ISSN No:-2456-2165

Item Response Theory of Test Equating for Anchor Items

[1]Rishi Kumar Loganathan ; [2]Dr. Zahari Suppian ; [3]Dr. Siti Eshah Mokhsein
Sultan Idris Education University, Perak, Malaysia

Abstract:- The item response theory is based on the latent response theory that presents a mathematical relationship between the latent traits and manifestations. It helps in measuring different constructs with the help of analysis, scoring tests, questionnaires, and similar evaluative instruments. The item response theory includes different approaches such as equating process, a multistage approach based, and others to perform tests while considering a common scale as a measure of standard. The main aim of the research is to study the item response theory of test equating for anchor items with the help of empirical selected anchor items based on differential item functioning (DIF). It was found that item response theory plays a vital role in the identification and empirical selection of anchor items. The multistage differential item functioning strategy based on the item response theory is an effective and suitable identifying technique for the anchor items. It was also found that anchor contamination rates decrease as the sample size of the anchor items is increased.

Keywords:- *Item Response Theory, Anchor Class, Anchor Items, Anchor Selection Strategy, Two-Parameter Model (2PL).*

I. INTRODUCTION

The item response theory is also stated as the latent response theory that presents mathematical relationship between the latent traits and manifestations. The latent traits are the unobservable or attributes while the manifestations are the observed responses, outcomes or performances. This theory forms a link between the properties of item on an instrument and the underlying traits that are subjected to measurement. The item response theory suggests that the latent construct and the items of a measure are organized with each other in such a way that they formulate an unobservable continuum[1]. The item response theory of test equating is the based on the relationship between the performance presented by the individuals and performance level of test takers.

The equating process in the item response theory is the process of placing two or more scores that are taken from the two or more parallel performed tests while considering a common scale as a measure of standard[2]. This results in the effective comparison of the items selected for the testing. In this research the item response theory is applied to the anchor items and their empirical selection. This multistage approach based on the item response theory of test equating for the anchor items selection presents the more accurate differential item functioning that conventional approaches of anchor items selection. The outcomes of this research are true positive rates, false positive rates and familywise false positive rates. The results present that if multistage approach based on the item response theory is adopted for anchor items then there are lower anchor contamination rates as compared to the non-multi stage approach used for the anchor items[3].

II. THEORETICAL BACKGROUND

2.1 Differential Item Functioning based Item Response Theory

The differential functioning based on the item response theory occurs the anchor items selected for a test performs differently for a subgroup of test takers[4]. In this differential item functioning one equals the correct response and zero equals the zero equals the incorrect response. The differential item functioning for scored anchor items is presented in the equation below:

$$P(Y = 1 \mid \theta, G = R) \neq P(Y = 1 \mid \theta, G = F)$$

In the equation above the constant 1 corresponds to the true response in the testing mechanism and 0 equals the incorrect response. Also,
$P(Y = 1)$ states the probability of a correct response
The construct ability is represented by θ
The group membership for the test is G
The reference group is represented by R
The focal group is represented by F

The differential item functioning based on the item response theory can be either uniform or non-uniform. The infirmity and non-uniformity of the differential item functioning depends on the difference identified in the difficulty of anchor items selection and the discrimination of the anchor items. The difficulty of the anchor item selection depends on the challenging the anchor items n such a way that it provides the correct response. The discrimination of the anchor items is stated as the way in which anchor items are separated by test takers from the varying testing abilities from low ability to high ability of the test takers[5].

The anchor items stating the differential item functioning the signed area (SAR) between the reference curve and focal curve is presented in the equation presented below:

$$SAR = (1-c)(b_F - b_R)$$

In the equation above pseudo guessing parameter represented by c', the difficulty parameter for the reference group is b_R and the difficulty parameter for focal group is b_F. From the signed area, the average signed area is presented in the equation below:

$$ASAR = \sum_{i=1}^{n} \frac{SAR_i}{n}$$

The SAR_i is the signed area of the i-th anchor item selected while, i is the number of anchor item selected for the test.

The item response theory uses the latent variable to estimate the construct ability as compared to other theories that uses the observed data. There are three important parameters of the item response theory that are stated by the equation presented below:

$$p(Y=1)|\,\theta) = c_i + (1-c_i)\frac{\exp[a_i(\theta_i - b_i]}{1+\exp[\,a_i(\theta_i - b_i]}$$

In the equation above $p(Y = 1)|\,\theta)$the correctness of a response on the given anchor item$i$, give θ, the latent ability is represented by θ, the pseudo guessing parameter is represented by c_i, the item discrimination is represented by the a_i and item difficulty is represented by the b_i.

In the one parameter (1PL) model of the item response theory, the item discrimination a_i is held constant, pseudo guessing parameter is represented c_i is set to zero and only item difficulty represented by b_i is evaluated[6]. In the two-parameter model (2PL) of the item response theory pseudo guessing parameter is represented $c_i$is set to zero. The item discrimination a_i and item difficulty b_i are evaluated.

There are numerous methods to test for differential item functioning using the item response theory i.e., Lord's chi square test, likelihood ratio test, differential item functioning etc. [7]. The item response theory-based likelihood ratio is the effective technique for detecting the differential functioning. It examines the difference in the model fit between the constrained and less constrained model equation as presented below.

$$G^2 = 2\ln\left(\frac{L_c}{L_f}\right)$$

In the equation above G^2 represents the test statistics approximation with chi square approximation, the likelihood function is represented by L, the constrained model is represented by c and the less constrained model is presented by f. The Wald test for the differential item functioning using the item response theory is presented below[8]:

$$W_i = \frac{b_{i\,G=R} - b_{i\,G=F}}{\sqrt{var(b_{i\,G=R}) + (b_{i\,G=F})}}$$

In the above equation, W_i represented Wald's test, $b_{i\,G}$ represents the item difficulty for item i and group G. The var $b_{i\,G}$ is the variance of the item difficulty for the item $i$and group G.

2.2 Anchor Items

In this research the anchor item as stated as the items whose parameters are constrained to the equal groups. The non-anchor items are subjected to the variation of groups[9]. The functioning of the anchor items is presented in the equation presented below. This anchor model is the 1 parameter (1PL) of the item response theory

$$p(Y_{i\,G} = 1|\theta_G) = \frac{\exp[(\theta_i - b_{i\,G})]}{1+\exp[(\theta_i - b_{i\,G})]}$$

In the equation above $p(Y_{i\,G} = 1|\theta_G)$ represents the correctness of response on the anchor item i given θ for the group G. The latent ability of the anchor items is θ and G represents the group G. In the case of anchor items, the b parameter estimate for the reference group and the focal group is equal i.e.

$$b_{i\,G=F} = b_{i\,G=R}$$

However, in the case of the non-anchor estimates b parameter estimates for the reference and focal group are not same i.e.

$$b_{i\,G=F} \neq b_{i\,G=R}$$

2.3 Anchor Class

The anchor class states the length of the anchor as well as the overall approach used when identifying and selecting the anchor class[10]. The anchor class is commonly classified as:

- Constant anchors
- Iterative scale anchors

The constant anchor methods are utilized for determining the anchors from predetermined numbers of anchors. These methods of anchors are capable to identify the anchors as 1 anchor while leaving yielding a probability of selection of almost 20%. However, in the case of iterative scale anchors, there is not fixed limit of anchors, and it works on the iterative mechanism for proper selection. The iterative algorithm converges to a specific number yielding a stopping criterion for the selection of anchor items[11].The Wang method presented above works best for the iterative selection or identification of the anchor selection. The linear anchor items diverge after some iterations level but non-linear anchor does not converge to a specific number of anchor selection.

2.4 Anchor selection strategy

The anchor selection strategy refers to the mechanism though which anchor are chosen from a particular class of anchors. In literature there are different strategies proposed for the selection of anchors from a particular class[12].These strategies rank the potential anchor based on the differential item functioning index and statistically differential functioning tests. This can also be stated as that these there are certain algorithms mechanism that add and remove items to formulate the anchor class. This then presents as the type I and type II class anchors with the all-other anchors included in the type I anchors and single anchors included in the type II class of anchors. These classes are also stated as the All Items (AI) and single items (SA) anchors. These two approaches are then widely used for the identification of anchors as the All-items anchors and the single item anchor[13].

III. RESEARCH METHOD

The research is focused for the empirical selection of anchor items on the basis of differential item functioning (DIF) using the item response theory**.** The comparison analysis of different sets of the anchor items is presented on the basis of differential item functioning[10]. The study design for the selection of anchor items is made on the basis item response theory and required changes as per the requirement for the categorization of the anchor items. In this study design three hypotheses are prosed the justify for the identification of anchor items. First is that multistage anchors will have true positive rates. The second hypothesis for the anchor items selection is that the multistage anchors will have high false positive rates. The third hypothesis is that the family of multistage anchor will possess the true positive rates greater than 0.5[14]. The mean p-value threshold (MPT) and mean test statistic threshold (MTT) analysis are performed on all the classes of the anchor classes to proper identification and empirical selection of the anchor items[15]. In addition to this the DIF free anchors are also identified on the basis of these tests and some manipulated variables are suggested for the accurate classification and analysis of the anchor items.

3.1 Determining MPT and MTT

The mean p-value of anchor items is measured by the first of all testing these items for differential item functioning, using single anchor so that we can determine k-1 differential item functions[16]. The 'k' is the total numbers of items in the considered for the test. After that the mean p-values are ranked from maximum to minimum and then these man values are identified. Let, we have 20 mean p-values of the anchor items taken for a specific test, these values are first ranked and after arranging from high to low, the p-values of the anchor items are classified. In this research the 10th value of the ordered mean values was selected as mean test statistic threshold (MPT)of the anchor items. The maximum test statistic threshold (MPT) is selected in a similar fashion as that of MPT is selected from regular maximum to minimum arrangement. The total identified anchor items are ($[0.5 - k]$)test statistics[17].

3.2 DIF-free Anchors

These are two specific classes of anchors that are used to mirror the anchor length of C4 and IF class anchors. These DIF free anchors are utilized for the identification of the differential item functioning of the anchor class. The differential item functioning C4 anchor excludes the all percentages of DIF[18]. The other anchor is the IF-DIF free anchor and the anchor length variation of the IF-DIF anchor with the variation of DIF is presented in the table given below:

Percentage of DIF	Anchor Length
0	10
10	9
20	8
40	6

The chosen of IF-DIF anchor class is random while the choosing mechanism for theC4 free differential item functioning is uniform and selected on the basis of the p-vales of the anchor class of anchor class.

3.3 Manipulated Variables

The manipulated variables in this research are the sample size, percentage of DIF, Balance of DIF, number of replications and outcomes from the test performed from the identification and selection of the anchor items. The sample size of the comprises of the reference as well as focal groups[19].If large sample size of the anchor class is taken then it results in the large true positive rates and lower false positive rates. However, there are different tests present in literature through high rates of true positive rates can be obtained with small samples of anchor items. The percentage of differential functioning (DIF) taken in this research is 0%, 10%, 20% and 40%. In the case of balanced DIF, half of the anchor items favored the reference group and half of the anchor items favored the focal group[20]. If 'n' number of anchor items are standard error of the selection of anchor items is presented below:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

In the equation above, p is the proportion of a certain outcome and the total number of replications. The outcomes are the hypothesis test results performed n the identification and selection of the anchor items.

IV. RESULTS

The anchor items are classified on the basis of the differential item functioning using the item response theory. The result outcomes on the anchor items are the true positive rates of anchor items, false positive rates of anchor items, familywise false positive rates of anchor items, the anchor contamination rates, the familywise anchor contamination rates and the observed anchor lengths for IF selection method of the anchor items.

4.1 True Positive Rates

The true positive rates for the anchor items are higher in the case of large sample size of anchor class taken for testing. These high positive rates in observed in all types of selection strategies of anchor items i.e. ,IF anchor selection methods versus C4 anchor selection methods and the balanced DIF conditions versus the unbalanced DIF conditions. There is no difference observed in the true positive rates in the case of multistage, non-multistage and DIF free methods of selection on anchor items. The only difference observed is 14% in the case of empirical selection of anchor items.

		Balanced DIF			One-Sided DIF		
		Sample Size Per Group					
Percentage of DIF	Anchor Method	500	700	1000	500	700	1000
10	C4-DIF-free	0.5	0.69	0.8	0.4	0.7	0.78
	C4-SA(MPT)	0.5	0.7	0.83	0.4	0.7	0.8
	MS[C4-SA(MPT)]	0.5	0.68	0.82	0.4	0.7	0.8
	IF-DIF-free	0.5	0.72	0.84	0.4	0.7	0.82
	IF-SA(MTT)	0.5	0.74	0.84	0.4	0.7	0.82
	MS[IF-SA(MTT)]	0.5	0.72	0.84	0.4	0.7	0.82
20	C4-DIF-free	0.49	0.66	0.81	0.47	0.67	0.78
	C4-SA(MPT)	0.49	0.68	0.83	0.43	0.65	0.79
	MS[C4-SA(MPT)]	0.48	0.68	0.82	0.43	0.63	0.79
	IF-DIF-free	0.5	0.69	0.85	0.49	0.69	0.82
	IF-SA(MTT)	0.52	0.7	0.85	0.45	0.66	0.8
	MS[IF-SA(MTT)]	0.51	0.7	0.84	0.46	0.66	0.81

4.2 False Positive Rates

The false positive rates under all the controlled test conditions are calculated to be 0.1 and 0.2. In the case of 40% differential item functioning, the false positive and false negative rates are found to be almost same. The false positive rates are calculated for both balanced DIF as well as unbalanced DIF. It is found that these rates significantly identify anchor items from the anchor methods.

		Balanced DIF			One-Sided DIF		
		Sample Size Per Group					
Percentage of DIF	Anchor Method	500	700	1000	500	700	1000
10	C4-DIF-free	0.02	0.02	0.02	...	...	...
	C4-SA(MPT)	0.02	0.02	0.02	...	...	...
	MS[C4-SA(MPT)]	0.02	0.02	0.01	...	...	...
	IF-DIF-free	0.01	0.01	0.01	...	...	...
	IF-SA(MTT)	0.02	0.02	0.02	...	...	...
	MS[IF-SA(MTT)]	0.02	0.02	0.02	...	...	...
20	C4-DIF-free	0.02	0.02	0.02	0.02	0.02	0.02
	C4-SA(MPT)	0.02	0.02	0.02	0.02	0.02	0.02
	MS[C4-SA(MPT)]	0.02	0.02	0.02	0.02	0.02	0.02
	IF-DIF-free	0.01	0.01	0.01	0.01	0.01	0.01
	IF-SA(MTT)	0.02	0.02	0.02	0.02	0.02	0.02
	MS[IF-SA(MTT)]	0.02	0.02	0.02	0.02	0.02	0.02

4.3 Familywise False Positive Rates

In the identification and selection of anchor items using the differential item functioning based item response theory, it is found that the familywise false positive rates are lower as the percentage of differential item functioning (DIF) increases[21]. However, in the mechanism of identification of anchor items the familywise false positive rates are found to be same in the case of multistage DIF as well as non-multistage DIF.

		Balanced DIF			One-Sided DIF		
		Sample Size Per Group					
Percentage of DIF	Anchor Method	500	700	1000	500	700	1000
10	C4-SA(MPT)	0.016	0.007	0.003	0.019	0.006	0.001
	MS[C4-SA(MPT)]	0.016	0.008	0.003	0.017	0.008	0.001
	IF-SA(MTT)	0.016	0.005	0.002	0.017	0.006	0.002
	MS[IF-SA(MTT)]	0.016	0.006	0.002	0.017	0.006	0.002
20	C4-SA(MPT)	0.031	0.016	0.003	0.041	0.016	0.006
	MS[C4-SA(MPT)]	0.031	0.014	0.002	0.042	0.018	0.006
	IF-SA(MTT)	0.03	0.01	0.005	0.043	0.014	0.009
	MS[IF-SA(MTT)]	0.032	0.012	0.004	0.04	0.018	0.006

4.4 Anchor Contamination Rates

The anchor contamination rates decrease as the sample size of the anchor items is increased. The weight of contamination rates was under one-sides 40% of the differential item functioning. The multistage DIF is recommenced for the anchor contamination rates determination with 40% DIF.

4.5 Familywise Anchor Contamination Rates

The familywise anchor contamination rates also decrease with the increase in sample size. These rates are higher in the case of one-sided differential item functioning as compared to the balanced differential item functioning[22]. Also, the percentage of the familywise anchor contamination rates increases as the sample size is increased.

		Balanced DIF			One-Sided DIF		
		Sample Size Per Group					
Percentage of DIF	Anchor Method	500	700	1000	500	700	1000
10	C4-SA(MPT)	0.063	0.028	0.01	0.075	0.025	0.005
	MS[C4-SA(MPT)]	0.063	0.033	0.01	0.065	0.03	0.005
	IF-SA(MTT)	0.15	0.045	0.018	0.16	0.06	0.018
	MS[IF-SA(MTT)]	0.158	0.055	0.02	0.158	0.055	0.015
20	C4-SA(MPT)	0.12	0.063	0.01	0.158	0.063	0.02
	MS[C4-SA(MPT)]	0.123	0.053	0.008	0.165	0.073	0.025
	IF-SA(MTT)	0.258	0.093	0.048	0.358	0.123	0.075
	MS[IF-SA(MTT)]	0.263	0.103	0.033	0.323	0.153	0.053

V. LIMITATIONS

The research proposed is limited to the conditions of simulations. The results presented above are subjected to change with the change in the sample size and population size of the test sample.

VI. CONCLUSION

It is concluded that the item response theory plays a vital role for the identification and empirical selection of anchor items. The multistage differential item functioning strategy based on the item response theory is effective and suitable identifying technique for the anchor items. The results of this proposed strategy are subjected to the hypothesis true positive rates, false positive rates and familywise false positive rates and the item response theory is the most effective technique for the empirical selection of anchor items.

REFERENCES

[1]. DeMars, C.E., Alignment as an alternative to anchor purification in DIF analyses. Structural Equation Modeling: A Multidisciplinary Journal, 2020. **27**(1): p. 56-72.

[2]. Belzak, W. and D.J. Bauer, Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. Psychological Methods, 2020.

[3]. Himelfarb, I., A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. Journal of Chiropractic Education, 2019. **33**(2): p. 151-163.

[4]. Vermunt, J.K. and J. Magidson, How to perform three-step latent class analysis in the presence of measurement non-invariance or differential item functioning. Structural Equation Modeling: A Multidisciplinary Journal, 2020: p. 1-9.

[5]. Bauer, D.J., W.C. Belzak, and V.T. Cole, Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. Structural Equation Modeling: A Multidisciplinary Journal, 2020. **27**(1): p. 43-55.

[6]. Fernandes, S.M. and A.C. Bornia, Reporting on supply chain sustainability: Measurement using item response theory. Corporate Social Responsibility and Environmental Management, 2019. **26**(1): p. 106-116.

[7]. Knoll, M.A. and C.R. Houts, The financial knowledge scale: An application of item response theory to the assessment of financial literacy. Journal of Consumer Affairs, 2012. **46**(3): p. 381-410.

[8]. Kolbe, L. and T.D. Jorgensen, Using restricted factor analysis to select anchor items and detect differential item functioning. Behavior Research Methods, 2019. **51**(1): p. 138-151.

[9]. Iscen, A., et al. Mining on manifolds: Metric learning without labels. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[10]. Gübes, N. and S. Uyar, Comparing Performance of Different Equating Methods in Presence and Absence of DIF Items in Anchor Test. International Journal of Progressive Education, 2020. **16**(3): p. 111-122.

[11]. Robitzsch, A. and O. Lüdtke, A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. Psychological Test and Assessment Modeling, 2020. **62**(2): p. 233-279.

[12]. Lorenzi, D.M., Selecting Anchor Items in Differential Item Functioning: A Case Study. 2020, Fordham University.

[13]. Huelmann, T., R. Debelak, and C. Strobl, A Comparison of Aggregation Rules for Selecting Anchor Items in Multigroup DIF Analysis. Journal of Educational Measurement, 2020. **57**(2): p. 185-215.

[14]. O'Neill, T.R., J.L. Gregg, and M.R. Peabody, Effect of sample size on common item equating using the dichotomous rasch model. Applied Measurement in Education, 2020. **33**(1): p. 10-23.

[15]. 15. Casper, W., et al., Selecting response anchors with equal intervals for summated rating scales. Journal of Applied Psychology, 2020. **105**(4): p. 390.

[16]. Kopf, J., A. Zeileis, and C. Strobl, Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. Educational and psychological measurement, 2015. **75**(1): p. 22-56.

[17]. Furter, R.T. and A.C. Dwyer, Investigating the classification accuracy of Rasch and nominal weights mean equating with very small samples. Applied Measurement in Education, 2020. **33**(1): p. 44-53.

[18]. Pohl, S. and D. Schulze, Assessing group comparisons or change over time under measurement non-invariance: The cluster approach for nonuniform DIF. Psychological Test and Assessment Modeling, 2020. **62**(2): p. 281-303.

[19]. Lu, R. and H. Guo, A Simulation Study to Compare Nonequivalent Groups With Anchor Test Equating and Pseudo-Equivalent Group Linking. ETS Research Report Series, 2018. **2018**(1): p. 1-16.

[20]. Chun, S., et al., MIMIC methods for detecting DIF among multiple groups: Exploring a new sequential-free baseline procedure. Applied psychological measurement, 2016. **40**(7): p. 486-499.

[21]. Thissen, D. Similar DIFs: Differential item functioning and factorial invariance for scales with seven (“plus or minus two”) response alternatives. in The Annual Meeting of the Psychometric Society. 2016. Springer.

[22]. Ye, M., X. Lan, and P.C. Yuen. Robust anchor embedding for unsupervised video person re-identification in the wild. in Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[23]. Craig, B., The empirical selection of anchor items using a multistage approach. 2017.