International Journal of Educational Studies and Policy (IJESP)

Volume: 6, Issue: 1, May 2025

Comparison of Classification Accuracy and Consistency Indices Under the Item Response Theory*

Nurşah Yakut[1], Emine Önen[2]

ABSTRACT

In educational settings, individual diagnostic and placement decisions are made based on several measures, and classification accuracy indicates how accurate these decisions are. In this study, the effectiveness of Lee's, Guo's, and Rudner's methods in assessing classification accuracy and consistency were examined under Dichomotous IRT models in terms of different sample sizes and test lengths. The data were generated using the 'irtoys' package in R Studio. Classification accuracy and consistency indices and bias values related to these indices were calculated using the 'cacIRT' package. As the number of items increased, the classification accuracy and consistency indices showed a remarkable difference; for Kappa values calculated using Lee's method and FP and FN rates calculated using Guo's method, higher bias values were observed. Rudner indices were observed to have lower "absolute values of the bias" than other methods. In terms of classification decisions, it is considered that Rudner's method would work better when applied to large sample sizes.

Keywords: Classification accuracy, classification consistency, Item Response Theory

DOI: https://doi.org/10.5281/zenodo.15075438

Article Info:

Received: 04.12.2024 **Accepted:** 20.03.2025 **Article Type:** Research Article

Cite as: Yakut, N. & Önen, E. (2025). Comparison of classification accuracy and consistency indices under the Item Response Theory. *International Journal of Educational Studies and Policy, 6*(1), 90-105.

[1]Nurşah Yakut, Purdue University, College of Education, Learning Design & Technology, Curriculum & Instruction. nyakut@purdue.edu, ORCID: 0000-0002-2983-0329

[2]Corresponding Author: Emine Önen, Department of Educational Sciences, Division of Measurement and Evaluation in Education, Gazi Education Faculty, Gazi University, emineonen@gazi.edu.tr. ORCID: 0000-0002-0398-3191

*This manuscript was based on the master's thesis prepared by the first author under the supervision of the second author.

Introduction

The ultimate goal of many educational tests is to place individuals into appropriate categories based on their test performance and predetermined criteria. Based on test scores, diagnostic and placement decisions are made about individuals, and they are chosen and placed in several educational institutions. Since these decisions directly affect the lives of individuals, examining the accuracy and appropriateness of these decisions is of great importance for individuals and society (Cizek & Bunch, 2007). The two indices that can be used in assessing such classification and placement decisions in educational settings are classification accuracy (CA) and classification consistency (CC). Classification consistency refers to the degree to which individuals are consistently classified in the same category based on repeated or parallel measures. Classification accuracy provides information on the accuracy of classification decisions. Classification accuracy measures how well individuals' classifications based on observed test scores agree with their "true classifications," assuming the individuals' "true classifications" are known (Lee et al., 2000). Based on Item Response Theory (IRT) and Classical Test Theory (CTT), various methods and techniques have been developed to examine classification consistency and classification accuracy. For some CTT-based or IRT-based methods, researchers need to administer the same test or a parallel form of the test. In practice, only one form of the test is usually used. In examining classification accuracy, the “true” scores of individuals are not known, and they should be estimated. For these reasons, methods for examining classification consistency and accuracy based on a single administration were developed. The methods developed based on the CTT are based on the observed test scores and assume that the cut-scores are on the raw score scale (Cohen, 1960; Hanson & Brennon, 1990; Huynh, 1976; Lee et al., 2009; Livingston & Lewis, 1995; Subkoviak, 1976). Although the methods developed under the IRT framework are based on IRT models, the distributional assumptions underlying these methods are different. While in some methods, the classification decisions are based on a raw score scale, in others, these decisions are based on a latent trait scale. The advantage of IRT-based methods is that they provide flexibility to researchers (Guo, 2006; Lee, 2010; Rudner, 2001). The most commonly used IRT-based methods based on single administration are Rudner’s method, Lee’s method, and Guo’s method (Lathrop & Cheng, 2013; Lee, 2010).

Rudner’s method

Rudner’s (2001) method assumes that the cut scores and the individuals’ test scores are both mapped on the continuous theta scale (θ). Another assumption is that theta scores estimated based on individuals’ item responses are directly compared to theta cut scores placed on the continuous theta scale. If there are K categories of classification, K-1 theta cut scores ($\lambda 1 < \lambda 2 <$ ... $< \lambda K-1$) are set on the theta scale. To estimate Rudner-based indices, as lower and upper limits, two additional theta scores ($\lambda 0 = -\infty$ and $\lambda K = \infty$) are included. If the individual’s estimated theta ($\hat{\theta}_i$) is higher than or equal to $\lambda k-1$ and smaller than λk, this individual is classified into category k. The probability for classifying into category k is given as:

$$\Pr(K = k|\theta_i) = \Phi\left(\frac{\lambda_k - \theta_i}{se(\theta_i)}\right) - \Phi\left(\frac{\lambda_{k-1} - \theta_i}{se(\theta_i)}\right) \quad (1)$$

ϕ is the “cumulative normal distribution function” with mean θ_i, and the standard error of the estimate is se(θ_i).

The “true” classification of the individuals is unknown. The classifications based on observed test scores could be treated as the best estimation of the “true” classification. The probability of an individual being classified into their “true” category is called conditional classification accuracy ($\gamma\theta_i$). This equals the area between the two theta cut-scores under the normal distribution curve. The area under the normal distribution curve in the relevant regions can be calculated, and conditional false positive and false negative rates can be obtained (Rudner, 2005). The original Rudner method was developed just to calculate classification accuracy indices. As an extension of this method, a method for calculating classification consistency indices was suggested (Wyse & Hao, 2012). Assume that a parallel test form is administered to individuals, in which the same cut-scores are employed as the administered form of the test, and the items comprising the test have the same IRT item parameters. Using Equation 1, the probabilities of the same individual with the same latent trait score (θ_i) being classified into each category can be computed. Conditional classification consistency ($p\theta_i$) is calculated by using the following equation under the assumption that test forms are independent:

$$p_{\theta_i} = \sum_{k=1}^{K} [\Pr(K = k|\theta_i)]^2 \qquad (2)$$

K represents the number of classification categories. The probability of an individual with ability θ_i being classified into category k based on the administered test form ((Pr(K =k|θ_i)) is computed. This probability is calculated for each k category, and the conditional classification consistency is obtained by summing these probabilities (Rudner, 2001; 2005).

Guo’s method

Guo's (2006) method was developed by extending Rudner's method. The indices are calculated using the likelihood functions of individuals’ ability estimates, so Guo’s method does not require the normality assumption. It is based on the assumption that observed test scores and cut-scores are both mapped on the continuous theta scale. The lowest and the highest theta cut scores are replaced with some relatively large values, as demonstrated in Guo’s (2006) original study. Conversely, in Rudner's method, positive and negative infinities are considered the borders of cut scores. In Guo’s method, a set of theta points (e.g., 100 points) is created with equal intervals between the cut-scores. The expected probability of an individual with any θ being classified in category K could be computed as:

$$p_{ic}=\frac{\sum_{\theta=\kappa_{ci}}^{\kappa_{ci+1}} L\left(u_{1i}, u_{2i}, \dots, u_{ji}\middle|\theta\right)}{\sum_{h=1}^{C+1}\ \sum_{\theta=\kappa_{h}}^{\kappa_{h+1}} L\left(u_{1i}, u_{2i}, \dots, u_{ji}\middle|\theta\right)} \qquad (3)$$

To compute this probability, the sum of the likelihood functions from category C to the succeeding category (C+1) for a set of theta points (e.g., 100 points) created with equal intervals between the cut-scores should be calculated. Then, the sum of all cross-category likelihood functions is computed (Guo, 2006). Guo (2006) only developed a method for calculating classification accuracy. The extension of this method for calculating classification consistency was proposed by Wyse and Hao (2012). The rationale behind this method is like the rationale of the extension of Rudner’s method. Assume that a parallel test form is administered, consisting of the same items as those in the administered form of the test and employing cut-scores the same as the ones employed in the administered test form. In this case, the probability of being classified in each category in both forms of the test would be the same for an individual who gives the same responses to items in both test forms. The conditional classification consistency (pi) could be computed by assuming that two test forms are independent:

$$p_i = \sum_{k=1}^{K} \left[Pr(K = k | u_{i1}, u_{i2}, \dots, u_{iJ}) \right]^2 \tag{4}$$

The conditional classification consistency is the sum of “the normalized probability of being classified into category k” Pr (K = k|ui1, ui2, ..., uiJ) across categories (Guo, 2006; Wyse & Hao, 2012).

Lee’s method

In Lee’s method, the cut-scores are taken in the raw-score metric and compared with the raw scores of individuals. Consider that respondents are classified into k categories based on their scores on a test of dichotomous items. The probability of obtaining each raw score depending on the individual's $\hat{\theta i}$ and item parameters is computed. The raw score (x) is a function of the response pattern to test items. There would be many item response patterns giving x raw score, and the probabilities of all these patterns should be summed. While “true” ability is estimated in theta scale, the classification decisions are based on the raw score scale. For “true” classification, $\hat{\theta}$i is transformed into the expected or “true” raw score($\tau\hat{\theta}$i) with a test characteristic curve. $\tau\hat{\theta}$i, as measured on a raw score scale and compared to raw cut-scores for true classification, can be considered the best estimate of the examinee's true 'ability.' Once transformed, cut-scores are rounded to the nearest possible raw score. The following equation can be used to compute the probability of being classified into the “true” category k:

$$\Pr(K = k|\widehat{\theta_\iota}) = \sum_{x=c_{k-1}}^{c_{k-1}} \Pr(X = x|\widehat{\theta}_\iota) \tag{5}$$

The term Pr(K = k|$\hat{\theta}$i) is computed using Equation 5. The above-mentioned conditional indices give information for a particular individual or a theta level. However, they are generally not preferred for examining overall classification consistency and classification accuracy for decisions made for a group (Chau, 2018; Lee, 2010).

Comparison of Rudner’s method, Guo’s method, and Lee’s method

There are some similarities and differences between these three methods. In Rudner’s method, there is an assumption that ability estimates are normally distributed. However, in Guo’s method, no single distribution assumption is accepted, and expected probabilities are used in calculating the indices. In Lee’s method, cut-scores are used on the raw score scale, and for the decisions regarding classification, the raw scores of individuals are compared with the raw cut-scores. In Rudner’s and Guo’s methods, the observed test scores and cut-scores are assumed to be mapped onto theta scale. In Lee’s method, the “true ability” of the individual is estimated in theta scale, whereas classifications are made in the raw score scale. The assumptions underlying these methods can lead to different classification accuracy and consistency estimates (Guo, 2006; Lee, 2010; Rudner, 2001). In studies examining classification accuracy and consistency indices using IRT-based methods, it has been observed that model misspecification (Chau, 2018; Lathrop & Cheng, 2013), sample size (Chau, 2018; Martineau, 2007; Sen & Cohen, 2020), test length (Chau, 2018; Wyse & Hao, 2012), choice of model (Lathrop & Cheng, 2013; Lee et al., 2002), cut scores (Chau, 2018), ability distribution, and ability estimator (Wyse & Hao, 2012) affect these indices. Wyse and Hao (2012) proposed two new classification consistency indices based on IRT. They compared Rudner-based indices and Guo’s indices with “IRT-Recursive-Based Indices.” The researchers noted that Guo's method generally resulted in the highest overall classification accuracy and classification consistency rates, followed by Rudner's and then Lee's methods. Only a few studies have been found in which the methods of Rudner, Guo, and Lee have been comparatively examined (Chau, 2018; Lathrop& Cheng, 2013; Wyse & Hao, 2012). In one study

(Chau, 2018), these three methods were considered together. However, the researcher only examined the effectiveness of these three methods in assessing CA and CC by generating data in a simulation condition in which the 60-item test was applied to a sample of 5000 individuals, and the cut-off score was fixed at 1. In real test applications, tests with fewer items are widely used. In addition, Chau (2018) considered a large sample size in his study. In educational settings, tests such as the English Proficiency test, which are used to determine the proficiency levels of individuals, are frequently administered to smaller samples/groups. In these applications, decisions are made about individuals based on different cut-off scores corresponding to different proficiency levels. This indicates a need to investigate which of these three methods leads to more accurate results in assessing classification accuracy and classification consistency under which conditions for different cut-off scores in relatively smaller samples. In their article, Diao and Sireci (2018) explained Rudner's, Lee's, and Guo's methods and the package programmes used to calculate CA and CC indices using these methods. However, researchers have not examined the effectiveness of these methods in assessing classification accuracy and classification consistency based on real data or simulation data. In Lathrop and Cheng's (2013) study, the effectiveness of Lee's and Rudner's methods in just assessing classification accuracy under different sample sizes, test lengths, and cut score conditions was examined, but Guo's method was not taken into consideration. However, according to the authors' review of the literature, there is no comprehensive study examining the effectiveness of Rudner's, Lee's, and Guo's methods for different cut-off scores when tests with fewer items are administered to smaller samples (both for correct IRT model specification and model misspecification) to assess classification accuracy and classification consistency. Therefore, in the current study, as an extension of the previous research, it was aimed to examine the effectiveness of Rudner's, Lee's, and Guo's methods in assessing classification accuracy and classification consistency when short and medium-length tests were administered to small and medium-sized samples for different cut scores. In this context, it is considered that the findings of this study will guide educational practitioners and researchers about which of these methods would be more appropriate to use in examining classification accuracy and classification consistency in real testing settings under these conditions. Accordingly, the effectiveness of these methods was comparatively examined under the 1-Parameter logistic model (1PLM), 2-Parameter logistic model (2PLM), and 3-Parameter logistic model (3PLM) in terms of different sample sizes and test lengths. Also, the effect of the cut-score was examined by calculating the classification accuracy and consistency indices based on three different cut-scores (-0.75, 0, 0.75). Additionally, the accuracy of classification accuracy and consistency indices were examined by calculating absolute values of the bias (related to these indices) in the study (Wang & Wang, 2001).

Method

Simulation study

Since this study aims to compare the classification accuracy and consistency indices under different conditions, simulation data sets were generated and analyzed. For this reason, this is a simulation study. The dichotomous data was simulated in the R software environment under 1PLM, 2PLM, and 3PLM, based on two different test lengths (15 items and 30 items) and two different sample sizes (500 and 1000). Reviewing the literature (Lathrop & Cheng, 2014; Lee, 2010; Wyse & Hao, 2012), it is seen that large samples (n≥1000) are used to examine classification accuracy and consistency. However, no study examined how effective Guo's, Lee', and Rudner's methods were in assessing classification accuracy and consistency with small samples. Therefore, in this study, to examine how these methods work in smaller samples, the effectiveness of these methods was examined under the conditions of a sample size of 500 and a sample size of 1000. In the previous studies, 15-item tests are considered short tests, and 30-item tests are considered long tests (Chen et al., 2013; Minchen & de la Torre, 2018; Terzi & de la Torre, 2018). Accordingly, 15 items and 30 items were chosen for this study. This study simulated data sets separately for the 12 simulation conditions using the "irtoys" package (Partchev, 2017) in the R Studio program. The simulation conditions are presented in Table 1. For data simulation, a total of 500 and 1000 values from the normal distribution (N(0,1)), in the range of (-4, +4) for the ability parameters were randomly drawn in the R software. These values are considered as “true” theta values.

Table 1. Simulation conditions.

Simulation condition	Data generating IRT model	Sample size	Number of items
1	1PLM	500	15
2	1PLM	500	30
3	1PLM	1000	15
4	1PLM	1000	30
5	2PLM	500	15
6	2PLM	500	30
7	2PLM	1000	15
8	2PLM	1000	30
9	3PLM	500	15
10	3PLM	500	30
11	3PLM	1000	15
12	3PLM	1000	30

The following procedure was followed in determining the item parameters to generate data for each simulation condition: (1) Values for the a-parameter were drawn from a uniform distribution as U[0.5, 2.0] to represent medium and high discrimination levels (Kingsbury &Weiss, 1980); (2) the values for the b-parameter were drawn from the normal distribution as N(-0.5, 1.5) to be close to the values in the actual administration and (3) the values the c-parameter were drawn from the normal distribution as N(0.20, 0.05), again considering an actual administration (Thompson, 2009). These values are considered as “true” item parameters. Data sets were generated under 1PLM using 15 random values drawn for the b-parameter for the first and third simulation conditions. Data sets were generated under 2PLM using random values drawn for the b-parameter and a-parameter for the fifth and seventh simulation conditions. Data sets were generated under 3PLM using random values drawn for the b-parameter, a-parameter, and c-parameter for the ninth and eleventh simulation conditions. For even-numbered simulation conditions, data sets were generated using 30 random values drawn for the a-parameter, b-parameter, and c-parameter data sets under the related IRT models.

Validity and reliability evidence

Exploratory Factor Analysis (EFA) was applied to each generated data set by using the “psych” package in the R Studio program (Revelle, 2015). For each data set, single-factor solutions were obtained (eigenvalues between 3.00 and 9.63). Factor loadings were observed to be higher than 0.25(factor loadings between λ=.25 and λ=.78). These findings were considered evidence for construct validity. Marginal reliability coefficients calculated for each generated data set (varied between .647-.842) had an acceptable level of reliability (Md Desa, 2012).

Data analysis

It was first examined whether the IRT assumptions were met for the simulated data sets. Results of a series of EFAs also indicated that the unidimensionality assumption was met for each generated data set. For the local independence assumption, the Q3 statistic (Yen, 1984) for all item pairs under 1PLM, 2PLM, and 3PLM was calculated using the “sirt” package (Robitzsch, 2020) in the R Studio environment. This assumption was met for each item pair. For the overall model fit for each model, model fit statistics were calculated using the “mirt” package (Chalmers, 2020) in the R studio environment. Each data set was tested under the IRT model on which it was generated. To examine the sensitivity of classification consistency and accuracy indices to model misspecification, each data set was also tested under two other IRT models different from the IRT model on which data generation was based. The calculated model fit statistics and error values (Comparative Fit Index: CFI, Tucker Lewis Index: TLI, Root Mean Square Error of Approximation: RMSEA, Standardized Root Mean Square Residual: SRMR) were presented in the supplementary file (Supplementary File, S-1). The model better fit the data when the data sets were tested under IRT models on which data generation was based.

The classification accuracy indices, consistency indices, and absolute values of the bias were estimated using the package ‘cacIRT (Lathrop, 2020). False Negative (FN) and rates False Positive (FP), Classification Accuracy (CA), Classification Consistency (CC), and Kappa (κ) values were calculated based on each method. These values were calculated for each simulation condition, under each IRT Model, and for three different cut-scores (-0.75, 0.00,0.75). Absolute values of the bias were calculated to assess the accuracy of the classification consistency and classification accuracy indices. All these values were calculated and compared for both correct model specification and model misspecification. In Rudner and Guo's methods, the "true" theta values used in data generation are compared with cut-scores (“true” classification). Based on the data sets generated, a classification is made by comparing the "theta" values estimated (with Maximum Likelihood estimation) under the relevant IRT model with the cut scores. CA is obtained as the proportion of agreement between these two classifications. FP was the rate of simulees being incorrectly classified in a category above the “true” classification. FN, the rate of incorrectly being classified in a lower category, was also calculated based on these two classifications. In Lee's method, "true" thetas and cut scores are transformed into raw score scales and then compared (true classification). Based on the generated data, observed raw scores are obtained by summing up the item scores for each simulee. Another classification is made by comparing these scores with raw cut-scores. Then, CA, FP, and FN rates are calculated based on these two classifications. To calculate the CC values, which indicate the degree to which individuals are classified in the same category in repeated or parallel measurements, a second data set is generated using the same “true” thetas and “true” item parameters (as a parallel form of the first test). In Rudner and Lee's method, a classification is made by comparing thetas estimated based on the second data set and theta cut-scores. In Lee's method, classification is made by transforming thetas estimated based on the second data set into a raw score scale and comparing them with the raw cut-scores. Regardless of the method used, CC values were calculated as the proportion of agreement between classifications based on the data sets generated. Cohen's Kappa coefficient refers to the proportion of consistent classifications above and beyond what would be expected by chance. Assuming two simulation data sets as measures obtained from independent test forms, Cohen's kappa coefficient was calculated by using the following equation and the marginal rates of the simulees classified based on these measures:

$$p_c = \sum_{k=1}^{K}(p_{.k})(p_{k.}) \tag{6}$$

Here pc is the degree of consistency obtained by chance between the two test forms; $p{\cdot}k$ and pk, are the marginal proportions of individuals classified in the category k in both test forms (Cohen, 1960; Chau, 2018; Lathrop & Cheng, 2013).

Results

Initially, each data set was tested under the IRT model on which it was generated. The overall CA and CC rates, Kappa values, and FP and FN rates were calculated to examine the classification consistency and accuracy based on Guo's, Rudner's, and Lee's methods. The results are given in Table 2. The absolute values of the bias for these indices were also calculated and presented in the Supplementary file (See Supplementary file, S-2). The findings obtained when the data sets that were generated under 1PLM were tested under 1PLM (without taking into account the cut-scores) indicate that there are differences for these indices due to the increase in the number of items: (1) In the overall CA and overall CC rates, and FP rates and Kappa values calculated based on Rudner's method, (2) in the overall CA and overall CC rates and Kappa values calculated based on Guo's method, (3) in the Kappa values calculated based on Lee's method. While an increase was observed in overall CA and CC rates and Kappa values, a decrease was observed in FP rates. Considering the cut-scores, it is noteworthy that at each cut-score, the overall CA and overall CC rates and the Kappa values calculated based on each method increase due to the increase in the number of items. It was found that the change in FN and FP rates depending on the number of items differed according to the cut-score. This differentiation has changed between methods. When the effect of sample size is examined without taking into account the cut-scores, there is no remarkable difference due to the increase in sample size for the overall CA and CC rates, Kappa values, FP, and FN rates calculated based on Rudner's, Guo's and Lee's methods. When the cut-score was taken into account, there were differences according to the cut-scores in the changes that occurred due to the sample size. These differences changed depending on the number of items and the methods used to estimate these indices. Due to the increase in the number of items, there was a decrease in the absolute values of the bias regarding the overall CC rates and (regarding) the Kappa values calculated based on Rudner's and Guo's methods. In Lee's method, it was found that the absolute values of the bias calculated for the kappa values and FN rates were lower in 30-item simulation conditions. Due to the increase in sample size, a notable difference was observed only for the absolute value of the bias for the overall CA rate calculated based on Rudner's method. The absolute value of the bias for the overall CA rate calculated based on Rudner’s method was higher for simulation conditions where the sample was large (n =1000).

Table 2. Classification accuracy and consistency Indices and Kappa values for data generating IRT model.

Simulation condition	**Cut-score**	**Rudner's method**					**Guo's method**					**Lee's method**				
		CA	**FP**	**FN**	**CC**	*κ*	**CA**	**FP**	**FN**	**CC**	*κ*	**CA**	**FP**	**FN**	**CC**	*κ*
1	-0.75	0.901	0.474	0.051	0.858	0.580	0.855	0.019	0.126	0.824	0.595	0.848	0.046	0.106	0.788	0.464
1	0	0.859	0.073	0.067	0.799	0.599	0.855	0.081	0.064	0.796	0.592	0.818	0.081	0.101	0.753	0.451
1	0.75	0.885	0.086	0.030	0.842	0.519	0.807	0.186	0.007	0.813	0.578	0.901	0.690	0.030	0.859	0.321
2	-0.75	0.931	0.041	0.028	0.903	0.723	0.902	0.024	0.074	0.866	0.675	0.893	0.035	0.072	0.848	0.622
2	0	0.899	0.042	0.059	0.855	0.711	0.881	0.053	0.066	0.832	0.664	0.863	0.062	0.075	0.810	0.593
2	0.75	0.914	0.060	0.026	0.882	0.648	0.863	0.126	0.011	0.851	0.640	0.919	0.031	0.051	0.897	0.484
3	-0.75	0.895	0.045	0.060	0.851	0.561	0.840	0.019	0.140	0.814	0.576	0.837	0.041	0.122	0.780	0.457
3	0	0.864	0.077	0.060	0.806	0.612	0.857	0.086	0.057	0.799	0.599	0.829	0.078	0.094	0.766	0.475
3	0.75	0.890	0.083	0.028	0.850	0.545	0.817	0.176	0.006	0.825	0.599	0.904	0.064	0.031	0.864	0.352
4	-0.75	0.924	0.033	0.043	0.892	0.696	0.885	0.020	0.094	0.858	0.658	0.888	0.035	0.077	0.847	0.604
4	0	0.904	0.060	0.037	0.865	0.730	0.886	0.071	0.043	0.841	0.681	0.870	0.086	0.071	0.818	0.619
4	0.75	0.915	0.050	0.035	0.880	0.650	0.875	0.109	0.016	0.854	0.652	0.914	0.037	0.049	0.886	0.498
5	-0.75	0.904	0.034	0.062	0.857	0.607	0.868	0.020	0.112	0.831	0.599	0.865	0.043	0.092	0.808	0.503
5	0	0.837	0.067	0.096	0.771	0.539	0.855	0.083	0.063	0.800	0.601	0.756	0.041	0.203	0.742	0.338
5	0.75	0.859	0.081	0.061	0.813	0.405	0.798	0.193	0.010	0.816	0.601	0.872	0.064	0.064	0.842	0.194
6	-0.75	0.923	0.034	0.042	0.893	0.689	0.894	0.018	0.088	0.868	0.674	0.895	0.026	0.079	0.857	0.585
6	0	0.903	0.051	0.046	0.865	0.729	0.890	0.052	0.059	0.846	0.693	0.852	0.033	0.115	0.817	0.634
6	0.75	0.937	0.038	0.026	0.910	0.740	0.900	0.085	0.016	0.877	0.693	0.910	0.055	0.036	0.871	0.612
7	-0.75	0.905	0.029	0.066	0.861	0.609	0.870	0.017	0.113	0.841	0.615	0.863	0.044	0.094	0.808	0.513
7	0	0.836	0.065	0.099	0.769	0.536	0.857	0.081	0.062	0.803	0.605	0.746	0.039	0.215	0.742	0.305
7	0.75	0.854	0.081	0.065	0.807	0.389	0.796	0.193	0.111	0.811	0.591	0.879	0.061	0.060	0.854	0.164
8	-0.75	0.916	0.036	0.048	0.885	0.669	0.885	0.018	0.096	0.864	0.670	0.890	0.030	0.081	0.851	0.567
8	0	0.907	0.049	0.044	0.871	0.741	0.893	0.050	0.057	0.851	0.702	0.856	0.031	0.113	0.821	0.641
8	0.75	0.938	0.036	0.026	0.912	0.742	0.901	0.083	0.015	0.878	0.695	0.909	0.056	0.036	0.870	0.617
9	-0.75	0.870	0.509	0.079	0.822	0.409	0.773	0.009	0.218	0.781	0.527	0.832	0.016	0.153	0.782	0.296
9	0	0.819	0.099	0.082	0.747	0.485	0.832	0.049	0.119	0.795	0.589	0.775	0.071	0.154	0.710	0.416
9	0.75	0.863	0.079	0.058	0.804	0.445	0.869	0.115	0.017	0.851	0.649	0.832	0.107	0.061	0.768	0.397
10	-0.75	0.920	0.034	0.046	0.889	0.667	0.886	0.017	0.097	0.863	0.661	0.879	0.024	0.092	0.845	0.582
10	0	0.885	0.047	0.067	0.834	0.668	0.890	0.053	0.057	0.845	0.690	0.859	0.047	0.094	0.810	0.600
10	0.75	0.881	0.059	0.060	0.839	0.511	0.870	0.113	0.016	0.858	0.666	0.927	0.027	0.046	0.916	0.260
11	-0.75	0.847	0.049	0.104	0.796	0.347	0.738	0.009	0.253	0.772	0.518	0.801	0.014	0.185	0.751	0.269
11	0	0.798	0.122	0.080	0.726	0.442	0.838	0.056	0.106	0.786	0.572	0.769	0.082	0.149	0.705	0.409
11	0.75	0.858	0.096	0.046	0.792	0.475	0.869	0.105	0.026	0.832	0.600	0.867	0.092	0.041	0.811	0.463
12	-0.75	0.913	0.038	0.048	0.875	0.620	0.884	0.014	0.102	0.861	0.659	0.881	0.025	0.094	0.842	0.568
12	0	0.875	0.055	0.070	0.819	0.637	0.886	0.057	0.056	0.839	0.679	0.849	0.049	0.103	0.799	0.571
12	0.75	0.882	0.060	0.058	0.838	0.499	0.868	0.117	0.015	0.855	0.656	0.937	0.027	0.035	0.925	0.233

Note. CA: Classification Accuracy, FP: False Positive Rates, FN: False Negative Rates, CC: Classification Consistency, *κ*: Kappa

Considering the findings obtained when the data sets that were generated under 2PLM were tested under 2PLM (without taking into account the cut-scores), some remarkable differences for these indices were observed depending on the increase in the number of items: (1) In the overall CA and CC rates, FN rates and the Kappa values calculated based on Rudner’s method, (2) in the overall CA and CC rates and the Kappa values calculated based on Guo’s method, (3) in the overall CC rates and the Kappa values calculated based on Lee’s method. An increase was observed in overall CA and CC rates and Kappa values, but a decrease was observed in FP rates. Taking into account cut-scores, at each cut-score, the overall CA and overall CC rates and kappa values calculated based on each method increased depending on the increase in the number of items. In contrast, a decrease was observed in the FP and FN rates calculated based on Lee's method. Also, there were differences according to the cut-scores in the changes that occurred due to the increase in the number of items in FN and FP rates calculated based on Rudner’s and Guo’s methods. These differences changed depending on the sample size. When examined in terms of the sample size without considering the cut-scores, no substantial difference was found depending on the increase in sample size for the CA, CC, FP, FN rates, and kappa values calculated based on all methods. When the cut-off score was considered, depending on the increase in the sample size, the overall CA, CC, FP, and FN rates calculated based on each method differed depending on the cut-scores. This differentiation was not systematic. Regarding the Kappa values, the differences were observed according to the cut-scores in the changes that occurred due to the sample size. These differences changed depending on the number of items and the methods used to estimate these indices. Depending on the increase in the number of items in all methods, a decrease was observed in the absolute values of the bias values regarding the overall CC rates and the Kappa values. The absolute values of the bias for FP rates calculated based on Rudner’s method and the absolute bias values for the overall CA rates calculated based on Guo’s method were found to decrease due to the increase in the number of items. Depending on the increase in the sample size, there was no significant difference in the absolute values of the bias regarding CA, CC, FP, FN rates, and kappa values (for any method).

Findings regarding 3PLM (without considering the cut-scores) indicated that overall CA and CC rates increased as the number of items increased in all methods. The kappa values based on both Rudner’s method and Guo’s method were found to be higher in 30-item simulation conditions in comparison to 15-item conditions. Besides this, due to the increase in the number of items, the FP rates calculated based on Rudner's method decreased. Considering the cut-scores, it was observed that at each cut-score, the overall CA rates calculated based on each method increased depending on the increase in the number of items. The changes in CC, FP, FN rate, and Kappa values depending on the number of items differed depending on the method and sample size. When the findings related to the effect of sample size were assessed without taking into account the cut-scores, it was understood that there was no substantial difference depending on the increase in sample size for the CA, CC, FP, FN rates, and kappa values (for any method). Considering the cut-scores (in all methods), the changes in CC, FP, FN rates, and Kappa values depending on the sample size differed depending on the increase in the number of items and the method. Depending on the increase in the number of items, it was observed that the absolute values of the bias regarding the CC, CA, FP rates, and Kappa values in Rudner’s method decreased substantially. In Guo’s method, the absolute values of the bias for CC rates and kappa values also decreased due to an increase in the number of items. However, in the Lee method, it was found that only the absolute value of the bias of the overall CC rate decreased. Due to the increase in sample size, no significant difference was found in the absolute values of the bias calculated based on any method. Then, the generated data sets were tested in case of model misspecification, and the absolute values of the bias were also calculated (See Supplementary File, S-3, S-4). The absolute values of the bias were interpreted without considering the cut-scores due to the page limitation. Regardless of the IRT model used in data generation, it was found that in the case of model misspecification, the absolute values of the bias related to classification accuracy and consistency indices tended to decrease as the number of items increased. However, there was no significant difference in these calculated absolute values of the bias depending on the sample size. In cases of model misspecification, the highest absolute values of the bias were obtained when the data set generated under 3PLM was tested under 1PLM. The lowest absolute values of the bias were obtained when the data set generated under 1PLM was tested under 2PLM, and the data set generated under 2PLM was tested under 3PLM.

Discussion and Conclusion

The effectiveness of Rudner's, Guo's, and Lee's methods in assessing classification accuracy and classification consistency were comparatively examined in terms of different test lengths and sample sizes under 1PLM, 2PLM, and 3PLM. Findings indicated that test length substantially affected classification accuracy, classification consistency indices, and the absolute values of the bias. As the number of items increased, the values of the indices increased, and the absolute values of the bias decreased. However, there was no remarkable difference in the classification accuracy and classification consistency indices and the absolute values of the bias of these indices, depending only on the sample size. In the case of model misspecification, test length appeared to have a notable effect on the absolute values of the bias for classification accuracy and classification consistency indices. A decrease in absolute values of the bias was observed due to the increase in the number of items. It was concluded that the absolute values of the bias did not differ substantially depending on the sample size. Lathrop and Cheng (2013) reported similar results in their study. As the number of items increased, absolute values of the bias decreased, but these values were not affected by the sample size. However, researchers stated that the standard errors decreased as the sample size increased.

In general, FP and FN rates are calculated based on Guo's method, and the overall CC rates and kappa values calculated based on Lee's method have the highest absolute values of the bias. For the classification accuracy and classification consistency indices calculated based on Rudner's method, lower absolute values of the bias were obtained compared to other methods. These findings were supported by the findings of the studies conducted by Chau (2018) and Lathrop and Chen (2013), which indicated that Rudner's method performed better. When these findings are assessed together with the findings of Martineau's (2007) study, it is considered that Rudner's method will be more appropriate for classification decisions for large sample groups. Rudner's method may be preferred for examining the classification accuracy and classification consistency in national exams administered in Turkey, as it performs well when applied to large sample groups and has lower absolute bias values than other methods.

In the case of model misspecification, the absolute values of the bias for the indices appear to differ, but there is no systematically significant difference. The classification accuracy and consistency indices had the highest absolute value of the bias when the data set generated under 3PLM was tested under 1PLM. This finding implies that considering the probability of answering correctly by chance could affect the accuracy of classification decisions based on test scores. It seems necessary to assess whether the model is correctly specified in modeling the responses to the items in such tests and in ability estimations based on these responses, especially in high-stakes tests used for the selection and placement of individuals in various educational institutions. In the case of model misspecification, kappa values calculated based on Lee's method had the highest absolute values of the bias in all simulation conditions. In the study conducted by Chau (2018), the values obtained in cases of model misspecification were similar to the values calculated under the correct model. Although these findings imply that these indices are not affected by model misspecification, Chau (2018) stated that in the case of model misspecification, the decisions made in practice and the evaluated decisions will not be the same as the ones being evaluated. In Chau's (2018) study, it was observed that the classification accuracy and consistency indices had high values when the location of the cut-scores had high test information. This study observed that the classification accuracy and consistency indices and the absolute values of the bias differed at different cut-off scores. However, there is no systematic differentiation. Additionally, when simulating data sets, ability parameters and item parameters drawn from specific distributions were used. Therefore, the findings can only be generalized to similar conditions.

As a result, national and international tests are used to determine the proficiency levels of individuals in a particular field. Accordingly, the accuracy of classification accuracy and classification consistency indices gain importance in such tests. In this study, only data sets based on dichotomously scored items are considered under simulation conditions, while it is of practical importance to investigate the accuracy of the classification accuracy and consistency indices for measures obtained from data sets based on multiple scored items.

Suggestions

Study findings indicated that sample size did not substantially affect the classification accuracy and consistency indices and the absolute bias values related to these indices. Although similar findings were obtained in the study of Lathrop and Cheng (2013), it was observed that the standard errors in that study decreased depending on the sample size. This means that the variability of the estimates of the classification accuracy decreases for larger sample groups. In this context, it is believed that it would be helpful to examine the effect of sample size on classification accuracy and consistency indices by studying with larger samples. While only dichotomous data sets have been considered in this study, it is of practical importance to investigate the accuracy of classification accuracy and consistency indices for measures obtained from polytomously scored items. In addition, in this study, ability parameters were drawn from the standard normal distribution for data simulation. Similarly, item parameters were drawn from the specific distributions, and data sets were generated based on these parameters. It is worth examining how different distributions that are used for data simulation can affect the accuracy of the classification accuracy and consistency indices. For example, a non-normal distribution for the ability parameter and distributions different from those used in this study for item parameters. Finally, in this study, model misspecification was considered the situation where the simulation data generated under one model was tested under another. However, model misspecification can be considered and examined differently in future research. For example, in future simulation studies, other situations that may lead to model misfit, such as multidimensionality of the data, can be examined, and model misspecification can be discussed in terms of different situations.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for this article's research, authorship, and/or publication.

Researchers’ contribution rate

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Ethics Committee Approval Information

An ethical approval form for the study was obtained from the Gazi University Ethical Committee, Document number and date: E-77082166-302.08.01-357512 11.05.2022

Acknowledgement

We would like to thank the Academic Writing Application and Research Center for the English language check of our article.

References

Chalmers, P. (2020). *Package 'mirt'*. [Computer software]. https://cran.r-project.org/ web/ packages/mirt/mirt.pdf.

Chau, L. H. (2018). *Evaluating the correctness of IRT-based methods in computing classification consistency and accuracy indices in model misspecification.* [Doctoral dissertation, University of British Columbia]. http://hdl.handle.net/ 2429/66984

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123-140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage Publication.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446000200010

Diao, H., & Sireci, S. G. (2018). Item response theory-based methods for estimating classification accuracy and consistency. *Journal of Applied Testing Technology, 19*(1), 20-25.

Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research, and Evaluation, 11*(1), 6.https://doi.org/10.7275/bxba-7466

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*(4), 345-359. https://doi.org/10.1111/j.1745-3984.1990.tb00753.x

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*(4), 253–264. https://doi.org/10.1111/j.1745-3984.1976.tb00016.x

Kingsbury, G. G., & Weiss, D. J. (1980). *A comparison of adaptive sequential, and conventional testing strategies for mastery decisions*. (ADA094478). https://apps.dtic.mil/sti/pdfs/ADA094478.pdf.

Lathrop, Q. N. (2020). *Package 'cacIRT'.* [Computer software]. https://cran.r-project.org /web/ packages/cacIRT/cacIRT.pdf.

Lathrop, Q. N., & Cheng, Y. (2013). Two approaches to the estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement, 37*(3), 226-241. https://doi.org/10.1177/0146621612471888.

Lathrop, Q. N., & Cheng, Y. (2014). A nonparametric approach to estimate classification accuracy and consistency. *Journal of Educational Measurement, 51*(3), 318-334. https://doi.org/10.1111/jedm.12048

Lee, W. C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*(1), 1-17. https://doi.org/10.1111/j.1745-3984.2009.00096.x

Lee, W. C., Brennan, R. L., & Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement, 33*(5), 374-390. https://doi.org/10.1177/0146621608321759

Lee, W. C., Hanson, B. A., & Brennan, R. L. (2000). *Procedures for computing classification consistency and accuracy indices with multiple categories.* https://www.act.org/content/dam/act/unsecured/documents/ACT_ RR2000-10.pdf

Lee, W. C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*(4), 412-432. https://doi.org/10.1177/014662102237797

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179-197. https://doi.org/10.1111/j.1745-3984.1995.tb00462.x

Martineau, J. A. (2007). An expansion and practical evaluation of expected classification accuracy. *Applied Psychological Measurement, 31*(3), 181-194. https://doi.org/10.1177/0146621606291557

Md Desa, Z. N. D. (2012). *Bi-factor multidimensional item response theory modeling for subscores estimation, reliability, and classification* [Doctoral dissertation, University of Kansas]. https://kuscholarworks.ku.edu/handle/1808/10126

Minchen, N., & de la Torre, J. (2018). A general cognitive diagnosis model for continuous-response data. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 30-44. https://doi.org/10.1080/15366367.2018.1436817

Partchev, I. (2017). *Package 'irtoys'*. [Computer software] .https://cran.r-project.org/ web/packages/irtoys/irtoys.pdf

Revelle, W. (2015). *Package 'psych'*. [Computer software]. https://cran.r-project.org/ web/ packages/psych/psych.pdf

Robitzsch, A. (2020). Package 'sirt'. [Computer software].https://cran.r-project.org/ web/ packages/sirt/sirt.pdf

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, 7*(14), 1-5. https://doi.org/10.7275/an9m-2035

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, 10*(13), 1-4. https://doi.org/10.7275/56a5-6b14

Sen, S., & Cohen, A. S. (2020). The impact of test and sample characteristics on model selection and classification accuracy in the multilevel mixture IRT model. *Frontiers in Psychology, 11*, 197. https://doi.org/10.3389/fpsyg.2020.00197

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement,* 265-276. https://doi.org/10.1111/j.1745-3984.1976.tb00017.x

Terzi, R., & De la Torre, J. (2018). An iterative method for empirically-based Q-matrix validation. *International Journal of Assessment Tools in Education, 5*(2), 248-262. https://doi.org/10.21449/ijate.407193

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*(5), 778-793. https://doi.org/10.1177/0013164408324460

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25*(4), 317-331. https://doi.org/10.1177/01466210122032163

Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement, 36*(7), 602-624. https://doi.org/10.1177/0146621612451522

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145. https://doi.org/10.1177/014662168400800201