Scholars Journal of Arts, Humanities and Social Sciences OPEN ACCESS
Abbreviated Key Title: Sch J Arts Humanit Soc Sci
ISSN 2347-9493 (Print) | ISSN 2347-5374 (Online)
Journal homepage: https://saspublishers.com/journal/sjahss/home

An Application of Item Response Model for the Development of Exam Banking: In Haramaya University, East Harerghe Zone, Oromia Region, Ethiopia

Mr. Usmael Abedella Hassen*, Mr. Abate Assefa Gubesa, Mr. Dejen Aberham Gugessa

Lecturer, Department of Psychology, College of Education and Behavioral Sciences, Haramaya University, Ethiopia

DOI: 10.36347/sjahss.2019.v07i09.006 | **Received:** 12.08.2019 | **Accepted:** 19.08.2019 | **Published:** 20.09.2019

***Corresponding author:** Mr. Usmael Abedella Hassen

Abstract **Original Research Article**

Recently developed testing theories and current technology make possible to develop a standard assessment and evaluation system for monitoring educational quality. In fact, this is achieved if the assessment system is supported by calibrated test item bank. However, its use is very limited in Haramaya University and assessment system has not been supported by test item bank due to the problems attributed to limited understanding of psychometric concepts for technical application of testing technology. This study assessed the practical application of 1-parameter *item response model. The applications of this model determine the extent to which our instrument fits the standard Rasch model is measuring standards abilities. It is also used to calculate item difficulty and student's ability parameters of common standards scale of measurements.* This *Study adopted introduction to psychology final examination consisting 40 multiple-choice test items. Responses for the tests were taken from 150 first year Haramaya university students' of 2018/19 fresh entry students. The score result data were entered into Winstepe control data setup, as per the response category codes (zero for wrong and one for right) and then converted to a text file. The data pattern had 50 columns: columns 1-10 were for the ID; columns 11-50 were answers for 40 test items. The data were analyzed using the (Winstepe Rasch-Model John M. Linacre 2014) Computer Programs and all the item difficulties were linked on the logit scale along with the student ability. Thirty –four test items fitted the measurement model with probability p>0.03.].* The item-trait interaction was not statistically significant at the 0.01 level, [Chi-square (df =1036 =1129.49, p =0.022]. indicating that dominant trait was measured and the test fits the standard item repose model, the item difficulties range from – 0.53. logits (SE= 0.30) to + 3.2 logits (SE=0.34) and the student ability measures range from – 1.11 logits to +4.0 logits. *It is recommended that teachers in higher education would be encouraged to learn more about Rasch measurement modeling, item banking and involved in the development of item bank for their own subject*.

Keywords: Item Response Model, Exam Bank, *and* Development of Exam Banking.

INTRODUCTION

In order to bring quality education in the given nations both teachers and test developers in higher education institutions expect availability and quick access to good quality test items. A large collection of good test items will help teachers to concentrate more on their teaching without having to spend much time on test construction. It could also ensure that only quality test items are used. When such a collection (popularly referred to as an item bank), consists of items measuring the same educational domain calibrated on to the common scale, it could help test developer in solving many of the practical testing problems [1]. The idea of item banking is associated with the need for making test construction easier, faster and more efficient. Concepts of item banking have also been connected with the movement toward both mastery learning and criterion referenced evaluation. Vander Linden [2], cited in viewed .Item banking as a new practice in test development, as a product of the introduction of education measurements a well-known Model of Item Response Theory (IRT) which covers different range of statistical tests used to analyze test scores across students that have different competency levels.

Therefore, if a massive collection of good items is available, it can shuffle off the burdens of the test developer. The quality of test used in schools, expected to be better than it could without an item bank. Test security is not the same problem as it is in a traditional testing situation. This can easily be

processed with specially developed computer program called *Rasch* Winstep which shows test item and patterns of student's response that fits the measurement model. In selecting item for future testing, item difficulty is considered and appropriate for the ability or competence level of students' .The model makes it possible to predict the likelihood of a correct answer to a given test item based on the knowledge of two variables: item difficulty versus person ability [3]. Using Rasch item response model, test users can create a common linier scale (logic scale usually scaled from around -4 to +4) up on which item difficulty parameter of each item in the test versus students' ability measure can be clearly located .The value on the scale interpreted as the item difficulty. Each item difficulty add together to give a precise measure of the test. Test takers who response to items can also place on to the same scale, and their scores interpreted as student's ability. As such, there is a direct connection between students' abilities level and difficulty level of the test involved (Fulcher and Davidson) cited in [4]. A calibrated item bank with Rasch measurement makes testing process more flexible and appropriate, because different groups of students can take different tests that are suitable to each of them and the results can still be compared on the same scale. Thus, Recently-developed testing theories and so calculated computer software application make possible fast and easy generation of the most appropriate items and suggested that, tests have to be calibrated, which means that some parameters of the evaluation items must be estimated to fit the requirements of the IRT model being used. Several studies have been carried out on the practical application of item response model contribute initiate test item bank. The requirements for an item bank are: 1) an adequate pool of test items, 2) an inventory of the abilities and content which each item intended to measure, 3) statistical data indicating the characteristics of each item as evidenced in test trialing (and 4) a theory or construct of ability which enables to interpret the meaning of scores on any test constructed from the banked items. Item bank with these requirements were proved to be well adapted for meeting the needs of criterion referenced evaluation and used us the standard base for monitoring and reviewing educational quality. However, the use of test item bank remains theoretical rather than practical in the assessment system of many Ethiopian Higher institutions including Haramaya University.

Research Questions: This study tried to answer the following research questions:

- The extent to which psychology test items through practical application of 1- parameter Rasch Measurement Model instrument fits the standard Rasch model in measuring student's abilities?
- How to calibrate the item difficulty and student's ability parameters on common standard scale of measurement?

Theoretical Frameworks for Application of Rasch Measurement Model

The development of IRT has opened new ways for test applications and research with tests Item response theory is a general term for a family of models, the item response or IRT models that share some fundamental ideas. These ideas are that IRT models persons' responses on individual items. The response of a person on a test item is conceived of as a function of person characteristics and item characteristics [5].

The latter function is commonly called item trace line, item characteristic function (ICF), or item characteristic curve (ICC). It specifies how the probability of a correct response to an item increases as the level of the trait increases. IRT consists of a class of mathematical models for which estimation procedures exist for model parameters (i.e., person and item parameters) and other statistical procedures for investigating to what extent the model at hand fits the data or persons' responses to a set of items.

IRT research and developments not only pervade scholarly journals, in the latest edition of the *Standards for Educational and Psychological Testing* [6], ample space is given to IRT in a one-dimensional model, the probability of correct response increases with increasing ability [5] The 1PL model is obtained if all item discrimination parameters are set equal to 1.0 Meredith and Kearns [7]. The investigation of model fit has boils down to statistical test of model fit and analysis of residuals in order to get the seriousness of the model violations for this reason, both statistical testing and the analysis of residuals are necessary in the study of model fit. Other chi-square statistics on the item and the test levels have also been proposed for the Rash model, based on the CML approach [8-10]. The residuals used for the computation of item fit statistics might also be plotted**.**

General Objective

The general objective of the study was to calibrate psychology test items through practical application of 1-parameter Rasch measurement model.

Specific Objectives

- Determine item parameter through conducting item analysis.
- The extent to which application of 1-parameter Rasch measurement model instrument fits the standard Rasch model in measuring student's abilities
- Calibrate item difficulty and students 'abilities parameters on common standard scale of measurement.

Description of the Study Area

The study area, East Hararghe, is located in the Eastern part of Oromia National Regional State,

Ethiopia. Its altitude ranges from 500 to 3,400 meters above sea level. It contains, three agro-ecological zones, highlands (elevations above 2,300m), midlands (elevations between 1,500 and 2,300m), and lowlands (elevations below 1,500m). The low lands occupy the largest area (62.2%), followed by midlands (26.4%) and highlands (11.4%) [11]. East Hararghe has 18 districts with a total population of 2,723,850, of which 1,383,198 are males and1, 340,652 of them are females. With an area of 17,935.40 km^2, East Hararghe has a population density of 151.87 per km^2. The majorities (90%) of the populations depend on agriculture in the rural area; 8.27% of them are urban inhabitants, and a further 1.11% is pastoralists [12]. Thus, this study has been conducted in one single and oldest public university which is called Haramaya University

METHODOLOGY

This study has taken practical applications of 1-parameter IRM model designed to estimate item characteristics and persons' abilities parameters. The data for the present study were obtained from the final examination of the course entitled 'introduction to psychology' which was conducted for the year I psychology students at Haramaya University during 2018/19 academic year. The sample data was taken from multiple choice test which assisted 40 items and was conducted to 150 students. Responses for the tests taken from 150 first year students of 2018/19 were entered into Winstepe control data set up, as per the response category codes (zero for wrong and one for right) and then converted to a text file. The data pattern had 50 columns: columns 1-10 were for the code test takers; columns 11-50 were answers for 40-test items. The data is analyzed using the Rasch Uni dimensional Measurement Model [13] Computer Programs. In Rasch analysis, the items are designed in a conceptual order by difficulty and this order was tested .The data for the items also fit the measurement model in order to create a linear scale and this was tested. To linearize these proportions, they are converted to log odds, or logits (usually from –4 to 4), by taking the natural log of the proportion incorrect for items or failures for persons. This transforms the proportions to a linear scale [14]. Logit scores (person ability and item difficulty) were calibrated on the same scale of standard units.

RESULTS AND DISCUSSION

Item analysis

The present study involved an analysis with the winstepe program tested 40 multiple choice introduction psychology questions items (N=150).The data were analysed using the Rasch Uni dimensional Measurement Model [13] Computer Programs. a conceptual difficulty order of items is tested. The data for the items have fit the measurement model in order to create a linear scale and this is tested. The person measures and item difficulties were calibrated on the standard units called logits (log odds of answering positively) (see Figure-1). In Figure-1, the student achievement measures are from low to high on the left hand side and the item difficulties are from easy to hard on the right hand side of the same linear scale.

The residuals were also examined; the residuals being the difference between the expected item score calculated according to the Rasch measurement model and the actual item score of the students. This was converted to a standardized residual score in the computer program. The global item fit residuals and global student fit residuals have a mean near zero and a standard deviation near one, when the data fit the measurement model. In this case, the global item and person fit residuals indicate a satisfactory, but not excellent, fit to the measurement model (see Table-1). The individual probability of fit of items to the measurement model was checked of the 40 items, 34 fitted the measurement model with probability $p>0.03$. Six Item are beyond the range, these items should, accordingly be taken out from the item list. However, in the present research which is to demonstrate the item banking procedure, considering the number of the items is small and the percentage does not seem fatal to the analysis (6 out of 40 items or 15 % of the whole test), they were left as they are in this list.

Item Trait Test-of-fit

The item-trait test of fit examines the consistency of the item difficulties across the student measures along the scale. This determines whether there was agreement among the students as to the difficulties of all items along the scale. The item-trait interaction was not statistically significant at the 0.01 level [Chi-square (df =1036 =1129.49, p =0.022].This means that a dominant trait was measured and that overall fit to the measurement is acceptable, but not excellent. Tests of acceptance" are concerned with whether what is observed meets empirical requirements. It is typical in Ranch analysis that the probability of the chi-square is 0.0, because empirical data rarely fit a theoretical ideal.

Table-1: Summary of fit statistics for Psychology achievement test (40 items, N=150)

	ITEM	STUDENTS
Number	**40**	**150**
Location Mean	**0.45**	**1.00**
Standard deviation	**.83**	**1.11**
Fit statistic Mean	**0.99**	**0.98**
Fit statistic Standard deviation	**.17**	**.19**
Standard error of Measure S.E	**.22**	**.14**
Raw score-to-measure correlation	**-1.00**	**.99**
item raw score-to-measure correlation = -1.00 **item-trait interaction chi square =1129.49, d.f= 1036 , prob.=.0222** **global root-mean-square residual (excluding extreme scores): .4485** **capped binomial deviance = .2546 for observations** **person separation index =.89** **power of test-of fit: good (based on the separation index**		

Notes on Table 1

- The item means are constrained to zero by the measurement model.
- Numbers are given to two decimal places

Targeting

In winstepe IRM software application, the Item difficulty parameter and students ability is expressed in the logit scale and their relationships are presented in the Item-Person Map (IPM), in which both types of information can be evaluated simultaneously. The distribution of students and items along the latent variable is laid vertically with the most able students and most difficult item at the top. On the left hand side are examinee proficiency values whereas those on the right hand side are item parameter values. This IPM can tell us the "big picture" of both items and students.

In this analysis the item difficulties range from – 0.53. Logits (SE= 0.30) to + 3.2 logits (SE=0.34) and the student measures range from – 1.11 logits to +4.0 logits. There are some students (19%) whose s abilities are more than +3.2 logits and less than – 0.53 logits and hence not 'matched' against an item location on the scale. In Figure-1, there are no items matching persons at either the lowest end (-0.64 to –1.5 logits) or the highest end (+3.21 to +4.0 logits) of the scale, indicating the improvements that are needed for the test. That is, both easy items and hard items need to be added to improve the targeting of the items for these students. There are very few students who found these test items easy and approximately, 25% who found who found them hard. The item difficulties were appropriate for the rest of the students, approximately 72% 0f students. The examinees on the upper left are said to be better or smarter than the items on the lower right, which mean that those easier items are not difficult enough to challenge those highly proficient students. On the other hand, the items on the upper right outsmart examinees on the lower left, which implies that these items are beyond their ability level. The finding of this study shows that the examinees overall are better than the exam.

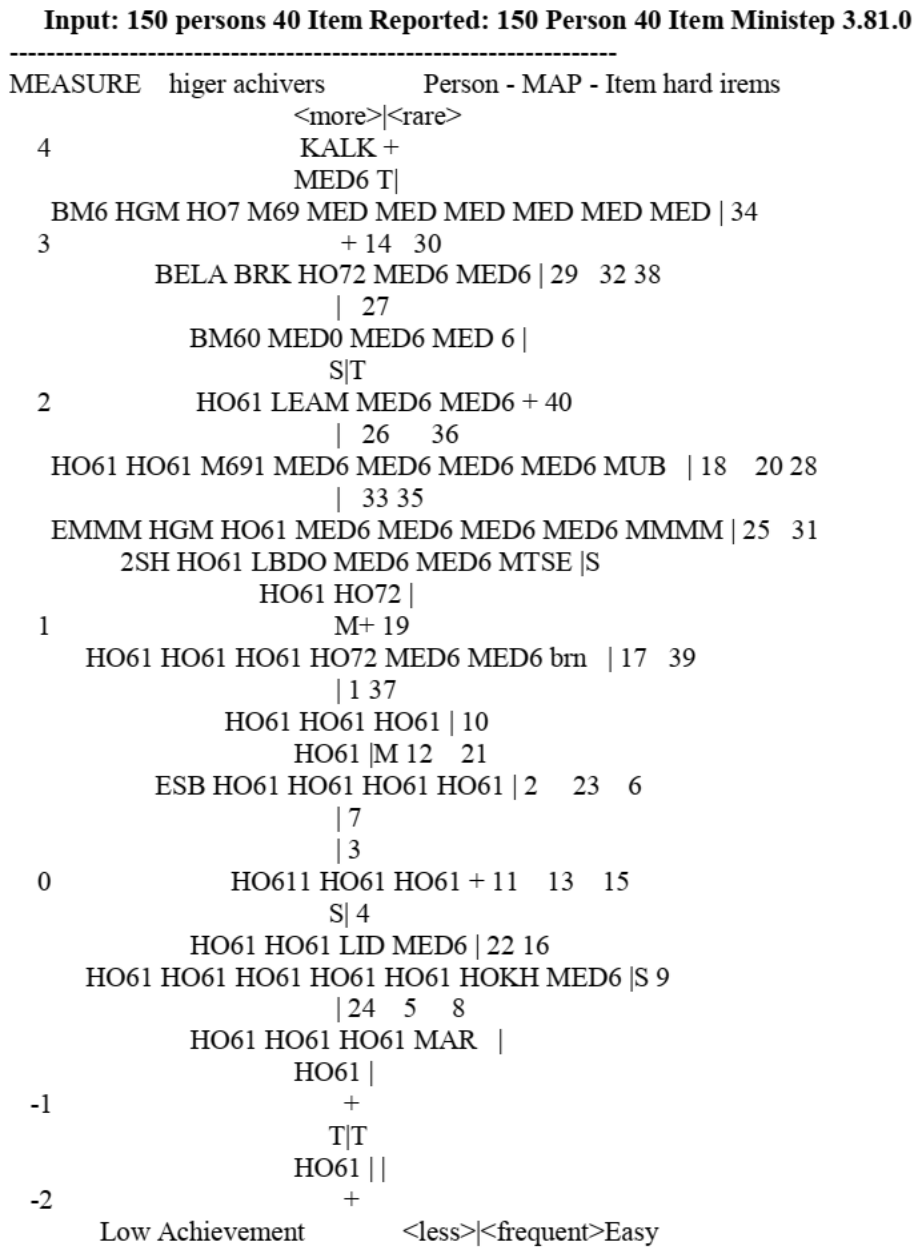
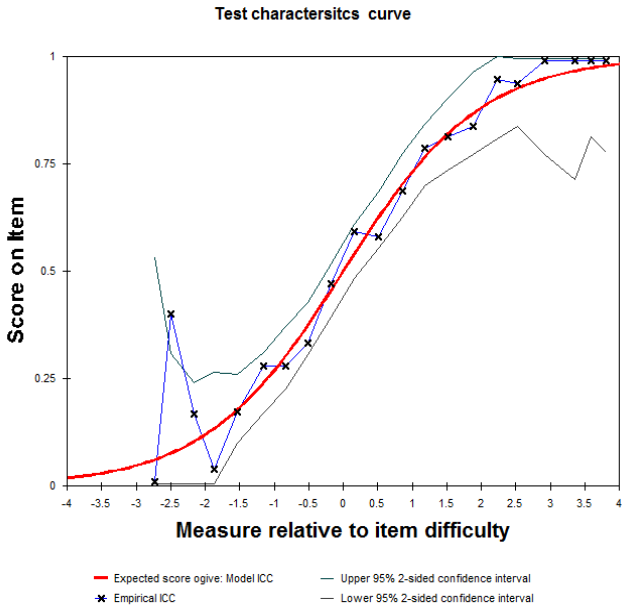


Fig-1: Person measures of achievement and item difficulty map for psychology test (N=150, I=40)

Test Characteristic Curves

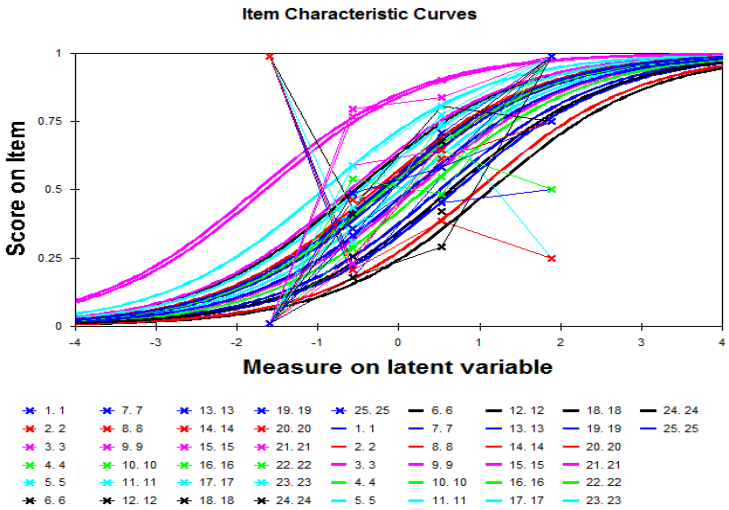
Displays the Rasch-model test characteristic curve (TCC), showing the relationship between raw scores and Rasch measures on the complete test of all active items. EXP. is the expected value of the correlation when the data fit the Rasch model with the estimated measures. This shows the joint display of the expected and empirical ICCs. The boundary lines indicate the upper and lower 95% two sided confidence intervals (interpreted vertically). When an empirical point lies outside of the boundaries, then some model source of variance maybe present in the observations.

The solid red "model" line is generated by the relevant Rasch model. For a test of dichotomous items, these red curves will be the same for every item. The empirical blue line is interpolated between the average ratings in each interval along the variable, marked by x. The empirical ("x") x- and y-coordinates are the means of the measures and ratings for observations in the interval. The upper green lines (and the lower gray line) are at 1.96 model standard errors above (and below) the model “red line”, i.e., forms a two-sided 95% confidence band around the model red line. The vertical distance of these lines from the red line is determined by the number of observations in the interval.



Item Characteristic Curve

The item characteristic curves for good-fitting items of the test are shown on Figure 2.b. The line indicates the expected score of test across student's ability, ranging from the lowest to highest ability groups, for each observed mean measure (dots) of a student ability group. When the observed scores closely follow the curve of expected values, the group is performing as expected on the items (as shown for items stated at the center).



14. 14

The position of the curves also gives a good item difficulty of each item. for the difficulty items, The curve starts to rise on the right side of the plot(high total test scores).this indicates that the probability of answering difficult item correctly is low for most students rise only for outstanding students. For the easier items, the curve starts to rise on the left side of the plot (low total test scores); the probability of answering an easy item correctly is high for everyone but the very bottom group of examinees.

CONCLUSION

Item banking is the assessment archive that is achieved by Practical Applications of Item Response Theory. The most basic model in IRT is the one-parameter logistic Rasch model (1PL). The item parameter that is estimated in 1PL model is the difficulty parameter is scaled using a distribution with a mean of 0.0 and standard deviation of 1.0. Note that student ability is also scaled on a normal distribution so that the mean is 0.00 and standard deviation is 1.00. They are converted to log odds, or logits (usually from

–4 to 4), by taking the natural log of the proportion incorrect for items or failures for persons. The present study involved an analysis with the win step program tested 40 psychology items (N=150). The major findings of this study are summarized as: The individual probability fit of items to the measurement model was then revealed that 34 items fitted the measurement model with probability p>0.03.

The item-trait interaction was not statistically significant at the 0.01 level [Chi-square (df =1036 =1129.49, p =0.022] indicating that a dominant trait was measured and the test fits the standard item response model .the item difficulties range from – 0.53. logits (SE= 0.30) to + 3.2 logits (SE=0.34) and the student ability measures range from – 1.11 logits to +4.0 logits. The position of the curves a so gives a good indication of item difficulty of each item .for the most difficulties items, the curve starts to rise on the right side of the plot(high total test scores).this indicates that the probability of answering difficult item correctly is low for most students rise only fore outstanding students. For the easier items, the curve starts to rise on the left side of plot.

Implications for Students and Teachers

With regard to teachers, testing with calibrated item bank is likely to be accurate in assessing individual student‘s ability in any tested situation. The teachers can use it with individual students or groups without worrying about cheating in the examinations. Each examinee does different test items and a different number of items. This depends upon an individual‘s ability.

In addition, data gained from the test can be used for many purposes, such as, to follow up an individual‘s learning progress, to diagnose deficiencies in each student. Student‘s weaknesses in any subject matter can consequently be remedied. calibrated item bank provides teachers with an efficient and authentic assessment of student‘s learning. It is recommended that universities teachers would be encouraged to learn more about Rasch measurement modeling, item banking and involved in the development of item bank for their own subjects.

Implications for University Academic Staffs

In relation to institutional networking, it would be useful for members of the network (academic staffs) to access the item banks available through Computerized net work. The network could develop a bank containing tests for different subject areas. This can be done by establishing one center as the item bank, equipped with a central computer, while other member faculities in the network can access the bank through the networking computers in their faculities. This can save time and school resources in preparing tests and conducting examinations whenever it is needed. Regarding the development of the test items, teachers in every school network could cooperate to construct, try out, analyzing, and selecting qualified items to store in the item bank. If this process is continuously done, the item bank will become large with thousands of well-calibrated items by difficulty equated on the same scale or learning assessment.

REFERENCE

1. Kiive E, Fischer K, Harro M, Harro J. Platelet monoamine oxidase activity in association with adolescent inattentive and hyperactive behaviour: a prospective longitudinal study. Personality and individual differences. 2007 Jul 1;43(1):155-66.
2. van der Linden JJ. The sites and services approach reviewed: solution or stopgap to the Third World housing shortage?. Gower Publishing Company; 1986.
3. Chalapat K, Timonen JV, Huuppola M, Koponen L, Johans C, Ras RH, Ikkala O, Oksanen MA, Seppälä E, Paraoanu GS. Ferromagnetic resonance in ϵ-Co magnetic composites. Nanotechnology. 2014 Nov 14;25(48):485707.
4. Fulcher G, Davidson F. Language testing and assessment. New York: Routledge; 2007 Jan.
5. Holland DC, Eisenhart MA, Eisenhart MA. Educated in romance: Women, achievement, and college culture. University of Chicago Press; 1990.
6. Aera AP. Standards for educational and psychological testing. New York: American Educational Research Association. 1999.
7. Meredith W, Kearns J. Empirical Bayes point estimates of latent trait scores without knowledge of the trait distribution. Psychometrika. 1973 Dec 1;38(4):533-54.
8. Andersen EB. Conditional inference and models for measuring. Mentalhygiejnisk forlag; 1973.
9. Kelderman H. Loglinear Rasch model tests. Psychometrika. 1984 Jun 1;49(2):223-45.
10. Molenaar IW. Some improved diagnostics for failure of the Rasch model. Psychometrika. 1983 Mar 1;48(1):49-72.
11. Tolossa D, Tafesse T. Linkages between Water Supply and Sanitation and Food Security. 2008.
12. Krebs CP, Lindquist CH, Warner TD, Fisher BS, Martin SL. The campus sexual assault (CSA) study. Washington, DC: National Institute of Justice, US Department of Justice. 2007 Dec.
13. Linacre JM, Wright BD. WINSTEPS Rasch-Model computer program. 2014.
14. Bode RK, Wright BD. Rasch measurement in higher education. InHigher education: Handbook of theory and research 1999 (pp. 287-316). Springer, Dordrecht.