Science Journal of Applied Mathematics and Statistics
2026, Vol. 14, No. 5, pp. 113–118
https://doi.org/10.11648/j.sjams.20261405.11

SciencePG
Science Publishing Group

Research Article

Item Response Theory Model Based on the Softmax Function

Vladimir Bratishchenko*

Department of Mathematical Methods and Digital Technologies, Baikal State University, Irkutsk, Russia

Abstract

Quite often, test performance is measured on polytomous scales. The Partial Credit Model (PCM) used in such cases applies latent difficulty parameters for each score point gradation and a single test-taker ability parameter. To describe the score probability distribution more accurately, we propose to use ability parameters for each assessment gradation. For this purpose, we propose to describe the distribution by a multivariate sigmoid function—a softmax function. The maximum likelihood method was used to obtain equations linking test scores and the model's latent parameters. Numerical methods were used to solve the equations. If all the scores of the test-taker (test item) are the same, then the corresponding latent parameter tends to infinity, and the probability to one or zero. In such cases, the parameter is limited to the range [-20, 20]. The method has been tested on IQ testing data. The analysis of variance showed that the test-takers and test items had no influence on the deviation of scores from mathematical expectations. The proposed model and PCM were compared in terms of prediction accuracy. The ability parameters of the models' are collated. To assess the quality of the score prediction, confusion matrices were obtained and ROC curves were constructed. For the proposed model, the score prediction accuracy was 69%, while the PCM prediction accuracy was 57%. This shows that the softmax model more accurately describes the scores. The model can be used to probabilistically describe test scores, study the characteristics of test assignments, and predict academic performance.

Keywords

Item Response Theory, Models with Latent Parameters, Scores in Polytomous Scales, Probability Distribution of Scores, Softmax Function

1. Introduction

Item Response Theory (IRT) offers [1-4] a method for describing the influence of various factors on score items that is fundamentally different from regression models, namely, in the form of parameters of the probability distributions of scores. IRT models have been successfully applied to processing data from tests, questionnaires [5], rating assessments [6-8], medical data [9, 10], counseling research [11].

Like any ingenious idea, IRT generates new models that use IRT principles to account for the specifics of the data. They use beta distribution [12], division of respondents into classes [13] and Dirichlet processes [14]. IRT models are used together with neural networks to assess the ability of test takers [15]. The paper [16] examines the tendency of respondents to choose the extreme or average versions of the response scale. In this paper, a model is proposed that takes into account such inclinations.

*Correspondence: Vladimir Bratishchenko (vbrat56@mail.ru)

Received: 27 July 2026; **Accepted:** 9 November 2026; **Published:** 25 November 2026

Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

The scores may be obtained by polytomous scales. To process such data, a Partial Credit Model (PCM) is used, in which the score probability distribution relates to the test-taker's ability parameter and the difficulty parameters corresponding to the steps involved in obtaining the score. In some cases, test-takers show no evidence of some score gradations. For example, there are no failures in well-prepared students. Nevertheless, in such cases, the model can predict significant probabilities for unobserved scores. Prediction accuracy can be expected to increase by introducing ability parameters for each score gradation. This paper proposes such a model.

2. Methodology

To describe the probability distribution of the score, we propose to apply the softmax function. This function is a multivariate analogue of the logistic function and is widely used to describe the probability distributions of classes determined by neural networks. We assume that the probability distribution of the score is described by the softmax function in the following way:

$$p_{ijk} = P(X_{ij} = k) = \frac{\exp(\theta_{ik} - \delta_{jk})}{\sum_{h=0}^{K} \exp(\theta_{ih} - \delta_{jh})}, \quad (1)$$

where X_{ij} is the random variable of the i-th test taker's score obtained for the j-th test item,

$k = 0, \dots, K$ is the gradation number in the polytomous scale,

K is the maximum gradation number in the scale,

θ_{ik} is the ability of the i-th test taker to achieve the k-th graduation,

δ_{jk} is the difficulty of achieving the k-th graduation on the j-th test item.

The logarithm of the likelihood function

$$\lambda = \sum_{i,j,k} \left(I_{ijk} (\theta_{ik} - \delta_{jk}) - ln\left(\sum_{h=0}^{K} exp(\theta_{ih} - \delta_{jh}) \right) \right)$$

where I_{ijk} is the k-th gradation indicator for X_{ij}, - allows getting equations for calculating parameters based on scores.

In general, different test-takers may have various sets of test items. In such cases, summation will be performed by different index sets. From here on, the following conventions will be used: summation by i, j implies using all N scores, summation by i implies using test-takers' scores for the j-th item, summation by j implies using the scores for the i-th test-taker, and summation by k implies using all score gradations.

To determine the λ maximum, we equate the partial derivatives to zero

$$\frac{\partial \lambda}{\partial \theta_{ik}} = \sum_j \left(I_{ijk} - \frac{\exp(\theta_{ik} - \delta_{jk})}{\sum_{h=0}^{K} \exp(\theta_{ih} - \delta_{jh})} \right) = \sum_j (I_{ijk} - p_{ijk}) = 0, \quad (2)$$

$$\frac{\partial \lambda}{\partial \delta_{jk}} = \sum_i \left(-I_{ijk} + \frac{exp(\theta_{ik} - \delta_{jk})}{\sum_{h=0}^{K} exp(\theta_{ih} - \delta_{jh})} \right) = \sum_i (-I_{ijk} + p_{ijk}) = 0. \quad (3)$$

To solve such equations in IRT, the tangent method is applied:

$$x^{(l+1)} = x^{(l)} - \frac{f(x^{(l)})}{f'(x^{(l)})}.$$

The left sides of equations (2) and (3) are used as $f(x)$, and the second derivatives of the log likelihood function are used as $f'(x)$

$$\frac{\partial^2 \lambda}{\partial \theta_{ik}^2} = -\sum_j \left(\frac{\exp(\theta_{ik} - \delta_{jk})}{\sum_{h=0}^{K} \exp(\theta_{ih} - \delta_{jh})} \left(1 - \frac{\exp(\theta_{ik} - \delta_{jk})}{\sum_{h=0}^{K} \exp(\theta_{ih} - \delta_{jh})} \right) \right) =$$

$$= -\sum_j \left(p_{ijk} (1 - p_{ijk}) \right),$$

$$\frac{\partial^2 \lambda}{\partial \delta_{jk}^2} = -\sum_i \left(\frac{\exp(\theta_{ik} - \delta_{jk})}{\sum_{h=0}^{K} \exp(\theta_{ih} - \delta_{jh})} \left(1 - \frac{\exp(\theta_{ik} - \delta_{jk})}{\sum_{h=0}^{K} \exp(\theta_{ih} - \delta_{jh})} \right) \right) =.$$

$$= -\sum_i \left(p_{ijk} (1 - p_{ijk}) \right).$$

To determine the initial values, we assume that the probabilities of single indicator values are proportional to the frequency of such observations.

$$\frac{1}{m_i} \sum_j I_{ijk} = \exp(\theta_{ik}^{(0)}) \text{ or } \theta_{ik}^{(0)} = ln\left(\frac{1}{m_i} \sum_j I_{ijk} \right),$$

where m_i is the number of scores of the i-th test taker,

$$\frac{1}{n_j} \sum_i (1 - I_{ijk}) = exp(\delta_{jk}^{(0)}) \text{ or } \delta_{jk}^{(0)} = ln\left(\frac{1}{n_j} \sum_i (1 - I_{ijk}) \right),$$

where n_j is the number of scores of the j-th item.

A characteristic feature of the calculation procedure is the possibility of obtaining a zero value under the logarithm sign or infinite growth of the latent parameter during iterative calculations. This situation arises if all scores are the same or there are no scores for a certain gradation. In such cases, the latent parameters are limited to certain minimum and maximum values (in the computer program for calculating the parameters, -20 and +20 are suggested).

The parameters are determined up to a certain common bias,

which depends on the initial parameter values. The bias does not affect the probability distribution (1), since the probabilities are calculated based on the difference between the latent parameters of ability and difficulty.

3. Application

To test the model, an IQ test was selected[1]. The dataset includes test results from 3,194 respondents. The data was posted on the website of OpenPsychometrics.org, a nonprofit organization dedicated to educating the public about psychology and collecting data for psychological research. The test included term association tasks. Each question had eight possible answers, of which three to five were correct. The score was calculated as follows: one point was added for each correct answer and subtracted for each incorrect answer. Thus, the final score ranged from -3 to 5.

The scores of 266 test-takers, whose test scores were identical, were excluded from the original set of scores as in such cases the latent parameter of ability tends to infinity, and the probability of the corresponding gradation tends to one. Excluding such scores does not affect the results of estimating other latent parameters. One observation of a -3 score was also excluded as statistically insignificant. As a result, the scores ranged from -2 to 5. This reduced the number of scores analyzed to 20,596.

For each score item, formula (1) determines the probability and mathematical expectation. For many test takers, only 2-3 gradations were present in the responses across the entire scale from -2 to 5. For them, the probability of missing gradations was zero, and the corresponding hidden parameters θ_{ik} had a value of -20. Similarly, if there was a complete lack of gradation among the score items of a task, then the difficulty parameter δ_{jk} gets a value of +20, and the probability of such a gradation turns out to be close to zero. Due to the set of latent parameters for each gradation, the softmax model can be used to process score items from different scales. For example, in tasks 1-5 of the data set under study, the maximum score is 4. The model determines zero probabilities for unused gradations.

For comparison, the data set was processed using the PCM and softmax models. A histogram of the deviations of scores from expected values is shown in Figure 1. The softmax model histogram is highlighted in gray, while the PCM histogram is outlined in black. A density plot of the normal distribution is superimposed on the histogram for comparison. The deviations failed the chi-square normal distribution test. Analysis of variance showed no effect of test takers or tasks on the deviations. Both models correctly predict the expected values of scores, accounting for statistical error.

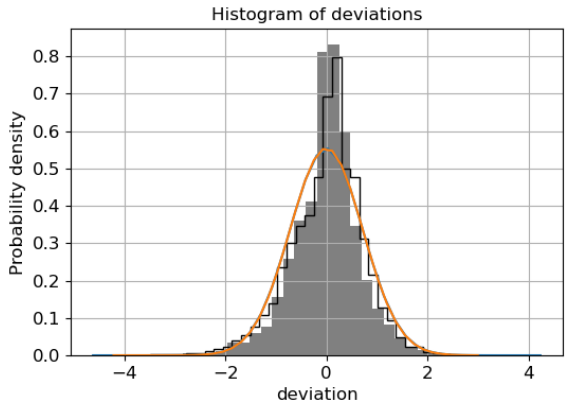


Figure 1. *Histogram of deviations of scores from expected values.*

The models were applied to forecast the scores. The forecast value was determined by the maximum probability. The Confusion matrix for the softmax model is shown in Table 1. The percentage of correctly forecast scores was 69.21%.

Table 1. *Confusion matrix for the softmax model.*

Gradations		Forecast								Total
		-2	-1	0	1	2	3	4	5	
Actual	-2	19		2	1	3	2			27
	-1		104	2	64	12	14	1		197
	0	3	14	234	83	120	59	7		520
	1	2	37	66	1697	543	366	35	2	2748
	2	2	17	86	434	3615	1179	162	18	5513
	3	1	8	37	298	999	5345	666	23	7377
	4		1	3	21	168	692	3027	35	3947
	5						3	20	143	166

Gradations	Forecast								Total
	-2	-1	0	1	2	3	4	5	
Total	27	181	430	2598	5460	7660	3918	221	20495

The Confusion matrix for PCM is shown in Table 2, the percentage of correctly predicted scores was 57.37%. The highest percentage of PCM errors is associated with the extreme gradations. This is due to the single test-taker ability parameter in PCM, while the softmax model applies ability parameters to each score gradation.

Table 2. *Confusion matrix for the PCM.*

Gradations		Predicted								Total
		-2	-1	0	1	2	3	4	5	
Actual	-2		7		18	2				27
	-1	3	31	14	126	17	5	1		197
	0		18	6	267	176	52	1		520
	1		52	34	1156	1104	382	20		2748
	2				558	2887	2011	57		5513
	3				127	1260	5169	818	3	7377
	4				7	107	1333	2489	11	3947
	5						39	107	20	166
Total		27	181	430	2598	5460	7660	3918	221	20495

For the PCM model, it is defined as

$$ln\left(\frac{P\{X_{ij}=k\}}{P\{X_{ij}=k-1\}}\right) = \alpha_i - \beta_{jk},$$

where β_{jk} is the threshold for the k-th answer option: the larger it is, the larger the ability parameter α_i must be in order to choose the correct answer. For the softmax model the equation will be as follows:

$$ln\left(\frac{P\{X_{ij}=k\}}{P\{X_{ij}=k-1\}}\right) = ln\left(\frac{p_{ijk}}{p_{ij(k-1)}}\right) = \left(\theta_{ik} - \theta_{i(k-1)}\right) - \left(\delta_{jk} - \delta_{j(k-1)}\right).$$

Comparing the last two formulas allows us to establish whether the PCM and the softmax model's parameters correspond to each other:

$$\alpha_i = \theta_{ik} - \theta_{i(k-1)},$$

$$\beta_{jk} = \delta_{jk} - \delta_{j(k-1)}.$$

The difference $\theta_{ik} - \theta_{i(k-1)}$ must be the same for all k gradations for the softmax model to fully correspond to the PCM. Parameter evaluation results show that this is not the case. A fragment of the ability parameter evaluation is presented in Table 3. Significant variations in the differences $\theta_{ik} - \theta_{i(k-1)}$ for various k indicate that the softmax model accounts for the differences in test-taker ability which allows obtaining different gradations of scores and more accurately describes the probability distributions of test item scores.

Table 3. Ability parameters for randomly selected test takers.

θ_{i1}	θ_{i2}	θ_{i3}	θ_{i4}	$\theta_{i2}-\theta_{i1}$	$\theta_{i3}-\theta_{i2}$	$\theta_{i4}-\theta_{i3}$
-2,004	-1,123	-0,711	-2,463	0,880	0,412	-1,751
-1,824	-0,603	-1,242	-2,763	1,221	-0,640	-1,520
-2,004	-1,123	-0,711	-2,463	0,880	0,412	-1,751
-2,004	-1,123	-0,711	-2,463	0,880	0,412	-1,751
-2,224	-1,322	-1,201	-0,974	0,903	0,121	0,227

The ROC curve plots (Figures 2 and 3) also demonstrate significantly smaller prediction spread for the softmax model. Single AUC (areas under the curve) values are due to the small number of observations at the corresponding score gradation.

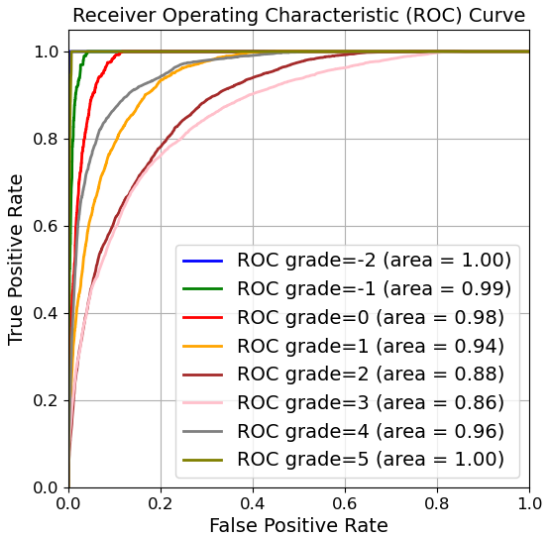


Figure 2. ROC curves for score forecast using a softmax model.

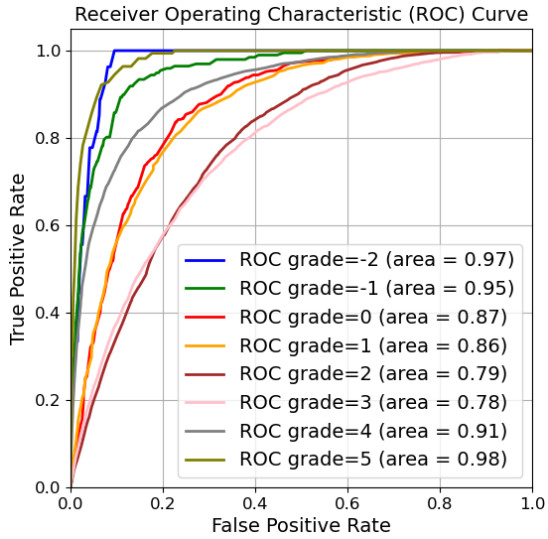


Figure 3. ROC curves for predicting scores using a PCM.

4. Conclusion and Discussion

The softmax model offers a detailed description of test-taker ability, which improves modeling accuracy. The resulting probability distributions of test scores can be used to calculate statistical characteristics of test results and forecast academic performance. Replacing step-by-step evaluation of test scores with a probability distribution can more accurately describe the measurements when grading is not used, for example, when processing questionnaire results. When the number of parameters increases, a larger number of observations is needed for their reliable estimation.

Abbreviations

AUC	Area Under the ROC Curve
IQ	Intelligence Quotient
IRT	Item Response Theory
PCM	Partial Credit Model
ROC	Receiver Operating Characteristic

Author Contributions

Vladimir Bratishchenko: Conceptualization, Data curation, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press, 1980. 199 p.

[2] Wright B. D., Masters G. N. Rating scale analysis: Rasch measurement. Chicago, Illinois: MESA Press; 1982. 223p.

[3] Masters G. N. A rasch model for partial credit scoring. Psychometrika 47, 149-174 (1982). https://doi.org/10.1007/BF02296272

[4] Chen, Y., Li, X., Liu, J., Ying, Z. Item response theory—A statistical framework for educational and psychological measurement. Statistical Science, 40(2), 167-194.

[5] Qiu, X., de la Torre, J., Wang, Y.-G. Wu, J. Item Response Theory Models for Polytomous Multidimensional Forced-Choice Items to Measure Construct Differentiation. Educational Measurement: Issues and Practice, 43: 157-168. https://doi.org/10.1111/emip.12621

[6] Hu B., Zhou Y., Wang J., Li L., Shen L. Application of Item Response Theory to Collaborative Filtering. In *Advances in Neural Networks* - ISNN 2009. Berlin, Heidelberg: Springer, 2009. P. 744-753. (Lecture Notes in Computer Science, Vol. 5551). https://doi.org/10.1007/978-3-642-01507-6_86

[7] Bergner, Y., Halpin, P., Vie, J.-J. Multidimensional Item Response Theory in the Style of Collaborative Filtering. Psychometrika, *87*(1), 266–288. https://doi.org/10.1007/s11336-021-09788-9

[8] Silva, W., Spalenza, M., Bourguet, J. de Oliveira, E. Towards a Tailored Hybrid Recommendation-based System for Computerized Adaptive Testing through Clustering and IRT. In Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020) - Volume 1, pages 260-268. https://doi.org/10.5220/0009419902600268

[9] Gilbert, J. B., Himmelsbach, Z., Soland, J., Joshi, M., Domingue, B. W. Estimating heterogeneous treatment effects with item-level outcome data: Insights from Item Response Theory. Journal of Policy Analysis and Management, 44, 1417–1449. https://doi.org/10.1002/pam.70025

[10] Horie T, Kurisu K., Inada S., Kawahara K., Matsuyama Y., Kikuchi H., Yamamoto Y., Yamauchi T., Yoshiuchi K.. Development of computer adaptive tests to assess the psychological status of individuals with an eating disorder or type 2 diabetes. Biopsychosoc Med. 2025 Feb 14; 19(1): 2. https://doi.org/10.1186/s13030-025-00325-z

[11] Cook, R. M., & Wind, S. A. (2026). Making Meaning of Item Response Theory in Counseling Research. Measurement and Evaluation in Counseling and Development, 59(2), 109–113. https://doi.org/10.1080/07481756.2026.2639942

[12] Li S, Shin HJ. An Extended Two-Parameter Logistic Item Response Model to Handle Continuous Responses and Sparse Polytomous Responses. Psychometrika. 2025; 90(5): 1705-1731. https://doi.org/10.1017/psy.2025.10044

[13] Bartolucci F, Farcomeni A, Scaccia L. A Nonparametric Multidimensional Latent Class IRT Model in a Bayesian Framework. Psychometrika. 2017; 82(4): 952-978. https://doi.org/10.1007/s11336-017-9576-7

[14] Miyazaki K, Hoshino T. A Bayesian Semiparametric Item Response Model with Dirichlet Process Priors. Psychometrika. 2009; 74(3): 375-393. https://doi.org/10.1007/s11336-008-9108-6

[15] Iskakova, A., Salykova, O., Didarbekova, N., Ivanova, I.; Akmoldina, A., Zhumadillayeva, A. A Hybrid Neural-IRT Framework for Addressing Cold-Start Challenges in Computerized Adaptive Testing. Computers 2026, 15, 132. https://doi.org/10.3390/computers15020132

[16] Bolt, D. M., & Meng, L. (2025). IRT-based response style models and related methodology: Review and commentary. British Journal of Mathematical and Statistical Psychology, 00, 1–29. https://doi.org/10.1111/bmsp.70006

1 Alpha-Version Fullscale IQ Test Responses. https://www.kaggle.com/datasets/lucasgreenwell/alphaversion-fullscale-iq-test-responses/data