Journal of Current Science and Technology, September-December 2022 Vol. 12 No. 3, 582-591
ISSN 2630 -0656(Online)


Cite this article: Loganathan, R. K., Suppian, Z., & Mokshein, S. E. B. (2022, September). A survey on different IRT model evaluation software's for test equating and MCQ assessment using IRT (1PL–4 PL). *Journal of Current Science and Technology*, *12*(3), 582-591. DOI:

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

# A survey on different IRT model evaluation softwares for test equating and MCQ assessment using IRT (1PL–4 PL)

Rishi Kumar Loganathan*, Zahari Suppian, and Siti Eshah Binti Mokshein

Department of Human Development, Universiti Pendidikan Sultan Idris, Tanjong Malim, Perak, 35900, Malaysia

*Corresponding author; E-mail: rishiephdizone@gmail.com

Received 1 November 2021; Revised 17 December 2021; Accepted 27 December 2021;
Published online 26 December 2022

## Abstract

Item Response Theory- IRT is a measurement framework utilised in psychological and educational design and evaluation assessments based on rating scales, instruments, achievement tests, and others that measure mental traits. It is increasingly popular among the academic field and review to evaluate cognitive and non-cognitive measurement. In higher education assessment tasks, multiple-choice questions are generally used since it is simple and easily scored with better coverage of the instructional information. However, statistical evaluation is necessary to ensure the high-quality items is utilised as an inference basis. Hence, this paper reviews the IRT models, equating 1PL-4PL, assumptions, and various other models. Also, the powerful software utilised in IRT model evaluation, such as XCalibre and jMetrik software, is reviewed and presented. Further, the Multiple Choice Questions- MCQs assessment using IRT is analysed focusing on the item parameters and performed comparatives study. This study concluded that IRT is the better framework exploited by different researchers in evaluating the educational and psychological data for an assessment concerning high-quality items identification for the anchor items selection utilised for the test equating approach.

***Keywords*****:** *1PL-4PL; IRT; jMetrik; test equating error; XCalibre.*

## 1. Introduction

In the psychometrics field, Item Response Theory- IRT was initially proposed for the ability assessment purpose. In education, it is used widely for evaluating and calibrating the items in questionnaires, tests and other measures for scoring the subjects on their attitudes, abilities or other underlying traits. The IRT equating is essential across various test forms, the common measurement for generating comparable scores and measuring common content in a common measurement scale (Hambleton, Swaminathan, & Rogers, 1991). IRT equating comprises two steps such as i) true and observed scores and ii) on a similar measurement scale, the person and or item parameters are placed. While equating coefficient estimation performing, the random error has indexed by equating standard errors. In the IRT test equating, the common mistakes reporting has been subjected to standard practice like delta and bootstrap methods, which are widely utilised. Highly computationally intensive complexities have been seen in the bootstrap method, whereas for the expressions of standard errors, the delta method is more based on the complicated mathematical formulas derivations. In some cases, both approaches are not practically flexible where the IRT models, equating methods, and complex linking designs have been utilised (Brown, & Abdulnabi, 2017; Uduafemhe, Uwelo, John, & Karfe, 2021).

### 1.1 Test equating errors

Test equating study is performed on the generated response items in which the scoring of the common items is automatically performed in 500 samples (Almond, 2014). Tucker linear equating and

linear logistic equating approaches have been utilised. There exists only one test equating study with the mixed format of automated scoring. Further, another study (Olgar, 2015) used open-ended and 30 multiple-choice items. Linear logistics equating used by (Almond, 2014; Olgar, 2015). Also, equating tests concentrated on (Uysal, & Doğan, 2021) with automated scoring and constructed-response items in large numbers. A single test equating method (Uysal, & Doğan, 2021) was performed in this study using the IRT and CTT concerning test correlating techniques with automatic scoring. More constructed-response items are presented in larger-scale tests. Several steps are presented for better evaluation to measure the higher difficult skills like reasoning, higher-order and critical thinking. And accurately, these items are scored. Test equating research on the limited automated scores of constructed items response is inadequate. Hence, this (Uysal, & Doğan, 2021) focused on two purposes: investigating the equating errors and conditions in automated scoring and constructed item responses effect evaluated by the computerised scoring while analysing the equating errors.

Test equating procedures are performed based on IRT and CTT- Classical test theory. It is impossible to provide invariance and equality assumptions when the CTT is utilised. Individual abilities can be independently recognised when IRT is used for items. It doesn't affect the unique abilities when the tests are easy or difficult. When they satisfy local independence and uni-dimensionality, one-dimensional IRT models are utilised—single latent ability measures one-dimensional tests. Based on the abilities, independence of answers is associated with local independence provided to items from participants. IRT is commonly based on the probability of replying to one item appropriately. The probability is generated by item difficulty and individual ability. This item response function is denoted in graphical form, and it is referred to as the test characteristic curve. There is no requirement to equate IRT operations since IRT shows invariance characteristics. The abilities can be identified directly when the calibration operations are performed on item parameters.

In real applications, this is not possible. When the anchor item design is utilised in non-equivalent groups, items parameters are attained from two different tests and samples from various populations. For separate calibrations, these situation is called. It is important to place item parameters when separate calibrations are executed on a general scale. For obtaining linear transformation, two values are measured using the equivalent approaches. Through several methods, slope A and B methods are obtained, which helps identify the individual's ability levels in various test forms.

The study's major contribution involves a review of the IRT models with the test equating 1PL-4PL, assumptions and various other models. Also, the significant software utilised in IRT model evaluation, such as XCalibre and jMetrik software, is reviewed and presented.

Section 2 describes the IRT models and different IRT software such as XCalibre and jMetrik. Further, the performed comparative analysis is in section 3. Finally, the study is concluded in Section 4.

## 2. IRT models and IRT software's for identification of high-quality items

2.1 IRT for test equating using IRT (1PL–4 PL)

Further, the random imputations based multiple imputation method (Zhang, 2020a) are developed for the item parameters and standard errors estimating for the transformation coefficients' IRT parameter scale. Compared with the bootstrap method and delta method, the multiple imputation method is lesser computationally intensive, and also it has not based on the derivations of the complicated formulas. Real and simulation data have been utilised in this (Zhang, 2020a), comparing and examining the multiple imputation method performances with delta and bootstrap method for 2PL IRT model in the perspective of non-equivalent groups of common item equating design. Furthermore, for the transformation coefficients of the IRT parameter scale, the standard errors are determined in many simulated conditions by using the multiple imputation method.

Different IRT models are utilised. Three 3 PL- logistic parameter model is defined as,

$$P_{ji}=c_i+(1-c_i)\left(\frac{exp\left(a_i(\theta_j-b_i)\right)}{(1+exp\left(a_i(\theta_j-b_i)\right)}\right) \quad (1)$$

From Eq. (1), for examinee j, correct response probability is Pji to item i, ability parameter is θj, for item i discrimination parameter is ai, location or difficulty parameter is bi, and the pseudo-guessing parameter is ci. If ci=0, then 2-PL

(two-parameter logistic model) is obtained and expressed as,

$$P_{ji} = exp\,(a_i(\theta_j - b_i)) / (1+exp(a_i(\theta_j - b_i))) \quad (2)$$

The similar discrimination parameter ai=1 associated with all items are assumed, and 1- PL (one-parameter logistic model) is obtained. Multi-dimensional IRT models are generated for the tests and measure several dimensions simultaneously (Sansivieri, Wiberg, & Matteucci, 2017).

Two IRT model categories are comprised of polytomous and dichotomous models. The polytomous model dealt with essay test items, and the dichotomous models dealt with objective test items. The IRT and the reporting scale provided have overcome the classical test theory shortcomings, which is independent of the specific test items choice (Nathaniel, Edougha, & Odjegba, 2019). Alternative procedure for the standard errors obtaining has considered as multiple imputation method, according to this (Zhang, 2020b) study for the equating coefficients of IRT true score in the non-equivalent groups equating design's common items based on 3-PL model. Several set of parameter values of imputed item has been utilised. The multiple imputation method performances have been investigated based on real and simulation data compared with delta and bootstrap methods. Results show delta method is similar to multiple imputation method results.

In addition to that, the four-parameter logistic IRT – 4PL model is presented as the fourth parameter model by Barton and Lord, which allows the student with greater ability to miss the simple item without underestimating their abilities. Furthermore, the 4PL IRT model has been analysed as an error correction concept by relating it with the 3PL IRT model based on two administration test conditions. Also, the studies show that the 4PL model assists the examinees to recover from previous errors without measurement efficiency and precision diminishes. Specifically, the upper asymptote accommodates students with a greater ability to make an inconsiderate error on the previous item. Also, the 4PL IRT model is a strong mechanism against unusual responses (Battauz, 2020; Liao, Ho, Yen, & Cheng, 2012).

IRT- Item responses theory evaluate and regulate the items from IRT on surveys, same instruments and tests in the psychology and education field. However, it is more popular in marketing and health. For research, data management and evaluation, a statistical analysis software tool is widely considered, which performs statistical analysis and is used in different industries. IRT procedure introduction across different industries have been evaluated, and the test items are calibrating, connecting and scoring analysed with separate software. The IRT procedure, namely PROC IRT with small-scale standard evaluation, has been introduced (Matlock Cole, & Paek, 2017).

Moreover, data simulation of IRT is favouring the IRT equating techniques. Pseudo groups and tests have been generated for the equating results generated related to the data simulation of IRT. Large sample single group rule and equating are set as true or criterion equating global and local indices have been identified. Kernel equating shows stability and accuracy in IRT observation score in the design of random equivalent groups results. With anchor test design based on non-equivalent groups, the IRT observed score exhibited random errors and the lowest systematic in the equating procedures. As a shorter test, the errors have been decreased, and for equating, the larger sample has been utilised and can ignore it. True equating has been spotted in the design of random equivalent groups (Wang, Zhang, & You, 2020).

The test length and sample size effects have been evaluated on the estimation of item parameters in developing a test using the IRT 3 uni-dimensional dichotomous models. Utilised fifty items were with a real language test performed on the 6,288 students. Three test lengths dataset comprised of 10, 20 and 30, and 9 various sample sizes have been obtained. Estimation of Item parameter accuracy has been manipulated based on the sample size, test length and model variables of IRT under various conditions. Accurate item parameters are estimated in three unidimensional dichotomous Item Response Theory models based on the employed model and test length (Sahin, & Anil, 2017).

Consequently, the multi-dimensional IRT equating accuracy has been compared among the separated calibrated MOSE and concurrent calibrated MOSE procedure with the test characteristic function based on Monte-Carlo simulation. It is performed mainly for a mixed-format test with a basic structure with a non-equivalent group with a common item score and

anchor test design. The score equating accuracy has been considered based on the standard error of equating variance co-efficient- CVSE. While the CVSE value is compared with SMOSE and CMOSE procedures, the interaction among the common item properties and MOSE procedures affects the value of CVSE. Compared with SMOSE, CMOSE shows minimised CVSE value. However, for CMOSE, while relating with common item proportions, the results show that the values of CVSE among the proportions show no variations (Panidvadtana, Sujiva, & Srisuttiyakorn, 2021).

The equivalence of mathematics is examined based on junior high school scores. Through the Haebara method, the IRT with 2-PL analysis has been performed. For the test, equating R and IRT packages have been utilised while performing the estimation step. The try out questions of package 1 (2099 respondents) and 2 (2068 respondents) relationship have not been similar. After the measurement and estimation of the regression equation, it shows that mathematics national examination package two questions contain more complexities than package 1. Similar or fair scores are provided based on the package 1 and 2 test equating to the 2-PL method through the Haebara method. Common item equating design followed. However, the test packages completely differ. Therefore, the examiners should work on the same tests to properly detect test-takers ability (Elvira, & Sainuddin, 2021).

The Monte Carlo simulation detects IRT score classification accuracy and in case of model misspecification, sum score based equating methods are followed. For the test equating four kinds of equating methods were utilised such as sum-score based equating, haebara method, Kernel equating and IRT score based equating. Test settings have been conducted based on the Swedish scholarship aptitude test, German university psychology exam, and Australian citizenship tests. Higher classification accuracy from IRT score based equating technique and single cutoff scores. Reduced classification accuracy has been seen from model misspecification. IRT equating methods used test agencies because of the higher capacity. If test items are multi-dimensional, the number of proficiency levels have been minimised (REWIND project, n.d.).

Delta scoring – Dscoring utilised in large-scale measurements at the National Centre for Saudi Arabia Assessment in the educational measurement field. Based on DSM range (0-1), the D scores are obtained, indicating measurement ability by binary test items defined by the examinee. This article identified whether D-scale is an interval scale related to IRT ability scores concerning interval through axioms testing of ACM- additive conjoint measurement. Conjoint checks are considered a testing approach using the Bayesian method to analyse the hypotheses violated in the IRT dataset. Fewer violations are seen under the D scores compared with IRT scores concerning the ACM ordering axioms. DSM exhibited the dependable D-scale concerning the intervals essential property (Domingue, & Dimitrov, 2021).

The item fit statistics and model are compared in a mixed format with respect to response items and multiple choices. This study was performed in turkey for 2351 students for the 4th-grade science test in 35 schools. The items are calibrated concurrently and separately by using various IRT models. The calibration method effect on item fit and model examined on the real data. While logistic parameter models PL-1, 2 and 3 are used for calibrating the binary coded items, generalised partial credit model for open-ended calibration and graded response model. Further, the polytomous and dichotomous models are applied. Finally, the graded response model and the 3PL combination is used for the best-fit statistics results (Himelfarb, 2019).

Multi-dimensional IRT models were further developed in the early 1990s, and for item dependency, common traits are considered. Mathematical logic with 1 PL – 3 PL models is presented. Compared with traditional testing, IRT shows an advantage, like it defines a scale for the latent variable underlying in which test items are measured. By the single latent traits, unidimensional test responses are assumed by the IRT to refer to the test taker's ability. This kind of trait is not observed straight, and it is generated using the test items' observed responses (Himelfarb, 2019). IRT models are presented. The ability measures relationship is presented through the item response and instrument. A person with provided ability can respond in the correct manner given by IRF. Lesser chance is only given if the person possesses the lower ability and vice-versa. Birnbaum and Rasch IRT models are presented in this study for the dichotomous parameters and the four theoretical parameters. IRT models application is presented for the survey analysis and social data

set. For analysing the Rasch model and IRT, the R package is presented. The most common models are covered for the polytomous and dichotomous data. Latent ability and test item performances relationship has been provided and also addressed the ability information of the examinee (Brzezińska, 2020).

2.2 IRT software

This study focuses on the significant IRT software as X-Calibre and jMetrik.

*2.2.1 X-Calibre*

The chemistry achievement test assessed in this (Bichi, Embong, Talib, Salleh, & bin Ibrahim, 2019) article and the generated item statistics are related to using IRT and CTT methods. Using the 530 students descriptive survey has adopted. The software, namely ITEMAN and XCALIBRE, have been utilised for item analysis performance. The problematic items (32.5 %) and good items (67.5 %) are identified from the results. From IRT and CTT models, the item statistics were derived using higher correlation for discrimination and item difficulty, respectively. Through a standardised process, the assessment test has been performed. Same and related results were exhibited from the two contexts, which presented as reliable and effective. This study suggested that evaluation and item development have been incorporated because of the superiority in reducing the measurement errors and reliability examination. Regarding the rash analysis, the remaining items are evaluated in measuring the ability of a single dimension. Item reduction with missing data eliminated based on Rasch analysis. The item parameters have been estimated using the XCalibre software 4.2 with respect to expectation maximisation algorithm- EM by using the marginal maximum likelihood. If the IRT model fit into information, the items are calibrated using the maximum likelihood estimation (Wiebe et al., 2019).

*2.2.2 Jmetrik software*

For the calculation of test scaling, non-parametric IRT applications, statistics, DIF, IRT equalisation and linking, estimation of reliability, test scaling, IRT models and Rasch measurement models, jMetrik software has been utilised. In the use of non-parametric and parametric IRT applications, jMetrik 4 exhibited higher importance. The non-parametric IRT procedures are simply saved in colour as .png or .jpg files. The information is examined by the non-parametric characteristic curves and further evaluating the relationship among the correct responses and latent traits by a simple and rapid tool. Every item's real difficulty and significant discrimination interpretation are considered as a drawback of the non-parametric characteristic curves. For quantifying the properties, items comparing or two various groups comparison the parametric IRT makes it simpler.

Two evaluation options are offered by jMetrik with respect to the parametric IRT. For the Rasch model, maximum likelihood estimation, rating and credit scale models are used by the software. By the parameter of item difficulty and threshold, a partial credit model is generated. The partial credit model special case is considered as a rating scale model based on the parameters of the threshold. For the item, threshold and individual parameters, instead of the Newton-raphson method, jMetrik utilises the proportional curve-fitting algorithm (Meyer, & Hailey, 2012). The fit statistics goodness are computed by the software for the individuals, items and estimation of the parameter. Statistics of scale quality like reliability and separation are measured within software scope. For the multi-category and two-category IRT models, marginal maximum likelihood estimation-MMLE is utilised from jMetrik comprised with generalised partial credit model- GPCM, 3-PLM and 4-PLM. For individual characteristics scoring, three options are made such as expected a posteriori- EAP, maximum likelihood and maximum a posteriori- MAP. In addition to analysis results, the output table is created based on software options, utilised as procedure inputs like scales linking and so on (Gökhan, Güzeller, & Eser, 2019).

For the IRT analysis, results depicting jMetric exhibited two options. The first technique gives the information and standard error functions and item characteristic curves for the whole test and all items individually. The output tables are comprised of information used by software in which the suitable IRT model is selected automatically, and the graphics are produced rapidly. The next technique is to give the item maps in result analysis. Within the Rasch measurement scope, item mapping is common, and the individual skills estimation and item parameter distribution are illustrated with respect to two histograms with the

general axis. With respect to the quality of match assessment among the items and individuals, this method is helpful. More precise individual skill estimation is obtained from more or lesser items. Thus for the test and test development selection, both methods are considered tools and guides.

Under the IRT and CTT data analysis, jMetrik software is performed under a single roof without additional software. Significantly jMetrik software handled the multi-category and two-category items. Analysis of the differential item function has been done by IRT. Analysis can be performed simply by using the pop-up windows. Through commands, every analysis is performed. jMetrik is open-source software installed free of cost. The obtained parameters in jMetrik are the same as PASCALE, BILOG, and IRTPRO. IRT programs are inspiring the theoretical foundation of this software and provide a higher advantage in reporting and interpreting the analysis results. Compared with other programs, jMetrik analysis is performed in a short duration. Without the creation of a database and if it is not defined in other IRT programs, analysis cannot be performed (Gökhan et al., 2019).

## 3. IRT software comparative study

In various products, IRT is utilised, the insights and results are obtained on pupils and tests. In this section, the comparative analysis of certain significant software's utilised for IRT is presented (Garza, & Fiore, n.d.).

**Table 1** Comparative Analysis

| Software | Description | Benefits | Limitations |
|---|---|---|---|
| J-Metrik | J. Patrick Meyer established a free psychometric software, namely JMetrik. In 2006, its origins were realised, and in 2009 it was released. JMetrik allows the user to navigate the software easily, and it is based on the graphical interface, and necessary evaluations are performed. | JMetrik is free software with a psychometric analysis special offer. It has been portable on various platforms since it was written in Java. Instead of a file-based model, it depends on the database model, and it leads to results and computations in a similar place. Thus the project management is easier. The R package is finally provided, which allows product and R interactions. | For data importing incorrect format, it needs certain work since the software is based on a database model. Any command-line interfaces are not provided, and they can be very helpful during the automation process. |
| IRTPRO | For test scoring and calibration of the item, IRTPRO is considered a new application. | This IRTPRO software made a way to execute the .irtpro command. For the application, these significant features are highly useful. | Due to the error, it is impossible in installing the software. For full usage, activation is required, and also, the software is not free. |
| Xcalibre | Assessment systems corporation was established by X-Calibre. For performing the IRT parameters estimation based on reports which are user friendly, the software is designed as a window application. | The user interface of Xcalibre is highly spontaneous in use with easy input files structure. It further allows the user in IRT potentialities exploitation without additional works. Software support is helpful and includes an exhaustive manual. Theoretical knowledge is needed in considering the evaluations. For final results interpretation, the report with all information is submitted. The CSV files can be used and elaborated again. | Significant consideration should be given to the license. From the command line, the software cannot be executed. With various question sets, the software is used without reloading item specifications and results every time. |

The IRT model practical consequences have been examined by various studies. Also, the equating misfit and examinees classification performances and its categories with respect to large scale assessment programs have been studied with mixed-format test data. Three factors considered are IRT scaling choice techniques, IRT model choice and examinees abilities change/growth amount among two administration years. The important model consequences misfit is differentiated over IRT scaling and choice methods. Fixed common item parameter and separate

calibration with connection procedure shows high sensitivity for misfitting and potential against different ability shifts among the adjacent administrations related with the stocking and lord characteristic curve- SL and mean/sigma methods- MS. For equating conversion, recovering SL shows lesser sensitivity and against the ability shifts and when misfit of substantial degree present, the MS shows the least potential. Thus IRT model misfit consequences have been addressed, and also the specific IRT models validity has been examined. The IRT models applications have improved with psychological and educational test data (Zhao & Hambleton, 2017).

## 4. MCQs Assessment using IRT

In the educational assessment, MCQs- Multiple Choice questions are generally utilised in higher secondary education since they are simple, easy and accurate, saving time and manpower. However, some of the studies stated that MCQs concentrates only on what student remember and are not based on the student's understanding, course-related knowledge and analysis. It is used in standardised tests. The instructional contents are covered and easily scored. Based on student experiences, it focuses on the assessment approach. Generally, three approaches are utilised for the MCQs quality assessment. First is the traditional method, namely process control comprised of four steps in MCQ items writing, content defining, selecting format and style, writing the items and options and five items are applied for item quality assessment. These five items are clearly mentioned questions, error-free questions, feasible distractors, better explanations and correct specified answers. But these kinds of approach is about prevention more compared with evaluation. The second approach is focused on the classical test theory- CTT. This method is about true scores with the sum of examinees with respect to unobserved random error and theoretical ability, same as individual examinees scores in the test.

(observed score = true score + error) (3)

When there is a change in examinees class, assigned similar items with various discrimination and difficulty values, the reason for the change is not based on student population abilities distribution. The third method is IRT- Item response Theory. Correct responses are plotted and fitted in smooth lines, which is S-shaped refer as Item Characteristic Curve (Cruz, Freitas, Macedo, & Seabra, n.d.). Based on the examinee's distribution, the item parameter in the IRT model is estimated. Item parameters such as difficulties, discrimination and guessing (?) show variation in different examinees. By the item parameters linear variance, one item parameters for various examinee teams are associated.

Discrimination parameter (a) item and difficulty parameter (b) can be used to assess the quality of MCQs in IRT. The difficulty parameter is the point that equals examinees' abilities, in which the probability of answering items is 50%. This research (Jia, He, & Zhu, 2020) shows that the examinees are selecting the guessing method has not associated with the type of exam or question. MCQs disadvantages are termed as students who answer correctly by guessing and however, they do not know the answers in real. The 2PL model, which has no guessing parameter, consistently has the best fit for MEU, MEC, HEU, and HEC. The feature of the most fitting model was used to interpret MCQs exams. Oppositely, in most MCQ exams, students resolve the items with their abilities. The limitation of this research is the use of simple models. For revealing the final mathematics examination items quality in statistical form, this research (Kusumawati, & Hadi, 2018) has been performed. 2-PL model employed based on IRT. The medium level is 60 %, and difficult items are 40 % of 35 items. The trigonometric calculation is considered as most difficult material. Item discrimination index percentage shows that very low items are 8.57%, low items are 51.43%, medium items are 31.43%, higher items are 5.71%, and very high items are 2.86%.

**Table 2** Comparative analysis of MCQs assessment by IRT and other models

| S.No | Author | High-quality items | Parameters | Students target | Assessment tool | Suggestions |
|---|---|---|---|---|---|---|
| 1 | (Mehta, & Mokhasi, 2014) | A hundred First-year MBBS students took the MCQs test | Difficulty index, Discrimination index and Distractor effectiveness | Department of Anatomy | Item analysis | Item analysis helps tremendously to achieve better teaching, better |

| S.No | Author | High-quality items | Parameters | Students target | Assessment tool | Suggestions |
|---|---|---|---|---|---|---|
| | | comprising of fifty questions. | | | | learning and in the long term, better tests |
| 2 | (Quaigrain & Arhin, 2017) | 50 MCQs were administered as an end of semester examination | Difficulty index (p-value) and discrimination index (DI) with distractor efficiency (DE) | 247 first-year students | Kuder–Richardson | Items having average difficulty and high discriminating power with functional distractors should be integrated into future tests to improve the quality of the assessment |
| 3 | (Azevedo, Oliveira, & Beites, 2019) | MCQ contained in a bank of questions, implemented in Moodle | The Difficulty and Discrimination Indexes | Data set of students' grades from tests | Item Response Theory (IRT) and Classical Test Theory (CTT) | The analysis also showed that the bank of questions presents some internal consistency and, consequently, some reliability. Groups of questions with similar features were obtained, which is very important for the teacher to develop tests as fair as possible |
| 4 | (Benedetto, Cappelli, Turrin, & Cremonesi, 2020) | newly generated multiple-choice questions | The difficulty and the discrimination | Cloud academy dataset | R2DE (which is a Regressor for Difficulty and Discrimination Estimation), IRT | This model enables a reduction in the number of questions that have to be removed from the set of assessment items due to either a too low discriminative power or too high difficulty. |
| 5 | (Jia et al., 2020) | determine the characteristics of MCQs and factors | Item difficulty and discrimination | four samples of different sizes from the US and China in secondary and higher education | IRT | high-quality items can be used as bases of inference in middle and higher education |

## 5. Conclusion

IRT models have been utilised for several decades in the field of educational, psychological assessment with focusing on reliable outcomes. Also, it is a cornerstone in the evaluation and supports in improving the test quality and producing scales, test equating and administration. Also, it is helpful in understanding the functioning of differential items, adaptive testing computerisation, test scoring and interpretation. This study performed a review of the IRT models with the test equating 1PL- 4PL, IRT software utilised in IRT model evaluation such as XCalibre and jMetrik software. Further, the MCQs assessment using IRT is analysed focusing on the item parameters and performed comparatives study. Finally, this study concludes that IRT is a better framework for producing reliable, credible and valid results for the identification of high-quality items.

## 6. References

Almond, R. G. (2014). Using automated essay scores as an anchor when equating constructed response writing tests. *International Journal of Testing*, *14*(1), 73-91. DOI: https://doi.org/10.1080/15305058.2013.816309

Azevedo, J. M., Oliveira, E. P., & Beites, P. D. (2019). Using Learning Analytics to evaluate the quality of multiple-choice questions: A perspective with Classical Test Theory and Item Response Theory. *The International Journal of Information and Learning Technology*. DOI:

https://doi.org/10.1108/IJILT-02-2019-0023

Battauz, M. (2020). Regularized Estimation of the Four-Parameter Logistic Model. *Psych*, *2*(4), 269-278. DOI: https://doi.org/10.3390/psych2040020

Benedetto, L., Cappelli, A., Turrin, R., & Cremonesi, P. (2020, March). R2DE: a NLP approach to estimating IRT parameters of newly generated questions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. DOI: https://doi.org/10.1145/3375462.3375517

Bichi, A. A., Embong, R., Talib, R., Salleh, S., & bin Ibrahim, A. (2019). Comparative analysis of classical test theory and item response theory using chemistry test data. *International Journal of Engineering and Advanced Technology*, *8*(5), 1260-1266. DOI: 10.35940/ijeat.E1179.0585C19

Brown, G. T., & Abdulnabi, H. H. (2017, June). Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. In *Frontiers in Education*. Frontiers Media SA. DOI: https://doi.org/10.3389/feduc.2017.00024

Brzezińska, J. (2020). Item response theory models in the measurement theory. *Communications in Statistics-Simulation and Computation*, *49*(12), 3299-3313. DOI: https://doi.org/10.1080/03610918.2018.1546399

Cruz, J., Freitas, A., Macedo, P., & Seabra, D. (n.d.). Quality of Multiple Choice Questions. Retrieved from https://core.ac.uk/download/pdf/231952247.pdf

Domingue, B., & Dimitrov, D. (2021). A comparison of IRT theta estimates and delta scores from the perspective of additive conjoint measurement. DOI: 10.35542/osf.io/amh56

Elvira, M., & Sainuddin, S. (2021). *Equating Test Instruments Using Anchor to Map Student Abilities Through the R Program Analysis.* In *Proceedings of the International Conference on Engineering, Technology and Social Science (ICONETOS 2020).* DOI: https://doi.org/10.2991/assehr.k.210421.095

Garza, P., & Fiore, M. (n.d.). *Historical Data Analysis and Item Response Theory for Calibration of Question Items*. Retrieved from https://webthesis.biblio.polito.it/7456/1/tesi.pdf

Gökhan, A., Güzeller, C. O., & Eser, M. T. (2019). JMETRIK: Classical Test Theory and Item Response Theory Data Analysis Software. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(2), 165-178.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2): London, US: Sage publications. DOI: https://doi.org/10.2307/2075521

Himelfarb, I. (2019). A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. *Journal of Chiropractic Education*, *33*(2), 151-163. DOI: https://doi.org/10.7899/jce-18-22

Jia, B., He, D., & Zhu, Z. (2020). Quality And Feature Of Multiple-Choice Questions In Education. *Problems of Education in the 21st Century*, *78*(4), 576-594. DOI: https://doi.org/10.33225/pec/20.78.576

Kusumawati, M., & Hadi, S. (2018). An analysis of multiple choice questions (MCQs): Item and test statistics from mathematics assessments in senior high school. *REiD (Research and Evaluation in Education)*, *4*(1), 70-78. DOI: https://doi.org/10.21831/reid.v4i1.20202

Liao, W.-W., Ho, R.-G., Yen, Y.-C., & Cheng, H.-C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality: an international journal*, *40*(10), 1679-1694. DOI: https://doi.org/10.2224/sbp.2012.40.10.1679

Matlock Cole, K., & Paek, I. (2017). PROC IRT: A SAS procedure for item response theory. *Applied psychological measurement*, *41*(4), 311-320. DOI: https://doi.org/10.1177/0146621616685062

Mehta, G., & Mokhasi, V. (2014). Item analysis of multiple choice questions-an assessment of the assessment tool. *Int J Health Sci Res*, *4*(7), 197-202.

Meyer, J. P., & Hailey, E. (2012). A Study of Rasch, partial credit, and rating scale model parameter recovery in WINSTEPS and jMetrik. *Journal of Applied Measurement*, *13*(3), 248–258.

Nathaniel, E., Edougha, D., & Odjegba, G. (2019). Analysis of Past West Africa Examination Council/Senior Secondary Certificate Examination Physics Essay Test Items Parameters: Application of IRT Polytomous Model. *Journal of Education and Practice*, *10*(17), 29-34. DOI: 10.7176/JEP/10-17-04

Olgar, S. (2015). *The integration of automated essay scoring systems into the equating process for mixed-format tests* (Doctoral dissertation). The Florida State University, USA.

Panidvadtana, P., Sujiva, S., & Srisuttiyakorn, S. (2021). A Comparison of the Accuracy of Multidimensional IRT equating methods for Mixed-Format Tests. *Kasetsart Journal of Social Sciences*, *42*(1), 215–220-215–220. DOI: https://doi.org/10.34044/j.kjss.2021.42.1.34

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1301013. DOI: https://doi.org/10.1080/2331186x.2017.1301013

REWIND project. (n.d.). Video: Copy-move forgeries dataset. Retrieved from https://sites.google.com/site/rewindpolimi/downloads/datasets/video-copy-move-forgeries-dataset

Sahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, *17*(1), 321-335. DOI: https://doi.org/10.12738/estp.2017.1.0270

Sansivieri, V., Wiberg, M., & Matteucci, M. (2017). A review of test equating methods with a special focus on IRT-based approaches. *Statistica*, *77*(4), 329-352. DOI: https://doi.org/10.6092/issn.1973-2201/7066

Uduafemhe, M. E., Uwelo, D., John, S. O., & Karfe, R. Y. (2021). Item Analysis of the Science and Technology Components of the 2019 Basic Education Certificate Examination Conducted by National Examinations Council. *Universal Journal of Educational Research*, *9*(4), 862-869. DOI: https://doi.org/10.13189/ujer.2021.090420

Uysal, I., & Doğan, N. (2021). Automated Essay Scoring Effect on Test Equating Errors in Mixed-format Test. *International Journal of Assessment Tools in Education*, *8*(2), 222-238. DOI: https://doi.org/10.21449/ijate.815961

Wang, S., Zhang, M., & You, S. (2020). A Comparison of IRT Observed Score Kernel Equating and Several Equating Methods. *Frontiers in Psychology*, *11*, 308. DOI: https://doi.org/10.3389/fpsyg.2020.00308

Wiebe, E., London, J., Aksit, O., Mott, B. W., Boyer, K. E., & Lester, J. C. (2019, February). Development of a lean computational thinking abilities assessment for middle grades students. In *Proceedings of the 50th ACM technical symposium on computer science education*. DOI: https://doi.org/10.1145/3287324.3287390

Zhang, Z. (2020a). Asymptotic standard errors of equating coefficients using the characteristic curve methods for the graded response model. *Applied Measurement in Education*, *33*(4), 309-330. DOI: https://doi.org/10.1080/08957347.2020.1789142

Zhang, Z. (2022). Estimating standard errors of IRT true score equating coefficients using imputed item parameters. *The Journal of Experimental Education*, *90*(3), 760-782. DOI: https://doi.org/10.1080/00220973.2020.1751579