there is a demand for precise and top-notch assessment instruments in conclusive evaluations to enable personalized evaluation. Cognitive diagnostic evaluations have emerged as a promising approach in this regard.

Cognitive diagnosis is currently receiving significant interest from researchers and practitioners in health professions education, primarily due to its ability to diagnose individual traits and offer personalized treatment effectively [8]. Cognitive diagnostic assessments [9] are designed to offer diagnostic insights by providing detailed reports on an individual's traits [10]. A crucial feature of cognitive diagnostic assessments is their integration of cognitive psychology and cognitive diagnostic models (CDMs) in a unified framework, allowing researchers to evaluate both broad diagnostic data and specific criteria-level information about individuals within a specific assessment area [11].

This paper aims to use a cognitive diagnostic approach to evaluate the effectiveness of the Standardized Competence Test for Traditional Chinese Medicine Undergraduates (SCTTCMU) for fourth-year Traditional Chinese Medicine (TCM) students. While the primary focus is on the technical evaluation of TCM examinations, its implications for refining TCM practice, preserving cultural knowledge, and integrating sustainable healthcare solutions make it an important contribution to the broader discourse on sustainability. For detailed information, please refer to the Appendix A. To achieve this goal, the research agenda for CDMs includes: (a) conducting cognitive attribute modeling and assessing the fit of the model with the data, (b) determining calibrated item parameters and evaluating the reliability and validity of the national examination, and (c) offering practical recommendations and guidance for maintaining the quality of assessments in health professions education.

2. Materials and Methods

2.1. Participants

This study included 16,310 TCM undergraduates residing in 29 randomly chosen cities or provinces across China, with ages ranging from 17 to 55 years (median = 22.933 years, SD = 0.973). Participants who met certain exclusion criteria, such as having a total score of 0 or missing responses on items, were not included in the analysis. Recruitment of respondents commenced in May 2023, resulting in a sample consisting of 10,765 females (66.0%) and 5545 males (34.0%). These participants were distributed across the eastern (41.3%), central (32.6%), and western (26.1%) regions of China.

2.2. Measurement Tool

The SCTTCMU is a collaborative effort between the Certification Center for Chinese Medicine Practitioners of the National Administration of Traditional Chinese Medicine and the National Administry Committee on Teaching Traditional Chinese Medicine to Majors in Higher Education under the Ministry of Education of the People's Republic of China. It aims to evaluate students' comprehension of basic and clinical medical sciences by the end of their fourth year in a five-year TCM undergraduate program. The assessment comprises two components: objective structured clinical examinations (OSCEs) for clinical skills evaluation and computerized multiple-choice items to evaluate medical knowledge. This paper focuses on the medical knowledge section, which consists of 300 multi-choice items with dichotomous scoring, each presenting four distractors and one correct answer.

2.3. Overall Workflow

The psychometric evaluation of SCTTCMU was carried out using a 6-step process that encompasses the essential tasks illustrated in Figure 1. Each step builds upon the previous one, with guidelines for the necessary statistical thresholds provided in the figure.

Step 1 involves validating the Q-matrix and attribute hierarchy as the foundational analysis. Step 2 focuses on ensuring unidimensionality to enable the construction of CDT models. Step 3 requires researchers to assess the model's suitability, considering quantitative measures for model fit evaluation. Following the confirmation of CDT usage, Step 4 involves extracting item parameters and fitting information from the model. Steps 5 and 6 entail conducting test analysis, evaluating reliability and validity, and reporting screening scores. Detailed explanations for each step are provided below.

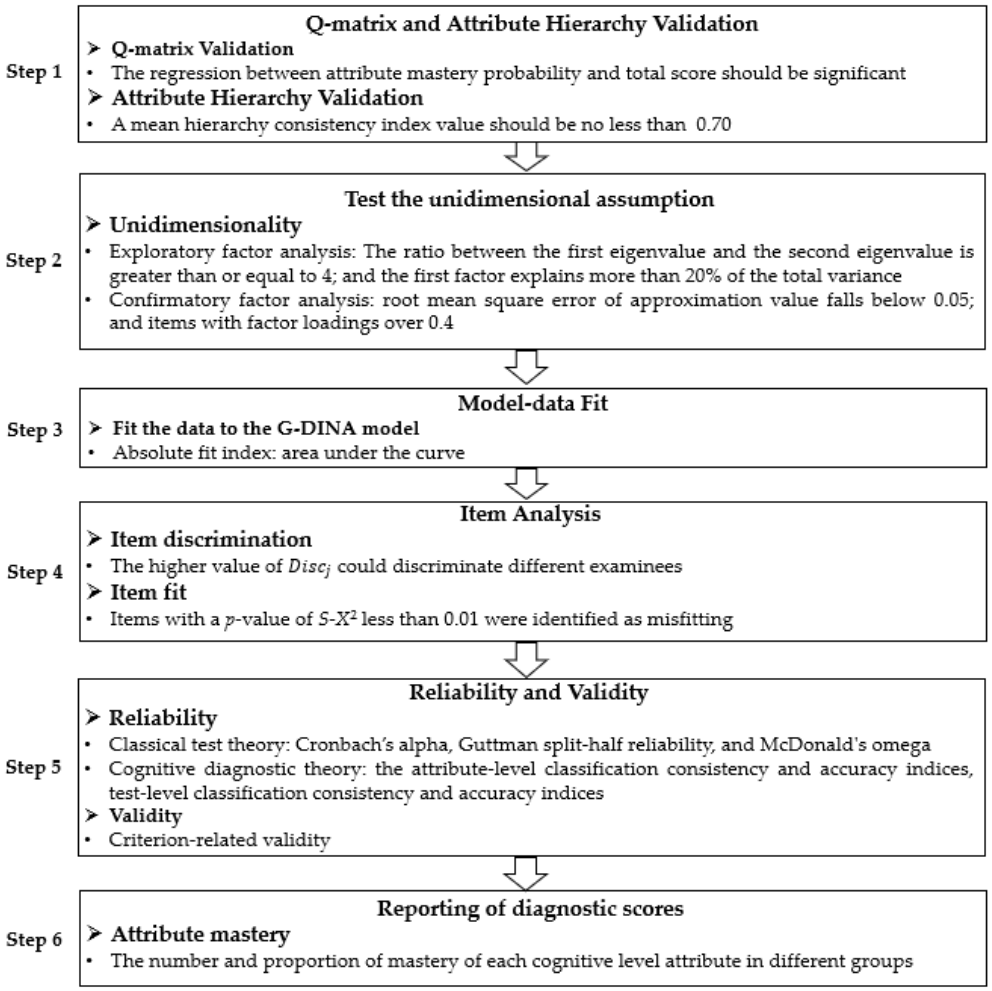

Figure 1. The process of assessing the quality of Traditional Chinese Medicine (TCM) examination using psychometric methods.

2.4. Cognitive Diagnosis Model: G-DINA

CDMs aim to create solid connections between people's responses to items and their patterns or profiles of attributes. There are various CDMs available, which are classified into two categories: simplified and saturated models, depending on their scope. It is possible to convert saturated CDMs into simplified ones under specific limitations or assumptions. The G-DINA (generalized deterministic input, noisy, "and" gate) [12] model is a notable example of a saturated model and is commonly applied in cognitive evaluations.

The G-DINA model uses a reduced vector $\boldsymbol{\alpha}_{lj}^{*} = \left(\alpha_{l1}, \ldots, \alpha_{lK_j^{*}}\right)'$ to illustrate the necessary attributes for an item, where $l = 1, \ldots, 2^{K_j^{*}}$ and $2^{K_j^{*}}$ indicates the unique attribute pattern count. The likelihood that respondents with the reduced attribute vector $\boldsymbol{\alpha}_{lj}^{*}$ would answer item j correctly is denoted as $P\left(X_j = 1 | \boldsymbol{\alpha}_{lj}^{*}\right)$. The G-DINA model is formally represented by the following formula:

$$P(X_j = 1|\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \ , \quad (1)$$

Equation (1) explains how the likelihood of a test taker with the attribute mastery pattern $\boldsymbol{\alpha}_{lj}^*$ answering an item *j* correctly is determined. The item intercept, δ_{j0}, indicates the probability of a test taker answering item *j* correctly even without mastering all the attributes being measured. δ_{jk} signifies the impact of attribute *k* on the probability of answering item *j* correctly. The term $\delta_{jkk'}$ represents the combined effect of attributes *k* and k' on item *j*, while $\delta_{j12\ldots K_j^*}$ indicates the collective interaction of all attributes on item *j*. The value of q_{jk} in the Q-matrix is a binary indicator (0 or 1), where $q_{jk} = 1$ means that item *j* assesses attribute *k*, and $q_{jk} = 0$ means it does not. In equation (1), $K_j^* = \sum_{k=1}^{K} q_{jk}$ represents the total attributes measured by item *j*. The parameter α_{lk} represents the mastery status of attribute *k* for the *l*th respondent's attribute mastery pattern—a binary variable where $\alpha_{lk} = 1$ signifies mastery and $\alpha_{lk} = 0$ signifies a lack of mastery.

2.5. Statistical Analysis

A 6-step procedure has been adopted to conduct the psychometric evaluation of the SCTTCMU, which includes essential tasks. The analyses were performed using the GDINA R package [13] and custom-written code in R [14]. Specific details for each step listed are provided below.

Step 1: Q-matrix and attribute hierarchy validation. The Q-matrix [15] is a critical element as it organizes items based on the cognitive attributes needed to solve them [16]. Typically, the Q-matrix is developed through input from domain experts, clinical theories, or empirical research findings [17]. It is essentially a binary matrix, with entries of 1 indicating that an attribute is measured by an item, while 0 denotes otherwise. To confirm the suitability of the Q-matrix, this research conducted a regression analysis to assess the likelihood of mastering candidate attributes using the CDM framework and the total SCTTCMU score [18].

Leighton and his team introduced the Attribute Hierarchy Method (AHM) and developed the Attribute Hierarchy Structure (AHS) to depict the cognitive model of interconnected tasks [19]. To assess the credibility of the attribute hierarchy, the Hierarchy Consistency Index (HCI) [20,21] was utilized. The HCI ranges from −1 to 1, with values closer to 1 indicating a better fit and values closer to −1 demonstrating a poorer fit.

Step 2: Testing the unidimensional assumption. Unidimensionality refers to the concept that a test measures a single primary underlying trait. This implies that responses to each question are influenced by only one main latent trait of the individual [7]. Both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were utilized to evaluate the assumption of unidimensionality. In EFA, unidimensionality is validated when the ratio of the first eigenvalue to the second eigenvalue is 4 or higher [22] and when the initial factor explains over 20% of the total variance [23]. A CFA involving a single factor was employed to assess unidimensionality. Two key indicators were considered: factor loading and root mean square error of approximation (RMSEA), determined using the weighted least squares method and variance-adjusted method. Following the guideline established by Browne and Cudeck [24], the fit of the model is considered close, fair, acceptable, mediocre, or poor if the RMSEA value falls below 0.05, between 0.06 and 0.08, between 0.09 and 0.10, or above 0.10, respectively. Items with factor loadings below 0.4 were omitted as such loadings could lead to misinterpretations [25].

Step 3: Model-data fit. Choosing the right CDM is an important step in drawing accurate conclusions. Evaluating how well a model fits the data is typically divided into absolute and relative fits. In this research, an absolute fit measure called the area under the curve (AUC) [26] was employed to assess the model-data fit. An AUC score above 0.5 suggests that the model's predictive ability is better than random chance, indicating a reasonably good fit for the model.

Step 4: Item analysis. Evaluating the psychometric properties is crucial when evaluating the SCTTCMU. The psychometric properties of each item in this study were determined by item discrimination and item fit based on the CDMs.

In this study, the G-DINA model was employed for estimating item parameters within the Q-matrix framework [12]. Parameter estimation was conducted using the Marginal Maximum Likelihood Estimation (MMLE) algorithm for estimating the G-DINA model parameters, ensuring quicker and unbiased convergence [27]. The analysis of data with cognitive diagnostic models involved the use of the $Disc_j$ index to examine item discrimination. This index is defined as the discrepancy between the probability of a correct response $P(X_j = 1|\boldsymbol{\alpha}_{lj}^* = 1)$ from examinees who have demonstrated all attributes related to item j and the probability of a correct response $P(X_j = 1|\boldsymbol{\alpha}_{lj}^* = 0)$ from those who have not exhibited any attribute associated with item j. Mathematically, this is expressed as:

$$Disc_j = P(X_j = 1|\boldsymbol{\alpha}_{lj}^* = 1) - P(X_j = 1|\boldsymbol{\alpha}_{lj}^* = 0), \quad (2)$$

$$P(X_j = 1|\boldsymbol{\alpha}_{lj}^* = 1) = 1 - P(X_j = 0|\boldsymbol{\alpha}_{lj}^* = 1) = 1 - s_j, \quad (3)$$

$$P(X_j = 1|\boldsymbol{\alpha}_{lj}^* = 0) = g_j. \quad (4)$$

In the equation above, s_j represents the likelihood of a test taker with all the necessary qualities providing an incorrect response, while g_j indicates the likelihood of a test taker without all the necessary qualities providing a correct response. $Disc_j$ served as a holistic indicator of slip and guessing parameters. A higher value of $Disc_j$ suggested improved item quality and the ability to differentiate between various test takers.

Additionally, the S-X^2 statistic was employed to evaluate item fitness [28] in this research, measuring the differences between observed and anticipated response rates. Items that exhibited a p-value of S-X^2 below 0.01 were recognized as not fitting well and were consequently excluded from the item pool.

Step 5: Reliability and validity analysis. In order to assess the reliability of the SCTTCMU, the study utilized Cronbach's alpha, Guttman split-half reliability coefficients, and McDonald's omega, all of which were derived from classical test theory (CTT). The study also evaluated attribute-level classification consistency and accuracy indices [29] as well as test-level classification consistency and accuracy indices [30] based on the CDM framework for the SCTTCMU. These reliability measures aimed to determine how effectively the CDM categorized candidates into the correct attribute profiles. Furthermore, criterion-related validity was established by examining the correlation between the likelihood of mastering attributes and each candidate's total score in the SCTTCMU.

Step 6: Screening score reporting. Through an analysis of the examinees' cognitive attributes, we can accurately determine the type of cognitive attributes and the current knowledge level of the examinees. This information is valuable for teachers to provide targeted remedial instruction based on the examinees' areas of need. By estimating the measure pattern model of examinees' cognitive attributes, this study identifies the cognitive attribute measures for both the overall and typical groups of examinees. Furthermore, it presents a spectrum illustrating the measurement pattern of the examinees' cognitive attributes.

3. Results

3.1. Descriptive Analysis

Figure 2 displays the distribution of total scores obtained by candidates. The total scores fell within a range of 3 to 289, with a median score of 193. Around 61.28% of candidates achieved scores surpassing the minimum passing score of 180.

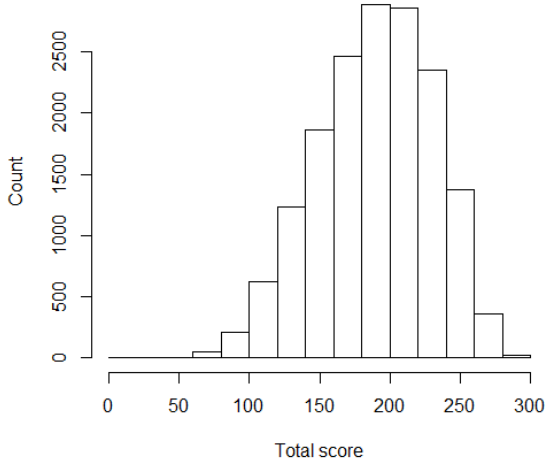

Figure 2. Total score distribution.

3.2. Q-matrix and Attribute Hierarchy Validation

Templin and Henson [31] highlighted the significance of the Q-matrix construction for ensuring the accuracy of diagnostic outcomes in cognitive diagnostic assessments. In this research, the task of defining which attribute is needed to respond to each item was overseen and coordinated by the Certification Center for Chinese Medicine Practitioners of the National Administration of Traditional Chinese Medicine. The Q-matrix of items comprises three columns, representing each of the three criteria. As indicated in Table 1, the Q-matrix demonstrates that every item evaluates a singular attribute criterion, with each criterion being gauged by an average of 90 items.

During the study, regression analysis was conducted, using the likelihood of mastery based on candidates' attributes as the independent variable and the total score of the SCTTCMU as the dependent variable. The findings from the regression analysis indicated that the regression equation was statistically significant ($p < 0.05$) with a coefficient of determination of 0.76, suggesting the reasonableness of the Q-matrix of the SCTTCMU. Additionally, the average HCI value for the SCTTCMU was observed to be 0.76. According to Wang and Gierl [32], an average HCI value around 0.70 signifies an unstructured attribute hierarchy, implying that the hierarchical relationships of the attributes were likely unstructured.

Table 1. Some item examples for Q-matrix.

Item	Q-Matrix		
	Memory (A1)	**Understanding (A2)**	**Application (A3)**
1	1	0	0
2	0	1	0
3	0	1	0
4	1	0	0
5	0	1	0
6	0	1	0
7	0	1	0
8	1	0	0
9	1	0	0
10	0	1	0

Note. The value of 1 in row *j* and column *k* of the Q-matrix indicates that item *j* measures attribute *k*, while a value of 0 indicates that item *j* does not measure attribute *k*.

3.3. Evaluation of Unidimensionality

In EFA, the ratio between the first eigenvalue and the second eigenvalue was 4.447 (i.e., greater than 4), and the first factor explained 31.9% of the total variance (i.e., higher than 20%). Results of the single-factor CFA indicated that the RMSEA value was 0.08, indicating that the single-factor model was fair or acceptable; all factor loadings were above 0.4. Therefore, the items of the SCTTCMU were considered unidimensional and suitable for the next phase of analysis.

3.4. Model-Data Fit

This study utilized the G-DINA model within the CDM framework to conduct a model-data fit analysis on the original data from the SCTTCMU. The findings revealed an AUC index of 0.97, indicating a strong predictive performance of the G-DINA model compared to random guessing and demonstrating its outstanding classification accuracy. Essentially, the G-DINA model demonstrated a strong alignment with the SCTTCMU dataset.

3.5. Item Analysis

Figure 3A displays the estimated item discrimination parameters of the G-DINA model. The findings indicate that all items have a discrimination parameter greater than or equal to 0, suggesting their ability to effectively differentiate between candidates who have mastered the cognitive attributes being evaluated and those who have not. Moreover, no item was identified as a poor fit for the G-DINA model, as evidenced by a *p*-value above 0.01 in the $S\text{-}X^2$ statistic (refer to Figure 3B).

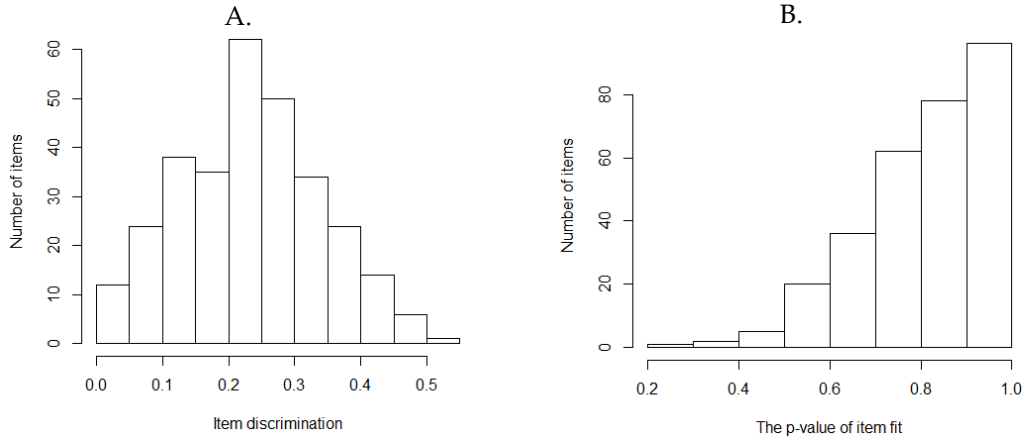

Figure 3. Analysis of items on the Standardized Competence Test for Traditional Chinese Medicine Undergraduates (SCTTCMU) within the context of CDMs. Note. (**A**) The distribution of item discrimination. (**B**) The distribution of *p*-values for item fit.

3.6. Reliability and Validity

The reliability of the SCTTCMU was assessed using both the CTT and CDM frameworks. In the CTT framework, the SCTTCMU demonstrated a Cronbach's alpha coefficient of 0.97, a Guttman split-half coefficient of 0.94, and a McDonald's omega coefficient of 0.97. The CDM approach offered a different method to evaluate the reliability of the three attribute criteria. The results from the CDM framework indicated that the classification consistency reliability of the attributes ranged from 0.97 to 0.99, with an average of 0.98, while the classification accuracy reliability of the test was 0.95. Moreover, the SCTTCMU exhibited strong criterion-related validity, as the mean probability of attribute mastery based on the DCMs had a significant correlation of 0.87 ($p < 0.001$) with the total SCTTCMU score. These findings collectively suggest that the SCTTCMU demonstrates

robust reliability and validity across both the traditional CTT and newer CDM frameworks.

3.7. Reporting of Diagnostic Scores

The data in Table 2 displays the number and percentage of proficiency in each cognitive level attribute among all candidates, including those who passed and failed the SCTTCMU. The findings reveal that candidates who passed the SCTTCMU demonstrate a higher level of mastery in each cognitive level attribute compared to those who failed, validating the accuracy of the estimated proficiency levels.

Table 2. Mastery of each attribute by different groups of candidates.

Cognitive Level	All Candidates				Passing Group				Failed Group			
	0		1		0		1		0		1	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
Memory (A1)	7693	47	8617	53	1499	15	8495	85	6194	98	122	2
Understanding (A2)	7598	47	8712	53	1354	14	8640	86	6244	99	72	1
Application (A3)	7161	44	9149	56	1313	13	8681	87	5848	93	468	7

Note. Element 0 signifies that the candidate has not mastered attribute *k*, whereas element 1 denotes the candidate's mastery of attribute *k*. *N* and % represent the number and percentage of participants, respectively.

Furthermore, Figure 4 displays specific details of the two individuals' data regarding the three attribute criteria as evidenced in their score reports, illustrating the distinct insights offered by the CDM. Despite both individuals achieving an equivalent overall score (180) in the SCTTCMU, the probability of each of them meeting the attribute criteria greatly differed due to unique individual characteristics. This showcases how cognitive diagnostic analysis can offer more detailed and personalized feedback compared to conventional test results.

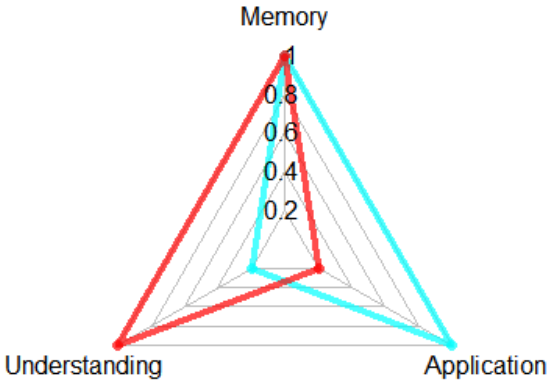

Figure 4. Spectrum of attribute measures for two candidates scoring the same 180.

4. Discussion

The purpose of this research was to assess the feasibility of a nationwide summative test (i.e., SCTTCMU) for the first time using DCMs. Specifically, the G-DINA model was utilized to examine the effectiveness of the SCTTCMU. The study then calculated the psychometric characteristics of the SCTTCMU and generated a detailed score report for individual respondents. The findings of the study revealed that each item of the SCTTCMU had a strong ability to discriminate between individuals who had the attributes of an item

and those who did not. Furthermore, the analysis demonstrated that the SCTTCMU not only exhibited strong reliability but also displayed high criterion-related validity. These compelling results suggest that the SCTTCMU serves as a high-quality assessment tool in the field of health professions education.

In comparison to previous research on health professions education assessment, this study has significant implications. First, it demonstrates how a thorough analysis utilizing the CDMs can effectively evaluate the quality of national examinations. This study offers detailed guidance for researchers on how to conduct, interpret, and report cognitive diagnosis analysis, allowing for potential replication. The CDM not only assesses the psychometric properties of tests and items but also provides diagnostic insights at both the general and attribute levels. Second, CDMs, as specialized latent class models, have distinct characteristics. Unlike unrestricted latent class models, CDMs apply specific constraints to score predetermined dichotomous latent variables or attributes outlined in the Q-matrix. These constraints define the attributes necessary for positive responses to each item, thus unveiling individuals' general response patterns. Consequently, we believe that incorporating feedback based on cognitive diagnosis modeling in SCTTCMU results can be a valuable tool for learners, educators, and other stakeholders in addressing cognitive attribute gaps, surpassing mere remediation of content knowledge deficiencies.

Therefore, the key findings of this research focus on the evaluation and screening of the skills of health professional applicants: (1) The CDM is highlighted as an efficient statistical method for psychological research in this article, considering the new psychometric approach being employed. Research into ability profiles could be improved with the help of the dichotomous attribute nature of the CDM, which enables researchers to better grasp the specific components of traits. (2) The SCTTCMU serves as an enhanced assessment tool in this study, offering both broad diagnostic data and specific insights into how each person aligns with diagnostic criteria. Detailed information at the criteria level can help tailor personalized treatments for future health professionals, potentially enhancing the efficacy of these interventions. (3) This research establishes a model for forthcoming score reporting in national exams, furnishing candidates and healthcare educators with pertinent details to enhance their teaching and learning practices. Furthermore, it can serve as a valuable reference for other global assessment agencies.

While the results showed promise, several limitations need to be taken into account in future research. Firstly, this study utilized a data-driven Q-matrix calibration procedure. Employing an exploratory Q-matrix calibration method in the future could enhance the accuracy of calibration. Secondly, the analysis in this study relied on a single model. To improve the analysis of test data, future studies could explore using a mixed-model approach, where different cognitive diagnostic models are selected for each item. Thirdly, it is important to acknowledge that both the baseline observation data and the Q-matrix used in the SCTTCMU were dichotomous. However, for researchers wanting to investigate interval scale variables with CDMs, there are various approaches available. These include converting Likert-type scale items into dichotomous format, as suggested by Templin and Henson [31], or converting scores above 0 to 1 while retaining a score of 0 [33]. Additionally, the seq-GDINA model [34] offers a direct method for handling Likert scale data. Fourth, the SCTTCMU's 300 items could potentially be burdensome for participants. Future studies could explore implementing computerized adaptive tests to conduct more efficient assessments with fewer items. Finally, recent advancements in cognitive diagnostic modeling such as facet and hierarchical rater models could incorporate rater effects, which might be beneficial for practical assessment tools like OSCEs and other assessments [35].

5. Conclusions

The article presented how the SCTTCMU assessment was adapted within the CDM framework, and the results demonstrated that the SCTTCMU is an enhanced TCM assessment. This tool has the potential to offer extensive diagnostic information, allowing

researchers to evaluate each person's symptom profile. This detailed diagnostic data may enhance the efficiency of teaching and learning processes.

Author Contributions: Conceptualization, L.X., Z.J. and Y.H.; formal analysis, L.X.; funding acquisition, Z.J.; methodology, L.X., Z.J. and Y.H.; project administration, Z.J.; software, Y.H.; validation, L.X.; writing—original draft, L.X.; writing—review and editing, Z.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Certification Center for Chinese Medicine Practitioners of the National Administration of Traditional Chinese Medicine under Grant TA2022002, National Natural Science Foundation of China for Young Scholars under Grant 72104006, China Postdoctoral Science Foundation under Grant 2023M740082, Certification Center for Chinese Medicine Practitioners of the National Administration of Traditional Chinese Medicine under Grant TC2023005, Peking University Health Science Center Medical Education Research Funding Project under Grant 2023YB24, and Beijing Social Science Foundation under Grant 23JYC019.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by Biomedical Ethics Committee of Peking University (IRB00001052-22070).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Linkages between Research Contributions and Sustainability

1. Promoting Sustainable Healthcare Practices: By employing cognitive diagnostic approaches to evaluate the national Traditional Chinese Medicine (TCM) examination, the study enhances the accuracy in assessing practitioners' competence. This ensures a higher standard of TCM practice, which often involves natural therapies and preventive measures. Integrating effective TCM into healthcare systems can reduce reliance on resource-intensive Western medicine, contributing to a more environmentally sustainable model of healthcare delivery.

2. Fostering Knowledge Sustainability: TCM embodies centuries-old wisdom and practices. Validating examinations using advanced methodologies sustains and propagates this knowledge, preserving cultural heritage and promoting a diverse global medical knowledge base. This knowledge sustainability is crucial for maintaining traditional practices that can complement modern medicine, offering alternatives with potentially fewer ecological footprints.

3. Supporting Economic Sustainability in Healthcare: A reliable evaluation system for TCM professionals can stimulate the growth of the TCM industry, creating jobs and economic opportunities aligned with green and sustainable practices. As TCM gains wider acceptance, it can contribute to local and regional economies, fostering sustainable economic development through the provision of health services and natural product markets.

4. Encouraging Sustainable Resource Use: Ensuring TCM practitioners are thoroughly evaluated and competent encourages responsible sourcing and cultivation of medicinal plants and herbs. This approach supports sustainable agriculture and forestry practices, protecting ecosystems and biodiversity—fundamental components of environmental sustainability.

5. Enhancing Community Health and Resilience: TCM often focuses on holistic health and preventive care, which can enhance population health and resilience. Healthier communities are more capable of adapting to social and environmental changes, a key aspect of sustainability. By validating TCM examinations, the study indirectly contributes

to building more resilient societies that can better face challenges like pandemics and environmental crises.

References

1. Downing, S.M.; Yudkowsky, R. Introduction to assessment in the health professions. In *Assessment in Health Professions Education*; Routledge: Oxfordshire, UK, 2009; pp. 21–40.
2. Yudkowsky, R.; Park, Y.S.; Downing, S.M. (Eds.) *Assessment in Health Professions Education*; Routledge: New York, NY, USA, 2019; p. 26.
3. Bloch, R.; Norman, G. Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Med. Teach.* **2012**, *34*, 960–992. https://doi.org/10.3109/0142159X.2012.703791.
4. De Champlain, A.F. A primer on classical test theory and item response theory for assessments in medical education. *Med. Educ.* **2010**, *44*, 109–117. https://doi.org/ 10.1111/j.1365-2923.2009.03425.x.
5. DeVellis, R.F. Classical test theory. *Med. Care* **2006**, *44*, S50–S59. https://doi.org/10.1097/01.mlr.0000245426.10853.30.
6. Brennan, R.L. Generalizability theory. *Educ. Meas. Issues Pract.* **1992**, *11*, 27–34.
7. Embretson, S.E.; Reise, S.P. *Item Response Theory*; Psychology Press: London, UK, 2013.
8. Collares, C.F. Cognitive diagnostic modelling in healthcare professions education: An eye-opener. *Adv. Health Sci. Educ.* **2022**, *27*, 427–440. https://doi.org/10.1007/s10459-022-10093-y.
9. Roberts, M.R.; Gierl, M.J. Developing score reports for cognitive diagnostic assessments. *Educ. Meas. Issues Pract.* **2010**, *29*, 25–38. https://doi.org/10.1111/j.1745-3992.2010.00181.x.
10. Jang, E.E. Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Lang. Test.* **2009**, *26*, 31–73. https://doi.org/10.1177/0265532208097336.
11. Lee, Y.W.; Sawaki, Y. Application of three cognitive diagnosis models to ESL reading and listening assessments. *Lang. Assess. Q.* **2009**, *6*, 239–263. https://doi.org/10.1080/15434300903079562.
12. de la Torre, J. The generalized DINA model framework. *Psychometrika* **2011**, *76*, 179-199. https://doi.org/10.1007/s11336-011-9207-7.
13. Ma, W.; de la Torre, J. *R Package GDINA, Version 0.9.9.8*; GDINA: The Generalized DINA Model Framework; R Core Team: Vienna, Austria, 2016.
14. R Core Team. *R: A Language and Environment for Statistical Computing [Computer Software Manual]*; R Core Team: Vienna, Austria, 2016.
15. Tatsuoka, K.K. Rule space: An approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.* **1983**, *20*, 345–354.
16. Birenbaum, M.; Kelly, A.E.; Tatsuoka, K.K. Diagnosing knowledge states in algebra using the rule space model. *ETS Res. Rep. Ser*. **1993**, *24*, 442–459. https://doi.org/10.5951/jresematheduc.24.5.0442.
17. de la Torre, J.; van der Ark, L.A.; Rossi, G. Analysis of Clinical Data from a Cognitive Diagnosis Modeling Framework. *Meas. Eval. Couns. Dev.* **2018**, *51*, 281–296. https://doi.org/10.1080/07481756.2017.1327286.
18. Kang, C.; Xin, T.; Tian, W. Development and Validation of Diagnostic Test for Primary School Arithmetic Word Problems. *Exam Res.* **2013**, *6*, 24–43.
19. Leighton, J.P.; Gierl, M.J.; Hunka, S.M. The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *J. Educ. Meas.* **2004**, *41*, 205–237. https://doi.org/10.1111/j.1745-3984.2004.tb01163.x.
20. Leighton, J.P.; Cui, Y.; Cor, M.K. Testing expert-based and student-based cognitive models: An application of the attribute hierarchy method and hierarchy consistency index. *Appl. Meas. Educ.* **2009**, *22*, 229–254. https://doi.org/10.1080/08957340902984018.
21. Cui, Y.; Leighton, J.P. The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *J. Educ. Meas.* **2009**, *46*, 429–449. https://doi.org/10.1111/j.1745-3984.2009.00091.x.
22. Reeve, B.B.; Hays, R.D.; Bjorner, J.B.; Cook, K.F.; Crane, P.K.; Teresi, J.A.; Thissen, D.; Ravicki, D.A.; Weiss, D.J.; Hambleton, R.K.; et al. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med. Care* **2007**, *45*, S22–S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04.
23. Reckase, M.D. Unifactor latent trait models applied to multifactor tests: Results and implications. *J. Educ. Stat.* **1979**, *4*, 207–230. https://doi.org/10.3102/10769986004003207.
24. Browne, M.W.; Cudeck, R. Alternative ways of assessing model fit. In *Testing Structural Equation Models*; Bollen, K.A., Long, J.S., Eds.; Sage: Newbury Park, CA, USA, 1993.
25. Nunnally, J.C. *Psychometric Theory*, 2nd ed.; McGraw-Hill: Hillsdale, NJ, USA, 1978.
26. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.
27. George, A.C.; Robitzsch, A.; Kiefer, T.; Groß, J.; Ünlü, A. The R package CDM for cognitive diagnosis models. *J. Stat. Softw.* **2016**, *74*, 1–24. https://doi.org/10.18637/jss.v074.i02.
28. Orlando, M.; Thissen, D. Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Appl. Psychol. Meas.* **2003**, *27*, 289–298. https://doi.org/10.1177/0146621603027004004.

29. Wang, W.; Song, L.; Chen, P.; Meng, Y.; Ding, S. Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *J. Educ. Meas.* **2015**, *52*, 457–476. https://doi.org/10.1111/jedm.12096.
30. Iaconangelo, C.J. Uses of Classification Error Probabilities in the Three-Step Approach to Estimating Cognitive Diagnosis Models. Ph.D. Dissertation, Rutgers University-School of Graduate Studies, New Brunswick, NJ, USA, 2017.
31. Templin, J.L.; Henson, R.A. Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* **2006**, *11*, 287. https://doi.org/10.1037/1082-989X.11.3.287.
32. Wang, C.; Gierl, M.J. Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *J. Educ. Meas.* **2011**, *48*, 165–187. https://doi.org/10.1111/j.1745-3984.2011.00142.x.
33. Lee, Y.S.; Park, Y.S.; Taylan, D. A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *Int. J. Test.* **2011**, *11*, 144–177. https://doi.org/10.1080/15305058.2010.534571.
34. Ma, W.; de la Torre, J. GDINA: The Generalized DINA Model Framework. R Package Version 0.13.0. 2016. Available online: http://cran.rproject.org/package=GDINA (accessed on 20 April 2024).
35. Li, X.; Wang, W.C.; Xie, Q. Cognitive diagnostic models for rater effects. *Front. Psychol.* **2020**, *11*, 525. https://doi.org/10.3389/fpsyg.2020.00525.