Theoretical Assumptions of Polytomous Scored Items and the Suitability of Their Logistic Weighted Model

Jian Xiaozhu, Dai Buyun

2026-08-23
ChinaRxiv

Abstract

Previous polytomous scoring models use multiple difficulty parameters, which cannot accommodate the diverse scoring-point structures of polytomously scored items and cannot reflect the weighting effect of scores. This paper elaborates two basic measurement characteristics of polytomously scored items in ability tests: the existence of multiple types of scoring-point structures and the score-weighting effect of polytomous scoring. Correspondingly, it proposes the theoretical hypothesis of the average difficulty parameter and the theoretical hypothesis of the score-weighting effect. The Logistic weighted model includes an average difficulty parameter and a score-weighting parameter, corresponding respectively to the theoretical hypothesis of the average difficulty parameter and the theoretical hypothesis of the score-weighting effect. It can accommodate multiple types of scoring-point structures and non-continuous scoring forms in polytomously scored items, and can also express the score-weighting effect. The theoretical hypothesis of the score-weighting effect for polytomously scored items will help improve the measurement theory system of item response theory.

Full Text

Theoretical Assumptions of Polytomous Scored Items and the Suitability of Their Logistic Weighted Model

Jian Xiaozhu [1] Dai Buyun [2]

1 (School of Public Policy and Administration, Nanchang University, Nanchang, Jiangxi, 330036)

2 (School of Psychology, Jiangxi Normal University, Nanchang, Jiangxi, 330000)

Abstract: Previous polytomous scoring models have used multiple difficulty parameters, which cannot accommodate the various score-point structures of polytomous scored items and also cannot reflect the score-weighting effect. This paper expounds two basic measurement characteristics of polytomous scored items in ability tests: the existence of multiple types of score-point structures, and the presence of a score-weighting effect in polytomous scoring. Correspondingly, it proposes the theoretical assumption of an average difficulty parameter and the theoretical assumption of a score-weighting effect. The Logistic weighted model contains an average difficulty parameter and a score-weighting parameter, corresponding respectively to the theoretical assumption of the average difficulty parameter and the theoretical assumption of the score-weighting effect. It can fit the various score-point structure types of polytomous scored items and discontinuous scoring forms, and can also express the score-weighting effect. The theoretical assumption of the score-weighting effect for polytomous scored items will help improve the measurement-theory system of item response theory.

Keywords: polytomous scored items; average difficulty parameter; score-weighting parameter; Logistic weighted model

* This paper is one of the research outcomes of the National Natural Science Foundation of China regional project "Theoretical Assumptions, Measurement Techniques, and Application Tests of the Logistic Weighted Model in Psychological and Educational Measurement" (No.: 32360204).

The Measurement Characteristics and Measure-Theory Hypotheses of the Polytomously Scored Item and the Polytomous Model

Xiaozhu Jian[1], Buyun Dai[2]

1 (Department of Psychology, School of Public Policy and Management, School of Economics and Management, Nanchang University; Nanchang, Jiangxi, 330036, China)

2 (School of Psychology, Jiangxi Normal University; Nanchang, Jiangxi, 330036, China)

Abstract:

Previous polytomous models also cannot be effectively applied to polytomous items in aptitude tests, and thus constitute a technical bottleneck in educational and psychological measurement under item response theory. The appropriate and effective polytomous model for polytomous items remains both a research objective and a research difficulty. This article summarizes the limitations and deficiencies of previous polytomous models. It analyzes the nature and characteristics of polytomous items in aptitude tests. First, in the scoring form of a polytomous item, there are multiple structural types of scoring points; second, the full score of the polytomous item reflects the importance of the examination content and has the function of score weighting. In a polytomous item, the values of the scoring points are different, and score values have the characteristics of both continuity and discontinuity. According to the two basic measurement characteristics of the polytomous item, the mean-difficulty-parameter hypothesis and the score-weighting-function hypothesis are proposed respectively, and their theoretical basis is discussed from many angles. This is a breakthrough in the measurement theory of the polytomous item. The weighted-score parameter of the weighted-score logistic model (WSLM) can conform to the mean-difficulty-parameter hypothesis and the score-weighting-function hypothesis in the aptitude test. The WSLM can be better suited to the six types of scoring-point structures in the aptitude test, and the weighted-score logistic model also fits discontinuous scores. WSLM is the proper way to solve the bottleneck problem of polytomous measurement under item response theory. However, the Graded-Response Model, General Partial Credit Model, and other polytomous models

are not suitable for the aptitude test.

Key words polytomous item; mean-difficulty parameter; weighted-score parameter; weighted-score Logistic model;

1 Introduction

Under modern psychological and educational measurement theory (primarily item response theory), previous researchers have proposed many polytomous scoring models or methods. However, in practical measurement applications, ability tests with polytomously scored items involve multiple item types, complex and diverse scoring-point structures, and discontinuous scoring formats. As a result, previous polytomous scoring models often cannot be effectively applied to ability tests in real measurement situations. This is one of the bottlenecks in the measurement technology for polytomously scored test items.

1.1 Polytomous scoring models or methods proposed by previous researchers

van der Linden (1997, 2016) summarized multiple models previously proposed by researchers that are applicable to polytomously scored test items, including the Graded-Response Model (GRM) proposed by Samejima (1969); the Rating Scale Model (RSM) proposed by Andersen (1977); the Nominal Response Model (NRM) proposed by Bock (1972); the General Partial Credit Model (PCM) proposed by Masters (1982); the General Partial Credit Model (GPCM) proposed by Muraki (1992); and several other polytomous scoring models. Nering and Ostins (2010) also summarized previous polytomous scoring models or methods in their monograph and discussed, in separate chapters, polytomous scoring models such as GRM, NRM, PCM, and RSM, while also briefly listing several other less commonly used polytomous scoring models or methods. In recent years, measurement methods for polytomously scored items have included the D-scoring method proposed by Dimitrov and Luo (2019). The D-scoring method adapts the polytomous scoring format in ability tests by virtually constructing a dichotomous scoring format. The research results of Dimitrov and Luo (2019) showed that the D-scoring method did not differ significantly from the GRM in ability parameters and item difficulty parameters. In China, Xiao Hanmin, Du Wenjiu, and Zhang Tingting (2011) proposed an analysis method for polytomously scored items in ability tests, namely test item sample space theory, which explores the relationships among item score levels and among polytomously scored items from the perspective of the knowledge-relationship nodes in polytomously scored items. However, test item sample space theory methods are often applicable only to certain special polytomous item types and special testing contexts in ability tests.

Some researchers have summarized and classified previous polytomous scoring models from the perspective of the form and structure of measurement models.

Thissen and Steinberg (1986) divided dichotomous and polytomous scoring models under item response theory (IRT models containing item parameters) into three categories: classifying the GRM as a difference model, classifying the RSM, NRM, and PCM models as divide-by-total models, and classifying measurement models containing guessing parameters as left-addition models. Bockenholt and Meiser (2017) divided polytomous scoring models into two model frameworks: threshold-based models and item response tree (IR tree models), and used the measurement models under these two frameworks to analyze and compare the evaluation of response styles in psychological scale rating data. Tutz (2020) analyzed the mathematical formal structures of several previous scoring models and argued that dichotomous scoring models are a basic component of polytomous scoring models, further dividing them into several types, such as cumulative, sequential, adjacent-category, and category-response models. This paper holds that the above researchers' summarized classifications are often based on the mathematical forms used to construct the mathematical models, rather than on the measurement properties of polytomously scored items, thereby departing from the measurement foundations of polytomously scored items in practical tests.

1.2 Analysis of the causes of measurement-technology bottlenecks in polytomously scored items

Of course, the fact that many previous researchers have proposed so many polytomous scoring models and conducted numerous studies indicates that polytomously scored items'

Measurement models have attracted the attention of many researchers, but they also indirectly indicate the complexity inherent in measuring polytomously scored items. Although previous researchers have proposed many polytomous scoring models or methods, no polytomous-scoring measurement model can be comprehensively and effectively applied to polytomously scored items in ability tests. This is a bottleneck problem in measurement techniques for polytomously scored items.

The fact that previous polytomous-scoring measurement models have not been widely applied in empirical testing can be summarized as arising from the following two aspects:

First, in empirical tests, the measurement and scoring forms of polytomously scored items are complex: there are multiple types of score-point structures, and there is also a score-weighting effect, that is, the phenomenon of discontinuous scoring forms. This has led previous polytomous-scoring models (or methods) to be unable simultaneously to accommodate the multiple score-point structural forms and discontinuous scoring forms of polytomously scored items.

Second, the modeling ideas adopted by previous researchers when proposing polytomous-scoring models have limitations, including two points: (1) When constructing mathematical models, previous polytomous-scoring models have often used relational expressions involving multiple difficulty parameters to represent polytomous scoring forms, thereby confounding two different measurement properties of polytomously scored items in ability tests: score-weight properties and difficulty-parameter properties. For example, the graded response model proposed by Samejima (1969) and the partial credit model proposed by Masters both use multiple item difficulty parameters to express the grade relationships of polytomous scoring. Jian Xiaozhu and Dai Buyun (2022) discussed from multiple perspectives the use of multiple continuously increasing difficulty parameters under the graded response model to represent polytomous scoring forms; this form of expression cannot effectively embody the score-weighting effect of polytomously scored items in ability tests. (2) Previous polytomous-scoring models have been constructed only according to a particular special form of polytomously scored items or a particular scoring context, "ignoring" the fact that, in ability tests, polytomously scored items have multiple types of score-point structures and discontinuous scoring forms. For example, the partial credit model (PCM) proposed by Masters (1982) applies only to polytomously scored items in which the polytomous scoring form increases successively (incrementally), such as scores of 0, 1, 2, and 3, with no discontinuous score in between; the two-parameter Logistic testlet response model (2-PLBTM) proposed by Bradlow, Wainer, and Wang (1999) applies only to the special context of polytomously scored items in which one large item contains multiple subitems.

1.3 Research Ideas of This Paper

The study by Jian Xiaozhu, Dai Buyun, and Dai Haiqi (2016) mainly discussed the model-construction ideas of the Logistic weighted model, the feasibility of model-parameter estimation, and the good fidelity of simulation in test-simulation research, and used empirical test data for empirical estimation. Different from Jian Xiaozhu et al. (2016), this paper further discusses the issue from two perspectives: (1) From the perspective of measurement theory, it analyzes the specific connotations of the measurement-theory assumptions for polytomously scored items (the average difficulty-parameter theoretical assumption and the score-weighting-effect theoretical assumption). The average difficulty parameters of the Logistic weighted model reflect the average-difficulty theoretical assumption, and the weight parameters reflect the score-weighting-effect theoretical assumption; Logistic weighted models can better express the meaning of the score weights of polytomously scored items in empirical tests. The proposal of the theoretical assumption of score-weighting effects for polytomously scored items will help enrich and improve the measurement-theory system of item response theory. By contrast, the model structures of previous polytomous-scoring models have been unable to effectively express the score-weighting effect of polytomous scoring. (2) From the perspective of test-data forms, previous polytomous-scoring models have often been constructed according to one or two types of polytomous-scoring categories, and therefore cannot accommodate bottleneck problems such as the simultaneous existence of multiple score-point types in polytomously scored items in tests, nor can they accommodate discontinuous scoring forms. Compared with other polytomous-scoring models, Logistic

The weighted model, in the simplified structural form of the average difficulty parameter, can accommodate complex and diverse scoring-point structural types, and can also accommodate multilevel-scoring test items with discontinuous scoring forms. This is an advantage of the Logistic weighted model.

The logic of the argument in this paper will unfold in detail from the above two perspectives. First, it will describe two basic measurement characteristics of multilevel-scoring test items in ability testing: the existence of multiple types of scoring-point structures, the role of score weights in multilevel scoring, and the existence of discontinuous scoring forms. Second, on the basis of these two basic measurement characteristics of multilevel-scoring test items, two corresponding measurement-theoretical assumptions will be proposed. Third, it will analyze whether previous multilevel-scoring models can be applied to the two basic measurement characteristics of multilevel-scoring test items (multiple types of scoring-point structures and discontinuous scoring forms), and show that the Logistic weighted model is constructed on the basis of the two measurement-theoretical assumptions of multilevel-scoring test items. The simulation study by Jian Xiaozhu et al. (2016) demonstrated that the Logistic weighted model can fit multilevel-scoring test items (under continuous scoring forms), and the simulation study by Jian Xiaozhu and Dai Buyun (2025) demonstrated that the Logistic weighted model can fit multilevel-scoring test items with discontinuous scoring forms. Therefore, the Logistic weighted model can better solve the measurement-technical bottleneck problems that exist in multilevel-scoring test items. Finally, the applicability of the Logistic weighted model will be discussed and summarized; the advantage of the Logistic weighted model is that it expresses the measurement characteristics of multilevel-scoring test items in a concise measurement-model form.

2 Two Basic Measurement Characteristics of Multilevel-Scoring Test Items in Ability Testing

2.1 The first measurement characteristic of multilevel-scoring test items: the existence of multiple types of scoring-point structures

The types of multilevel-scoring test items in ability testing include objective item types such as single-choice items, multiple-choice items, fill-in-the-blank items, and ordering items, as well as subjective item types such as short-answer items, essay items, computational-solution items, and application items. The forms and names of these item types are mainly named from the perspectives of how examinees respond and what the response content is. When expressing multilevel-scoring test items with mathematical models, it is not necessary to consider the item-type forms and names of multilevel-scoring test items; instead, the focus should be on analyzing the structural relations among the scoring points within multilevel-scoring test items. In the past, when researchers proposed multilevel-scoring models, they analyzed only the testing situations in which multilevel-scoring test items had continuous scoring forms, and they gave examples of one or two types of scoring-point structures for multilevel-scoring test items. When Samejima (1969) proposed the GRM model, she listed three examples of ability tests. Each scoring point in these three examples was worth 1 point, which made the scoring of these multilevel-scoring test items continuous. The scoring-point structural relations in these three examples belonged respectively to two types of scoring-point structures: parallel relations and stepwise progressive relations. When Masters (1982) proposed the partial credit model, he listed three multilevel-scoring test items: (1) the first was a mathematics item, in which the scoring points had a stepwise progressive form; (2) the second was a screen-picture item, in which a circle was drawn and points were awarded according to the degree of agreement between the examinee' s response content and the answer, belonging to a content-progressive form; (3) the third was a geography item, in which points were also awarded according to the degree of agreement between the examinee' s response content and the answer, belonging to a content-progressive form. It should be noted here that, in previous studies such as Samejima (1969) and Masters (1982), when these multilevel-scoring examples were listed, the scoring-point structural relations of the examples were not discussed.

On the basis of the types of multilevel-scoring test items listed by previous researchers, as well as the various possible types of scoring-point structures in multilevel-scoring test items in ability tests found in actual tests, this paper, according to the logical relations among the scoring points of multilevel-scoring test items, classifies the scoring-point structural forms of multilevel-scoring test items in ability testing into the following six forms (or types):

Type 1: simple additive form, that is, the multilevel scoring of a multilevel-scoring test item is in a simple additive form. In academic proficiency ability tests, the commonly used item type is the single-choice item, and the multilevel scoring of single-choice items often adopts a simple additive form. For example, in a Chinese-subject ability test, selecting

for an item of Type 1 is 2 points, with scoring levels of 0 and 2; in a mathematics subject-ability test, the full score for a Type 1 multiple-choice item is 3 points, with scoring levels of 0 and 3. These all belong to a simple weighted form and are non-continuous scoring forms.

The second category is the parallel form, that is, the multiple scoring points of a polytomously scored item are in a parallel relationship. In other words, whether a score is obtained at each scoring point of a polytomously scored item does not affect the score obtained at other scoring points. For example, in a matching cloze item in an English subject-ability test, each subitem requires selecting one word from a set of alternatives to fill in the blank; each correctly answered blank is scored 1 point, and the score for the entire item is the sum of the scores for all subitems. At this point, the possible scores for this polytomously scored item are: 0, 1, 2, ⋯⋯. If each correctly answered blank in this item is scored 2 points, then the scoring form of the polytomously scored item is non-continuous, and the possible scores are: 0, 2, 4, ⋯⋯.

The third category is the dependent form, that is, there is a certain dependent relationship among the multiple scoring points of a polytomously scored item. For example, in reading-comprehension items in Chinese and English tests, one reading passage is accompanied by 5 subitems, and there may well be certain dependencies among the responses to these 5 subitems. If each subitem is scored 1 point, then the possible scores for this polytomously scored item are 0, 1, 2, 3, 4, and 5, which is a continuous scoring form. If each subitem is scored 2 points, then the possible scores for this polytomously scored item are 0, 2, 4, 6, 8, and 10, which is a non-continuous scoring form. Some studies have found that the dependencies among polytomously scored items are often relatively weak, so the dependencies can be ignored and the items can be directly regarded as having parallel relationships (Sireci, Thissen, & Wainer, 1991).

The fourth category is the stepwise progressive form, that is, in a polytomously scored item, responses must be made progressively in sequence according to the response steps: one must first complete the preceding scoring point before proceeding to the next scoring point. For example, a polytomously scored item includes 3 scoring points in a stepwise progressive relationship. Completing the first scoring point earns 1 point; completing the second scoring point earns another 1 point (2 points in total); and completing the third scoring point earns another 1 point (3 points in total). Thus, the item is scored in a continuous scoring form with scores of 0, 1, 2, and 3. The examples of polytomously scored items listed in Samejima' s (1969) paper are examples of the stepwise progressive form. If the test developer applies score weighting, then completing the first scoring point earns 2 points; completing the second scoring point earns another 4 points (6 points in total); and completing the third scoring point earns another 3 points (9 points in total). The scores are then 0, 2, 6, and 9, which is a non-continuous scoring form.

The fifth category is the content-progressive form, that is, in a polytomously scored item, polytomous scoring proceeds progressively according to the degree to which the examinee' s response content matches the reference answer. For example, in the Similarities subtest of the Wechsler Intelligence Scale, the scoring rule is: if the examinee does not identify any similarity between two things, the response is scored 0; if the examinee only generally mentions some partial or secondary attributes shared by the two things, the response is scored 1; if the examinee can appropriately use a concept to summarize the similarity between the two things, the response is scored 2. This is therefore a continuous scoring form of 0, 1, and 2. If score weighting is applied to this item, it may become a non-continuous scoring form of 0, 2, and 4.

The sixth category is the mixed form, in which a single polytomously scored item contains two or more types of polytomous scoring structures. For example, a polytomously scored item contains 6 subitems: the first 3 subitems are in a parallel relationship, while the last 3 subitems are in a progressive relationship, meaning that one must answer subitem 4 correctly before doing subitem 5, and one must answer subitem 5 correctly before doing subitem 6. Then this polytomously scored item includes two scoring-point structures, the parallel form and the stepwise progressive form, and therefore belongs to the mixed form.

In practical applications, a single ability test often contains many polytomously scored items, and the scoring points of polytomously scored items ···

The structural type may be one or more of the above six scoring forms.

2.2 Measurement Characteristic 2 of Polytomously Scored Items: Polytomous Scoring Has a Score-Weighting Effect

Within the theoretical framework of item response theory, previous researchers have almost never discussed the score-weighting effect of polytomously scored items. Only Jian Xiaozhu et al. (2016), in their study, distinguished between two attributes of polytomously scored items: item difficulty and the degree of importance of the item' s content examined. Below, from the two perspectives of setting the full-score value when constructing polytomously scored items and the scoring-point rules for polytomously scored items, the measurement characteristic that polytomous scoring has a score-weighting effect in ability tests is discussed:

(1) Analyze the score-weighting effect at the item level of polytomously scored items from the perspective of setting full-score values according to the importance of the test content reflected by the polytomously scored item. The importance of the content examined by a test (the content examined by an item) is designed at the time of test construction by setting the full-score value of the item or by increasing the number of items. When the total score of a test is fixed, the basic approach adopted by test designers in item writing and test construction is to increase the importance of the content module examined in two ways: the first is to directly increase the full-score value of the items testing the content, that is, to increase the score weight at the item level while keeping the number of items in that content module unchanged; the second is to increase the number of items in the content module examined while keeping the full-score value of the items in that content module unchanged. Both methods can increase the proportion of score weight of the content module in the total test score, thereby increasing the importance of that content module in the overall test. Of course, the two methods can be applied simultaneously. Take the English section of the National College Entrance Examination as an example. Using changes in the content modules examined in the English section of the National College Entrance Examination to illustrate changes in item full-score values and changes in the proportion of content modules, and their role in strengthening guidance for English-testing ability. Taking the

1983 National College Entrance Examination English test as an example (Liu Qingsi, 2017), English knowledge items (including knowledge of pronunciation, grammar, and vocabulary) were worth 0.5 points or 1 point; the reading-comprehension section had 14 sub-items, each worth 1 point. Later, in order to reflect the importance of reading-comprehension ability and writing ability in English-language learning, during the period from 1989 to 1993 (including the subsequent National College Entrance Examinations), the value of reading-comprehension items in the English test was raised to 2 points per sub-item, and the number of items was increased by 20 sub-items. That is to say, by both increasing the full-score value of the items and increasing the number of items, the weight proportion of reading-comprehension items in the whole test was greatly increased. In short, increasing the full-score value of a certain polytomously scored item will increase the weight proportion in the test of the polytomously scored item and of the content module it examines, thereby increasing the importance of the content examined by that polytomously scored item; conversely, greatly reducing the full-score value of a certain polytomously scored item will reduce the weight proportion in the test of the polytomously scored item and of the content module it examines, thereby weakening its importance. Therefore, the full-score value of a polytomously scored item has a score-weighting effect. The above discusses the meaning of score weight at the item level; in the Logistic weighting model of this paper, the score weight corresponds to the score weight at the item level.

(2) Analyze the score-weighting property of polytomously scored items at the within-item level from the perspective of the scoring-point rules for polytomously scored items. When scoring polytomously scored items, it is necessary to score each scoring point and then accumulate the scores. (a) For parallel or interdependent polytomously scored items, the scores of multiple scoring points that are in a parallel relationship (or an interdependent relationship) within a polytomously scored item are accumulated to obtain the total score of that polytomously scored item. For example, reading-comprehension item types in Chinese and English belong to parallel polytomously scored items. (b) For stepwise-progressive polytomously scored items, it is necessary to assign different score values to each scoring point according to each step, and then to follow the ⋯

The total score of a polytomously scored item is obtained by successive accumulation. For example, in the scoring of integrated item types in mathematics and physics subject-ability tests, assigning scores according to the progress of steps belongs to the stepwise progressive type. (c) For polytomously scored items of the content-progressive type, at each score point different score values must be assigned according to differences in the quality of the response content; if there are multiple score points, these are then accumulated to obtain the total score of the polytomously scored item. For example, the scoring of discussion-type items in history and politics subject-ability tests, in which scores are assigned according to the depth and breadth of the examinee' s response content, belongs to the content-progressive type. Regardless of the type of score-point structure of a polytomously scored item in a test, each score value at each score point of the polytomously scored item plays the same contributory role for that polytomously scored item, and can likewise make the same contribution to the total test score (the examinee' s ability). Therefore, from the perspective of accumulating scores from each score point, the polytomous scoring of polytomously scored items has a score-weighting effect.

In actual tests, a polytomously scored item often contains multiple score points (of course, a polytomously scored item in a simple weighted form contains only one score point); moreover, the score value of each score point may differ: the score value of a score point may be 1 point, or it may be 2 points or more. If a certain polytomously scored item has even one score point worth 2 points or more, then discontinuous scoring may occur in the total score of that polytomously scored item (within the scoring range from 0 points to the full score of the item). That is to say, the score-weighting effect of polytomous scoring in polytomously scored items may, during scoring, appear in the form of continuous scoring, or it may appear in the form of discontinuous scoring. The occurrence of discontinuous scoring in polytomously scored items is one of the external manifestations, in scoring, of the score-weighting effect of polytomous scoring in polytomously scored items.

3 Proposing Two Measurement-Theory Hypotheses for Polytomously Scored Items

On the basis of the two basic measurement characteristics (testing phenomena) of polytomously scored items in ability tests, this paper proposes two corresponding measurement-theory hypotheses: the average-difficulty-parameter theoretical hypothesis, and the score-weighting-effect theoretical hypothesis. This represents a theoretical breakthrough in the measurement-theory assumptions of polytomously scored items, and provides a unified theoretical-hypothesis explanation for testing phenomena such as the multiple types of score-point structures and discontinuous scoring forms of polytomously scored items in ability tests. Thus, it may be considered that the theoretical hypothesis of the score-weighting effect of polytomously scored items will help improve the measurement-theory system of item response theory.

3.1 Measurement-Theory Hypothesis 1: The Average-Difficulty-Parameter Theoretical Hypothesis for Polytomously Scored Items

Whether under classical measurement theory or item response theory, when researchers propose a measurement model (or measurement method) to analyze test data, they need to presuppose certain assumptions in order to simplify test information and ignore certain test information. Under classical measurement theory, when calculating the score of a certain polytomously scored item or calculating the total test score, the following three preliminary theoretical assumptions are needed: (1) the content modules and ability dimensions of the test are, overall, unidimensional; differences among items in the content modules and ability dimensions examined are ignored, and only under this premise can the scores of all items be accumulated to calculate the total test score; (2) each score obtained at different score points within a polytomously scored item has the same score-weighting effect; differences among score points are ignored, and only under this premise can an examinee' s scores at different score points on a polytomously scored item be accumulated into the total score of that polytomously scored item; (3) each score obtained on items of different difficulty has the same score-weighting effect; information about the differing difficulty and discrimination of each item in the test is ignored,

Only under this prerequisite assumption can the examinee' s scores on items of different difficulty be accumulated to obtain a total test score. Similarly, under item response theory, including the logistic weighted model, there are analogous prerequisite theoretical assumptions: (1) the testing objectives and the ability dimensions measured by all items in the test are, overall, unidimensional; (2) each score at different scoring points within the same polytomously scored item has the same score-weighting effect.

On the basis of these two prerequisite theoretical assumptions, and in view of the measurement characteristic that polytomously scored items have multiple categories of scoring-point structures (a measurement phenomenon), a theoretical hypothesis regarding the average difficulty parameter of polytomously scored items is proposed under item response theory. That is, the difficulty characteristic of a polytomously scored item is simplified and expressed by using an average difficulty parameter; in other words, under the logistic weighted model, an average difficulty parameter is used to represent the difficulty of a polytomously scored item. The theoretical basis is discussed below from three aspects.

First, from the perspective of the correspondence between classical test theory and item response theory, in classical test theory the average difficulty of a polytomously scored item is expressed by the index of pass rate P. The pass rate P is the ratio of the mean score to the full score, and in essence it is also a difficulty mean parameter. The calculation of the score pass rate for polytomously scored items under classical test theory uses only the index of score pass rate, without considering the diversity of item types of polytomously scored items or the complex diversity of scoring-point structures, nor the differences among examinee groups at each scoring point of the polytomously scored item. Therefore, previous methods for calculating score rates under classical test theory have also embodied the assumption of an average difficulty parameter. Under an ideal normal distribution, for dichotomously scored items, there is an equivalence relationship between the item difficulty in item response theory and the difficulty (pass rate p) in classical test theory (Qi Shuqing, Dai Haiqi, Ding Shuliang, 2002; Lord & Novick, 1968). Similarly, for polytomously scored items, the difficulty parameter under item response theory can also be derived and expressed by using the average difficulty parameter (pass rate p) calculated under classical test theory, namely

$$b_j = z_j / r_{b_j} \qquad (1)$$

where $r_{b_j} = r_{pbj}\sqrt{p_j(1-p_j)}/y(z_j)$, z_j is the standardized normal score transformed from the average difficulty (average score rate p) of the polytomously scored item, r_{pbj} is the point-biserial correlation coefficient between the score on the polytomously scored item and the total test score, and $y(z_j)$ is the value of the normal distribution density function.

Second, from the perspective of mathematical statistical methods, the average difficulty parameter of a given polytomously scored item can effectively represent the difficulty characteristic of that polytomously scored item. Whether under classical test theory or item response theory, the difficulty of a polytomously scored item can be represented by an average difficulty index (difficulty parameter).

Third, from the perspective of simplification through conceptual generalization, the multiple and complex scoring-point structures of polytomously scored items in ability tests require necessary simplified generalization. An average difficulty parameter can be used to concisely describe and express the difficulty of a polytomously scored item.

3.2 Measurement Theory Hypothesis 2: Theoretical Hypothesis on the Score-Weighting Effect of Polytomously Scored Items

The multiple scoring levels of polytomously scored items in ability tests have the measurement characteristic of a score-weighting effect. Therefore, in measurement theory and measurement models, a theoretical hypothesis regarding the score-weighting effect of polytomously scored items should be proposed, using either an arithmetic cumulative weighting form or a geometric multiplicative weighting form to express the score-weighting effect of polytomously scored items. Within the measurement-theoretical framework of item response theory, when previous researchers proposed polytomous scoring models

there has been no theoretical discussion of the score-weighting effect of polytomous scoring items. This paper further discusses the theoretical basis for the assumption that polytomous scoring items have a score-weighting effect in ability tests from the following three aspects.

First, from the perspective of the measurement objective expressed by score weights in test design. Polytomous scoring items adopt a polytomous scoring format; through polytomous score weighting, they reflect the measurement objectives of an ability test and the direction of importance of the test content. This is also an embodiment of educational training objectives, guiding the key directions of students' ability development and knowledge learning.

Second, from the perspective of the correspondence between classical test theory and item response theory. Under classical test theory, the magnitude of the scores of polytomous scoring items has a weighting effect. Correspondingly, under item response theory, polytomous scoring items likewise have a score-weighting effect. Whether under classical test theory or item response theory, the basic measurement meaning of score weighting should be similar; only the definition and calculation methods of score weighting differ under different measurement theories.

Third, from the perspective of mathematical and statistical methods. Under classical test theory, in ability tests the score-weighting effect is embodied by arithmetically summing the examinee' s scores on polytomous scoring items to calculate the total test score (an arithmetic weighting form). Correspondingly, under item response theory, the scores on polytomous scoring items embody score weights, and it is necessary to use a geometric product form to compute the examinee's ability likelihood function equation (a geometric weighting form), thereby embodying the score-weighting effect.

4 Suitability Analysis of Polytomous Scoring Models for Two Measurement Characteristics of Polytomous Scoring Items

Semejima (1995) proposed several criteria for evaluating polytomous scoring models. The first criterion is that the mathematical model-construction principles and assumptions of the polytomous scoring item should conform to the psychological phenomena reflected by the test. This paper argues that the polytomous scoring models (or methods) discussed in the introduction of this paper do not embody the score-weighting effect of polytomous scoring items when constructing measurement models, and therefore these polytomous scoring models do not effectively express the score-weighting effect of polytomous scoring items in ability tests.

Table 1 summarizes the applicability of several previous polytomous scoring models and the logistic weighted model to various score-point structures of polytomous scoring items. The scope of the summary in Table 1 is limited to ability tests (including subject-ability tests and academic achievement tests), and is premised on the assumption that the full-score values assigned to polytomous scoring items in ability tests have a score-weighting effect. Moreover, the judgments "not suitable," "to be demonstrated," and "suitable" in Table 1 are summary conclusions based on the publication status of previous research papers and on certain consensuses among previous researchers regarding these polytomous scoring models.

Table 1 Fit of polytomous scoring models to two basic measurement characteristics of polytomous scoring items in ability tests

Two measurement characteristics of polytomous scoring items	Two measurement characteristics of polytomous scoring items	GRM	GRM acceleration model	RSM	NRM	PCM	WSLM
Whether scoring is continuous	Score-point structural form	GRM	GRM acceleration model	RSM	NRM	PCM	WSLM
Continuous scoring form	Simple weighted form (nonexistent)						
Continuous scoring form	Parallel form	formally suitable	formally suitable	formally suitable	/	/	suitable
Continuous scoring form	Sequential form	formally suitable	formally suitable	formally suitable	/	/	suitable
Continuous scoring form	Stepwise progressive form	formally suitable	formally suitable	formally suitable	/	formally suitable	suitable
Continuous scoring form	Content-progressive form	formally suitable	formally suitable	formally suitable	/	formally suitable	suitable
Continuous scoring form	Comprehensive form	formally suitable	formally suitable	formally suitable	/	formally suitable	suitable

Two measurement characteristics of polytomous scoring items	Two measurement characteristics of polytomous scoring items	GRM	GRM acceleration model	RSM	NRM	PCM	WSLM
Non-continuous scoring form	Simple weighted form	/	to be demonstrated	/	/	/	suitable
Non-continuous scoring form	Parallel form	/	to be demonstrated	/	/	/	suitable
Non-continuous scoring form	Sequential form	/	to be demonstrated	/	/	/	suitable
Non-continuous scoring form	Step-wise progressive form	/	to be demonstrated	/	/	/	suitable
Non-continuous scoring form	Content-progressive form	/	to be demonstrated	/	/	/	suitable
Non-continuous scoring form	Comprehensive form	/	to be demonstrated	/	/	/	suitable

Note: (1) "/" indicates not suitable; (2) "to be demonstrated" indicates that researchers have not yet conducted analysis and discussion; (3) when the continuous scoring form is also the first category, the simple weighted form is in fact a two-level scoring form of 0 and 1.

In Table 1, (1) when previous researchers proposed these polytomous scoring models, they had already proposed examples of polytomous scoring items suitable for these models and conducted analyses and demonstrations using test data. If one looks only at the form of polytomous score categories, when polytomous scoring is continuous, models such as GRM, the GRM acceleration model, and RSM can be suitable in mathematical form; therefore, the term "formally suitable" is used in the summary table. However, in ability tests, and under the premise that the full score value of polytomous scoring items has a score-weighting function, previous polytomous scoring models such as GRM, the GRM acceleration model, RSM, and PCM do not embody the score-weighting function. That is, they cannot reflect the measurement objective and premise of the score-weighting function in ability tests, and therefore cannot "directly" and effectively fit and reflect the score-weighting function of polytomous scoring items in ability tests. Whether polytomous scoring models such as GRM, the GRM acceleration model, RSM, and PCM can realize the score-weighting function of polytomous scoring items in ability tests indirectly, through score conversion, external weight processing outside the model, and other methods, remains to be further demonstrated and explored.

(2) "/" indicates not suitable, mainly on the basis of the theories on which these polytomous scoring models are constructed and the mathematical expression structures of the models; it is inferred that these polytomous scoring models cannot reflect the score-weighting function of polytomous scoring items in ability tests, or are not suitable for non-continuous…

scoring forms. The inapplicability of these models is based on the consensus of previous researchers, and previous researchers have never demonstrated that such polytomous scoring models are suitable for this type of polytomously scored item. For example, previous polytomous scoring models (including GRM, RSM, PCM, and 2-PLBTM) are not suitable for non-continuous polytomous scoring items. Tutz (2020) also stated that these earlier polytomous scoring models focus only on the scoring ordinality of polytomously scored items; that is, previous polytomous scoring models are suitable only for continuous scoring forms of polytomously scored items in ability tests, and cannot accommodate non-continuous scoring forms.

(3) The "to be verified" entry for the GRM acceleration model in Table 1 means that, although in terms of data form it can fit non-continuous scoring data, the GRM acceleration model was proposed by Samejima (1995), and neither Samejima nor other researchers have subsequently carried out research on the GRM acceleration model. At present, no researcher has conducted a study of test data in which the GRM is applied to non-continuous scoring, nor has any study discussed whether the GRM acceleration model can reflect the score-weighting effect of polytomously scored items in ability tests. This paper argues that, in form, the GRM acceleration model can accommodate non-continuous scoring forms; however, on the one hand, the GRM acceleration model uses multiple difficulty parameters to represent the polytomous score-category form of polytomously scored items, and on the other hand it uses a geometric weighting form

to weight a certain score category. Jian Xiaozhu and Dai Buyun (2022), from the perspective of ability estimation, argued that the GRM' s use of a dual representation form leads to the contradictory phenomenon in which the ordering of examinees' ability estimates is inverted relative to the magnitude of their observed scores. Therefore, the dual representation form of the GRM acceleration model cannot effectively embody the score-weighting effect of polytomously scored items.

(4) The last column of Table 1 shows that the Logistic weighted model can accommodate various score-category structures of polytomously scored items, as well as non-continuous scoring forms. The test-simulation studies of Jian Xiaozhu et al. (2016) and Jian Xiaozhu and Dai Buyun (2025) provide evidential support that the Logistic weighted model can accommodate non-continuous scoring forms of polytomously scored items. From the perspective of the mathematical-model form, the Logistic weighted model includes an average difficulty parameter and a score-weighting parameter, which respectively correspond to the theoretical assumption concerning the average difficulty parameter of polytomously scored items and the theoretical assumption concerning the score-weighting effect. In the construction logic of the Logistic weighted model, the average difficulty parameter b can parsimoniously express the item difficulty of various score-category structural types of polytomously scored items; that is, the average difficulty parameter b is the embodiment of the theoretical assumption concerning the average difficulty parameter of polytomously scored items. The full-score value m_j and the examinee score $u_{\alpha j}$ jointly express the geometric weighting form of polytomously scored items, reflecting the score-weighting effect of polytomously scored items (and also accommodating non-continuous scoring forms); that is, the item full-score value (score-weighting parameter) is the embodiment of the theoretical assumption concerning the score-weighting effect of polytomously scored items. Therefore, the Logistic weighted model achieves a unified theoretical explanation and measurement-model expression for such testing phenomena as the multiple score-category structural types and non-continuous scoring forms (score-weighting forms) of polytomously scored items in ability tests. From the perspective of the measurement meaning and the type of measurement data for polytomously scored items, the Logistic weighted model can better fit polytomously scored items in actual tests.

5 Discussion and Conclusion

5.1 Review of the Advantages and Disadvantages of the Logistic Weighted Model

The Logistic weighted model is applicable to the various complex score-category structures of polytomously scored items in ability tests; it has advantages, but also shortcomings. Its shortcomings are manifested as follows: based on the two prerequisite assumptions of the theoretical assumption concerning the average difficulty parameter and the theoretical assumption concerning the score-weighting effect, the Logistic

The weighted model is relatively concise in terms of its mathematical-model construction, and can be applied fairly well to various structural forms of polytomously scored items and to discontinuous scoring formats; at the same time, it can also accommodate multiple types of score-point structures. However, the Logistic weighted model also has limitations (or shortcomings): the model uses average difficulty to express the difficulty of polytomously scored items, ignoring differences in the structural relationships among the score points of such items. Meanwhile, the geometric weighting form of this model indicates that every score value within a polytomously scored item has the same score-weighting effect, thereby ignoring differences in response patterns among score points, as well as differences among the score values obtained at each score point.

At the same time, to investigate the influence of changes in the maximum score of polytomously scored items on the applicability of the Logistic weighted model, a test-simulation design was conducted. The number of examinees was 1,000; after logarithmic transformation, discrimination followed a normal distribution, and difficulty followed a normal distribution. The simulated maximum scores of all polytomously scored items were gradually increased. According to the test-simulation results, when all items in a test were assigned the same maximum score, as the maximum score increased to 14 or 15 points, the simulation-recovery indices for item difficulty parameters had Bias less than 0.05 and RMSE less than 0.15, indicating good recovery performance for the difficulty parameters. However, when the maximum score was 14, 15, or higher, the simulation-recovery indices Bias and RMSE for item discrimination parameters were slightly larger, and exceeded 0.20, relatively higher than the initial values in the test simulation. The specific factors leading to the increase in the recovery indices for the discrimination of polytomously scored items remain to be further explored.

5.2 Discussion of the Applicable Contexts of the Nominal Response Model and the Graded Response Model

The Logistic weighted model simplifies and omits some measurement information through a relatively concise mathematical measurement model, but it can accommodate the complex and diverse score-point structures of polytomously scored items in ability tests. If more response information for polytomously scored items in a test needs to be reflected, then a relatively complex measurement model must be constructed; however, this will also narrow the range of testing contexts to which the measurement model applies. For example, the nominal response model proposed by Bock (1972) can conduct a detailed analysis of the response information for each option in multiple-choice items, making full use of examinees' response information on the test. However, the nominal response model uses multiple difficulty parameters, and its expression structure is relatively complex. Therefore, the range of testing contexts to which the nominal response model applies is very small: it is suitable only for a particular type of testing context, namely, one in which all items in the test are single-choice items and the maximum score for each single-choice item is the same. In other words, the nominal response model can incorporate more scoring information for polytomously scored items, but its mathematical expression for the measurement model is comparatively complex.

When personality tests and attitude tests are constructed, polytomously scored items do not have the measurement property of score weighting; therefore, the Logistic weighted model is not suitable for personality tests or attitude tests. In addition, in personality tests and attitude tests, the highest score (maximum score) for all items is often the same, and the measurement property of polytomous scores in personality tests is that of equally spaced data. In contrast, the GRM model expresses the corresponding score levels on the difficulty scale using multiple difficulty parameters; thus, the GRM model can be suitable for personality tests and attitude tests.

5.3 Summary

Under modern educational and psychological measurement theory, how to construct a measurement model that can better accommodate polytomously scored items in real testing contexts has long been one of the bottleneck problems in the measurement technology of polytomously scored items. This paper analyzed the reasons why previous polytomous scoring models have been difficult to adapt to polytomously scored items in ability testing, as well as their shortcomings. It then summarized two basic measurement characteristics of polytomously scored items in ability testing: first, there are multiple types of score-point structures, including six forms of score-point structures—simple weighting, parallel relationship, dependency relationship, stepwise progression, content progression, and mixed type; second, the full-score value of a polytomously scored item has a score-weighting effect, and there exists a non-continuous scoring form. This paper proposed two prerequisite assumptions under item response theory (the Logistic weighted model): (1) in a test, the measurement objective and the measured ability dimension measured by all items are, overall, unidimensional; (2) within a polytomously scored item, each score obtained at different score points has the same score-weighting effect. Based on these two prerequisite assumptions, and on the two basic measurement characteristics of polytomously scored items, this paper respectively proposed the corresponding theoretical assumption of average difficulty parameters and the theoretical assumption of score-weighting effects. This represents a theoretical breakthrough in the measurement-theory assumptions for polytomously scored items. The theoretical assumption of score-weighting effects will help improve the measurement-theory system of item response theory, and it also provides a theoretical-assumption basis for the theoretical construction and practical application of measurement models for polytomously scored items. The Logistic weighted model includes average difficulty parameters and score-weighting parameters, and may be considered as being constructed on the basis of the theoretical assumption of average difficulty parameters and the theoretical assumption of score-weighting effects. In summary, the Logistic weighted model can better accommodate multiple types of score-point structures of polytomously scored items, and can also accommodate non-continuous scoring forms; it will be a measurement model that can better solve the bottleneck problems in the measurement of polytomously scored items in real testing contexts.

References

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*(1), 69-81.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Bockenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology, 70*(1), 159–181.

Bradlow, E. T., Wainer, H., & Wang, X. H. (1999). A Bayesian random effects model for testlets]. *Psychometrika, 64* (2), 153–168.

Dimitrov, D. M., & Luo, Y. (2019). A note on the d-scoring method adapted for polytomous test items. *Educational and Psychological Measurement, 79*(3), 545-557.

Lord, F. N., & Novick, M. R. (1968). Statistical theories of mental test scores[M]. Reading, MA: Addison-Wesley. 1968.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement, 16*(1), 159–176.

Nering, L., & Ostins, R.(Eds.). (2010). *Handbook of polytomous item response models.* New York, New York Routledge. 2010.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4), 100-114.

Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika, 60*(4), 549-572.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3): 237–247.

Thissen, D., & Steinberg, L. . (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567-577.

Tutz, G. (2020). On the structure of ordered latent trait models. *Journal of Mathematical Psychology, 96*(4):10234..

https://doi.org/10.1016/j.jmp.2020.102346.

Van der Linden, W. J., & R. K. Hambleton, Eds. (1997). *Handbook of modern item response theory.* New York, Springer.

Van der Linden, W. (2016). *Handbook of Item Response Theory.* New York Springer.

Jian Xiaozhu, Dai Buyun, & Dai Haiqi. (2016). Theoretical construction and simulation analysis of the Logistic weighted model. *Acta Psychologica Sinica*, *48*(12), 1625–1630.

Jian Xiaozhu, & Dai Buyun. (2022). A discussion of the role of difficulty level under the graded response model in multilevel scoring formats. *China Examinations*, *30*(6), 57–67.

Jian Xiaozhu, & Dai Buyun. (2025). Simulation study of discontinuous scoring under the Logistic weighted model. *ChinaXiv Psychological Science Preprint Platform*.

[DOI:10.12074/202511.00062]. Link: https://psych.chinaxiv.org/abs/202511.00062

Liu Qingsi. (2017). Forty years of the English subject in the Gaokao. *China Examinations*, *25*(2), 13–19.

Qi Shuqing, Dai Haiqi, & Ding Shuliang. (2002). *Principles of modern educational and psychological measurement*. Beijing: Higher Education Press.

Xiao Hanmin, Du Wenjiu, & Zhang Tingting. (2011). Unification of polytomous scoring models based on item steps. *Acta Psychologica Sinica*, *43*(12), 1462–1467.

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