DOI URL: https://doi.org/10.63944/vmc.NHE.2
Vol.1 No.2

Study on Optimizing Senior High English Assessment via Multi-dimensional Item Set Effect

Zhiyi Wu *

Faculty of Education, Sichuan Normal University, China | 840850191@qq.com

*** Corresponding Author: Zhiyi Wu | 840850191@qq.com**

Abstract: Currently, English reading, as a core module in the middle school English assessment system, accounts for a high proportion usually exceeding 70%. Addressing the limitations of traditional models in English reading assessment, this study proposes the Multidimensional Partial Credit Testlet Model (MPCTM) based on the Partial Credit Model (PCM), by integrating the multidimensional ability space, multi-level scoring mechanism, and testlet random effects. Results of the simulation study show that the error and bias of MPCTM in parameter estimation are significantly lower than those of traditional models. Findings from the empirical study confirm that MPCTM has better explanatory power and goodness of fit in complex assessment scenarios; in the future, it is necessary to increase the proportion of mixed question types and expand its interdisciplinary applications.

Keywords: Partial Credit Model; Multidimensional Ability; Local Dependence; English Reading Assessment; Markov Chain Monte Carlo

1. Introduction

At present, English reading evaluation in middle schools is facing two challenges. On the one hand, most of the current Item Response Theory models are only applicable to the 0-1 scoring data, which greatly limits the application of the Item Response Theory models in practice, and it is difficult to capture the fine-grained ability signal contained in the multi-level scoring data of the simple answer questions.[1] On the other hand, the topic based on the same reading text usually has local dependence due to sharing context. Therefore, the set of items sharing the same stimulus in the test is referred to as question Testlet, and the interdependence among the responses of question groups caused by the common stimulus is referred to as question Testlet random Effect.[2] Because the IRT model is generally based on the assumption of local independence, when this assumption can not be satisfied, the dependent IRT model should be used.[3]

Therefore, early researchers made breakthrough explorations in two technical dimensions. In the innovative dimension of scoring mechanism, Graded Response Model (GRM) proposed by Samejima (1969) has pioneered the modeling of multi-level scoring field. Based on the parametric strategy of cumulative probability threshold, the modeling of multi-level scoring is realized by setting the progressive difficulty critical value.[4] To this end, David Andrich (1982) put forward a partial scoring model Partial Credit Model (PCM) based on the Rasch model. The Step Probability used by PCM allows different steps to set the difficulty independently and is more flexible for practical application of step-by-step scoring or partial scoring than GRM.[5]

Although the innovation of GRM and PCM in scoring mechanism has extended the application boundary of IRT, it has not solved the deep-seated structural dependence problem in a large number of assessments. For this reason, in the field of project response theory, the Local Dependence caused by the random effect of question groups seriously affects the stability of project parameter estimation. The problem Testlet Response Theory (TRT) expanded and proposed by Wainer and Kiely has the significance of inheriting and starting from the beginning.[6] Then the Crossed Random Effects Model (CREM) developed by DeMars (2006) realized a new method breakthrough. The research shows that CREM can improve the question group variance interpretation rate and significantly reduce the estimation standard error of the item difficulty parameter.[7] Then, based on the Rasch framework, Zhan Peida et al. (2014) identified the multi-dimensional question group attribution structure of the project through the judgment matrix, separated the subjects' ability (θ), project difficulty (b_i) and multi-dimensional random effect (γ_1, γ_2,...,γ_m) by combining the question group random effectΓ matrix.[8]

The methodological veins in the field of project response theory evolve in parallel in the field of Cognitive Diagnosis Theory (CDT). In the field of CDT, the T-CDM model reconstructs the potential response mechanism of the DINA model, introduces the question group interference factor into the item difficulty parameter, and realizes the statistical control of the question group deviation in the diagnostic framework for the first time.[9] However, the model has the collinearity problem of random effect and attribute mastery probability. In order to solve this problem, a two-parameter separation estimation strategy is proposed for the combined question group cognitive diagnosis model (JT-CDM).[10] What is more breakthrough is the 2-tier-LCDM mixed model proposed by Cai (2010). The problem of parameter space explosion caused by high-order interaction effect is solved successfully by studying the adaptive sampling technology of MCMC algorithm.[11] In recent years, aiming at the complex diagnosis demand of multi-level scoring and random effect interlacing of question groups, the Polytomous Cognitive Diagnosis Testlet Model (PCDTM) proposed by Zhou Wenjie et al. (2023) embeds the multi-dimensional question group random effect matrix and dynamic judgment matrix while preserving the core framework of attribute diagnosis, and effectively solves the parameter confusion problem of traditional diagnosis model in mixed question groups.[12]

Although the existing research has made progress in multilevel scoring model and multidimensional extension, its application still focuses on the fields of mathematics and science, and fails to effectively solve the triple interaction problem of multidimensional, multilevel and group random effects. Therefore, the paper introduces the Multidimensional Partial Credit Testlet Model (MPCTM) into the field of language evaluation, and proposes to model the multilevel scoring, multidimensional ability and local dependence, aiming at the adaptability and practical value of the Multidimensional Partial Credit Model (MPCM) in middle school English reading evaluation. Firstly, the paper constructs a MPCTM framework suitable for middle school English reading evaluation, and solves the separation limitation of traditional models in multi-dimensional ability and random effect modeling. Secondly, generate simulation data to simulate the research; Using the English reading test data of a middle school in a province, this paper makes an empirical study on the advantages of MPCTM measurement accuracy. Finally, the application value of the model in educational practice is explored, and the research direction of future educational evaluation is prospected.

2. Development of Multidimensional Partial Credit Model

2.1. PCM

PCM is an important part of multi-level scoring model in IRT theory. Its core idea is to parameterize the "step difficulty" of each scoring grade, allowing the difficulty of different scoring grades to change independently, that is, each scoring grade, for example, from 0 to 1, or from 1 to 2, is regarded as an independent step, and the probability is determined by the capability parameter (θ) of the subject n and the difficulty threshold (β_{jv}) of the topic in the step.[13]

For question j, Masters (1982) assumes that there are M_j steps, i.e., there are M_j rating levels (m=0,1,...,M_j−1), and the subject n capacity is, then the probability of scoring m is:[13]

$$P_{njm}(\theta)=\frac{exp\left[\sum_{v=0}^{m}\alpha_j(\theta-\beta_{jv})\right]}{\sum_{c=0}^{M_j-1}exp\left[\sum_{v=0}^{c}\alpha_j(\theta-\beta_{jv})\right]} \qquad （1）$$

It can be seen from the formula (l) that α_j refers to the discrimination parameter in question j, d_{jm} refers to the difficulty threshold of score from v-1 to v in question j, and β_{j0} is specified as the reference point, θ refers to the position of the subject n in the one-dimensional capability space, that is, the comprehensive reading capability. Suppose that the distribution is $\theta \sim N$ (0,1), and c refers to the index variable for traversing all possible scores, in order to ensure $P_{njm}(\theta)$=1.

2.2. MPCM

The PCM are extended to MPCM, the single-dimensional capability is extended to the multi-dimensional capability vector (θ=$(\theta_1,\theta_2,...,\theta_D)^T$), and the load coefficient (α_{jd}) of the title in multiple dimensions is introduced. In multidimensional models, the difficulty term is replaced by a linear combination of multidimensional capabilities: $log\,(P(Y\geq v|\theta))=\alpha_j(\sum_{d=1}^{D}\alpha_{jd}\theta_d-b_{jv})$. α_{jd} is the load coefficient of the topic in dimension *d*, which represents the contribution weight of the topic response in this dimension. Replace with in a single-dimensionalmodel to get a multidimensional *log* expression: $log\,(P(Y\geq v|\theta))=\eta_j-b_{jv}=\sum_{d=1}^{D}\alpha_{jd}\theta_d-b_{jv}$.

The multidimensional partial scoring model GPCM is obtained by substituting the cumulative probability difference formula, which can be described as:

$$P_{njm}(\theta)=\frac{exp\left[\sum_{v=0}^{m}(\sum_{d=1}^{D}\alpha_{jd}\theta_d-b_{jv})\right]}{\sum_{c=0}^{M_j-1}exp\left[\sum_{v=0}^{c}(\sum_{d=1}^{D}\alpha_{jd}\theta_d-b_{jv})\right]} \tag{2}$$

The meaning of each parameter in the model is similar to that of formula (1) and formula (2). *D* represents the total number of ability dimensions. According to *China English Ability Rating Scale*, English comprehensive reading ability is divided into three-dimensional ability of students' vocabulary, reasoning and discourse. Similar to the one-dimensional model, the absolute difficulty parameter indicating the v-th scoring level of the topic means that the subject *n* needs to meet the multi-dimensional capability threshold simultaneously to obtain the comprehensive difficulty of the v-th scoring level of the topic. Suppose that in the three-dimensional capability test, if $\alpha_{j1}\theta_1+\alpha_{j2}\theta_2+\alpha_{j3}\theta_3>b_{jv}$, the subject *n* may obtain this rating. Compared with the one-dimensional model, b_{jv} not only reflects the ability of a certain dimension, but also needs to meet the threshold of the weighted combination of other dimensions even if the ability of a certain dimension is prominent.

2.3. MPCTM

Based on the MPCM, a random effect of group MPCTM was added, and Testlet Effect was used to modify the local dependence between the topics in the same group. Its core is to introduce γ_k for each group, adjust the difficulty of the topic. Taking the random effect γ_k of the question group as the adjustment factor of the difficulty item, we get: $\eta_j=\sum_{d=1}^{D}\alpha_{jd}\theta_d+\gamma_k-b_{jv}$. The GPCTM is represented by the following formula:

$$P_{njm}(\theta,\gamma_k)=\frac{exp\left[\sum_{v=0}^{m}(\sum_{d=1}^{D}\alpha_{jd}\theta_d-b_{jv}+\gamma_k)\right]}{\sum_{c=0}^{M_j-1}exp\left[\sum_{v=0}^{c}(\sum_{d=1}^{D}\alpha_{jd}\theta_d-b_{jv}+\gamma_k)\right]} \qquad （3）$$

In formula (3), the random effect of a certain question group is specified as zero, that is $\gamma_1=0$ or $\sum_{k=1}^{K}\gamma_k=0$, set to prevent confusion with the overall difficulty parameter, and the random effect γ_k of a question group obeys the normal distribution $\gamma_k \sim N\ (0,\sigma_\gamma^2)$. Suppose that the topic belongs to the topic group k, and the response probability is affected by the random effect of the topic group, it can be expressed as follows: $log\,(P(Y \geq v|\theta,\gamma_k))=\sum_{d=1}^{D}\alpha_{jd}\theta_d-b_{jv}+\gamma_k$.The random effect of the topic group $\gamma_k \sim N\ (0,\sigma_\gamma^2)$ is a random intercept term, which acts on all the topics in the topic group k. If $\gamma_k < 0$, the difficulty is reduced.

3. MCMC parameter estimation

In this paper, *R* language mirt package is used and MCMC algorithm is called for parameter estimation. To set each condition combination, 100 repeated experiments are required to improve the stability of the results. The number of chains is set as four, the length of each chain is 2000, the interval is one, the first 1000 times of preheating, and the average number of convergence results of parameters after 3000 times is taken as the result of parameter estimation. If the R of all estimated parameters is less than 1.05 or 1.1, the parameters converge substantially.[14] With reference to the settings of Wei Dan et al. (2017), the prior distribution of the parameters to be estimated is set as: $\gamma_k \sim N(0,\sigma_\gamma^2)$, $\theta \sim N(0,1)$, $\alpha \sim N(\mu_\alpha,\sigma_\alpha^2)$, $b \sim N(\mu_b,\sigma_b^2)$.[15]

4. Simulation Research

4.1. Research design

Three-factor crossover experiment design was used to systematically investigate the performance difference between MPCTM and the multi-dimensional divisional scoring model (MPCM). The independent variables include: True model as the core independent variable, distinguishing the MPCM model with no problem group effect from the MPCTM model with problem group effect; The sample size is set to three levels, namely N= 200, 500 and 1000, to test the stability of the model under different data scales; Two levels are set for the number of questions, namely 15 and 30. Among them, there are ten secondary scoring questions and five multi-level scoring questions in the 15th test. The first test group is 1-5 questions, the second test group is 6-10 questions and the third test group is 11-15 questions. The test consists of twenty secondary scoring questions and ten multi-level scoring questions. The test group is 1-5, the test group is 6-10, the test group is 11-15, the test group is 16-20, the test group is 21- 25 and the test group is 26-30. The research design can systematically test the adaptability of the model to the complex test structure by controlling the combination of the number of test groups (three groups and/6 groups) and the type of test (secondary/multi-level scoring).

4.2. Parameter setting

When generating the capability parameters under the project IRT framework, it is usually assumed that the potential capability of the subjects n is a continuous variable and follows the standard normal distribution. Set the number of subjects (N= 200, 500, 1000) to generate multidimensional capability parameters θ_i. If the number of dimensions is K=3, the capability values of subjects n in each dimension are independent and obedient: $\theta_{nk}{\sim}N(0,1)$.[16][17] Study and set the K-dimensional effect of each question group. Here, the multidimensional capability parameter and the random effect γ_k parameter of each question group obey the multivariate normal distribution, representing the covariance ∑ matrix, that is, the identity ∑ matrix with diagonal being one and non-diagonal being zero, and the variance of each dimension is one. The diagonal matrix is as follows:[8]

$$\sum=\begin{bmatrix}\sigma_{\gamma 1}^2 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma_{\gamma 6}^2\end{bmatrix}$$

In addition, the load parameters are α_j set in the study, and each topic is only investigated in one dimension, that is, it is denoted as 1, and vice versa. Set the difficulty threshold parameter d_{jm}, and generate difficulty thresholds M_j-1 for each score level M_j. For a topic with group random effect, the adjusted threshold parameter is $d_{jm}^{*}=d_{jm}+\alpha_j^T\gamma_k$, which is the adjusted difficulty threshold. By superimposing the interaction between group random effect γ_k and topic load, the local dependence between topics and the load vector of the topic are described. The formula $\alpha_j^T\gamma_k$ is the point product of the topic load vector and the topic group random effect vector. While the questions without random effect of question group are not adjusted and d_{jm} continue to be used.

Finally, when simulating the answer of the subject, the parameter values of the subject, the question and the question group are substituted into the formula (2) (3), and the answer probability P_{njm} of m points obtained by the subject n in the item is calculated, and the final answer result of the subject in the question is generated according to the corresponding probability; Next, a random number is generated: $r_{nj}(0\leq r_{nj}\leq 1)$, and the score is defined as the highest score w corresponding to a cumulative probability less than or equal to: [8]

$$r_{nj}\leq\sum_{m=0}^{w}P_{njm}$$

4.3. Evaluation index

The estimation deviation of RMSE quantization ability parameter, difficulty threshold and load parameter is studied: $RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i-\theta_i)^2}$. At the same time, calculate the deviation (Bias) to detect the systematic error. The formula is: $Bias=\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i-\theta_i)$, where $\hat{\theta}_i$ is the estimated value and θ_i is the true value. Aiming at the random effect of the question group, the variance recovery rate of the question group and the cross-dimensional correlation recovery were investigated.

4.4. Result of simulation research

4.4.1. Parameter estimation of tested capability

It can be seen from Table 1 that when the MPCTM is a true model, the estimation accuracy of the tested ability parameters is significantly better than that of the MPCM model because it can effectively use the question group effect information, while the MPCM model does not consider the question group effect, and there is a systematic deviation in the estimation results. Although the deviation can be partially alleviated by the increase of the sample size, it cannot be completely eliminated. When MPCM is a true model, the overall estimation accuracy of MPCTM model is better, but there is an over-fitting phenomenon in small samples. However, when the sample size is larger, the MPCTM model can adjust and adapt to the data structure, and the estimation accuracy is close to the MPCM model, showing its adaptability and robustness. At the same time, the increase of the number of questions can improve the accuracy of the MPCTM model, while the compensation effect on the MPCM model is obvious in the case of small samples and weakens in the case of large samples.

Table 1. Capacity Parameter Estimation

N	Questions	MPCM_MPCM	MPCM_MPCTM	MPCTM_MPCM	MPCTM_MPCTM
200	15	0.412	0.396	0.502	0.421
500	30	0.285	0.276	0.358	0.302
1000	30	0.185	0.172	0.238	0.192

4.4.2. Problem Parameter Precision Estimation

As can be seen from Tables 2 and 3, when MPCTM is a true model, it is improved in difficulty parameter estimation accuracy by 10-15% compared to the MPCM model, and improved in load parameter estimation accuracy by 15-20%. This indicates that the MPCTM model can use the problem group effect information more effectively and optimize the problem parameter estimation, while the MPCM model does not consider the problem group effect. When the MPCM is a true model, the MPCTM model is better than the MPCM model as a whole, but its lifting range is less than the lifting range when the MPCTM is a true model. This indicates that when the MPCTM model processes the data without the question group effect, although the introduction of the question group effect parameter has a certain improvement in the estimation accuracy, the effect is not as significant as when the MPCTM model processes the data with the question group effect. At the same time, the effect of increasing the sample size on the improvement of parameter estimation accuracy is greater than that of increasing the number of topics, which reflects that in complex models, the increase of sample size can reduce the uncertainty of parameter estimation more effectively, while the increase of the number of topics has a relatively limited effect on the improvement of parameter estimation accuracy.

Table 2. Estimation Of Difficulty Threshold Precision

N	Questions	D (difficulty threshold) RMSE			
		MPCM_MPCM	MPCM_MPCTM	MPCTM_MPCM	MPCTM_MPCTM
200	15	0.532	0.518	0.621	0.558
500	30	0.385	0.372	0.462	0.398
1000	30	0.285	0.268	0.358	0.298

Table 3. Precision Estimation Of Load Parameters

N	Questions	A (load parameter) RMSE			
		MPCM_MPCM	MPCM_MPCTM	MPCTM_MPCM	MPCTM_MPCTM
200	15	0.382	0.365	0.435	0.382
500	30	0.265	0.251	0.325	0.268
1000	30	0.185	0.172	0.238	0.192

4.4.3. Random Effect Estimation

It can be seen from Table 4 that when MPCTM is a true model, the increase of sample size significantly increases the variance recovery rate of question groups from 87.6% of small sample size to 95.6% of large sample size, and the increase of the number of questions brings about an additional 2-3% increase, which indicates that the MPCTM model can accurately capture the effect of question groups, and the cross-dimensional correlation estimation value tends to zero with the increase of sample size and the number of questions, highlighting the advantage of maintaining dimensional independence in multi-dimensional data; When MPCM is a true model, the recovery rate of MPCTM model to the problem group variance is lower than that when MPCTM is a true model, but it still increases with the increase of sample size, and the cross-dimension correlation is closer to zero. This indicates that the

MPCTM model can better adapt to the data by virtue of the structural flexibility when processing the data with no problem group effect, and can more effectively maintain the dimension independence, which reflects the adaptability and robustness of the MPCTM model as a whole.

Table 4. Random Effect Estimation Of Question Groups

N	Questions	Variance Recovery	Avg Correlation	*P*
200	15	0.876	0.0320	<0.05
200	30	0.892	0.0280	<0.05
500	15	0.915	0.0210	<0.05
500	30	0.928	0.0175	<0.05
1000	15	0.942	0.0128	<0.05
1000	30	0.956	0.0093	<0.05

In conclusion, the MPCTM model is superior to the MPCM model in estimating the capability parameters, topic parameters and group random effects. When MPCTM is a true model, it can effectively use the problem group effect information, optimize the parameter estimation, and increase the sample size and number of problems to further improve the estimation accuracy and result stability. When MPCM is a true model, MPCTM model can adapt to data structure and provide accurate estimation by virtue of adaptability and robustness. In addition, Rhat<1.1 under all conditions indicates that the parameters converge well, and the stability of the results increases with the increase of the sample size and the number of topics. The MPCTM model shows significant advantages when dealing with complex test structures, and provides more reliable parameter estimation for multidimensional capability evaluation.

5. Empirical Study

5.1. Research purpose

In this study, MPCM and MPCTM models were used to analyze the final joint examination data of high school English in 2024. The English joint examination is a mixture of two-level and multi-level questions, including fifty-eight English reading questions, accounting for 73%. It examines three English reading attributes of 6804 students: Basic vocabulary, reasoning logic and language discourse. Among them, there are 6 test groups, each of which belongs to the ability dimension and the distribution of specific question groups (see Table 2), so the study forms the random effect of multi-dimensional question groups among projects.

Table 5. Distribution Of Positive Research Ability Dimensions To Question Groups

	Title
Lexical dimension	1, 2, 12, 21
Inference dimension	3,4,5,6,7, 8, 13, 14, 16, 17, 18, 19, 20, 22, 23
Discourse dimension	9, 10, 11, 15
Testlet1	1, 2, 3
Testlet2	4, 5, 6, 7
Testlet3	8, 9, 10, 11
Testlet4	12, 13, 14, 15
Testlet5	16, 17, 18, 19, 20
Testlet6	21, 22, 23

5.2. Evaluation index

This paper studies the fitting effect of the evaluation model of multi-level information criterion system in empirical data. Firstly, the basic goodness of fit is evaluated based on the log-likelihood (logLik) of the model[18], and then the prediction accuracy and parameter quantity are weighed by the Akaike information criterion (AIC)[19], and the penalty effect on the model complexity is enhanced by the Bayesian information criterion (BIC).[20] According to the characteristics of Markov chain Monte Carlo (MCMC) parameter estimation[14], the Deviance information criterion (DIC) and its cross-model difference value (ΔDIC) are calculated synchronously to evaluate the fitting effect and model improvement degree of the model in the empirical data. This multi-dimensional evaluation strategy, which combines classical likelihood theory and Bayesian paradigm, systematically reveals the fitting effect of the model in empirical application through the four-dimensional analysis framework of "basic fitting-prediction efficiency-complexity control-parameter effectiveness."

5.3. Result

The model fitting results are shown in Table 6. The analysis shows that MPCTM is superior to MPCM in all information criteria (AIC, BIC, DIC) and log likelihood values, and the ΔDIC difference reaches the threshold of statistical significance. The results show that the MPCTM model can more accurately capture the complex characteristics of the data and improve the efficiency of parameter estimation and the fitness of the model in test scenarios with multi-level scoring characteristics and group dependence structure.

In the study, MPCTM estimated the random effect variance of six test groups as $\sigma_{\gamma 1}^2$=0.017, $\sigma_{\gamma 2}^2$=1.199, $\sigma_{\gamma 3}^2$=0.988, $\sigma_{\gamma 4}^2$=1.054, $\sigma_{\gamma 5}^2$=0.602, $\sigma_{\gamma 6}^2$=0.237, where the random effect of the first test group was very small, and this test group may not exist. The remaining five test groups produced moderate or high degree of random effect of the test group. It is because MPCM neglects the random effect of the group in the English test that the model fitting error is increased.

Therefore, when test data involves polytomous scoring mechanisms, exhibits local item dependence within testlets, or spans multiple proficiency dimensions, the MPCTM can more precisely elucidate item response mechanisms, providing a more accurate methodological paradigm for handling such complex data. Furthermore, to further validate the advantages of the MPCTM, the study conducted preliminary comparisons with a typical cognitive diagnostic model (DINA). While the DINA model focuses on dichotomous mastery of specific attributes, the MPCTM integrates testlet effects within a continuous multidimensional ability space, making it more suitable for formative assessment and continuous proficiency tracking. In contrast, by preserving the continuity of multidimensional abilities while effectively controlling bias from local dependence through testlet random effects, the MPCTM demonstrates broader applicability.

Table 6. Fitting Results Of Empirical Research Model

	AIC	BIC	logLik	DIC	ΔDIC	*P*
MPCM	200632.2	200987.2	-100264	200530.5	0	<0.01
MPCTM	190355.3	190833.1	-95107.6	190215.3	-10315.2	<0.01

5.4. Empirical application expansion

Latent Class Analysis is a statistical method of mining hidden groups behind data based on observation variables. The study uses LCA to classify students into four categories: Vocabulary expertise (category 1), reasoning advantage (category 2), balanced development (category 3) and prominence (category 4). Category 1 scores high in the lexical dimension, category 2 leads in the inference dimension, category 3 is relatively balanced in each dimension, and category 4 scores highest in the textual dimension.[21]

Based on the four distinct student ability profiles identified through Latent Class Analysis (LCA), this study proposes corresponding tiered instructional intervention strategies: For students with lexical expertise, the focus is on enhancing their inferential skills in reading comprehension and discourse processing training; for those with reasoning strengths, the emphasis lies on addressing vocabulary gaps while further intensifying training in logical reasoning tasks; for students excelling in discourse comprehension, efforts are concentrated on strengthening foundational vocabulary and logical coherence within texts; meanwhile, balanced learners are provided with comprehensive tasks aimed at holistic skill advancement. These targeted interventions contribute to establishing an integrated "diagnosis-intervention-tracking" closed-loop teaching mechanism, thereby furnishing scientific data support for the realization of personalized instruction personalized instruction. This kind of stratified teaching strategy provides the method support for breaking the "one-size-fits-all" teaching mode, and highlights the experimental value of data-driven education reform.

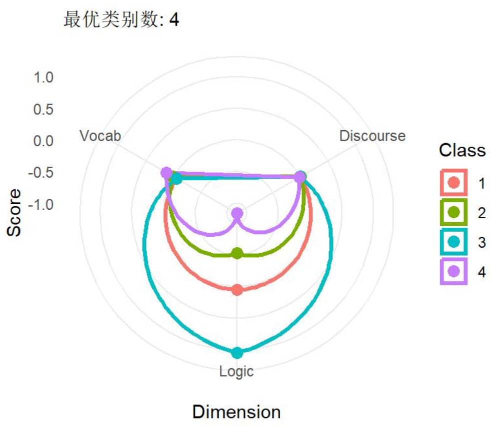

Figure 1. Potential Category Capability Profile Radar Map

6. Research Conclusion and Discussion

6.1. Study conclusion

This paper introduces the random effect parameters of multidimensional random group into MPCTM, and systematically explores the methodological effectiveness of MPCTM in middle school English reading evaluation through simulation and empirical analysis, and draws the following conclusions:

(1) The ability of parameter estimation was significantly improved: MPCTM showed better parameter recovery ability than traditional MPCM in the presence of random effect, and its error control and error correction effect were significant.

(2) The residual distribution is obviously improved: By introducing group random effect, MPCTM can effectively alleviate the disturbance of local dependence on the model, the residual distribution is closer to the theoretical assumption, and the systematic deviation phenomenon is significantly reduced.

(3) The model has better adaptability: In the empirical study, MPCTM showed better fitting performance in the actual English reading evaluation data, and showed better performance in multiple information criteria than the control model.

(4) Quantitative Value of Random Effects in Question Groups: The model successfully identifies local dependencies within most question groups, provides theoretical support for the discourse sharing effect in English reading assessment, and reveals the differences in the effects of different question groups on the assessment results.

6.2. Discussion and Future Outlook

6.2.1. Discussion

By integrating multi-dimensional capability parameters, complex scoring mechanism and group random effect, MPCTM provides a framework of accuracy and adaptability for middle school English reading assessment. The model is suitable for the complex scenarios of multi-level scoring, local dependency and multi-dimensional ability, such as the mixed question type of simple answer and choice question in the question group based on the same reading text. Its multi-dimensional ability parameters can delineate the complex structure of students' reading ability in the dimensions of vocabulary, reasoning and discourse. In addition, the adaptability of the model in empirical research shows that it can not only serve for large-scale standardized evaluation, but also provide ability diagnosis support for school-based personalized teaching, especially for the dynamic educational scene that needs to take into account the continuous tracking of ability and the topic relevance analysis.

In order to give full play to the practical effectiveness of MPCTM, the teaching application strategy can be adjusted. Firstly, when designing evaluation tools, we should increase the proportion of open multi-level scoring questions, add analytical simple answer and logical sorting questions to enhance the analytical ability of the model to high-level cognitive ability. Secondly, combined with the results of students' ability classification generated by LCA, we develop a closed-loop and targeted teaching scheme of "diagnosis-intervention-tracking," customize the vocabulary expansion training module for students with weak vocabulary, or design the interdisciplinary logical strengthening task for the dominant group of reasoning. In addition, it is suggested to incorporate process data such as response time and cognitive load into the model extension framework to dynamically capture learners' ability evolution path and cognitive resource allocation model, and to promote the transformation of assessment tools from static diagnosis to dynamic learning support.

6.2.2. Future outlook

In the future, we can explore the following aspects: (1) The research will deepen the design and data optimization of English reading questions, and enhance the analytical ability of the model to complex cognitive ability by increasing the proportion of multi-level score open questions; Dynamic covariates such as eye tracking, response time and cognitive load were introduced to construct a time-sensitive random effect model. （2）The paper expands the interdisciplinary assessment, explores the applicability of MPCTM in scientific text reading, cross-language logical reasoning and so on, and promotes the transformation of the model from language assessment to comprehensive literacy assessment. （3）Deep ploughing model fusion and technological innovation, combined with cognitive diagnostic theory (CDT) or deep learning framework, study the mixed model with both attribute diagnosis and random effect correction. The efficiency improvement path of Bayesian optimization algorithm in parameter estimation is further explored.

References

[1] Wang, D. X., & Guo, Y. Y. (2022). Development and application of multilevel scoring IRT model for fusion reaction time. Psychological Exploration, 42, 269-278.

[2] Zhan, P. D., Wang, W., & Wang, L. J. (2013). Topic group response theory of new development of item response theory. Advances in Psychological Science, 21, 265-280.

[3] Tu, D. B., Cai, Y., & Paint, S. G. (2009). The new development of item response

theory—The realization of problem group model and its parameter estimation. Psychological Science, 32, 143-145.

https://doi.org/10.16719/j.cnki.1671-6981.2009.06.026

[4] Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika, 34(S1), 91-97.

[5] Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. Psychometrika, 47(1), 105-113.

[6] Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24(3), 185-201.

[7] DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. Journal of Educational Measurement, 43(2), 145-168.

[8] Zhan, P. D., Wang, W., Wang, L. J., et al. (2014). Multi-dimensional problem group effect Rasch model. Journal of Psychology, 46, 108-122.

[9] Zhan, P. D., Li, X. M., Wang, W., et al. (2015). Cognitive diagnosis model of multi-dimensional problem group effect. Journal of Psychology, 47, 689-701.

[10] Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. British Journal of Mathematical and Statistical Psychology, 71(2), 262-286.

[11] Cai, L. (2010). A two-tier full-information item factor analysis model with applications. Psychometrika, 75(4), 581-612.

[12] Zhou, W. J., Tong, W. W., & Guo, L. (2023). Multilevel score cognitive diagnostic question group model. Applied Psychology, 29, 470-479.

https://doi.org/10.20058/j.cnki.cjap.022163

[13] Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149-174.

[14] Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology, 66(1), 8-38.

[15] Wei, D., Liu, H. Y., & Zhang, D. H. (2017). Multidimensional problem group response model: Application extension of multidimensional random coefficient multiple logistic model. Journal of Psychology, 49, 104-114.

[16] Embretson, S. E., & Reise, S. P. (2013). Item response theory for psychologists. Psychology Press.

[17] Baker, F. B., & Kim, S. H. (2017). The basics of item response theory using R (Vol. 10). Springer.

[18] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222(594-604), 309-368.

[19] Akaike, H. (1974). A new look at the statistical model identification problem. IEEE Transactions on Automatic Control, 19(6), 716-723.

[20] Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461-464.

[21] McCutcheon, A. L. (1987). Latent class analysis (Vol. 64). Sage.