A Simulation Study on Non-continuous Scoring under the Logistic Weighted Model

Authors: Jian Xiaozhu, Dai Buyun, Jian Xiaozhu, Dai Buyun

Date: 2025-11-05T00:00:00+00:00

Abstract

Through test simulation studies of polytomously scored items under discontinuous scoring, the results demonstrate that the bias and root mean square error (RMSE) of item parameters under the Logistic weighted model are relatively small, indicating that the Logistic weighted model can simulate discontinuous scoring scenarios and achieve accurate item parameter estimation for polytomously scored items under discontinuous scoring. Based on the principle of chi-square testing, a new chi-square test statistic Q5 is proposed. In test simulation contexts, the goodness-of-fit statistics Q1, Q4, and Q5 under the Logistic weighted model are all below the chi-square critical value, indicating that the test data for polytomously scored items under the Logistic weighted model can achieve effective fit with the model.

Full Text

Preamble

Simulation Study of Non-Continuous Category Scoring Under the Logistic Weighted Model

Jian Xiaozhu[1], Dai Buyun[2]
[1](School of Public Policy and Administration, Department of Psychology, Nanchang University, Nanchang 330031, China)
[2](School of Psychology, Jiangxi Normal University, Nanchang 330022, China)

Abstract

This study investigates the performance of the Logistic Weighted Model (WSLM) for polytomous items with non-continuous scoring categories through simulation. Results demonstrate that the WSLM produces relatively small bias and root mean square error (RMSE) in item parameter estimation, indicating

that the model can effectively handle non-continuous scoring scenarios and achieve accurate parameter recovery for polytomous items. Based on the principles of chi-square testing, a novel fit statistic, Q5, is proposed. Under simulated testing conditions, the fit statistics Q1, Q4, and Q5 for polytomous items under the WSLM all fall below the critical chi-square values, suggesting that the WSLM demonstrates effective model-data fit for polytomous test data.

Keywords: item response theory; weighted-score logistic model; polytomous items; model fit

This research was supported by the National Natural Science Foundation of China Regional Project "Theoretical Assumptions, Measurement Techniques, and Applied Testing of the Logistic Weighted Model in Psychological and Educational Measurement" (Grant No. 32360204).

Introduction

Numerous polytomous item response models have been developed within the framework of item response theory. Van der Linden (1997, 2016) reviewed several models applicable to polytomous items, including Samejima's (1969) Graded Response Model (GRM), Andersen' s (1977) Rating Scale Model (RSM), and Masters' (1982) Partial Credit Model (PCM). Nering et al. (2010) also summarized these polytomous models in their handbook, with dedicated chapters discussing GRM, NRM, PCM, and RSM. However, previous research on these polytomous models has exclusively addressed continuously scored items. To date, no published studies have examined the applicability of existing polytomous models to non-continuous scoring scenarios, and the consensus among researchers is that these conventional models are unsuitable for polytomous items with non-continuous scoring categories.

Jian et al. (2016) proposed and validated the Logistic Weighted Model for polytomous items, but their simulation studies only examined continuously scored items. They did not specifically investigate the performance of the WSLM under non-continuous scoring conditions nor evaluate model-data fit. The present study addresses this gap by conducting simulation studies to evaluate parameter recovery and model fit for non-continuously scored polytomous items under the WSLM.

2.1 Limitations of Existing Chi-Square Formulas for Model Fit Testing

Model fit testing in item response theory involves chi-square tests comparing expected and observed response patterns. Several chi-square test statistics have been proposed for evaluating model fit in IRT, all derived from Pearson' s fundamental chi-square formula. Elliott et al. (1973) proposed a formula for testing model fit of item characteristic curves (for item i):

$$Q_1 = \sum_{j=1}^{10} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})}$$

In this formula, examinees are divided into 10 ability groups when calculating Q1 for a given item, with degrees of freedom determined by the number of groups (9 for dichotomous models). Here, N_j represents the number of examinees in group j, O_{ij} is the observed proportion of correct responses in group j for item i, and E_{ij} is the expected proportion, calculated as:

$$E_{ij} = \frac{1}{N_j} \sum_{k=1}^{N_j} P_i(\hat{\theta}_k)$$

where $P_i(\hat{\theta}_k)$ is the expected response probability for item i based on the ability estimate $\hat{\theta}_k$ for examinee k in group j. Subsequently, Yen (1981) proposed an improvement to Elliott et al.' s formula, resulting in:

$$Q_1 = \sum_{j=1}^{10} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})}$$

These formulas, which divide examinees into 10 groups, yield 9 degrees of freedom with a critical value of 16.92, regardless of group size. However, a significant limitation is that these statistics were designed exclusively for dichotomously scored items.

2.2 Problems with Existing Model-Data Fit Chi-Square Statistics

Consider a test administered to 2,000 examinees divided into 10 groups of 200 each. With E_{ij} ranging from 0.01 to 0.99, suppose the discrepancy between E_{ij} and O_{ij} could be 0.03 (small), 0.05 (medium), or 0.08 (large). Table 1 presents the chi-square values for a single group under these conditions.

[TABLE:1 should appear here]

For difficult items (e.g., $b > 2$), E_{ij} values near 0.01 or 0.02 commonly occur in low-ability groups. Even with a small discrepancy of 0.03, the resulting chi-square value for a single group exceeds 18, surpassing the critical value of 16.92 for 9 degrees of freedom. Similarly, for easy items, E_{ij} values near 0.98 or 0.99 in high-ability groups produce the same problem. The Q1 statistic exhibits a symmetric pattern centered at $E_{ij} = 0.5$, decreasing as E_{ij} approaches 0.5 from either direction, but the chi-square values near 0.01 and 0.99 (18.18) are 25 times larger than those near 0.50 (0.72).

For Q4, when E_{ij} approaches 0.98 or 0.99, the corresponding chi-square values become extremely small. As shown in Table 3 , Q4 decreases monotonically with increasing E_{ij}, with a 100-fold difference between values at $E_{ij} = 0.01$ and $E_{ij} = 0.99$. This extreme variation renders Q4 ineffective for stable assessment of model fit across the ability continuum.

According to Pearson' s chi-square requirements, groups with expected frequencies less than 5 should be excluded from calculation. Similarly, when applying Q1 and Q4 to empirical data, cases where $E_{ij} < 0.01$ or $E_{ij} > 0.99$ should be omitted from analysis, regardless of actual group size.

2.3 Improvement of Chi-Square Formulas for Model-Data Fit Testing

The preceding analysis reveals that Q4 exhibits problematic monotonic decrease with increasing E_{ij}, while both Q1 and Q4 produce inflated values when E_{ij} is extremely small or large. To address these issues, we propose an improved chi-square statistic based on fundamental chi-square principles. The original Q4 formula aligns with Pearson' s basic principle, whereas Yen' s (1981) Q1 includes an additional $(1-E_{ij})$ term in the denominator. To mitigate the excessive inflation problem, we propose taking the square root of the denominator, resulting in:

$$Q_5 = \sum_{j=1}^{10} \frac{N_j(O_{ij} - E_{ij})^2}{\sqrt{E_{ij}(1 - E_{ij})}}$$

This new statistic, designated Q5, uses the same notation as Q1. Table 4 presents Q5 values for the same conditions examined previously. The results show that the Q5 value at $E_{ij} = 0.01$ is only 5 times larger than at $E_{ij} = 0.50$, and at $E_{ij} = 0.05$ it is merely 2.1 times larger. Compared to Q1 and Q4, Q5 substantially reduces the disparity in chi-square values across different E_{ij} levels, providing a more robust fit index.

2.4 Model Fit Calculation Formulas Under the Logistic Weighted Model

The Q5 formula described above applies to dichotomous items. For polytomous items under the WSLM, we must calculate Q5 across multiple score categories. The chi-square formula for exactly u points is:

$$Q_{5i} = \sum_{u=1}^{m} \sum_{j=1}^{10} \frac{N_j(O_{iju} - E_{iju})^2}{\sqrt{E_{iju}(1 - E_{iju})}}$$

where m is the maximum score for item i. When $m = 1$, this reduces to the dichotomous formula. Here, N_j is the number of examinees in group j, O_{iju} is the observed proportion scoring exactly u points, and E_{iju} is the expected proportion. The degrees of freedom for Q_{5i} are calculated as: (1) for continuously scored polytomous items with maximum score m, $df = 10m - 1$; (2) for non-continuously scored items with k actual score categories ($2 \leq k \leq m$), $df = 10k - 1$.

The WSLM has two forms of item characteristic functions, leading to two corresponding chi-square formulas. The formula for u points or above is:

$$Q_{5i} = \sum_{u=1}^{m-1} \sum_{j=1}^{10} \frac{N_j (O_{iju}^{\geq} - E_{iju}^{\geq})^2}{\sqrt{E_{iju}^{\geq}(1 - E_{iju}^{\geq})}}$$

where $O_{iju}^{\geq}$ and $E_{iju}^{\geq}$ represent observed and expected proportions for scores of u or above. Since $E_{ijm} = 1$, the term for $u = m$ equals zero, allowing this simplification. Corresponding Q1 and Q4 formulas for polytomous items can be similarly derived for both exact and cumulative scoring.

3.1 Simulation Algorithm for Continuous Scoring

The simulation procedure under the WSLM for continuous scoring involves five steps:

First, generate N discrimination parameters a from $N(0.7, 0.2)$, truncated to $[0.2, 1.5]$; N difficulty parameters from $N(0, 1)$, truncated to $[-4, 4]$; and M examinee ability parameters from $N(0, 1)$, truncated to $[-4, 4]$.

Second, for each polytomous item j with mean difficulty b_j and maximum score F_j, calculate the expected probability of scoring u points or above for an examinee with true ability θ:

$$P_{ju}^{\geq}(\theta) = \frac{1}{1 + \exp[-a_j(\theta - b_j + w_{ju})]}$$

where w_{ju} represents the weighted-score parameter.

Third, generate a random number r for each examinee-item combination. Compare r with the cumulative probabilities to determine the simulated score: if $r > P_{jF_j}^{\geq}$, the score is F_j; if $r < P_{j1}^{\geq}$, the score is 0; otherwise, the score is u where $P_{ju}^{\geq} < r \leq P_{j(u+1)}^{\geq}$.

Fourth, estimate item parameters from the simulated response matrix using the WLogistic program (available at: https://pan.baidu.com/s/1OY27D5aB-AasezyMvTOlrQ?pwd=bmkc).

Fifth, compare estimated parameters with true values to calculate bias, absolute bias (ABS), and RMSE.

3.2 Simulation Algorithm for Non-Continuous Scoring

For non-continuous scoring, we modify the continuous scoring algorithm by introducing a post-simulation adjustment process. After generating continuous scores, we apply non-continuity rules based on "leap segments"and "leap points."

The adjustment rule defines a critical probability for score modification:

$$LP = \frac{\text{Upper bound of leap segment} - \text{Current score}}{\text{Number of leap points} + 1}$$

If a random number $Rnd \geq LP$, the score changes to the upper bound; otherwise, it changes to the lower bound. For example, a 10-point item might have scores {0, 1, 2, 5, 10}, creating two leap segments: {3,4} (2 points, upper bound=5) and {6,7,8,9} (4 points, upper bound=10). We limit maximum leap segment size to 4 points, as score increments exceeding 5 points are rare in practice.

Specific rules apply based on leap segment size. For a 2-point item scored only as {0,2}, scores of 1 are modified: with probability $(2-1)/(1+1) = 0.5$, they become 2; otherwise 0. For a 4-point item scored as {0,1,2,4}, scores of 3 are similarly adjusted using the upper bound 4 and lower bound 2.

Test designs incorporate mixed continuous and non-continuous items. For items with maximum scores of 2-5 points, the first half of items use non-continuous scoring while the second half use continuous scoring. For tests with varied item types (maximum scores of 1,2,3,5,8,10), non-continuous designs exclude intermediate scores (e.g., for 8-point items, scores 2,3 and 5,6,7 are omitted). Simulation conditions include 1,000 and 5,000 examinees with 50 replications.

3.3 Results for Non-Continuous Scoring Simulations

Table 2 presents the simulation results for non-continuous scoring conditions.

[TABLE:2 should appear here]

Comparing these results with Jian et al. (2016) reveals: (1) for 2-point and 3-point items, recovery indices are nearly identical to continuous scoring conditions; (2) for 4-point, 5-point, and multiple-score conditions, ABS and RMSE values are slightly larger but remain comparable to previous studies. These values also align closely with those reported by other researchers for similar sample sizes and test lengths, indicating good parameter recovery for non-continuous scoring.

Further analysis of the 4-point and 5-point conditions from Jian et al. (2016) shows that while most item parameter estimates closely match their generating

values, a few items exhibit large discrepancies, particularly those with extreme difficulty parameters ($|b| > 2$). Additionally, non-continuously scored items with maximum scores of 4 or 5 points (scored only as 0/4 or 0/5) show larger bias and RMSE than 2-point or 3-point items (scored as 0/2 or 0/3).

4.1.1 Design for Continuous Scoring Fit Analysis

To evaluate model fit under continuous scoring, we simulated six test conditions with maximum scores of 1, 2, 3, 4, 5, and multiple values, each totaling 100 points. Each condition used 1,000 examinees and one replication. Average chi-square values were computed across all items.

4.1.2 Results and Discussion for Continuous Scoring

Table 3 presents the fit statistics for continuous scoring conditions.

[TABLE:3 should appear here]

For 100 dichotomous items, the average Q1 value for cumulative scores is 7.215 and Q4 is 3.375, closely matching Yen' s (1981) findings. All chi-square values are substantially below their respective critical values, indicating excellent model-data fit. As maximum scores increase, Q1, Q4, and Q5 values increase correspondingly due to greater computational complexity and degrees of freedom. Across all conditions, $Q1 > Q4 > Q5$ consistently holds. Notably, chi-square values for exact scores are much larger than for cumulative scores, suggesting that cumulative aggregation reduces discrepancies between expected and observed proportions.

Figures 1-4 [FIGURE:1-FIGURE:4] illustrate expected probabilities and observed proportions for a sample item (Item 11, mean difficulty = -0.363), showing close alignment with theoretical WSLM predictions.

[FIGURE:1-FIGURE:4 should appear here]

4.2.1 Design for Non-Continuous Scoring Fit Analysis

Fit analysis for non-continuous scoring employed five test conditions with maximum scores of 2, 3, 4, 5, and multiple values, each totaling 100 points. For the 2-5 point conditions, the first half of items used non-continuous scoring while the second half used continuous scoring. Each condition used 1,000 examinees and one replication.

4.2.2 Results and Discussion for Non-Continuous Scoring

Table 4 presents the fit statistics for non-continuous scoring conditions.

[TABLE:4 should appear here]

All average chi-square values are below their critical values, indicating good overall model-data fit. The cumulative score chi-square values (Q1, Q4, Q5) are substantially smaller than exact-score values across all conditions, demonstrating satisfactory fit when aggregating across score categories.

For non-continuous items, degrees of freedom must be calculated based on actual score categories. For example, in the 5-point condition where the first 10 items are scored only as {0,5}, $df = 2 \times 10 - 1 = 19$ (critical value = 30.14); for the remaining 10 continuously scored items, $df = 6 \times 10 - 1 = 59$ (critical value = 77.93). Table 5 provides detailed results for the 20-item, 5-point condition.

[TABLE:5 should appear here]

Two items exceed the Q1 critical value and one exceeds the Q4 critical value, but all items fall within the Q5 critical value. This pattern aligns with McKinley and Mills (1985), who reported that some proportion of items typically exceed critical values in simulation studies.

Conclusion

By extending the continuous scoring simulation algorithm to incorporate non-continuous scoring, this study demonstrates that the WSLM achieves small bias and RMSE for item parameters under non-continuous conditions, comparable to results for dichotomous items. The proposed Q5 statistic addresses limitations of existing fit indices by providing a more robust measure across the ability continuum. Simulation results show that Q1, Q4, and Q5 values remain below critical chi-square thresholds for both continuous and non-continuous scoring, confirming that the WSLM provides adequate model-data fit for polytomous test data.

References

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

Elliott, C. D., Murray, D. J., & Saunders, R. (1977). Goodness of fit to the Rasch model as a criterion of test unidimensionality. Manchester: University of Manchester.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

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

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, 9(1), 49-57.

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

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

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

Van der Linden, W. (2016). *Handbook of item response theory*. New York: Springer.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5(2), 245-262. https://doi.org/10.1177/014662168100500212

Jian, X., Dai, B., & Dai, H. (2016). Theoretical construction and simulation analysis of the Logistic weighted model. *Acta Psychologica Sinica*, 48(12), 1625-1630.

Correspondence: Jian Xiaozhu, E-mail: jianxiaozhu2003@126.com; Dai Buyun, E-mail: biweijianpsy@qq.com

Appendix: Video Tutorial for WSLM Simulation and Parameter Estimation Software

The video tutorial for the WSLM simulation and parameter estimation software is available at:
Link: https://pan.baidu.com/s/1dRnV9Vqu7rGCjPax0YIGew?pwd=djtc
Access code: djtc

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