Estimation of college students' ability on real analysis course using Rasch model

[1]Isnaini; [*2]Wikan Budi Utami; [3]Purwo Susongko; [4]Herani Tri Lestiani

[1,2]Mathematics Education Department, Universitas Pancasakti Tegal

Jl. Halmahera Km. 1, Mintaragen, Kec. Tegal Tim., Kota Tegal, Jawa Tengah 52121, Indonesia

[3]Department of Natural Science Education, Universitas Pancasakti Tegal

Jl. Halmahera Km. 1, Mintaragen, Kec. Tegal Tim., Kota Tegal, Jawa Tengah 52121, Indonesia

[4]Mathematics Education Department, Institut Agama Islam Negeri Syekh Nurjati Cirebon

Jl. Perjuangan By Pass Sunyaragi, Kota Cirebon, Jawa Barat 45132, Indonesia

[*]Corresponding Author. E-mail: wikan.piti@gmail.com

Abstract

This study is aimed at estimating the difficulty level of essay tests and the accuracy of students' ability in Real Analysis essay test using the Rasch model with the QUEST program and R 3.0.3 package eRm program. The population in this study was all students of the Department of Mathematics Education, Universitas Pancasakti Tegal in the academic year 2016/2017, who were enrolled in the Real Analysis course. The data were analyzed using the R 3.0.3 package eRm program and QUEST program. The students' ability was obtained from the result of the course final exam of the first Real Analysis course. The analysis shows that: (1) by using Rasch model for partial credit scoring, the difficulty level shows that 100% of essay questions in Real Analysis final exam is categorized as difficult, (2) the estimation of students' ability in Real Analysis course using Rasch Model with CML method is better than the estimation of students' ability using Rasch Model with JML approach.

Keywords: *estimation of ability, level of difficulty, Rasch Model, Item Response Theory*

Introduction

One important component in the formation of quality human resources is education. The most important factor to be able to compete globally in the 21st century is education. According to Mardapi (2012, p. 12), efforts to improve the quality of education can be pursued through improving the quality of learning and the quality of the assessment system. Thus, in the process of education in Higher Education, for example in learning mathematics must strive to implement the learning process and assessment as well as possible. A good process of learning mathematics can certainly be done by providing flexibility for students to develop and explore their abilities.

Today, education in Indonesia is still considered very low, especially for mathematics. Even though mathematics is the main science taught from elementary school to university. This indication can be seen from the low student achievement in each academic year. Ironically, mathematics is a subject that is not liked. Many students are afraid of mathematics. For them, math is like a frightening enemy they want to avoid. Schwartz (2005, p. 1) suggests the basic success of mathematics education is to support the development of intelligence in mathematics from a variety of life conditions. Student's mathematical skills in living conditions at the School can be seen when students take the test. The implementation of the test is basically to assess the success of students during the learning process.

The test is very necessary so that the educator in this case the lecturer can know the student's learning achievement after being given the subject matter in the learning process. Therefore, making a good test needs to be pursued by considering the ability of students, so that the tests carried out as a measuring tool to test student achievement can reflect/describe the true abilities of students.

Students of the Mathematics Education program at Universitas Pancasakti Tegal all this time consider the most difficult subjects to be Real Analysis. Real Analysis comprises deductive and axiomatic topics. Previous observation on the performance of students of Universitas Pancasakti revealed the students' ability in this course is relatively low. It is indicated by their ability to prove a convergent sequence yet, they found it difficult in solving some problems related to convergent sequence as there are many theorems are included.

Student learning evaluation activities are one of the important tasks that must be done by lecturers. In the field of education, evaluation of student learning achievements is conducted to determine the progress of students in the curriculum that has been taught. One effort to evaluate students is to give examinations in the middle of the semester and at the end of the semester. However, sometimes giving questions that are too difficult or too easy causes it to be difficult for lecturers to distinguish students' abilities. Therefore, an analysis of exam questions is needed in the hope that the exam results present the ability of students.

Evaluation is a series of activities in improving the quality, performance, or productivity of an institution in carrying out its program. Through evaluation, information about what has been achieved and which have not will be obtained, then this information is used to improve a program. According to Tyler (1950), evaluation is a process of determining the extent to which educational goals have been achieved. According to Griffin and Nix (1991), evaluation is a judgment on the value of the measurement results or implications of the measurement results. Tyler emphasizes the achievement of the objectives of a pro-

gram, while Griffin and Nix emphasize the use of assessment results. Thus, the focus of evaluation is a program or group, and there is a judgment element in determining the success of a program (Mardapi, 2012, p. 4).

The form of real analysis subject evaluation is the midterm and the final semester examination. The test is in the form of a description test, the advantages of the description form test are easy in the preparation. This form of description will also train students in expressing opinions both systematically and logically (Buckley, Winkel, & Leary, 2004). A lecturer will be able to find out where the weaknesses of the students are in the material that has been taught so that they will give input on what things must be improved. Scoring on the description form tests takes a long time and is relatively more difficult so the form of the description test is difficult to use for large-scale tests. An assessment will be meaningful if the results can be used to improve the quality of the learning process. An assessment will be meaningful if the results can be used to improve the quality of the learning process (McMillan, 2005).

The existence of the midterm and final semester exams in the Real Analysis course is to evaluate the ability of students. Some theories and models that can be used to analyze test items are the ones with the Rasch Model.

In this study, Rasch model was employed to analyze test items. According to Imaroh, Susongko, and Isnani (2017), the items parameter does not depend on the sample. Further, Ningsih and Isnani (2010) revealed the different reliability levels of essay test items analyzed using Item Response Theory model (1PL, 2PL, 3PL) and Rasch model.

The concept of objective measurement in the social sciences and the assessment of education, according to Wright and Mok (2004), must have five criteria, namely: (1) producing linear measurements with equal intervals, (2) exact estimation process, (3) identifying inaccurate (misfits) or uncommon items (outliers), (4) able to handle missing data, (5) produce measurements that are independent of the parameters studied. Of the five conditions, so far only the Rasch model can fulfill these five conditions. The quality of

measurements in the assessment of education carried out with the Rasch model will have the same quality as the measurements made in the physical dimension in the field of physics (Sumintono & Widhiarso, 2015). In measuring modern test theory, the Rasch model is seen as the most objective measurement model. The use of the Rasch model in measuring education has advantages in specific objectivity and the stability of high grain parameter estimates (Wu & Adams, 2007).

The main characteristic of the Rasch Model is that this model considers all responses of a test taker regardless of the sequence in solving the problems. It means that the level of difficulty of each test item is not necessarily in consecutive order. The main advantage of the Rasch model is that the mental process used by participants in solving the problems is more accurate. Moreover, compared to other models (particularly classical test theory) this model has the ability to predict the missing data based on a systematic response pattern. This model has been applied to mathematics and reading tests, e.g., at the National Assessment of Educational Progress (NAEP) (Susongko, 2014). This model is also suitable for analyzing personality scale responses that have a multi-point scale.

Unlike the Rasch model which includes all responses without considering the sequence in solving the problems, the Gradation model requires sequential responses of the test takers from a low to a high category. In the Gradation model, the level of difficulty of each test item is arranged in sequence, while in classical test theory, the pattern of students' answers is not considered as classical test theory merely considers correct and incorrect answers. Gradation model is suitable for a course that requires regularities or sequential responses of each test item, such as mathematics, physics, and chemistry.

According to Lababa (2008), one of the oldest test theories about behavioral assessment is classical true-score theory. Classical test theory has an easy application. Moreover, it is a practical model to describe how measurement errors can affect the observed score.

Quantitative item analysis emphasizes the analysis of internal test characteristics through empirically obtained data. Internal characteristics include test item parameters which are the level of difficulty and discrimination power of a test.

Rasch model is a dichotomous scoring model that merely has two categories, namely the correct answer with a score of 1 and the incorrect answer with a score of 0. Currently, it has been developed more extensively in polytomous scoring. According to Retnawati (2014, p. 32), the polytomous scoring model is an item response model that has more than two scoring categories. In the Rasch model, it is assumed that all items have the same discrimination index (Isgiyanto, 2011).

To deal with polytomous data with various ranks, a new type of analysis of the Rasch model is developed, namely the Partial Credit Model. However, the main purpose of the Rasch model is to create a scale measurement at equal intervals. Meanwhile, as the raw scores are not shown in interval form, the scores cannot be used directly to interpret the students' ability. Rasch model requires both per person score data and per item score data. These two scores become the basis for estimating true scores that indicate the level of individual ability as well as the degree of difficulty of the test.

Rasch modeling uses both per person score data and per item score data. These two scores become the basis for estimating true scores that indicate the level of individual ability as well as the degree of difficulty of the test. The advantage of the Rasch Model compares to other models, particularly classical test theory, is the ability to predict the missing data, based on a systematic response pattern.

Some studies had been carried out related to the use of the Rasch Model in analyzing test items. A study by Kurniawan and Mardapi (2015) showed that the Rasch model provides complete information about test items, including its difficulty level. This study is aimed at estimating the difficulty level of the essay test on the first Real Analysis course by using the Rasch Model and describing the estimation of students' ability in Real Analysis course by using the Rasch Model, QUEST program, and R 3.0.3 package eRM program.

Method

This research is an explorative descriptive study of data sets of items and responses of participants in the semester's final examination of the real analysis subject in the academic year 2016/2017. This research is a post-hoc diagnosis that is described as a retrofitting approach (Gierl, 2007). The retrofitting approach is carried out through analysis of the items and item response data in the final semester exam in the real Analysis 2016/2017 academic year.

Some studies have implemented the Rasch model by involving 30 to 300 students as the sample (Bond & Fox, 2007; Keeves & Masters, 1999). The subject of this present study was 82 students of Mathematics Education Department of Universitas Pancasakti Tegal in the academic year 2016/2017 who took the first Real Analysis course.

The sampling technique used in this study is purposive sampling. It is one of the non-random sampling techniques where the researcher determines sampling by specifying specific characteristics suitable with the objectives of the study so that it is expected to answer the research problems. Based on the explanation of the purposive sampling, there are two things that are very important in using the sampling technique, namely non-random sampling and setting specific characteristics according to the research objectives by the researchers themselves.

The instrument used in this study was the final exam test on the first Real Analysis course. The test items include the introduction material, Real Numbers, Sequences and Series, and Limit (Bartle & Sherbert, 2000).

Rasch model was applied to analyze the collected data. This analysis resulted in a description of the difficulty level of the test items. By using the eRm package in R Program version 3.0.3, the analysis generated the estimation of item parameters on the exam of Real Analysis.

Measurement modeling explains the procedure of how to organize raw scores into more meaningful information. Moreover, it can utilize a mathematical model that can interpret raw scores into a score that provides more valid and accurate information. The analysis of raw scores leads to a new finding: the opportunity for students to correctly answer an item is the same as the comparison of students' ability and the difficulty level of the test items.

$$P_{i1}(\theta) = \frac{P_{i1}(\theta)}{P_{i0}(\theta) + P_{i1}(\theta)}$$
$$= \frac{exp(\theta_n - \delta_{i1})}{1 + exp(\theta_n - \delta_{i1})} \quad \cdots \quad (1)$$

(Bryan, 2004)

OCFs (Ogive Curve Function) become a prototype of Rasch model development for polytomous items. If i is a polytomous item with score category = 0, 1, 2,. . . , mi, then the probability of participant n with score x on item i is later described in Category Response Function (CRF), which is illustrated in the following equation (Glas & Verhelst, 1989):

$$P_{ix}(\theta) = \frac{exp\left[\sum_{j=0}^{x}(\theta_n - \delta_{ij})\right]}{\sum_{r=0}^{m_i}\left[exp\left[\sum_{j=0}^{x}(\theta_n - \delta_{ij})\right]\right]} \quad \dots (2)$$

Equation (2) can be elaborated by the number of categories in the test items. For example, if a scale has three categories of the score of 0, 1, and 2, then there will be a category (j) as many as three individual probability equations for each category. Probability in category 0 is:

$$P_{i0}(\theta) = \frac{1}{1 + exp(\theta_n - \delta_{i1}) + exp[(\theta_n - \delta_{i1}) + (\theta_n - \delta_{i2})]} \quad \dots (3)$$

Probability in category 1 is:

$$P_{i1}(\theta) = \frac{exp(\theta_n - \delta_{i1})}{1 + exp(\theta_n - \delta_{i1}) + exp[(\theta_n - \delta_{i1}) + (\theta_n - \delta_{i2})]} \quad \dots (4)$$

Probability in category 2 is:

$$P_{i2}(\theta) = \frac{exp(\theta_n - \delta_{i1}) + exp(\theta_n - ä_{i2})}{1 + exp(è_n - ä_{i1}) + exp[(è_n - ä_{i1}) + (è_n - ä_{i2})]} \quad \dots (5)$$

In the probability of category 0, there is a number 1 in the numerator since Rasch Model requires the following equation:

$$\sum_{j=0}^{0}(è - ä_{ij}) = 1 \quad \cdots \quad (6)$$

(Glas & Verhelst, 1989)

Findings and Discussion

The parameter of the difficulty level of test items has the same value interval as the parameter of participants' ability (θ), which is $b_{ij} = \theta$. The b_{ij} value ranges from $-\infty$ to $+\infty$. However, the values which are practically (or rationally) used are only between -4.0 to +4.0. It means that the more negative the difficulty level of an item or close to -4, the easier the problem. On the other hand, the more positive the difficulty level or approaching +4, the more difficult the problem (Naga, 2003, p. 224).

In case the parameter of the difficulty level of a test item meets $b_j \leq -2$, the item is then categorized as a very easy item. If it meets $-2 \leq b_j \leq 0$, the item is then categorized as an easy item. Furthermore, if it meets $0 < b_j \leq 2$ and $b_j \geq 2$, the item is then categorized as a difficult and very difficult item, consecutively (Hambleton, Swaminathan, & Rogers, 1991).

The analysis of the question number 1 showed that $\delta_{11} = 0.861$, $\delta_{12} = 0.374$, and $\delta_{13} = 0.45$. It implies that the difficulty level of the first, second, and third steps is included in the difficult category. In question number 2, the difficulty level of the first step is included in the difficult category ($\delta_{21}=1.731$), while the difficulty level of the second step is identified as very difficult ($\delta_{22}=2.787$). In question number 3, the results obtained were $\delta_{31}=1.149$ and $\delta_{32}= 1.796$, which suggest that the difficulty level of the first and second steps can be included in the difficult category. The analysis of question number 4 resulted $\delta_{41}=-0.363$ and $\delta_{42}=-0.963$. It indicates that the difficulty level in both steps is in included in the easy category.

The results showed that there are three categories (δ_{12}, δ_{21}, δ_{41}) which are identified as easy, one category (δ_{11}) is identified very easy, and six categories (δ_{22}, δ_{31}, δ_{32}, δ_{42}, b_{51}, and $b\delta_{52}$) are categorized as difficult. In general, the score of difficulty level of those items was 0.594, thus the four test items were identified as difficult.

It can be inferred from the aforementioned results that the final exam items of Real Analysis course are categorized as difficult for the participants, even though all topics in the questions had been discussed during the course. The value of the difficulty level of item varies (typically) from about -2.0 to +2.0. Item number 1 with sub-topic of the Completeness of Real Numbers was identified as a difficult item. Likewise, item number 2 and item number 3 with sub-topic of the Limit of a Sequence and the Theorems of Limit of a Sequence, respectively, were categorized as difficult items. On the contrary, item number 4 with sub-topic of the Theorems of Limit of a Sequence was identified as an easy item. To make it clearer, Figure 1, Figure 2, and Figure 3 present the questions in the test and the sample of student's answers.

From the students' answers which are presented in Figure 1, Figure 2, and Figure 3, it can be foreseen that the student was incapable to solve the problems number 1, 2, and 3 systematically, because of the incapacity in understanding some theorems and definetions which are related to the problems. The students could not recognize and analyze the relation between the theorems and definitions.
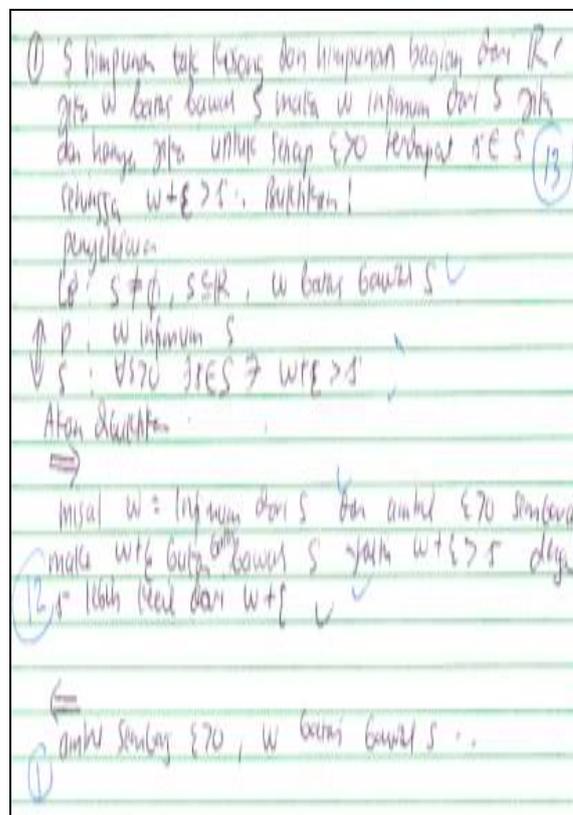

Figure 1. Student's answer on Problem 1

Figure 2. Student's answer on Problem 2

Figure 3. Student's answer on Problem 3

It is presented in Figure 4 that in the fourth problem, the student seemed to comprehend the topic. The theorems related to sequences and series were analyzed before the implementation for solving a problem. It can be seen from the sample in which the student could use the theorems systematically as suggested in solving the problem.

Figure 4. Student's answer on Problem 4

The result of the analysis showed that the ability of the test participants was quite diverse. In fact, merely a small number of students can solve questions number 1, 2, and 3 correctly. Most of the students could not determine specific theorems and definitions to solve the problems, especially in the second and third problems. In contrast, most of the students already understand the theorems used to solve the fourth problem, which are the sequences and series theorems, even though they faced a difficulty to analyze the theorems.

The estimation of the students' ability is presented in the interval scale (-3, +3). The category score in Rasch Model shows the number of the required steps to solve an item correctly. A high score indicates a good ability category. On the contrary, a low score indicates a low category of ability as well. The output of the estimation of ability parameter obtained from QUEST program and the package eRM with partial credit modeling or

Rasch Model is used to illustrate the comparison between the students' ability estimated using the Joint Maximum Likelihood (JML) approach with the package eRM and those estimated using the Conditional Maximum Likelihood (CML) approach with the QUEST program.

In JML approach, the students' ability could not be expressed in score 0 and score 100. Meanwhile, in CML approach, the students' ability can be expressed in score 0 (approximately a value of -3.09) and score 100 (as approximately a value of 85). Therefore, it can be inferred that Rasch Model using CML approach is more suitable than Rasch Model using JML approach to estimate the students' ability in understanding the subject-matter.

The result of analysis meets the OutfitMSQ criteria if the value is 0.035 < OutfitMSQ < 3.239. The analysis resulted a value of 0.5 < OutfitMSQ < 1.5, thus it fulfills the range of OutfitMSQ. The criteria of INFIT MNSQ is 0.5 < MNSQ <1.5. According to the mean value and the standard deviation of Rasch model, the CML approach with the package eRM is eligible since the mean and the standard deviation meets the criteria. On the contrary, the JML approach with Quest program is less appropriate as indicated by the mean and the standard deviation that do not meet the criteria.

In conclusion, the result of analysis on the estimation of students' ability reveals that the estimation of students' ability using Rasch model with CML approach and eRm program is more accurate than the estimation of students' ability using Rasch model with JML approach and QUEST program. Similarly, based on OutfitMSQ, Rasch model using CML approach with eRm program has better performance than Rasch model using JML approach with Quest program.

Conclusion

Based on the results and discussions, it can be concluded that the essay test items on the first Real Analysis course that have been tested to the students of Mathematics Education Department, Universitas Pancasakti Tegal can be classified as a good test. Besides, the students' ability can be estimated precisely

by using Rasch Model with CML approach and eRm package. The estimation of participants' ability was quite diverse. A small number of students can solve questions number 1, 2, and 3 correctly despite these questions were classified difficult. Meanwhile, most of students already understand the theorems used to solve the fourth problem. The students are capable to apply the theorems systematically to solve the fourth problem.

References

Bartle, R. G., & Sherbert, D. R. (2000). *Introduction to real analysis*. New York, NY: John Wiley & Sons.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Buckley, K. E., Winkel, R. E., & Leary, M. R. (2004). Reactions to acceptance and rejection: Effects of level and sequence of relational evaluation. *Journal of Experimental Social Psychology*, *40*(1), 14–28. https://doi.org/10.1016/S0022-1031 (03)00064-7

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the Rule-Space model and Attribute Hierarchy method. *Journal of Educational Measurement*, *44*(4), 325–340. https://doi.org/10.1111/j.1745-3984. 2007.00042.x

Glas, C. A. W., & Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika*, *54*(4), 635–659. https://doi.org/10.1007/BF02296401

Griffin, P., & Nix, P. (1991). *Educational assessment and reporting: A new approach*. Sydney: Harcourt Jovanovich.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Imaroh, N., Susongko, P., & Isnani, I. (2017). Uji validitas tes ulangan akhir semester gasal mata pelajaran matematika (Studi deskriptif analisis dokumenter di SMP

Negeri Slawi tahun pelajaran 2016/2017). *JPMP (Jurnal Pendidikan MIPA Pancasakti)*, *1*(1), 80–89. https://doi.org/10.24905/jpmp.v1i1.792

Isgiyanto, A. (2011). Analisis data ujian nasional matematika berdasarkan penskoran model Rasch dan model Partial Credit. *Prosiding Seminar Nasional Penelitian, Pendidikan Dan Penerapan MIPA*, 43–52. Retrieved from https://eprints.uny.ac.id/7172/1/PM-7 - Awal Isgiyanto.pdf

Keeves, J. P., & Masters, G. N. (1999). Introduction. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment*. Amsterdam: Pergamon-Elsevier Science.

Kurniawan, D. D., & Mardapi, D. (2015). Penyetaraan vertikal tes matematika SMP dengan teori respons butir model Rasch. *Jurnal Evaluasi Pendidikan*, *3*(1), 12–25. Retrieved from http://journal.student.uny.ac.id/ojs/index.php/jep/article/view/1221/1093

Lababa, D. (2008). Analisis butir soal dengan teori tes klasik: Sebuah pengantar. *Iqra'*, *5*, 29–37. Retrieved from https://jurnaliqro.files.wordpress.com/2008/08/03-jun-29-36.pdf

Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Medika.

McMillan, J. H. (2005). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice*, *22*(4), 34–43. https://doi.org/10.1111/j.1745-3992.2003.tb00142.x

Naga, D. S. (2003). *Teori pengukuran*. Retrieved from http://dali.staff.gunadarma.ac.id/Downloads/folder/0.1

Ningsih, L. D., & Isnani, I. (2010). Studi komparatif tingkat reliabilitas tes prestasi hasil belajar matematika pada tes bentuk uraian dengan model penskoran GPCM (Generalized Partial Credit Model) dan Penskoran GRM (Graded Response Model). *Cakrawala: Jurnal Pendidikan*, *4*(8). https://doi.org/10.24905/cakrawala.v4i8.176

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.

Schwartz, S. L. (2005). *Teaching young children mathematics*. London: Praeger.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan*. Cimahi: Trim Komunikata.

Susongko, P. (2014). *Pengantar metodologi penelitian pendidikan*. Tegal: Universitas Pancasakti Tegal.

Tyler, R. (1950). *Basic principles of curriculum and instruction*. Chicago, IL: University of Chicago Press.

Wright, B., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 1–24). Maple Grove, MN: JAM Press.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement.