Accuracy and Bias of the Rasch Rating Scale Person Estimates using Maximum Likelihood Approach: A Comparative Study of Various Sample Sizes

Nurul Hafizah Azizan1,2,*, Zamalia Mahmud1 and Adzhar Rambl1

1Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
2Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kelantan Kampus Kota Bharu, 15050 Kota Bharu, Kelantan, Malaysia

Abstract. The focus of this article is to evaluate the maximum likelihood estimation (MLE) performance in estimating the person parameters in the Rasch rating scale model (RRSM). For that purpose, 1000 iterations of the Markov Chain Monte Carlo (MCMC) simulation technique were performed based on a different number of sample sizes and several number of items. The performance of MLE in estimating the person parameters according to the different number of sample sizes was compared through accuracy and bias measures. Root mean square error (RMSE) and mean absolute error (MAE) were used to examine the accuracy of the estimates, while bias in estimation was assessed through the mean difference of estimates and true values of the person parameters. The simulated survey data sets in this study were generated according to the RRSM under the assumption of normality was satisfied. Results from the simulation analysis showed that in comparison to the larger sample sizes, smaller sample sizes tend to produce higher RMSE and MAE. In addition, the maximum likelihood estimates of the person parameters in smaller sample sizes also recorded a higher value of the mean difference of the person estimates and its true values compared to larger sample sizes. Findings from this study imply that the use of the MLE approach in small sample sizes results in less accurate and highly biased person estimates across the number of items.

Keywords: Bias, Maximum likelihood estimation (MLE), Mean absolute error (MAE), Rasch rating scale model (RRSM), Root mean square error (RMSE)

1. Introduction

Rasch measurement model (RMM) is among the most commonly used psychometric model in social science studies. The dichotomous Rasch model (DRM) was the first model introduced by Rasch in 1960 [1]. In the beginning, the method was used in learning environment. Beginning from last decade, it is one of the essential tools used among teachers to validate test items and to examine students’ assessments in learning [2]. Research have been done to extend this model so that it can be used to measure polytomous responses. The Rasch model developed for polytomous responses is named as Rasch rating scale model (RRSM) and partial credit model (PCM). RMM is also widely employed in
development and validation of survey instrument. Researchers used RMM to validate scale items included in the instrument before it can be further used in data collection for actual surveys.

Unlike PCM, which allows mixture in a number of categories, RRSM, on the other hand, has additional assumption that a scoring of the ordered response categories must be equidistant [3]. As most of the survey instruments usually use equal response categories such as a 5-point Likert scale (e.g., strongly agree, agree, neutral, disagree and strongly disagree) across the items, RRSM is among the most popular methods employed by many research [4–8]. The RRSM model introduced by Andrich (1978) can be rewritten as in equation (1) [9–10], where according to this equation, the probability that person \( i \) with ability \( \beta_i \) to obtain a score of \( k \) (\( k = 0, ..., m \)) for the RRSM with discrete response is presented as follows:

\[
P(\beta_i, \delta_j + \tau_k) = \frac{\exp{\sum_{k=0}^{m} \left( (\delta_j + \tau_k) \right)}}{\sum_{r=0}^{m} \exp{\sum_{k=0}^{m} \left( (\delta_r + \tau_k) \right)}}
\]  

(1)

As shown in Equation (1), \( \delta_j \) represents the difficulty parameter for item \( j \), and \( \tau_k \) is a threshold parameter for category \( k \), that is the parameter for the transition from category \( k - 1 \), to category \( k \). Item \( j \) has \( m \) categories, and \( k \) is the count of the number of successfully completed categories for that item. In the present context, \( k \) is the count of gaps within text \( j \) that person \( i \) filled in correctly. The likelihood function of RRSM is required in order to proceed with the estimating of the parameters of interest. RRSM likelihood function given by [11] is shown in Equation (2):

\[
L = \frac{\exp{\sum_{i}^{N} \sum_{j}^{n} \sum_{k=0}^{m} \left[ (\delta_j + \tau_k) \right]}}{\prod_{i}^{N} \prod_{j}^{n} \left[ \sum_{k=0}^{m} \exp{\sum_{j=0}^{k} \left[ (\delta_j + \tau_k) \right]} \right]}
\]  

(2)

In RMM framework, there are two main parameters which need to be estimated. These parameters refer to the item and person measured. As shown in Equation (1), the item and person parameters are denoted as \( \beta_i \) and \( \delta_j \), respectively. As highlighted by Engelhard [12], both parameters can be estimated either using a non-iterative or iterative technique. For non-iterative technique, the graphical and PROX methods can be used to estimate the parameters. However, as RMM is a non-linear function, therefore, an iterative approach via maximum likelihood estimation techniques (e.g., joint maximum likelihood estimation, conditional maximum likelihood estimation and marginal maximum likelihood estimation) is more suitable to be used [13]. With these iterative techniques, the initial estimates (i.e., usually obtained through non-iterative methods) are improved consistently until the final estimates are obtained. Although conventional approaches like MLE techniques are popular in RMM framework, extra caution is needed particularly when estimating the parameters in a small number of samples.

This is because when Rasch model is fitted in relatively small sample sizes, less precise and less robust estimates with larger standard error of the parameter estimates could be produced, which result in less powerful fit analysis [14]. Generally, larger sample sizes are essentially required to ensure the accuracy of the estimates. The estimates were identical in the sample size ranging between 2000-3000, where the estimates tend to be very unstable in smaller numbers of samples [14]. The size of the sample has a significant effect on the accuracy of the estimates. Usually, a large number of samples is needed to get more accurate estimates; as the larger the sample sizes, the more accurate and less biased the estimates would be. The estimates can be biased and less accurate when sample sizes are small. Therefore, an adequate sample size is needed to ensure the accuracy of the estimates.

For that reason, researchers need to carefully plan and choose an adequate number of sample sizes as it gives an impact on the accuracy of the estimates in most statistical models including item response theory (IRT) [15]. To get more precise and accurate estimates, researchers need to follow the rules suggested by previous studies, which require them to collect more data and provide constraints when fitting the model in such a way that normality assumption is satisfied [13]. Besides that, it is also
essential for researchers to select appropriate techniques to be used in order to estimate the parameters of interest in their studies. Inappropriate estimation techniques tend to produce estimates which lack accuracy and can be potentially biased. High bias and inaccurate estimates can produce misleading statistical inferences which result in poor decision-making processes. Selection of adequate sample sizes and appropriate parameter estimation techniques are important issues in a research. The main purpose of this article is to evaluate and compare the maximum likelihood estimation performance in estimating RRSM person parameters for different sample sizes.

2. Methodology

2.1. Estimation and Simulation Analyses

This study aims to examine the accuracy and bias of person estimates in RRSM using MLE approach for various sample sizes. There is a number of software that can be used for the estimation and simulation purposes, including SAS, S-Plus, STATA and python software. In the present study, R programming software was used for data generations, estimation of person parameters, comparison and simulations analyses. At the first stage, the simulated survey data, according to the RRSM at 6-point Likert scale (where other scale points such as 5-point and 7-point can also be used), were generated based on the standard normal distribution. Next, the person parameters were estimated using MLE approach by using the TAM package provided in the R programming software. It should be noted that there are different packages that can also be used to fit the RRSM in R programming software including psychotools, sirem, eRm and mixRasch.

Table 1. Summary of Simulation Conditions

| n  | N   | n  | N   | n  | N   | n  | N   | n  | N   |
|----|-----|----|-----|----|-----|----|-----|----|-----|
| 5  | 500 | 10 | 500 | 15 | 500 | 20 | 500 | 25 | 500 |
| 5  | 300 | 10 | 300 | 15 | 300 | 20 | 300 | 25 | 300 |
| 5  | 100 | 10 | 100 | 15 | 100 | 20 | 100 | 25 | 100 |
| 5  | 50  | 10 | 50  | 15 | 50  | 20 | 50  | 25 | 50  |
| 5  | 30  | 10 | 30  | 15 | 30  | 20 | 30  | 25 | 30  |

To examine the effect of sample sizes on the accuracy and bias of the MLE approach in estimating the person parameters, criteria such as the number of respondents $N$ and the number of items $n$ were manipulated. For that purpose, five different sample sizes ($N = 500, 300, 100, 50, 30$) and six different number of items ($n = 5, 10, 15, 20, 25, 30$) were chosen. These two criteria were crossed to one another, to form a total of 30 separate simulation conditions. Details of simulation conditions are described in Table 1. The simulation analysis for estimating the person parameters was run up to 1000 iterations until the convergence criteria were reached.

2.2. Comparison Analysis

The accuracy and bias of the person estimates using MLE approach were compared across sample sizes $N$ and number of items $n$. In this study, the RMSE and MAE were used to assess the accuracy of the person estimates based on the formula given in Equation (3) and Equation (4), respectively:

\[ RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\beta_i - \hat{\beta}_i)^2}{N}} \]  
\[ MAE = \frac{\sum_{i=1}^{N} |\beta_i - \hat{\beta}_i|}{N} \]
In comparison, higher values of RMSE and MAE mean less accurate estimates. Meanwhile, bias in estimation of the person parameters was examined by calculating the mean difference of the estimates and their true values. The higher the value of this measure, the more biased the estimates would be. The formula of bias measure used in this study is given in Equation (5).

\[
BIAS = \frac{\sum_{i=1}^{N} (\hat{\beta}_i - \beta_i)^2}{N}
\]

Results of accuracy and bias in estimation for different sample sizes \( N \) using MLE approach in estimating the person parameters are explained in the next section.

3. Results and Discussion

Results of RMSE, MAE and bias measure of the maximum likelihood person estimates for each simulation condition is reported in Table 2. Overall, the results from this study revealed that the MLE approach produced less accurate and high bias results in smaller samples \( N \) across the number of items \( n \). As shown in Table 2, both RMSE and MAE increased with the decrease in the number of sample sizes, which implies that the MLE approach causes the accuracy of person estimates to be lower in smaller samples. In addition, the mean difference of the person estimates and their true values using the MLE approach are also much higher in smaller samples than larger ones. As can be seen from the results for the number of items \( n=5 \), the estimation’s accuracy and bias measures were slightly higher for smaller samples compared to larger ones; \((N=500): \text{RMSE}=0.4736, \text{MAE}=0.3662, \text{bias}=0.2253\); \((N=300): \text{RMSE}=0.4748, \text{MAE}=0.3667, \text{bias}=0.2269\); \((N=100): \text{RMSE}=0.4894, \text{MAE}=0.3803, \text{bias}=0.2434\); \((N=50): \text{RMSE}=0.5092, \text{MAE}=0.3962, \text{bias}=0.2669\); \((N=30): \text{RMSE}=0.5359, \text{MAE}=0.4218, \text{bias}=0.3031\). The results were found similar regardless of the number of items, where the maximum likelihood estimates of the person parameters becoming more biased and less accurate as sample sizes are reduced.

It can be seen from the findings presented in Table 2, results for \( n=30 \) also shows an increase in the values of RMSE, MAE and bias measures when the number of sample sizes were decreased from 500 to 30; \((N=500): \text{RMSE}=0.2164, \text{MAE}=0.1720, \text{bias}=0.0477\); \((N=300): \text{RMSE}=0.2226, \text{MAE}=0.1783, \text{bias}=0.0509\); \((N=100): \text{RMSE}=0.2473, \text{MAE}=0.2022, \text{bias}=0.0652\); \((N=50): \text{RMSE}=0.2749, \text{MAE}=0.2282, \text{bias}=0.0826\); \((N=30): \text{RMSE}=0.2883, \text{MAE}=0.2412, \text{bias}=0.0932\). Therefore, the results obtained from this study discovered that the maximum likelihood estimates of the person parameters were more biased and less accurate for relatively small values of \( N \) across the number of items \( n \).

| \( n \) | \( N \)  | RMSE  | MAE   | Bias  |
|-------|-------|-------|-------|-------|
| 5     | 500   | 0.4736| 0.3662| 0.2253|
|       | 300   | 0.4748| 0.3667| 0.2269|
|       | 100   | 0.4894| 0.3803| 0.2434|
|       | 50    | 0.5092| 0.3962| 0.2669|
|       | 30    | 0.5359| 0.4218| 0.3031|

\( n \) = Test length/Number of items; \( N \) = Number of respondents
RMSE = Root Mean Square Error; MAE = Mean Absolute Error
Bias = Mean difference of estimates and true values of the person measure
Table 2. Comparison of Estimation Performance for Person Parameter using Maximum Likelihood Estimation Approach (Continued)

| n  | Sample Size | RMSE | MAE | Bias |
|----|-------------|------|-----|------|
| 10 | 500         | 0.3372 | 0.2611 | 0.1141 |
|    | 300         | 0.3395 | 0.2628 | 0.1158 |
|    | 100         | 0.3534 | 0.2756 | 0.1267 |
|    | 50          | 0.3687 | 0.2890 | 0.1398 |
|    | 30          | 0.4008 | 0.3162 | 0.1690 |
| 15 | 500         | 0.2759 | 0.2148 | 0.0763 |
|    | 300         | 0.2798 | 0.2184 | 0.0787 |
|    | 100         | 0.2969 | 0.2336 | 0.0899 |
|    | 50          | 0.3223 | 0.2560 | 0.1081 |
|    | 30          | 0.3466 | 0.2775 | 0.1279 |
| 20 | 500         | 0.2443 | 0.1914 | 0.0599 |
|    | 300         | 0.2497 | 0.1964 | 0.0629 |
|    | 100         | 0.2719 | 0.2170 | 0.0763 |
|    | 50          | 0.2965 | 0.2394 | 0.0932 |
|    | 30          | 0.3261 | 0.2678 | 0.1152 |
| 25 | 500         | 0.2280 | 0.1799 | 0.0526 |
|    | 300         | 0.2328 | 0.1847 | 0.0552 |
|    | 100         | 0.2591 | 0.2091 | 0.0703 |
|    | 50          | 0.2813 | 0.2303 | 0.0858 |
|    | 30          | 0.3143 | 0.2607 | 0.1095 |
| 30 | 500         | 0.2164 | 0.1720 | 0.0477 |
|    | 300         | 0.2226 | 0.1783 | 0.0509 |
|    | 100         | 0.2473 | 0.2022 | 0.0652 |
|    | 50          | 0.2749 | 0.2282 | 0.0826 |
|    | 30          | 0.2883 | 0.2412 | 0.0932 |

n = Test length/Number of items; N = Number of respondents
RMSE = Root Mean Square Error; MAE = Mean Absolute Error
Bias = Mean difference of estimates and true values of the person measure

4. Conclusion and Recommendation
The main focus of this study was to compare the accuracy and bias in estimation of the person parameters using MLE approach for different sample sizes. The findings showed that the performance of MLE approach was poor using smaller number of samples compared to the larger ones. As presented in Table 2, the values of RMSE, MAE and bias measures increased when the number of samples were reduced for all number of items. In conclusion, the comparison analysis revealed that the MLE approach can result in more biased and less accurate estimates as the sample sizes decrease. Empirically, the results from this study strongly support the findings obtained by previous studies conducted using the dichotomous Rasch model [16], two-parameter logistic model [17], three-parameter logistic model [18] and with the estimation of item parameters in RRSM [19]. It is important to note that the selection of suitable estimation technique which produces lower bias and higher accuracy of estimates in small number of samples is essential to obtain precise statistical inferences, leading to more meaningful decision makings. Therefore, researchers should be more cautious in selecting the best estimation technique to be employed in such situation. In the future, aside from MLE approach, alternative techniques should be considered. Review of the literatures found that Bayesian estimation (BE) has
superior performance in estimating parameters in constraint with the number of samples. Results from previous studies revealed that BE is better than MLE (produces lower MSE) in estimating parameters for small sample sizes [15–18, 20–21].

5. Acknowledgment
We would like to thank the Ministry of Education (MOE) of Malaysia for the financial support for this research. This three-year project is fully funded under the Fundamental Research Grant Scheme (FRGS/1/2019/STG06/UiTM//02/8).

References
[1] Rasch G 1960 Probabilistic Models for Some Intelligent and Attainment Tests Copenhagen. Denmark: Institute of Educational Research
[2] Sumintono B 2018 Rasch Model Measurements as Tools in Assessment for Learning 1st International Conference on Education Innovation (ICEI 2017) Atlantis Press
[3] Andersen E B 1997 The rating scale model Springer New York NY
[4] Ravens-Sieberer U, Herdman M, Devine J, Otto C, Bullinger M, Rose M, and Klasen F 2014 The European KIDSCREEN approach to measure quality of life and well-being in children: development, current application, and future advances Quality of Life Research 23(3) 791–803
[5] Hong S, Kim B S, and Wolfe M M 2005 A psychometric revision of the European American Values Scale for Asian Americans using the Rasch model Measurement and Evaluation in Counseling and Development 37(4) 194–207
[6] Christensen K B, Makransky G, and Horton M 2017 Critical values for Yen’s Q 3: Identification of local dependence in the Rasch model using residual correlations Applied Psychological Measurement 41(3) 178–94
[7] Vollema M G, and Hoijtink H 2000 The multidimensionality of self-report schizotypy in a psychiatric population: An analysis using multidimensional Rasch models Schizophrenia Bulletin 26(3) 565
[8] Tondeur J, Van Braak J, Siddiq F, and Scherer R 2016 Time for a new approach to prepare future teachers for educational technology use: Its meaning and measurement Computers and Education 94 134–50.
[9] Andrich D 1978 A rating formulation for ordered response categories Psychometrika 43(4) 561–73
[10] Andersen E B 1995 Polytomous Rasch Models and their Estimation In Rasch Models 271–91 Springer New York NY
[11] Wright B D and Masters G N 1982 Rating Scale Analysis MESA press
[12] Engelhard Jr G 2013 Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences Routledge
[13] Linacre J M 1999 Estimation Methods for Rasch Measures Journal of Outcome Measurement 3 381–405
[14] Linacre J M 1994 Sample Size and Item Calibration or Person Measurement Stability Rasch Measurement Transactions 7(4) 328
[15] Finch H and French B F 2019 A Comparison of Estimation Techniques for IRT Models With Small Samples Applied Measurement in Education 32(2) 77–96
[16] Swaminathan H, and Gifford J A 1982 Bayesian Estimation in the Rasch Model Journal of Educational Statistics 7(3) 175–91
[17] Swaminathan H, and Gifford J A 1985 Bayesain Estimation in the Two-Parameter Logistic Model Psychometrika 50(3) 349–364
[18] Swaminathan H and Gifford J A 1986 Bayesian Estimation in the Three-Parameter Logistic Model Psychometrika 51(4)
[19] Azizan N H, Mahmud Z, and Ramblia A 2020 Rasch Rating Scale Item Estimates using Maximum Likelihood Approach: Effects of Sample Size on the Accuracy and Bias of the Estimates
International Journal of Advanced Science and Technology 29(4) 2526–31

[20] Burgos J G 2010 Bayesian methods in psychological research: The case of IRT International Journal of Psychological Research 3(1) 164–76

[21] Swaminathan H, Hambleton R K, Sireci S G, Xing D, and Rizavi S M 2003 Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates Applied Psychological Measurement 27(1) 27–51