ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Araştırma Makalesi (Research Article); 2022; 13(1); 15-22

Comparison of Item Response Theory Scaling Methods with ROC Analysis

Meltem Yurtçu *   Cem Oktay GÜZELLER **

Abstract

In this study, one-dimensional item response theory models were evaluated using different scaling methods. In this context, the equating errors and the area under the curve of four scaling methods (Stocking-Lord, Heabara, Mean-Sigma, Mean-Mean), and one, two, and three parameters logistic models (1PL, 2PL, and 3PL) in non-equivalent groups with anchor test (NEAT) design were examined. Additionally, the equating errors of the scaling methods and the results obtained from ROC analysis were compared. Qatar's and Australia's PISA 2012 mathematical literacy test data were used in the study. The minimum error was obtained from the Mean-Mean method with the 1PL model, and the maximum error was obtained from the Mean-Mean method with the 3PL model. Similar results were observed in all comparisons and supported each other. It is concluded that ROC analysis can be used to compare different conditions, methods and models.

Keywords: test equating, ROC analysis, scaling methods, AUC

Introduction

Evaluation is one of the fundamental parts of the education process. Although many different tools are used in the evaluation process, the most frequently used measurement tools, especially in large-scale evaluations, are tests. The evaluation of the tests requires more than one approach due to the administration conditions of large-scale exams. Different forms are used instead of a single form in these applications to ensure reliability. A proper comparison of these forms is crucial for a fair assessment of the scores.

Test equating is required to make the scores obtained from different test forms for the same purpose comparable (Kolen & Brennan, 2014). The selection of the equating design with the least error is an important step to get better results in the test equating process. Equating designs differ according to the groups who took the test and the method used to compare individuals' abilities. These designs are generally named single group design, equivalent groups design, and non-equivalent groups with covariates design. As it is impossible to administer the test forms to individuals with similar abilities, collecting additional information to support the score estimation process or using common variables may improve the accuracy of the estimation. It also contributes to evaluating the equating studies from many aspects (Branberg & Wiberg, 2011; Livingston & Lewis, 2009). This design, referred to as "Non-equivalent Groups with Covariates (NEC)," appeared in the literature with recent studies (Wiberg & Branberg, 2015).

Another important step in obtaining good results besides using appropriate test equating design is selecting the appropriate equating method. Equating methods are based on different theories and assumptions. In the literature, equating methods are classified under Classical Test Theory and Item Response Theory (IRT) (Hambleton & Swaminathan, 1985). IRT-based models determine the probability of the correct answer for a test item according to the ability parameter and item statistics (item difficulty and discrimination indices) (Gonzalez, 2014). The dimensions of IRT theories vary according to the dimension of the item they measure; they generally appear as one-dimensional with two categories. These IRT models are; one-parameter logistic (1PL), two-parameter logistic (2PL) and three-parameter logistic (3PL) models (Embretson, 1996, 1997; Embretson & Reise, 2000; Hambleton

[*] Assistant Professor, Inonu University, Faculty of Education, Malatya-Turkey,meltem.yurtcu@gmail.com, ORCID ID: 0000-0003-3303-5093

[**] Professor, Akdeniz University, Faculty of Tourism, Antalya-Turkey, cguzeller@gmail.com, ORCID ID: 0000-0002-2700-3565

To cite this article:
Yurtçu, M., & Güzeller, C. O. (2022). Comparison of item response theory scaling methods with ROC analysis. *Journal of Measurement and Evaluation in Education and Psychology, 13*(1), 15-22. https://doi.org/10.21031/epod.892079

Received: 06.03.2021
Accepted: 14.03.2022

& Swaminathan, 1985). After the model selection, a calibration process is conducted, and the equating method is determined. The calibration type is determined regarding the separate or simultaneous estimation of the item or ability parameters on a common scale (Petersen, et al., 1983; Hanson & Béguin, 2002). It is called simultaneous calibration if the data of the forms to be equated are in a single file, the mean and standard deviations are on the same scale (Brossman, 2010; Kolen & Brennan, 2014). It is defined as separate calibration if transferred to the same scale after being estimated separately in different forms. The most frequently used methods in separate calibration are the moment method (mean-mean, mean-deviation) and test characteristics curve (Heabara & Stocking-Lord) (Haebara, 1980; Stocking & Lord, 1983). Moment methods are based on the mean and standard deviations of the estimated parameters of the covariates. On the other hand, characteristic curves are based on the difference between the characteristic curves of the items.

The right equating method and design selection should fulfill certain assumptions and conditions. The right choices will contribute to making more correct decisions about individuals and their related processes. The results are compared by using Weighted Mean Square Error (WMSE) or Root Mean Square Error (RMSE) to find the equating method that gives the best result (Tian, 2011). The smallest errors indicate the most appropriate method (Kolen, 1988; Kolen & Brennan, 2014; Wang, 2006).

However, the lack of a practical calculation in choosing the method to be used in the equating process and the failure to calculate these error coefficients disturb researchers' test equating processes.

Besides, it is not easy to compare these equating methods because they are based on different assumptions coming from other theories (Wiberg & Gonzalez, 2016). This situation indicates the need for an analysis that shows the suitability of the comparison criteria or the method.

Receiver Operating Characteristic (ROC) analysis is a method that may help eliminate these limitations, and performance distributions for further analysis can be obtained with ROC charts (Flach, 2019). It is an analysis used to compare the models or to determine the most accurate estimation method when there are more than one, and to set the criterion value for a situation (Boduroğlu, 2017; Faraggi & Reiser, 2002; Hajian-Tilaki et al., 1997; Hajian-Tilaki, 2018; Heagerty et al., 2000; Jones & Rushton, 2019; Kılıç, 2013; Köksal, 2011; Krzanowski & Hand, 2009; Lasko et al., 2005; Pundir & Amala, 2015; Senaratna, Sooriyarachchim & Meyen, 2015; Swets, Dawes &Monahan, 2000). Roc analysis is based on sensitivity and selectivity.

Sensitivity (true positive) means that the test puts the individuals who possess a certain quality in this category; whereas selectivity (true negative) means that the test puts the individuals who do not have a certain quality in this category (Flach et al., 2003; Pepe et al., 2004; Pundir & Amala, 2015). Sensitivity and selectivity give the correct classification rate for the case under review (Karaismailoğlu, 2015). Higher sensitivity and a leftward shift of the ROC curve indicate a better test (Taşdemir & Çokluk, 2013).

The area under the ROC curve determines the suitability of the methods to be compared (Swaving et al., 1996). The area under the curve (AUC) is a brief measure of sensitivity and selectivity (Pardo & Franco-Pereira, 2016) and is the most preferred criterion for ROC analysis (Hanley & McNeil, 1982; Faraggi & Reiser, 2002; Krzanowski & Hand, 2009). AUC provides measurement to include all points in the area to be studied (Carrington et al., 2021). The maximum value for AUC is 1, and the minimum is 0.5 (Hosmer & Lemeshow, 2000; Karaismailoğlu, 2015; Kılıç, 2013; Köksal, 2011). A higher value indicates a more distinctive model (Forthofer et al., 2007).

Hence, many methods, models, or cases can be compared by calculating the area under the ROC curve. Many different methods and models are used in conducting test-equating studies. In addition to overcoming the difficulty of comparing the equated scores obtained from different models and methods, it is important to offer an analysis that will increase the usefulness of the computation used and provide convenience.

Purpose of the Study

In this study, ROC analysis, which allows comparing various models/methods, was used to compare the performance of the IRT scaling methods using NEAT design. It is aimed to compare them using ROC analysis as an alternative option of determining the model/method with the least error for the data set. RMSEs were calculated and compared to test the accuracy of model comparisons.

Method

This study compared IRT scaling models and methods using the ROC analysis. The results were compared with respect to the equating errors.

Data Set

The study group consists of 15-year-old students who participated in PISA 2012 exam. The study group was selected as two countries with different mathematical literacy scores. The PISA 2012 mathematics data was collected from Qatar, a non-OECD country, and Australia, an OECD country. Scaling was conducted for different booklets administered in these countries. In 2012, the math score of Australia was 504, and it was 19th in the PISA mathematics ranking, whereas the score of Qatar was 376 and ranked 63rd. First, the data was cleaned, and outliers were removed; as a result, the data of 617 students from Qatar and 647 from Australia were used to conduct the test equating process. Booklets 5 and 6 with an equal number of questions and common items were used in the analysis. Booklet 6 was used for Australia and Booklet 5 for Qatar. There were 36 questions in both booklets; 12 were common. 2 questions distorting the factor structure were excluded from the analysis. One common item (8th question in booklet 5 and 12th question in booklet 6) and some non-common items (36th question in booklet 5 and 27th question in booklet 6) were removed. Regarding partially scored questions, full scores were coded as 1, partial or incorrect scores were coded as 0. Therefore, partially scored items were evaluated on the binary scale. In the end, both booklets in the data set consisted of 34 items, 11 of which were common.

Data Analysis

The R program was used for all analyses of the process. Firstly, the one-dimensionality of the booklets was checked by exploratory factor analysis. Principal component analysis was used to test whether the booklets were unidimensional or not. As a result of the analysis, one item common in both booklets and two items that disrupt the factor structure were removed. The process was carried out with 34 items. "Psych" (Revelle, 2018) package were used for factor analysis. Booklet 5's KMO value was 0.95, and Bartlett value was 4718.5 (df = 33, p <0.05), whereas KMO value of Booklet 6 was 0.94 and Bartlett value was 1742.2 (df = 33, p <0.05). Accordingly, the sample sizes were sufficient (Çokluk et al., 2012), and the booklets could be factored in with 34 items. The factor loads of Booklet 5 items, filled by Qatar students, varied between 0.405 and 0.828. The single factor structure explained 40% of the total variance of this booklet. Regarding the factor loads of Booklet 6, they varied between 0.305 and 0.885, and the single factor structure explained approximately 51% of the total variance. The results obtained from the booklets indicated that the booklets could be equated, and they exhibit a single factor structure of 34 items.

After checking the factor structure, the booklets' item parameters and ability parameters were estimated using the "Irtoys" (Partchev, 2015) package. Item parameters of each booklet were estimated separately, and then the ability estimates were obtained.

Booklet 5 was taken as the basic test of the equating process, and scaling was carried out in NEAT design. "PLink"(Weeks, 2010) and "LTM" (Rizopoulos, 2018) packages were used for scaling. The scaling was performed with different models and methods, and RMSEs were compared with the results of ROC analysis.

In ROC analysis, true scores were obtained for each equating model and method, then AUCs were compared. The threshold of the ROC curve was set as 494, which is the PISA 2012 math average; individuals with PV1 math scores above 494 were considered successful, below 494 unsuccessful. "RandomForest" (Liaw & Wiener, 2018) and "pROC" (Robin et al., 2011) packages were used to conduct ROC analysis.

Results

RMSEs of binary scored data equating were obtained from one-dimensional IRT models with different parameters (1PL, 2PL, and 3PL) and four scaling methods (Stocking-Lord, Heabara, Mean-Sigma, and Mean-Mean). Errors obtained from the models in Theta transformation are given in Table 1.

Table 1

Equating Errors of the Models

Models	SL	HB	M-S	M-M
3PL	1.826823	1.9380969	1.939114	1.953558
2PL	1.809781	1.932047	1.762249	1.770353
1PL	1.634061	1.7329041	1.575568	1.553225

SL=Stocking-Lord, H=Heabara, M-S=Mean-Sigma, M-M=Mean- Mean

According to Table 1, the lowest amount of error is achieved by the 1PL model. Additionally, the SL scaling method based on the 3PL-IRT model had the least error, and it was followed by the M-S method for 2PL and M-M for 1PL. On the other hand, the scaling method with the highest error was the M-M for the 3PL model and HB for the 1PL and 2PL models. For the M-M method, the error increased as the number of parameters in the model increased.

After the RMSE was obtained, the ROC curves and AUCs were computed using the true scores of the same models. AUCs according to scaling methods and models are given below.

Table 2

AUCs and Confidence Intervals According to Scaling Methods and Parametric Models

		95% confidence intervals	
Variables	AUC	Lower Limit	Upper Limit
M-M			
3PL	.9829*	.9754	.9905
2PL	.9846*	.9773	.9919
1PL	.9876*	.9813	.9939
M-S			
3PL	.9832*	.9756	.9909
2PL	.9849*	.9778	.9920
1PL	.9875*	.9812	.9938
HB			
3PL	.9833*	.9758	.9909
2PL	.9834*	.9760	.9910
1PL	.9872*	.9814	.9939
SL			
3PL	.9848*	.9776	.9920
2PL	.9850*	.9778	.9921
1PL	.9874*	.9805	.9936

*p<.05, SL=Stocking-Lord, H=Heabara, M-S=Mean-Sigma, M-M=Mean-Mean

Regarding the p values of AUCs, the equated scores are significant for each model (p <0.05). This result shows that the scaling methods used in each model play an important role in estimating the true scores. For instance, the results of the M-M equating method according to three different models show that the highest AUC was achieved with the 1PL model; therefore, it makes better estimations than other models.

1PL model is followed by the 2PL, whereas the 3PL model produced more erroneous estimations than the two other models with the same equating method.

For the M-S equating method, the highest AUC is again achieved by the 1PL model. Like the M-M method, the lowest AUC comes from the 3PL model. Therefore, it can be said that the 1PL model can make better estimations with this method, and there may be fewer errors.

Regarding the curves obtained with the Heabara method, which is a method based on item characteristic curves, the model giving the highest AUC is 1PL. Therefore, it can be said that the 1PL model makes the best estimation with this equating method. 3PL model, on the other hand, makes more erroneous estimations.

Again, in the Stocking-Lord scale conversion method, the best estimation was made by the 1PL model and the most erroneous estimation by the 3PL model.

In addition, it is possible to see the method-model combination that gives the minimum error by comparing the AUCs of the scaling methods for each model. Regarding the first column of Table 2, among the methods used in the 3PL model, the Stocking-Lord method gave the highest area (.9848), whereas the lowest area is obtained from the M-M method (.9829). Therefore, it can be said for the 3PL model that the minimum error is obtained with Stocking-Lord and the highest error with the M-M method. Regarding the 2PL model, AUC is the highest for the Stocking-Lord method (.985) and the lowest for the Heabara method (.9834). For the 1PL model, the highest AUC is achieved with the M-M scaling method (.9876), whereas the lowest is achieved with the Heabara method (.9872).

The review of the results obtained from ROC curves, the comparisons of the models and methods, and the equating errors showed that the results supported each other. The M-M method and 1PL model had the lowest error, whereas the M-M method and the 3PL model had the highest error.

Discussion and Conclusion

Equating is a statistical process used to adjust the scores on the tests so that the scores coming from different forms can be used interchangeably (Kolen & Brenna, 2014). The test equating process is important in using the scores obtained from two or more test forms with similar content and difficulty. More than one theory and method are available for the test equating process, making it possible to question which method will give better results. Better estimates can be made by selecting appropriate models or methods that fulfill test equating assumptions.

In selecting the appropriate equating method, a comparison is made between the equating errors. However, the lack of a practical computational process or the failure to directly compare different theories (Wiberg & Gonzalez, 2016) is a challenge for equating processes. ROC analysis, which will overcome these difficulties, allows many models and methods to be compared.

In this study, binary-scored IRT models were used, and different scaling methods were evaluated with ROC analysis. The results were compared with the equating error to see the usability of the obtained results. In this study, the data of Qatar and Australia, which participated in the PISA 2012 math exam and ranked at different levels, were used. The analysis was performed based on the NEAT design for three different models (1PL, 2PL, and 3PL) and four scaling methods (Mean-Mean, Mean-Sigma, Stocking Lord, Heabara). The true scores from two different booklets were equated for all methods, and equating errors (RMSE) were calculated. The true scores were evaluated by calculating the area under the ROC curve and compared with RMSE.

In the light of the obtained equating errors, the 1PL model had the least error for all scaling methods. The M-M method had the best estimation among the methods used with the 1PL model. The moment methods are affected by the difference in item parameter estimations (Kolen & Brenna, 2014); therefore, a low number of parameters allows moment methods to make better estimations. For more parameters, moment methods should be replaced by item characteristic curve and the Stocking Lord methods. Especially in 3PL models, adding the c parameter to the calculations may lead to more accurate estimates for parameters a and b (Thissen & Wainer, 1982). The literature shows that the item characteristic methods give more relevant results than the moment methods (Baker & Al-Karni, 1991; Hanson & Béguin, 2002).

The AUCs obtained for the 1PL model in the ROC analysis are quite close. In addition, the final values support the results obtained with RMSE errors. Likewise, the AUCs of the ROC analysis obtained for the 3PL model support the RMSE results.

For the 2PL model, both RMSE and ROC analysis found that the scaling method with the highest error is the Heabera method. However, in RMSE, the least error is achieved with the M-S method; the ROC analysis, on the other hand, achieved it with the Stocking-Lord scaling method. The Stocking Lord and M-S values obtained by ROC analysis are very close, and it can be deduced that the error margin of ROC analysis is quite low. RMSE may vary according to sample sizes, equating methods (Aşiret &Sünbül, 2016), number of covariates (Gao et al., 2008), and the presence of different ability groups (Hanson & Béguin, 2002; Gialluca et al.,1984). They may give misleading results when IRT assumptions are not adequately fulfilled. From a different perspective, this result can be evaluated as ROC analysis based on nonparametric foundations can also make correct inferences.

It is believed that this research will contribute to the literature showing that different criteria can be used in the test equating process. These results show that ROC analysis can be used as an alternative to RMSE, which is frequently used to compare test equation methods. Although ROC analysis does not give a precise result, it can give an idea when comparing different situations, methods, and models. Especially in the comparison of the equating methods from different theories, ROC analysis is assumed to show the difference between methods more clearly. Future studies may make this comparison based on the equating design created according to equivalent ability levels, or a simulation can be run for different ability levels. In addition, comparisons should be made by repeating the study for values that can create criteria for ROC analysis.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Secondary data were used in this study. Therefore, ethical approval is not required.

References

Aşiret, S., & Sünbül, S. Ö. (2016). Investigating test equating methods in small samples through various factors. *Educational Sciences: Theory & Practice*, *16*(2), 647-668. https://doi.org/10.12738/estp.2016.2.2762

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147-162. https://www.jstor.org/stable/1434796

Boduroğlu, E. (2017). *The study of classification consistency of transition to higher education examination according to the cut-off scores obtained from different* [Master's Thesis, Mersin University]. https://tez.yok.gov.tr/UlusalTezMerkezi/

Branberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement*, *48*(4), 419-440. https://www.jstor.org/stable/41427533

Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., ... & Holzinger, A. (2021). *Deep ROC analysis and AUC as balanced average accuracy to improve model selection, understanding and interpretation.* IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2022.3145392

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve Lisrel uygulamaları*. Pegem Akademi.

Embretson, S. (1996) The new rules of measurement. *Psychological Assessment*, *8*(4), 341-349. https://doi.org/10.1037/1040-3590.8.4.341

Embretson, S. (1997). Multicomponent response models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern Item Response Theory* (pp. 305-321). Springer-Verlag.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists.* Lawrence Elbaum Associates.

Faraggi, D., & Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in Medicine, 21*, 3093-3106. https://doi.org/10.1002/sim.1228

Flach, P., Blockeel, H., Ferri, C., Hernandez-Orallo, J., & Struyf, J. (2003). Decision support for data mining: An introduction to ROC analysis and its applications. In D. Mladenić, N. Lavrač, M. Bohanec & S. Moyle (Eds.), *Data mining and decision support: Integration and collaboration* (vol. 745, pp. 81-90). Springer. https://doi.org/10.1007/978-1-4615-0286-9_7

Flach, P. (2019). Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. *Proceedings of the AAAI Conference on Artificial Intelligence, 33,* 9808–9814. https://doi.org/10.1609/aaai.v33i01.33019808

Gao, X., Zhu, R., Chen, H., & Harris, D.J. (2008, March 25-27). *Impact of anchor-item selections on IRT scale transformation and equating* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, New York.

Gialluca, K. A., Crichton, L. I., Vale, C. D., & Ree, M. J. (1984). *Methods for equating mental tests* (Report No. ED251512). ERIC. https://files.eric.ed.gov/fulltext/ED251512.pdf

Gonzalez, J. (2014). SNSequate: Standard and nonstandard statisticalmodels and methods for test equating. *Journal of Statistical Software, 59*(7), 1-30. https://doi.org/10.18637/jss.v059.i07

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*(3), 144-149. https://doi.org/10.4992/psycholres1954.22.144

Hajian-Tilaki, K. O., Hanley, J. A., Joseph, L., & Collet, J. P. (1997). A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Medical Decision Making, 17*(1), 94-102. https://doi.org/10.1177/0272989X9701700111

Hajian-Tilaki, K. (2018). Receiver operator characteristic analysis of biomarkers evaluation in diagnostic research. *Journal of Clinical and Diagnostic Research*, *12*(6), 1-8. https://doi.org/10.7860/JCDR/2018/32856.11609

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Kuluwer-Nijhoff Publisihing.

Hanley J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3-24. https://doi.org/10.1177/0146621602026001001

Heagerty, P., Lumley, T., & Pepe, M. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, *56*, 337-344. https://doi.org/10.1111/j.0006-341x.2000.00337.x

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). John Wiley & Sons

Jones, L., & Rushton, D. (2019, September 1-6). *Optimising geotechnical correlations using receiver operating characteristic (ROC) analysis*. The XVII European Conference on Soil Mechanics and Geotechnical Engineering (ECSMGE 2019), Reykjavik, Iceland.

Karaismailoğlu, E. (2015). *The use of time dependent roc curve for evaluation of the performance of markers during follow-up time* (Combined Doctoral Dissertation, Hacettepe University). https://tez.yok.gov.tr/UlusalTezMerkezi/

Kılıç, S. (2013). Klinik karar vermede ROC analizi. *Journal of Mood Disorders*, *3*(3), 135-140. https://doi.org/10.5455/jmood.20130830051624

Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practice*, *7*, 29-36.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3nd ed.). Springer.

Köksal, B. (2011). *Model selection with ROC curve estimation in regression analysis* [Master's thesis, Marmara University]. https://tez.yok.gov.tr/UlusalTezMerkezi/

Lasko, T. A., Bhagwat, J. G., Zou, K. H., & Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics, 38*(5), 404-415. https://doi.org/10.1016/j.jbi.2005.02.008

Liaw, A., & Wiener, M. (2018). *randomForest: Breiman and Cutler's Random Forests for classification and regression*. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Livingston, S. A., & Lewis, C. (2009). *Small-sample equating with prior information* (Report No. RR-09-25). ETS. https://files.eric.ed.gov/fulltext/ED507811.pdf

Pardo, M. C., & Franco-Pereira, A.M. (2017). Non parametric ROC summary statistics. *REVSTAT-Statistical Journal, 15*(4), 583-600. https://eprints.ucm.es/id/eprint/46564/1/PardoCarmen29.pdf

Pepe, M., Janes, H., Longton, G., Leisenring, W., & Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, *159*, 882-890. https://doi.org/10.1093/aje/kwh101

Pundir, S., & Amala, R. (2015). Evaluation of biomarker using two parameter bi-exponential ROC curve. *Pakistan Journal of Statistics and Operation Research, 11*(4), 481-496. https://doi.org/10.18187/pjsor.v11i4.992

Revelle, W. (2018). *psych: Procedures for personality and psychological research*. http://kambing.ui.ac.id/cran/web/packages/psych/psych.pdf

Rizopoulos, D. (2018). *ltm: Latent Trait Models under IRT*. https://cran.r-project.org/web/packages/ltm/ltm.pdf

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisace k F, Sanchez J, Müller M (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 77. https://doi.org/10.1186/1471-2105-12-77

Senaratna, D. M., Sooriyarachchim, M. R., & Meyen, N. (2015). Bivariate test for testing the EQUALITY of the average areas under correlated receiver operating characteristic curves (Test for comparing of AUC's of correlated ROC curves). *American Journal of Applied Mathematics and Statistics*, *3*(5), 190-198. https://doi.org/10.12691/ajams-3-5-3

Swaving, M., van Houwelingen, H., Ottes, F. P., & Steerneman, T. (1996). Statistical comparison of ROC curves from multiple readers. *Medical Decision Making*, *16*(2), 143-152. https://doi.org/10.1177/0272989X9601600206

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better decisions through science. *Scientific American*, *283,* 82-87. https://doi.org/10.1038/scientificamerican1000-82

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201-210. https://doi.org/10.1177 / 014662168300700208

Taşdemir, F., & Çokluk, Ö. (2013). Angoff (1-0), Nedelsky and examination of classification accuracies of a test by determination methods of limit values. *Mediterranean Journal of Humanities*, *3*(2), 241-261. https://doi.org/10.13114/mjh/201322482

Tian, F. (2011). *A comparison of equating / linking using the Stocking-Lord method and concurrent calibration with mixed-format tests in the non-equivalent groups common-item design under IRT* [Unpublished doctoral dissertation]. Boston College.

Thissen, D. M., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 397-412. https://doi.org/10.1007/BF02293705

Wang, T. (2006). *Standard errors of equating for equipercentile equating with log-linear pre-smoothing using the delta method* (Report No. 14). Center for Advanced Studies in Measurement and Assessment, Iowa.

Weeks, J. P. (2010). Plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software, 35*(12), 1-33. https://cran.r-project.org/web/packages/plink/vignettes/plink-UD.pdf

Wiberg. M., & Branberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement*, *39*(5), 349–361. https://doi.org/10.1177/0146621614567939

Wiberg, M., & Gonzalez, J (2016). Statistical assessment of estimated transformations in observed-score equating. *Journal of Educational Measurement.* *53*(1), 106–125. http://www.mat.uc.cl/~jorge.gonzalez/papers/TR/Assess_TR.pdf