ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(2); 118-127

The Relation of Item Difficulty Between Classical Test Theory and Item Response Theory: Computerized Adaptive Test Perspective

Eren Can AYBEK*

Abstract

This study aims to transform the calculated item difficulty statistics according to Classical Test Theory (CTT) into the item difficulty parameter of Item Response Theory (IRT) by utilizing the normal distribution curve and to analyze the effectiveness of this transformation based on Rasch model. In this regard, 36 different data sets created with catR package were studied. For each data set, item difficulty parameters and transformed item difficulty parameters were calculated and the correlation coefficients between these parameters were analyzed. Then, Computerized Adaptive Test (CAT) simulations were performed using these parameters. According to the simulation results, the correlation coefficients between the estimated theta values with both methods were high. Furthermore, in CAT simulations in which both parameters were used, especially in the samples which were over 250, it was found to have similar bias, RMSE values, and the average number of administered items.

Keywords: Item difficulty, classical test theory, item response theory, Rasch model

Introduction

A measurement tool can be developed based on Classical Test Theory (CTT) or Item Response Theory (IRT) (de Ayala, 2009). Tests are easy to develop under the CTT, yet it has some limitations. For example, a single standard error value for the entire test score can be calculated by using CTT; the item statistics depend on the examinees, and the true score estimates are based on the item set (Hambleton & Swaminathan, 1985). The studies show that an item that should be removed from the test according to CTT should also be taken out of the test according to IRT, which reveals the fact that CTT and IRT estimates are similar when deciding whether an item is good or bad (Çelen & Aybek, 2013). On the other hand, IRT comes to the fore for studies such as Computerized Adaptive Test (CAT), test equation and linking, and Differential Item Functioning (DIF), but loses its practicality for classroom assessment.

IRT models can be classified in different ways according to the dimension that is measured and the number of response categories. In addition to unidimensional IRT models in which an item measures one single dimension, there are also multidimensional IRT models in which an item can measure multiple dimensions (Reckase, 2009). In addition, there are some models such as Rasch, 1 Parameter Logistic (1PL), 2PL, 3PL, and 4PL models for dichotomous items (Hambleton et al., 1991); Nominal Response Model (NRM) (Bock, 1972); Partial Credit Model (PCM) (Masters, 1982); Generalized Partial Credit Model (GPCM) (Muraki, 1992); and Graded Response Model (GRM) for polytomous items (Samejima, 1996).

In the Rasch model, the probability of responding to an item correctly depends only on the item difficulty, *b*, parameter of that item, while the item discrimination, *a*, parameter is considered to be 1.00 for all the items. The Rasch and 1PL models are similar in that item discrimination is considered the

* Assoc. Prof., Pamukkale University, Faculty of Education, Denizli-Türkiye, erencan@aybek.net, ORCID ID: 0000-0003-3040-2337

To cite this article:
Aybek, E. C. (2023). The relation of item difficulty between Classical Test Theory and Item Response Theory: Computerized adaptive test perspective. *Journal of Measurement and Evaluation in Education and Psychology, 14*(2), 118-127. https://doi.org/10.21031/epod.1209284

Received: 23.11.2022
Accepted: 28.06.2023

same for all items; however, *a* parameter can take different values than 1.00 in 1PL model (de Ayala, 2009).

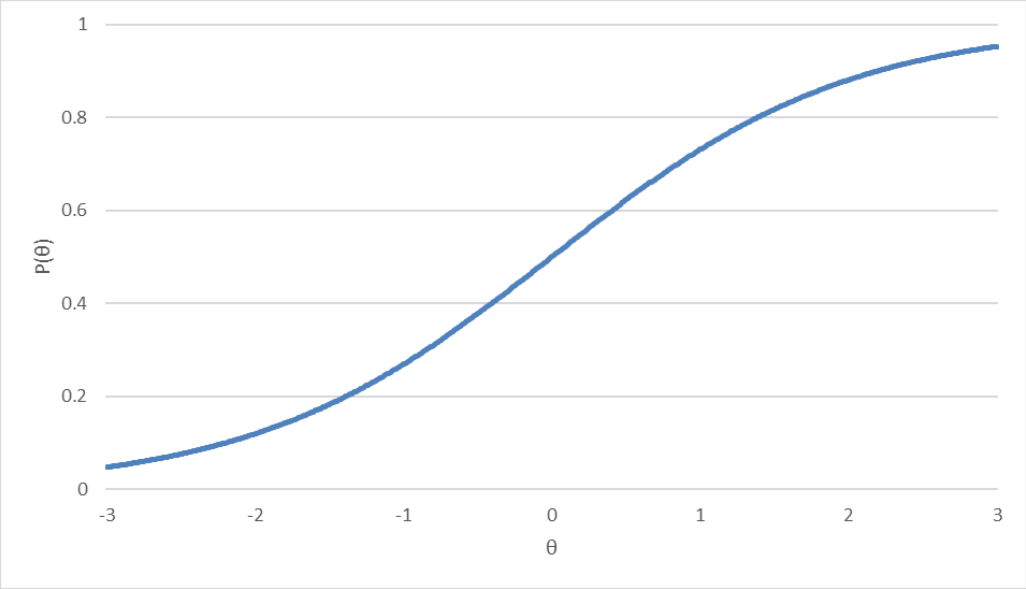
According to the Rasch model, the probability for an individual with a given (θ) ability level to respond correctly to an item whose difficulty parameter is *b* is calculated with the equation below (1) (Rasch, 1961):

$$p(\theta) = \frac{e^{(\theta-b)}}{1+e^{(\theta-b)}} \qquad (1)$$

The *b* parameter represents item difficulty and refers to the θ level at which the item is correctly answered with 50% probability. Although the theoretical limits for θ are between (-∞, ∞), they usually work within ranges such as [-3, 3] or [-4, 4]. When Equation 1 is applied, the probability of responding to an item correctly with $b = 0.00$ for all θ levels within the range [-3, 3] with an increment of 0.01 creates a curve as shown in Figure 1, and this curve is called the item characteristic curve.

When Equation 1 and Figure 1 are analyzed, another superiority of IRT can be recognized. Item parameters and the examinee's ability level are described on the same scale. As stated earlier, the *b* parameter for the item shown in Figure 1 is 0.00, which means that an examinee whose θ level is 0.00 responds to this item correctly with a probability of %50. In addition, when Figure 1 is carefully analyzed, it can be recognized that, as the θ level decreases, the probability of responding to an item correctly also decreases, and as it increases, the probability of responding to the item correctly increases, as well. In this context, the *b* parameter has similar limits as θ.

Figure 1.
A sample item characteristic curve for b=0 parameter in Rasch model



On the other hand, normal distribution is described as "*a theoretical distribution for a continuous variable measured for an infinite population"* (Crocker & Algina, 1986, p.24) and defined using the Equation 2 (Pitman, 1993):

$$Y = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}z^2} \qquad (2)$$

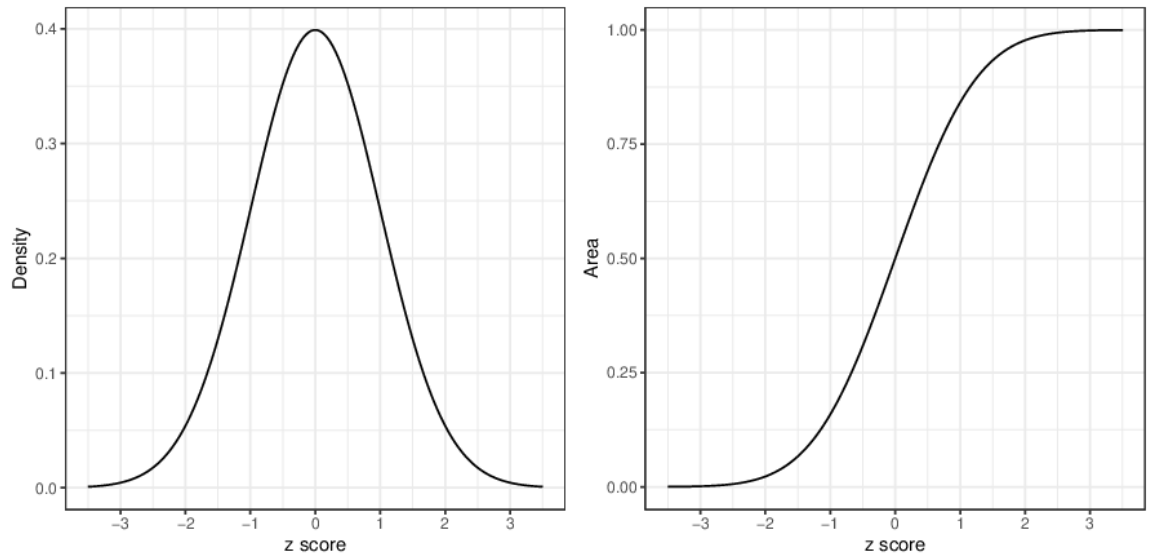
In this Equation, z stands for the standard z score and is obtained as $z = (x - \mu)/\sigma$ In addition, the area under the normal distribution curve can be obtained approximately by Equation 3 (Pitman, 1993):

$$\phi(z) \approx 1 - \frac{1}{2}(1 + c_1 z + c_2 z^2 + c_3 z^3 + c_4 z^4)^{-4} \quad (z \geq 0) \qquad (3)$$

c values in this equation as follows: c_1 = 0.196854, c_2 = 0.115194, c_3 = 0.000344, and c_4 = 0.019527. When z is below 0, $\phi(-z) = 1 - \phi(z)$ relation can be used by utilizing the symmetric characteristic of normal distribution curve. Accordingly, when the Equation 3 is used, the area under normal distribution curve for z = 1 constitutes approximately %84,3 of the whole area. Based on all this information, a normal distribution curve and the area under the normal distribution curve are given in Figure 2.

Figure 2.

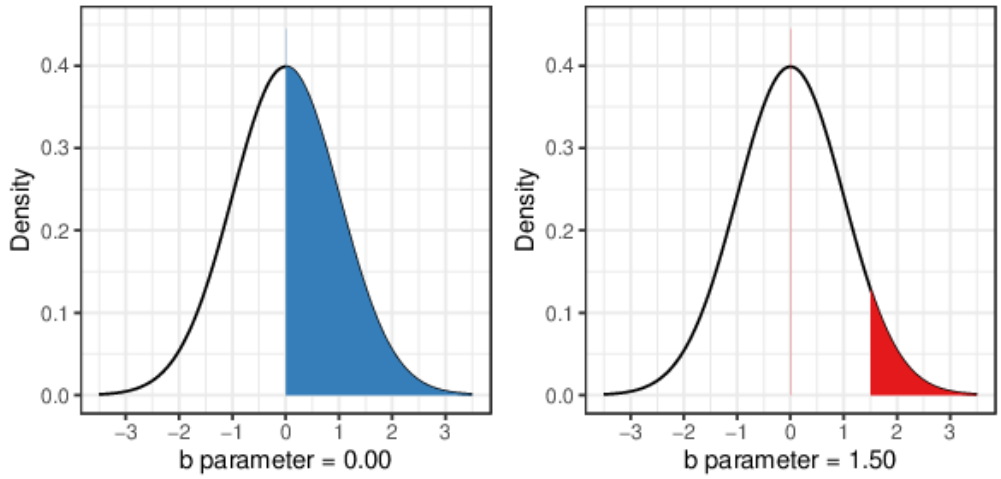
Normal distribution curve and the area under the normal distribution curve



The right side of Figure 2 shows the area under the normal distribution curve for different z scores. When this plot is analysed, its similarity with the item characteristic curve in Figure 1 is significant. Therefore, is it possible to interpret that an item with b = 0.00 is responded correctly by half of the group (p = 0.50 according to CTT) and an item with b = 1.50 is responded correctly by %6.5 (p = 0.065 according to CTT) of the group? Both cases are shown in Figure 3.

Figure 3.

Two sample b parameters on the normal distribution curve



As shown in the graphic in Figure 2, the view that the normal distribution curve can be used for an IRT-based conversion is not unrealistic. In fact, Lord (1980) focuses on the relation between CTT and IRT and he indicates that "*We can see from the figure that the item response function … is equal to a standardized normal curve area*" (p.31), and also adds that this approximation is "*not for practical use but rather to give an idea of the nature of the item discrimination parameter"* (p.33-34). Lord (1980) also describes the relationship between CTT item difficulty and discrimination statistics and IRT *a* and *b* parameters with mathematical proving. When all the items have the same discrimination (e.g., Rasch model), $b_j \approx \phi_j$ while b_j represents the IRT item difficulty parameter for *item j* and ϕ_j represents the area under the normal distribution curve at the point of CTT item difficulty statistics, p_j. On the other hand, if the items have different item discrimination, then $b_j \approx \phi_j \,/\, r_{jx}$ while r_{jx} represents the item-total biserial correlation or CTT item discrimination statistics. Lord (1980) also describes the relationship between IRT *a* parameter and CTT item discrimination statistics r_j as in Equation 4:

$$a_j = \frac{r_{jx}}{\sqrt{1 - r_{jx}^2}} \qquad (4)$$

In addition to that, even the equation of the area under the normal distribution curve looks very complicated, a simple function (e.g., NORM.DIST) in a spreadsheet software (Microsoft Excel, LibreOffice Calc, Google Sheet, etc.) can make the calculations.

Recent studies that support this perspective can be found in the literature. Kohli et.al. (2014) discussed the comparability of CTT and IRT based item parameters with underlying normal variable assumption. They found extremely high correlations between IRT and CTT based parameters, and these correlations are affected more by sample size rather than item pool size. Raykov and Marcoulides (2016) also shows the relationship and equivalence of CTT & IRT. They also recommend researchers combine the benefits of both test theories. A recent study by van der Ark and Smits (2023) suggests a new CAT method without using IRT, and they call it FlexCAT. Yet, their method is based on Latent Class Analysis (LCA), and it is still not very feasible for non-technical researchers.

Beyond the relationship between CTT and IRT in the manner of item parameters, how practical is the CTT to IRT transformation using this relationship for CAT applications? Due to the nature of CAT, we can estimate the trait level of the examinee with items that give more information about the examinee. In a typical CAT application, the next item to administer is selected based on the previous responses of the examinee, and the trait level can be estimated with much fewer items in contrast to conventional linear tests. This feature of the CAT makes it more convenient to schedule the test-taking time and place since not every examinee takes the same item set (van der Linden & Glas, 2002). Since CAT applications need a calibrated item bank, and the calibration process needs a large sample size, developing an item bank is not very feasible for a small-scale application. IRT calibration also needs expertise and cannot easily be implemented in the testing process for unfamiliar researchers to the IRT. The conversion of the item difficulty from CTT to IRT using the normal distribution curve mention above has potential for not only the development of CAT forms but also other applications based on IRT.

In this context, the research aims to evaluate the effectiveness of the transformation from the CTT-based *p* statistic to IRT-based *b* parameter using the normal distribution curve in terms of a CAT simulation. And due to fact that the focus of this study is on converting CTT item difficulty statistic to the *b* parameter, the present study is limited with the Rasch model since all the parameters but *b* are constant.

Methods

Data

In R (R Core Team, 2020), using the *genDichoMatrix* function of the catR package (Magis & Barrada, 2017; Magis & Raiche, 2012), 10, 50, 100, 250, 500 and 1000-item pools were created sequentially. The item pool was created according to the Rasch model, accordingly, the item discrimination parameter *a* was accepted as *a = 1.00*, the pseudo-guessing parameter as *c = 0.00*, and the asymptote parameter as *d = 1.00*. Therefore, only *b* parameters were generated using *genDichoMatrix*. Then, 10, 50, 100, 250, 500, and 1000 response patterns were generated for each item pool using the *genPattern* function included in the catR package. Therefore, 36 different response patterns have been studied, including a total of six item pools and six response patterns for each item pool. The rationale behind choosing these conditions, is due to test the performance of the conversion on the data from different sample sizes and item pools. For instance, a teacher may want to convert the item statistics calculated from the data obtained from a classroom as small as 10 and item pool as small as 10. But it is also important to see the performance of the conversion from the data from a larger sample and item pool. While generating response patterns, theta values and item parameters in the item pool were used. Theta values were obtained from a normal distribution whose mean score is 0 and standard deviation is 1.

Data Analysis

Since item parameters were generated according to IRT, item difficulties were first obtained according to classical test theory for data analysis. In this regard, item difficulty values were calculated by finding the means of each item in 36 response patterns. Then, those item difficulties were converted to standard z score using the following function below, and these scores were accepted as *b* parameter according to IRT. The item difficulty parameters obtained according to classical test theory are demonstrated with *p*; item difficulty parameters converted from classical test theory to item response theory with b_p, and the item difficulty parameters obtained according to the item response theory were indicated with *b*. The following function is used to obtain b_p:

$$b_p = 0 - qnorm(colMeans(var)) \qquad (4)$$

This function simply takes the *p* parameter as a percentage of the area under normal distribution and the z value corresponding to the percentage indicated by the *p* parameter as the *b* parameter according to IRT. In this function, *colMeans (var*) calculates the means of columns. In other words, the item difficulty value of each item is calculated. For example, in this function, if 0.065 is used instead of *colMeans (var),* 1.51 is obtained, and this value is in accordance with the example given after the Equation 4. Similarly, when 0.50 is written instead of *colMeans (var),* the function gives the output as 0. In other words, for $p = 0.50$, $b_p = 0$ is obtained.

Following the parameter transformations, *b*, *p*, and b_p parameters were obtained sequentiallyly, for each response pattern. At this stage, *Inf* and *-Inf* values were obtained during the b_p conversion, especially when the sample size was 10. To avoid errors in simulations, *Inf* values were changed into 6 while *-Inf* values were changed into *-6*.

A CAT simulation was conducted with both *b* and b_p parameters. In the simulation, Maximum a Posteriori (MAP) was used as an ability estimation method and Maximum Fisher Information (MFI) as item selection method. In the first item selection, theta was assumed as 0.00 and the simulation was terminated when the standard error value was below .40. The simulations were carried out via the *simulateRespondents* function included in the *catR* package.

According to the simulation results, when *b* and b_p were used, the average number of items used in the simulation, the correlation coefficients between the full-item estimated theta and theta levels estimated by CAT, and bias and RMSE values were compared and the seed value set as 26 for the item and response generation, and CAT simulations.

Results

According to the results of a total of 72 CAT simulations using the *b* and b_p parameters for a total of 36 data sets, correlations between theta values estimated using *b* and b_p parameters are shown in Figure 4.

Figure 4.

Correlation coefficients between theta levels estimated from CAT simulations using b and b_p

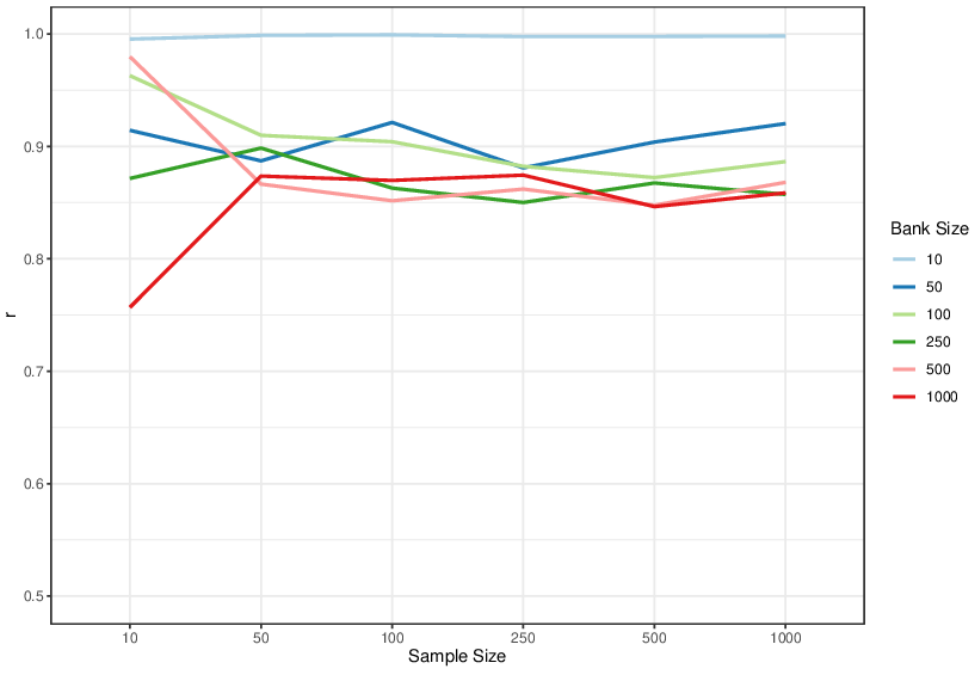
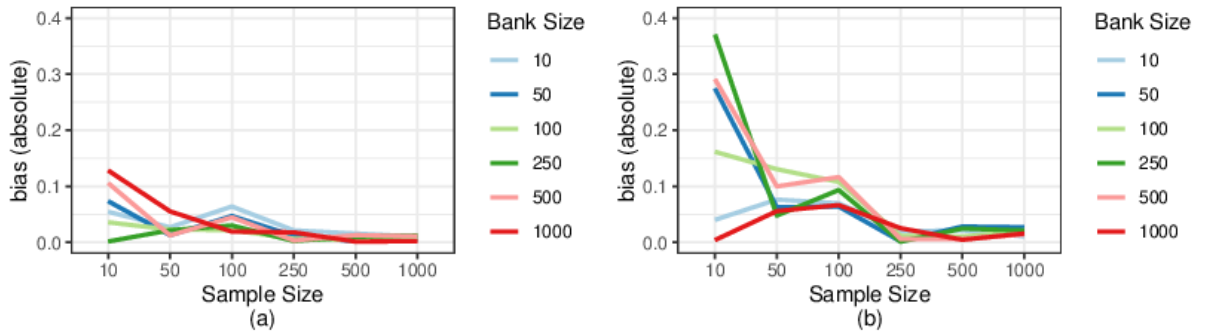


Figure 4 shows that the correlation coefficients were above .80 for almost all cases. When the item pool size was 10, the correlation coefficients for all sample sizes were close to 1.00. This is because the simulation cannot reach the .40 standard error used for the termination rule below 10 items. While the highest correlation coefficient was obtained with a pool of 50 items, other item pools were found to have around .85 correlation coefficients, especially in samples of 50 respondents and above. This indicates that the IRT-based *b* parameter or CTT-based b_p parameter can perform similar theta estimation in CAT simulation.

The bias values of theta estimates were analyzed for both *b* and b_p, and the values obtained for all item pools and sample sizes are presented in Figure 5. For clarity, bias values are analyzed as absolute values.

Figure 5.

Bias values from CAT simulations using b and b_p



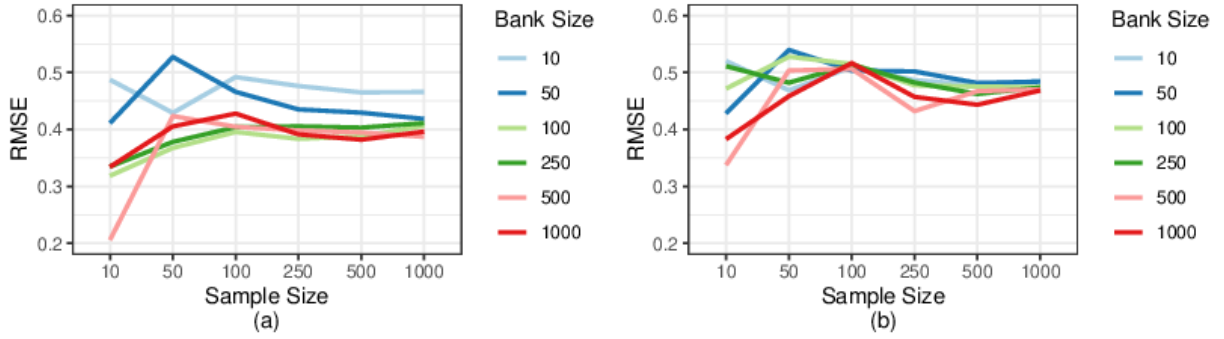
(a) with b parameter; (b) with bp parameter

Although it is seen that both methods have high bias values in small sample sizes, it is understood that the IRT-based parameter estimates with lower bias. Especially when the sample size is 250 and above, the bias value approximates to zero for the theta levels estimated by IRT-based parameters. A decrease in bias value because the sample size increased was also observed in the difficulty parameter obtained by CTT conversion. Similarly, when the sample size is 250 and above, the bias value drops below 0.05.

RMSE values of ability estimates were also analyzed for the whole item pool and sample sizes (see the plots in Figure 6).

Figure 6.

RMSE values from CAT simulations using b and b_p



(a) with b parameter; (b) with bp parameter

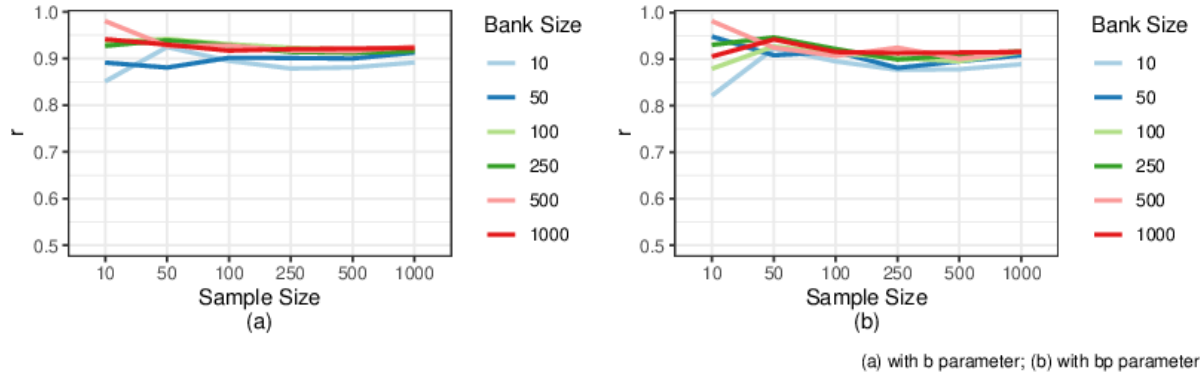
When RMSE values are analyzed, it is seen that, similar to bias, RMSE values of the ability level estimated by IRT-based difficulty parameter are lower. However, in cases where the sample size is 250 and above, it is seen that the RMSE value goes below .50 in both methods.

The correlation coefficients between the ability levels estimated by CAT simulations using both *b* and b_p parameters and the ability levels estimated from all items were analyzed and shown in Figure 7.

Figure 7 shows that the CAT simulations using IRT-based *b* and b_p converted from CTT have similar correlation coefficients between full-theta estimates and CAT estimates. Although the correlation coefficients obtained for 10 respondents vary according to the item pool sizes, it is seen that the correlation coefficients between all theta and estimated theta values with the sample sizes above 50 are around .90.

Figure 7.

Correlation coefficients between CAT simulations using b and b_p and theta levels estimated from all items

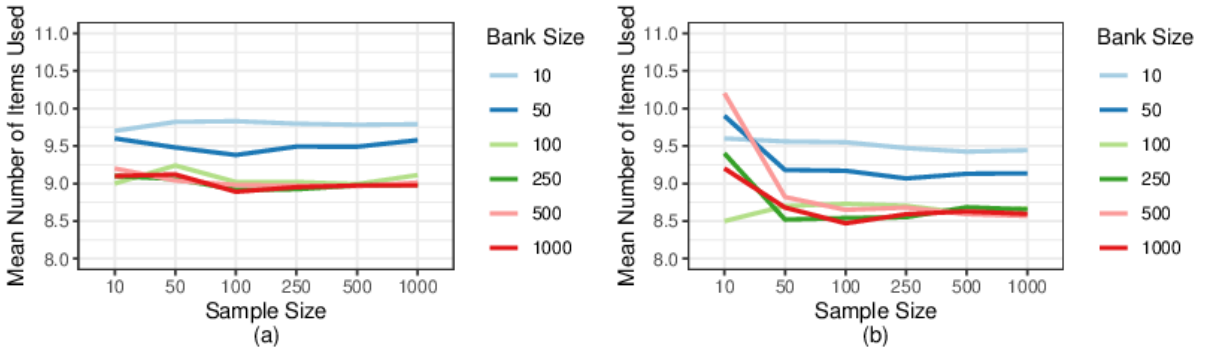


(a) with b parameter; (b) with bp parameter

The average numbers of items in which CAT simulations are terminated for both parameters are given in Figure 8.

Figure 8.

Average numbers of items in which CAT simulations are terminated using b and b_p parameters



(a) with b parameter; (b) with bp parameter

It is seen in Figure 8 that in the CAT simulation using the *b* parameter, the entire item pool is used when the item pool size is 10. On the other hand, in cases where the item pool size is 100 and above, it is seen that the simulation terminates with a similar number of items for both *b* and b_p parameters.

Discussion

This study aims to investigate the effectiveness of transforming CTT-based *p* statistic into IRT-based *b* parameter by utilizing the area under the normal distribution curve in the manner of CAT simulation.

It is seen that the converted b_p parameter has a higher bias and RMSE value than the *b* parameter in CAT simulation. However, it was found that bias and RMSE values in the simulations using b_p also decreased, especially when the sample size was 250 and above. On the other hand, while the correlation coefficients between the estimates were found to be around .85, the correlation coefficients between the ability levels estimated by CAT and the ability levels estimated from all items were found to be around .90 when both *b* and b_p parameters were used. In both cases, the simulation terminated with less than 10 items.

All these findings reveal the potential of b_p converted from CTT into IRT in IRT-based studies and supported by previous studies (Kohli et al. 2014; Raykov & Marcoulides, 2016). Simulation results are more effected by sample size rather than item pool size (except item size was 10) which is matched with the Kohli, Koran, & Henn (2014). Although the findings show that the b_p parameter is not as effective as the *b* parameter, the similarity of CAT simulation results is promising. Especially due to COVID-19 pandemic, the practicality of measurement and assessment processes in distance education has become even more important. In this process, tailored test solutions such as CAT are beyond being available to educators who are not particularly familiar with IRT.

In this context, it is expected that CAT applications can be developed by easily converting parameters from CTT to IRT with the proposed conversion. Practically, a teacher who applied a test to 250 students can convert the *p* statistic to *b* parameter and use the items in a CAT form. In addition to that, converted *b* parameters can be used to kickstart an operational CAT application, then make the IRT based calibrations as data grows. On the other hand, the data used in the research were produced in accordance with IRT assumptions with catR package. Investigating the performance of the b_p parameter where IRT assumptions are not met, as well as applying real data-based post-hoc CAT simulations, will provide a deeper understanding to see how effective the transformation is. In addition, the transformation applied in the research assumed that student ability is normally distributed. Further studies are required to be conducted on how violating this assumption may affect the b_p parameter and the results of the analysis.

Acknowledgement

This study is supported by Pamukkale University Scientific Research Projects Committee. Project No: 2021BSP008. The preliminary results of this study were presented in IACAT 2022, Frankfurt, Germany.

Declarations

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

Ethical Approval: The data was simulated; thus ethical approval is not required.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51. https://doi.org/10.1007/BF02291411

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Cengage Learning.

Çelen, Ü., & Aybek, E. C. (2013). Öğrenci başarısının öğretmen yapımı bir test ile klasik test kuramı ve madde tepki kuramı yöntemleriyle elde edilen puanlara göre karşılaştırılması. *Journal of Measurement and Evaluation in Education and Psychology*, *4*(2), 64-75. https://dergipark.org.tr/tr/pub/epod/issue/5800/77213

De Ayala, R.J. (2009). *The theory and practice of item response theory.* The Guilford Press.

Hambleton, R., & Swaminathan, R. (1985). *Fundementals of Item Response Theory.* Sage Pub.

Hambleton, R., Swaminathan, R., & Rogers, H.J. (1991). *Fundementals of Item Response Theory*. Sage Pub.

Kohli, N., Koran, J., & Henn, L. (2015). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educational and psychological measurement*, *75*(3), 389–405. https://doi.org/10.1177/0013164414559071

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Routledge.

Magis, D. & Barrada, J.R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software, Code Snippets*, *76*(1), 1-19. https://doi.org/10.18637/jss.v076.c01

Magis, D. & Raiche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8), 1-31. https://doi.org/10.18637/jss.v048.i08

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. https://doi.org/10.1007/BF02296272

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series, 1992*(1), i-30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Pitman, J. (1993). *Probability (6th Edition).* Springer.

R Core Team (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. URL https://www.R-project.org/.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 321-333). University of California Press.

Raykov, T., & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and psychological measurement*, *76*(2), 325–338. https://doi.org/10.1177/0013164415576958

Reckase, D. (2009). *Multidimensional Item Response Theory.* Springer.

Samejima, F. (1996). *Polychotomous responses and the test score*. The University of Tennessee.

van der Linden, W. J. & Glas, G.A.W. (2022). *Computerized Adaptive Testing: Theory and Practice.* Kluwer Academic Publishers.