

Comparison of Models for Simultaneous Estimation of Overall Score and Subscores: Estimation Accuracy, Reliability, and Classification Accuracy*

Ayşenur ERDEMİR**

Hakan Yavuz ATAR***

Abstract

This study compares the performance of the Multidimensional Item Response Theory (MIRT), Higher-Order IRT (HO-IRT), and Bifactor models for the simultaneous estimation of total and subscale scores in multidimensional tests. Using both simulated data and real data from an English proficiency exam, model performance was evaluated in terms of accuracy (RMSE), reliability, and classification accuracy. The simulation included 5,000 respondents, 120 items, and a four-dimensional structure, manipulating item format, test difficulty, and inter-dimensional correlation. Results indicated that MIRT consistently outperformed the other models, yielding the lowest RMSE and highest reliability and classification accuracy across conditions. HO-IRT also showed strong performance, while the Bifactor model underperformed, particularly in subscore estimation. Model performance was sensitive to test characteristics and dimensional relationships. Findings from the real data analysis supported the simulation results, underscoring the value of multidimensional modeling for diagnostic feedback and informed decision-making.

Keywords: *subscores, total score, MIRT, higher-order IRT, bifactor model, reliability, classification accuracy*

Introduction

Educational and psychological assessments often measure multiple abilities, resulting in inherently multidimensional structures (Reckase, 1985, 1997). Items are frequently designed to reflect both general and secondary abilities (Ackerman et al., 2003; DeMars, 2005), where total scores indicate general ability and subscores represent content-specific domains. Subscores are crucial for identifying student strengths and weaknesses and guiding instruction (DeMars, 2005; Haberman & Sinharay, 2010; Li, 2018; Reckase & Xu, 2015; Sinharay et al., 2011; Wedman & Lyren, 2015).

However, subscores often suffer from low reliability due to limited item numbers, posing a dilemma for diagnostic reporting. Reporting unreliable subscores may mislead rather than inform. Therefore, methods have been developed to improve subscore estimation accuracy and reliability (Chin, 2011). The *Standards for Educational and Psychological Testing* (2014) emphasize that when multiple scores are reported, their reliability and distinctiveness must be supported (Standard 1.14, p. 27).

Various approaches have addressed subscore estimation. Haberman (2008) and Sinharay (2010) evaluated the contribution of subscores using Classical Test Theory, while Brennan (2012) proposed a utility index. Other methods, such as subscore augmentation (Wainer et al., 2000) and collateral information (Kahraman & Kamata, 2004), also aim to enhance score accuracy. These approaches have been compared in terms of estimation accuracy and reliability (de la Torre & Patz, 2005; Fan, 2016; Shin, 2007; Sinharay, 2010; Wang et al., 2004; Yao & Boughton, 2007).

Item Response Theory (IRT) models the link between ability and responses. Model selection depends on item format, type, and number of latent traits (Hambleton & Jones, 1993). Unidimensional IRT

* This study was produced from the doctoral thesis prepared by the first author under the supervision of the second author.

**Dr. Police Academy, Institute of Forensic Sciences, Ankara-Türkiyeerdemiraysenur@gmail.com, ORCID ID: 0000-0001-9656-0878

***Prof. Dr., Gazi University, Faculty of Education, Ankara-Türkiye, hakanatar@gazi.edu.tr, ORCID ID: 0000-0001-5372-1926

To cite this article:

Erdemir, A. & Atar, H.Y. (2024). Comparison of Models for Simultaneous Estimation of Overall Score and Subscores: Estimation Accuracy, Reliability, and Classification Accuracy. *Journal of Measurement and Evaluation in Education and Psychology*, 16(4), 264-282. <https://doi.org/10.21031/epod.1748835>

Received: 24.07.2025

Accepted: 30.11.2025

assumes a single trait underlies responses (Hambleton et al., 1991), but this assumption is often violated in practice (Ackerman, 1994; Nandakumar, 1994; Reckase, 1985, 1997).

Multidimensionality can be planned (e.g., different content domains) or incidental (e.g., item formats, speededness). Planned multidimensionality may have a simple or complex structure. In simple structures, each item measures one trait (between-item multidimensionality); in complex structures, items measure multiple traits (within-item multidimensionality) (Reckase, 2009). In other words, a simple-structured test is a multidimensional test composed of two or more unidimensional subtests (Wang et al., 2004).

Many standardized tests, like the ACT, SAT, and Turkish national exams, include subtests. Applying unidimensional IRT to such tests may yield biased results by ignoring subtest differences (Ackerman, 1989; Ansley & Forsyth, 1985; Way et al., 1988). Although each subtest can be modeled separately, this may be inadequate when subtests are correlated or the number of items per subtest is small (Wang, 2012; Chalmers, 2012).

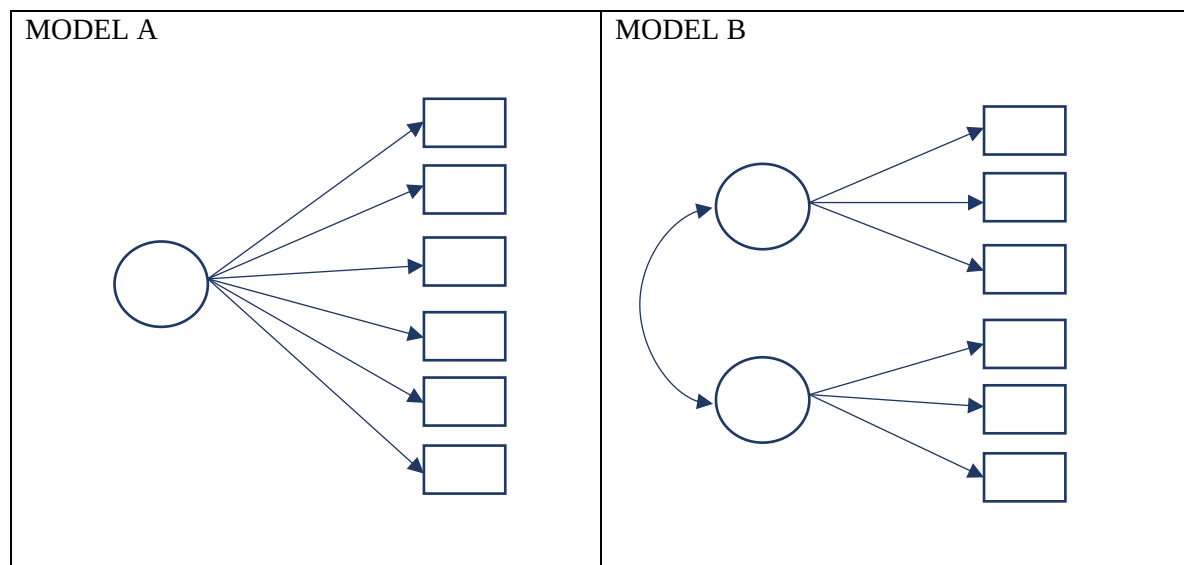
Multidimensional IRT (MIRT; Ackerman, 1994; Reckase, 1997, 2009) allows modeling of latent structures reflecting real test designs. For tests with simple structures, MIRT is recommended (de la Torre, 2008; Longabach, 2015; Sheng & Wikle, 2007). It improves estimation by accounting for interdimensional correlations (de la Torre & Patz, 2005). Studies have shown MIRT reduces bias in ability estimates (Chang & Davison, 1992; Wang et al., 2004).

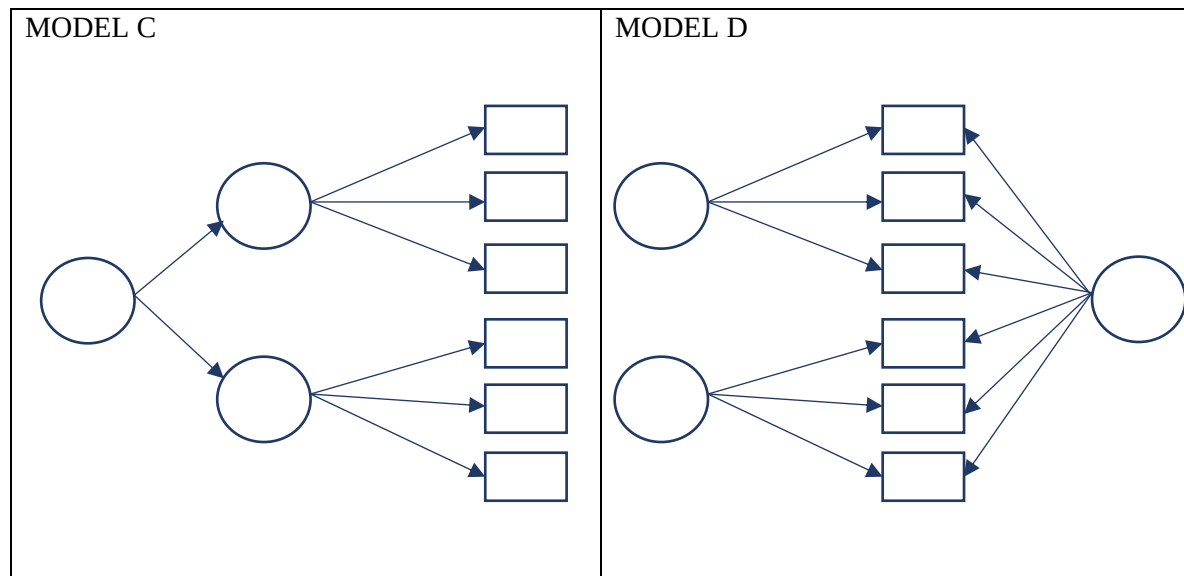
In assessments like TOEFL, both total scores and subscores are reported. Simultaneous estimation provides more detailed feedback (Yao, 2010). Researchers have emphasized the need to ensure subscore reliability (de la Torre & Song, 2009; Soysal & Kelecioğlu, 2018). Since simply averaging subscores to obtain total scores ignores subtest correlations, Yao (2010) proposed the maximum information method using MIRT. For the same purpose, de la Torre and Song (2009) introduced the Higher-Order IRT (HO-IRT) model, where a linear relation between subscores and the general factor is assumed. The Bifactor model (DeMars, 2013; Liu et al., 2019) represents another approach, where items are linked to both a general and a specific factor. In this model, specific factors are uncorrelated with each other and with the general ability.

Reise et al. (2010) illustrated various latent structures that may emerge in different test designs. Figure 1 shows diagrams of a unidimensional model and three multidimensional models, where circles represent latent traits and squares represent items.

Figure 1.

Model A: unidimensional model; Model B: correlated traits model; Model C: second-order model; Model D: bifactor model. (Adapted from Reise et al., 2010.)





In Model A, each item reflects a single latent trait—typical of unidimensional tests. Model B assumes correlated traits, with each trait measured by a distinct set of items. Although dimensions are related, this model lacks a general factor influencing all items. Model C adds a higher-order trait that explains correlations among lower-order traits. The general trait affects items indirectly through these first-order dimensions, assuming they stem from a common source. Model D, the bifactor model, includes both a general factor linked to all items and orthogonal specific factors linked to subsets of items. The general factor captures shared variance, while specific traits explain residual variance (Reise et al., 2007; Reise et al., 2010).

Few studies have examined joint estimation of total and subscale scores (e.g., de la Torre & Song, 2009; de la Torre et al., 2011; Liu et al., 2019; Soysal & Kelecioğlu, 2018; Yao, 2010). Among latent trait model parameters, person ability is the most crucial, as it informs classification decisions (Sass et al., 2008; Zhang, 2010). Given the binary nature of most test decisions (e.g., pass/fail), classification accuracy should be evaluated in model comparisons (Chin, 2011). Yet, little evidence exists on the classification performance of models estimating total and subscores together (except Yao & Boughton, 2007).

Most previous studies used either dichotomous (e.g., de la Torre & Hong, 2010) or polytomous data (e.g., Wang, 2012), while many real tests combine both formats. Thus, examining mixed-format tests is critical (Lane, 2005; Shin, 2007; Yao & Schwarz, 2006). Also, the influence of item difficulty on reliability is well-established (Allen & Yen, 1979; Paek, 2015). Accordingly, the present study evaluates model performance under different test difficulties and proportions of polytomous items.

This study compares three models which are MIRT, HO-IRT, and Bifactor regarding their ability to jointly estimate total and subscale scores, focusing on estimation accuracy, reliability, and classification under varying conditions. These include levels of test difficulty, interdimensional correlation, and proportions of polytomous items. Real data from a university-developed English proficiency exam were also used. As testing practices evolve, both total and diagnostic subscale scores become essential for feedback to learners and stakeholders. Evaluating classification performance is key to understanding each model's utility in informing educational decisions.

Purpose of the Study

The study aims to compare ability estimation models (MIRT, HO-IRT, and Bifactor) in terms of estimation accuracy, reliability, and classification accuracy, using both simulated and real data. This study addresses the following research questions:

1. In simulation data, do RMSE, reliability, and classification accuracy values for ability estimates from the three models vary across test difficulty, interdimensional correlation, and proportion of polytomous items for both total and subscale scores?

2. In real data, do the estimation error, reliability, and classification accuracy of ability estimates differ across the three models for total and subscale scores?

Method

This study aimed to compare different models for estimating overall and subscores based on both simulation and real data analyses. Accordingly, this is a simulation study that adopts a descriptive research design.

Simulation Study

In the simulation study, data were generated for a mixed-format test consisting of 120 items across four distinct dimensions, with a sample size of 5,000 examinees. Three factors were manipulated during data generation: test difficulty, the proportion of polytomously scored items, and the correlation among dimensions. For each combination of conditions, 100 replications were conducted. Table 1 presents a summary of the fixed and manipulated conditions, as well as key information regarding the data generation and analysis procedures used in the simulation study.

Table 1.

Summary of simulation conditions, data generation, and analysis procedures

Condition	Description
Sample size	5,000
Number of dimensions	4
Number of items	30 items per dimension; 120 items in total
Correlation among dimensions	$r = 0.2$, $r = 0.5$, and $r = 0.8$
Percentage of polytomously scored items	5%, 10%, 25%, and 50%
Test difficulty	Difficulty parameters following normal distributions with $sd = 1$ and means of -2, -1, 0, 1, and 2 (very easy to very difficult tests)
Discrimination parameter	$a_i \sim U [0.5, 2.5]$
Difficulty/location parameter	$b_i \sim N (-2, 1)$, $b_i \sim N (-1, 1)$, $b_i \sim N (0, 1)$, $b_i \sim N (1, 1)$, $b_i \sim N (2, 1)$
Guessing parameter	$c_i \sim \text{beta} (6, 16)$
Ability parameter	$\theta_i \sim MVN (0, \Sigma)$
Model	M-3PL for dichotomous items; M-2PPCM for polytomous items
Replication	100
Item response generation	SimuMIRT (Yao, 2003)
Estimation models	MIRT, HO-IRT, and Bifactor models
Ability estimation	BMIRT (Yao, 2003, 2016)
Estimation algorithm	MCMC - Metropolis-Hastings algorithm
Number of iterations	15000 (3000 burn-in)
Evaluation criteria	RMSE, reliability, and classification accuracy
Analysis method	Factorial mixed ANOVA

Simulation Procedures

Item Parameters: a_i , b_i , and c_i were generated using uniform, normal, and beta distributions, respectively. The item parameters were generated using the WinGen3 software (Han & Hambleton, 2007). To define the relationship between item responses and latent traits, the three-parameter logistic model (3PL) was used for dichotomously scored items, and the two-parameter partial credit model (2PPCM) was used for polytomously scored items. The polytomous items had four response categories (0–1–2–3).

Ability Parameters. For the generation of true ability parameters for each subtest, a multivariate normal distribution was used: $\theta_i \sim MVN (0, \Sigma)$. The pairwise correlations among subtests were set to be equal, with three levels of inter-dimensional correlation. The ability parameters were generated using SimuMIRT (Yao, 2003).

Data Generation. A total of 6,000 datasets were generated, based on 60 simulation conditions, defined by a full factorial design: 3 levels of correlation among dimensions $\times$ 4 levels of the percentage of polytomously scored items $\times$ 5 levels of item difficulty. The item response patterns were generated using SimuMIRT.

Test Format. The test followed a mixed-format structure, with polytomously scored items concentrated in the last subtests.

Model Estimation. For the estimation of item and ability parameters based on the generated data, the BMIRT software (Yao, 2003, 2016) was used, employing the Metropolis-Hastings algorithm under the Markov Chain Monte Carlo (MCMC) framework. Data analyses were conducted using resources provided by the Turkish National e-Science e-Infrastructure (TRUBA; for details, see: www.truba.gov.tr).

The models used for estimating overall score and subscores, and the corresponding procedures applied to the study data are summarized below:

- **MIRT:** A simple-structure MIRT analysis was conducted to estimate individuals' ability scores across the four subtests. The overall ability score for the test was obtained using the maximum information method, based on the four subtest scores (Yao, 2010; Yao, 2012).
- **HO-IRT:** This model assumes a linear relationship between the overall score and the subscores. Accordingly, the ability parameters for the general and specific dimensions were estimated simultaneously.
- **Bifactor Model:** In the Bifactor model, the subdimensions are orthogonal to each other and to the general dimension—that is, no correlations among dimensions are assumed. This model produces five ability estimates: one for the general factor and four for the specific factors. As in the HO-IRT model, the ability estimates for both the overall score and the subscores are provided in a single output.

Evaluation. The accuracy of the estimated overall scores and subscores was evaluated using Root Mean Square Error (RMSE), the correlation between the estimated and true ability parameters (reliability), and classification consistency and accuracy. For classification consistency and accuracy, Rudner's index (Rudner, 2001, 2005) was computed using the *cacIRT* package (Lathrop, 2015). Since the correlation between classification consistency and classification accuracy values obtained under each condition was found to be very close to 1.00, only classification accuracy was considered in the comparative analyses of model performance.

Statistical Analysis. The effects of the independent variables and estimation models were tested using factorial mixed ANOVA. Separate analyses were conducted for each evaluation criterion: RMSE, reliability, and classification accuracy. In the factorial mixed ANOVA for overall scores, the simulation conditions, which are correlation among dimensions, percentage of polytomously scored items, and test difficulty, were treated as between-subject variables, while the estimation model (MIRT, HO-IRT, and Bifactor model) was treated as a within-subject variable.

For the ANOVA involving four independent variables, interpretation was based on the highest-order significant interaction effect. When the four-way interaction was significant, simple three-way interactions were examined at each level of the fourth variable. If a three-way interaction was significant, simple two-way interactions were analyzed at each level of one of the variables. When two-way interactions were significant, simple main effects were investigated. Simple main effect tests compared the means of one variable across all levels of another variable.

In the absence of significant interaction effects, main effects for the relevant variables were examined. For all significant main effects and simple main effects, pairwise comparisons were conducted using the Bonferroni correction to determine which group differences were statistically significant.

A similar factorial mixed ANOVA was conducted for the analysis of subscores. Unlike the analysis for overall scores, score type (i.e., subtests) was included as an additional variable. In this design, the

simulation conditions were treated as between-subject variables; the estimation model (MIRT, HO-IRT, and Bifactor model) was treated as a within-subject variable; and subtests were included as a blocking variable.

Partial eta squared (η_p^2) was reported as the measure of effect size. In interpreting the η_p^2 values, 0.01, 0.06, and 0.14 were considered indicators of small, medium, and large effects, respectively (Cohen, 1988).

Real Data Analysis

At this stage of the study, the aim was to apply the proposed estimation methods to a real test dataset. For this purpose, a mixed-format dataset from an English proficiency exam developed and administered by the School of Foreign Languages at a public university in Türkiye was used. The dataset consists of responses from a total of 5,085 examinees. Information about the exam is presented in Table 2.

Table 2.
Structure and Scoring of the Real Dataset

Subtest	Number and Type of Items	Scoring	Total score
Listening (S1)	30 items (dichotomously scored)	0 – 1	30
Reading (S2)	33 items (32 dichotomous, 1 polytomous item)	0 – 1 0 – 6	38
Vocabulary (S3)	20 items (dichotomously scored)	0 – 1	20
Writing (S4)	2 items (polytomously scored)	0 – 6	12
Total test	85 items (mixed-format test)		100

Before conducting further analyses, the dimensional structure of the test was evaluated. The planned four-dimensional structure of the test was tested through confirmatory analyses. A Confirmatory Factor Analysis (CFA) was conducted to assess model fit, and the fit indices, CFI, TLI (both > 0.95), and RMSEA (< 0.05), indicated perfect levels of fit. In addition, the multidimensionality of the test was supported by the Poly-DETECT analysis (Jang & Roussos, 2007; Zhang, 2007), yielding the following values: Detect = 0.461, Assi = 0.327, and Ratio = 0.747. McDonald's ω reliability coefficients, calculated from the CFA output, were as follows: 0.951 for Listening, 0.962 for Reading, 0.963 for Vocabulary, and 0.975 for Writing.

Table 3 presents the inter-dimensional correlations based on raw scores. The correlations among subscore variables were found to be moderate to high. The correlations between the Writing subtest and the other subtests were relatively lower. Correlations between the overall score and the Listening, Reading, and Vocabulary subscores were high, whereas the correlation between the overall score and the Writing subscore was lower and in the moderate range compared to the others.

Table 3.
Correlations between Subscores and the Overall Score in the Real Dataset

	Listening	Reading	Vocabulary	Writing	Overall
Listening	1				
Reading	0.780*	1			
Vocabulary	0.651*	0.691*	1		
Writing	0.501*	0.525*	0.402*	1	
Overall	0.900*	0.932*	0.829*	0.643*	1

* $p < .01$

In the real data analysis, the performance of the models was evaluated in three ways:

- Estimation accuracy was assessed using the Conditional Standard Error of Measurement (CSEM) calculated for each individual. Lower CSEM values indicate more accurate ability estimates.

- Reliability was calculated using the Bayesian marginal reliability coefficient (de la Torre & Patz, 2005).
- Classification performance was evaluated using Rudner's indices of classification accuracy and consistency.

In the English proficiency exam, the cut score for passing was 60, and raw score analysis showed that approximately 65% of examinees passed. For all three models, the cut point was determined based on the estimated ability scores, such that 65% of the examinees were classified as passing, in alignment with the empirical distribution in the raw data.

Results

Simulation Results

Overall Score Analyses. In the simulation analyses, the values of RMSE, reliability, and classification accuracy obtained for the overall score were used to compare the performance of the MIRT, HO-IRT, and Bifactor models. A significant four-way interaction with a large effect size was found among the variables: model, test difficulty, percentage of polytomously scored items, and correlation among dimensions. Therefore, the findings were further examined through three-way and two-way interaction effects. The relevant ANOVA results and pairwise comparison tables are presented in the Supplementary Material. The results reported in this section are based only on statistically significant differences.

Overall, the highest performance was observed with the MIRT model, which yielded the lowest RMSE values, the highest reliability, and the best classification accuracy across conditions. The HO-IRT model followed, while the Bifactor model showed the lowest performance under most conditions. In particular, when the correlation among dimensions was low ($r = 0.2$), the Bifactor model exhibited a notable increase in RMSE and a decrease in both reliability and classification accuracy.

The mean RMSE, reliability, and classification accuracy values calculated across the levels of independent variables are presented in Table 4. This table displays model performance for ability estimation across three models (MIRT, HO-IRT, and Bifactor), under varying combinations of test difficulty and proportion of polytomously scored items, for each level of correlation.

Table 4.
Mean RMSE, Reliability, and Classification Accuracy for Overall Scores across Levels of Correlation, Test Difficulty, and Proportion of Polytomously Scored Items (for All Three Models)

r	Model	Difficulty	Mean RMSE				Mean reliability				Mean classification accuracy			
			Proportion of Polytomously Scored Items				Proportion of Polytomously Scored Items				Proportion of Polytomously Scored Items			
			5%	10%	25%	50%	5%	10%	25%	50%	5%	10%	25%	50%
r = 0.2	MIRT	Very difficult	.314	.273	.255	.185	.897	.918	.939	.966	.877	.884	.891	.939
		Difficult	.211	.197	.161	.128	.929	.944	.970	.978	.869	.889	.917	.951
		Moderate difficulty	.164	.149	.121	.098	.944	.956	.978	.982	.875	.891	.937	.927
		Easy	.164	.151	.128	.113	.942	.955	.972	.979	.880	.904	.942	.932
		Very easy	.202	.192	.179	.155	.934	.941	.957	.969	.893	.900	.922	.949
	HO-IRT	Very difficult	.605	.565	.571	.554	.815	.811	.799	.821	.722	.725	.726	.736
		Difficult	.453	.450	.461	.459	.864	.855	.796	.825	.736	.737	.738	.742
		Moderate difficulty	.366	.365	.414	.382	.882	.870	.798	.846	.741	.742	.743	.747
		Easy	.374	.405	.427	.444	.893	.853	.816	.834	.743	.743	.745	.747
		Very easy	.471	.485	.525	.537	.854	.853	.809	.824	.738	.741	.742	.746
	Bifactor	Very difficult	.794	.617	.714	.677	.582	.739	.612	.666	.830	.754	.779	.768
		Difficult	.478	.513	.763	.500	.831	.777	.486	.789	.746	.751	.828	.753
		Moderate difficulty	.499	.698	.729	.480	.746	.571	.491	.744	.765	.840	.837	.762
		Easy	.646	.434	.540	.466	.635	.819	.696	.818	.825	.767	.784	.790
		Very easy	.491	.700	.547	.575	.825	.614	.783	.789	.785	.847	.788	.815
r = 0.5	MIRT	Very difficult	.343	.303	.287	.213	.911	.928	.942	.969	.898	.907	.915	.947
		Difficult	.235	.217	.178	.148	.946	.956	.973	.981	.903	.914	.938	.956
		Moderate difficulty	.181	.165	.135	.110	.962	.969	.982	.987	.912	.923	.950	.950
		Easy	.185	.172	.149	.136	.961	.968	.979	.974	.916	.932	.953	.947
		Very easy	.235	.226	.217	.192	.951	.955	.965	.974	.924	.929	.944	.956
	HO-IRT	Very difficult	.521	.487	.493	.478	.879	.887	.879	.905	.825	.830	.831	.842
		Difficult	.415	.412	.412	.417	.924	.922	.899	.913	.846	.848	.850	.854
		Moderate difficulty	.365	.355	.376	.366	.940	.939	.910	.928	.857	.858	.860	.863
		Easy	.362	.379	.384	.404	.945	.932	.920	.916	.863	.863	.866	.868
		Very easy	.422	.430	.459	.469	.925	.926	.908	.916	.864	.866	.868	.872
	Bifactor	Very difficult	.542	.495	.512	.512	.861	.879	.866	.893	.829	.825	.827	.837
		Difficult	.420	.423	.463	.438	.920	.914	.865	.906	.832	.836	.841	.841
		Moderate difficulty	.378	.400	.424	.385	.931	.912	.880	.915	.840	.847	.852	.847
		Easy	.385	.386	.403	.419	.932	.927	.909	.912	.852	.851	.856	.864
		Very easy	.425	.451	.467	.492	.921	.911	.904	.911	.858	.866	.863	.877
r = 0.8	MIRT	Very difficult	.352	.328	.296	.231	.927	.935	.953	.976	.904	.909	.928	.954
		Difficult	.252	.242	.191	.167	.960	.962	.978	.985	.931	.936	.952	.963
		Moderate difficulty	.201	.196	.151	.123	.972	.973	.985	.990	.943	.943	.960	.966
		Easy	.208	.208	.168	.157	.972	.972	.983	.987	.946	.950	.962	.968
		Very easy	.255	.256	.239	.219	.962	.962	.970	.978	.946	.948	.960	.968

Table 4 continues

HO-IRT	Very difficult	.399	.376	.365	.323	.920	.927	.935	.963	.874	.880	.885	.901
	Difficult	.300	.298	.284	.265	.956	.957	.958	.971	.898	.901	.904	.910
	Moderate difficulty	.257	.253	.250	.223	.968	.968	.967	.977	.909	.909	.912	.916
	Easy	.258	.271	.257	.253	.969	.966	.969	.975	.911	.912	.915	.918
	Very easy	.307	.313	.322	.313	.957	.957	.956	.967	.908	.910	.913	.918
Bifactor	Very difficult	.400	.366	.369	.343	.920	.933	.934	.961	.878	.884	.886	.901
	Difficult	.301	.295	.296	.279	.956	.958	.954	.970	.899	.903	.905	.910
	Moderate difficulty	.260	.254	.261	.228	.967	.968	.964	.976	.910	.912	.915	.918
	Easy	.260	.265	.262	.264	.968	.967	.967	.974	.916	.916	.920	.923
	Very easy	.307	.308	.324	.329	.956	.958	.956	.966	.915	.919	.920	.925

As the correlation among dimensions increased, the RMSE values of the Bifactor and HO-IRT models decreased, while their reliability and classification accuracy values increased. In contrast, the RMSE values of the MIRT model slightly increased in parallel with the correlation level. Nevertheless, reliability and classification accuracy values increased across all three models as the correlation increased. When the correlation level reached 0.8, the differences among the models were substantially reduced.

In terms of test difficulty, tests of moderate difficulty yielded the best results for all models with respect to RMSE, reliability, and classification accuracy. As the test became easier or harder, model performance generally deteriorated. Especially under very difficult test conditions, RMSE increased for all models, while reliability and classification accuracy noticeably decreased. In terms of classification accuracy, higher accuracy rates were obtained as the test became easier.

As the proportion of polytomously scored items increased, the performance of the MIRT model was particularly positively affected; RMSE values decreased, and reliability and classification accuracy improved. For the HO-IRT and Bifactor models, the situation was more complex: depending on test difficulty or correlation level, different effects were observed. Therefore, it is difficult to identify a consistent pattern for these models.

In cases where the test was either very easy or very difficult, and the proportion of polytomously scored items was low, no significant difference was observed between the HO-IRT and Bifactor models in terms of classification accuracy. Similarly, under certain conditions with high interdimensional correlation, the differences between the models were also found to diminish. However, when test difficulty approached a moderate level, and the proportion of polytomously scored items increased, performance differences among the models became more pronounced, and the superiority of the MIRT model became evident.

In conclusion, the simulation analyses indicate that the MIRT model generally yielded more stable and superior results for overall score estimation across most conditions. The HO-IRT model outperformed the Bifactor model, particularly under conditions with high interdimensional correlation, moderate test difficulty, and a moderate-to-high proportion of polytomously scored items. In contrast, the Bifactor model showed only limited and condition-specific advantages.

Subscore Results. In the analyses based on subscores, the effects of the estimation models (MIRT, HO-IRT, and Bifactor) and simulation conditions (test difficulty, interdimensional correlation, and proportion of polytomously scored items) on each dependent variable (RMSE, reliability, and classification accuracy) were examined separately using factorial mixed ANOVA. In these analyses, subtests were included as a block variable in the model. When four-way interaction effects were statistically significant, the analyses proceeded by examining the corresponding three-way and two-way interactions at lower levels. In cases where two-way interactions were significant, the analyses were further detailed with simple main effects and pairwise comparisons. The relevant ANOVA results and pairwise comparisons are provided in the Supplementary Materials. Findings are reported only based on statistically significant differences. To guide the reader, the mean RMSE, reliability, and classification

accuracy values calculated according to the levels of the independent variables are presented in Table 5 and Table 6.

Table 5.

Mean RMSE, reliability, and classification accuracy values for subscores across levels of correlation, test difficulty, and proportion of polytomously scored items (for all three models)

<i>r</i>	Model	Difficulty	Mean RMSE				Mean Reliability				Mean classification accuracy			
			proportion of polytomously scored items				proportion of polytomously scored items				proportion of polytomously scored items			
			%5	%10	%25	%50	%5	%10	%25	%50	%5	%10	%25	%50
<i>r</i> = 0.2	MIRT	Very difficult	.525	.505	.492	.412	.849	.858	.864	.901	0.825	0.835	0.840	0.872
		Difficult	.429	.409	.393	.344	.902	.911	.914	.932	0.873	0.882	0.888	0.905
		Moderate difficulty	.381	.368	.348	.296	.924	.928	.933	.950	0.903	0.906	0.911	0.927
		Easy	.388	.379	.348	.317	.921	.924	.935	.945	0.911	0.914	0.923	0.932
		Very easy	.444	.423	.416	.371	.894	.905	.907	.925	0.900	0.909	0.913	0.928
	HO-IRT	Very difficult	.528	.507	.494	.414	.847	.857	.862	.900	0.826	0.836	0.842	0.873
		Difficult	.431	.410	.394	.345	.902	.910	.913	.931	0.874	0.883	0.888	0.905
		Moderate difficulty	.382	.369	.349	.297	.923	.928	.933	.950	0.904	0.906	0.912	0.927
		Easy	.389	.380	.349	.318	.921	.924	.934	.945	0.911	0.915	0.923	0.932
		Very easy	.446	.425	.418	.373	.893	.904	.906	.924	0.900	0.909	0.913	0.928
	Bifactor	Very difficult	.615	.576	.573	.514	.781	.812	.811	.846	0.785	0.799	0.803	0.827
		Difficult	.513	.500	.518	.464	.857	.864	.842	.881	0.830	0.838	0.837	0.849
		Moderate difficulty	.482	.504	.485	.433	.875	.853	.865	.897	0.855	0.850	0.855	0.868
		Easy	.506	.476	.462	.439	.858	.879	.885	.897	0.858	0.860	0.867	0.869
		Very easy	.523	.528	.506	.478	.851	.845	.861	.876	0.851	0.857	0.858	0.867
<i>r</i> = 0.5	MIRT	Very difficult	.504	.485	.472	.397	.861	.870	.875	.910	0.833	0.843	0.848	0.878
		Difficult	.416	.397	.380	.333	.908	.916	.920	.936	0.877	0.885	0.891	0.907
		Moderate difficulty	.370	.357	.337	.288	.928	.932	.937	.953	0.905	0.907	0.912	0.927
		Easy	.374	.365	.336	.307	.927	.930	.939	.939	0.912	0.915	0.923	0.933
		Very easy	.424	.405	.398	.358	.905	.914	.915	.932	0.902	0.911	0.915	0.929
	HO-IRT	Very difficult	.507	.488	.474	.399	.860	.869	.874	.909	0.834	0.844	0.849	0.879
		Difficult	.417	.398	.381	.334	.908	.915	.919	.936	0.878	0.886	0.891	0.907
		Moderate difficulty	.371	.358	.338	.288	.928	.932	.937	.952	0.905	0.907	0.912	0.928
		Easy	.375	.366	.337	.308	.927	.929	.939	.939	0.912	0.915	0.923	0.933
		Very easy	.426	.407	.400	.360	.904	.913	.915	.931	0.903	0.911	0.915	0.929
	Bifactor	Very difficult	.756	.744	.742	.717	.651	.663	.664	.690	0.754	0.760	0.766	0.787
		Difficult	.713	.707	.708	.690	.701	.707	.704	.723	0.788	0.796	0.801	0.810
		Moderate difficulty	.697	.697	.691	.673	.718	.718	.722	.740	0.810	0.815	0.817	0.828
		Easy	.701	.694	.687	.677	.714	.721	.728	.730	0.817	0.817	0.824	0.831
		Very easy	.715	.712	.707	.697	.698	.701	.707	.717	0.807	0.814	0.815	0.825
<i>r</i> = 0.8	MIRT	Very difficult	.446	.434	.412	.346	.895	.900	.910	.938	0.860	0.867	0.874	0.900
		Difficult	.369	.359	.335	.295	.929	.933	.940	.953	0.894	0.900	0.906	0.920
		Moderate difficulty	.331	.327	.301	.259	.943	.944	.951	.963	0.916	0.916	0.923	0.936
		Easy	.336	.335	.302	.279	.942	.942	.952	.959	0.921	0.923	0.931	0.940
		Very easy	.376	.368	.356	.324	.926	.930	.935	.947	0.915	0.920	0.926	0.938

Table 5 continues

	HO-IRT	Very difficult	.449	.437	.415	.349	.893	.899	.908	.937	0.860	0.867	0.873	0.899
		Difficult	.370	.360	.336	.296	.929	.932	.939	.953	0.894	0.900	0.906	0.920
		Moderate difficulty	.332	.328	.302	.260	.943	.944	.951	.963	0.916	0.916	0.923	0.935
		Easy	.337	.336	.303	.280	.942	.942	.951	.959	0.922	0.923	0.931	0.940
		Very easy	.378	.370	.358	.326	.926	.930	.934	.947	0.916	0.921	0.926	0.938
	Bifactor	Very difficult	.925	.922	.923	.925	.410	.415	.416	.431	0.714	0.718	0.724	0.746
		Difficult	.920	.921	.923	.923	.436	.440	.438	.450	0.739	0.748	0.752	0.766
		Moderate difficulty	.921	.923	.922	.922	.446	.446	.449	.463	0.759	0.762	0.767	0.783
		Easy	.921	.920	.921	.922	.442	.446	.452	.459	0.762	0.765	0.773	0.784
		Very easy	.919	.921	.920	.924	.431	.434	.436	.444	0.752	0.760	0.762	0.776

Table 6.

Mean RMSE, reliability, and classification accuracy values by estimation model across levels of correlation, proportion of polytomously scored items, and subtest

<i>r</i>	Model	Subtest	Mean RMSE				Mean Reliability				Mean classification accuracy			
			Proportion of polytomously scored items				Proportion of polytomously scored items				Proportion of polytomously scored items			
			%5	%10	%25	%50	%5	%10	%25	%50	%5	%10	%25	%50
<i>r</i> = 0.2	MIRT	S1	.454	.455	.458	.465	.893	.893	.891	.888	0.878	0.881	0.879	0.875
		S2	.461	.443	.451	.444	.886	.895	.890	.895	0.875	0.881	0.878	0.879
		S3	.453	.452	.441	.241	.886	.886	.893	.970	0.869	0.868	0.877	0.948
		S4	.365	.317	.247	.242	.928	.947	.967	.969	0.907	0.927	0.946	0.948
	HO-IRT	S1	.456	.457	.460	.467	.892	.892	.890	.887	0.879	0.881	0.880	0.876
		S2	.463	.444	.453	.445	.885	.894	.889	.894	0.876	0.882	0.879	0.880
		S3	.455	.455	.443	.242	.885	.885	.892	.970	0.870	0.869	0.877	0.948
		S4	.367	.317	.247	.243	.927	.946	.967	.969	0.907	0.927	0.946	0.948
	Bifactor	S1	.629	.640	.662	.578	.784	.781	.764	.821	0.782	0.779	0.767	0.820
		S2	.526	.510	.515	.528	.849	.858	.855	.849	0.846	0.852	0.853	0.840
		S3	.518	.518	.505	.381	.849	.849	.858	.923	0.841	0.841	0.851	0.880
		S4	.438	.399	.352	.376	.895	.914	.933	.925	0.874	0.890	0.904	0.883
<i>r</i> = 0.5	MIRT	S1	.432	.432	.434	.438	.900	.900	.899	.896	0.880	0.882	0.881	0.878
		S2	.443	.427	.434	.426	.895	.903	.900	.903	0.878	0.884	0.881	0.883
		S3	.433	.433	.422	.237	.896	.896	.902	.969	0.875	0.874	0.881	0.949
		S4	.362	.315	.248	.244	.932	.949	.969	.969	0.910	0.929	0.947	0.950
	HO-IRT	S1	.434	.434	.436	.440	.899	.899	.898	.895	0.881	0.883	0.882	0.879
		S2	.445	.428	.436	.428	.894	.902	.899	.902	0.878	0.884	0.882	0.884
		S3	.435	.435	.424	.238	.895	.895	.901	.969	0.876	0.874	0.882	0.949
		S4	.363	.316	.249	.245	.932	.949	.969	.969	0.911	0.929	0.948	0.950
	Bifactor	S1	.745	.751	.760	.754	.666	.662	.651	.657	0.781	0.782	0.780	0.789
		S2	.728	.720	.727	.734	.689	.697	.691	.683	0.793	0.799	0.799	0.797
		S3	.714	.716	.711	.629	.692	.690	.696	.773	0.794	0.795	0.802	0.843
		S4	.680	.657	.630	.647	.739	.759	.781	.767	0.813	0.825	0.837	0.835
<i>r</i> = 0.8	MIRT	S1	.381	.380	.377	.375	.923	.924	.925	.928	0.897	0.898	0.899	0.899
		S2	.389	.377	.377	.368	.920	.925	.925	.931	0.895	0.900	0.899	0.902
		S3	.387	.386	.376	.230	.921	.922	.927	.975	0.895	0.895	0.901	0.954
		S4	.329	.315	.234	.230	.944	.949	.972	.975	0.918	0.927	0.949	0.951

Table 6 continues

HO-IRT	S1	.383	.382	.379	.377	.922	.923	.924	.927	0.897	0.899	0.899	0.899
	S2	.391	.379	.380	.369	.919	.924	.924	.930	0.895	0.900	0.899	0.901
	S3	.389	.388	.377	.231	.921	.921	.926	.975	0.895	0.895	0.901	0.954
	S4	.330	.316	.235	.231	.944	.949	.972	.974	0.918	0.928	0.949	0.951
Bifactor	S1	.929	.935	.941	.951	.412	.406	.396	.384	0.737	0.740	0.740	0.744
	S2	.925	.925	.933	.942	.421	.426	.415	.406	0.742	0.747	0.749	0.752
	S3	.932	.937	.938	.901	.419	.413	.414	.505	0.741	0.743	0.746	0.796
	S4	.898	.889	.876	.899	.479	.500	.527	.502	0.760	0.772	0.787	0.793

Throughout the analyses, the lowest RMSE and highest reliability values were generally observed under moderate and easy test difficulty conditions, whereas the highest RMSE and lowest reliability were found in very difficult tests. This pattern was particularly evident in the MIRT and HO-IRT models. Additionally, as the proportion of polytomously scored items increased, RMSE decreased, and reliability improved. The Bifactor model, in contrast, generally yielded higher RMSE and lower reliability values compared to the other models.

When the correlation level among subtests was 0.2 or 0.5, RMSE decreased and reliability increased across all estimation models as test difficulty approached moderate or easy levels. Among the models, MIRT was most positively affected by this pattern. The mean RMSE and reliability values in Table 5 clearly reflect this trend. A similar pattern was also observed for the HO-IRT model; although the differences between the two models were statistically significant in many cases, their mean values were quite close. For the Bifactor model, overall performance remained lower than that of the other two models; however, its results improved when the proportion of polytomously scored items increased, particularly under moderate and easy test difficulty conditions. Nevertheless, its average performance remained inferior compared to MIRT and HO-IRT. That said, under a correlation level of 0.2, the RMSE and reliability values of the Bifactor model were closer to those of MIRT and HO-IRT than under higher correlation conditions.

When the correlation level increased to 0.8, the RMSE values of the Bifactor model increased markedly, and its performance further deteriorated compared to the other two models. Similar to the patterns observed at lower correlation levels, the MIRT and HO-IRT models continued to exhibit comparable performance. Across all proportions of polytomously scored items and levels of test difficulty, RMSE values decreased for both MIRT and HO-IRT models as the correlation increased. Similarly, except for the condition with 50% polytomously scored items, reliability values increased with rising correlation levels for these two models across all difficulty levels.

At the subtest level, the first three subtests (which consisted entirely of dichotomously scored items at 5%, 10%, and 25% levels of polytomously scored items) yielded similar RMSE and reliability values under the MIRT and HO-IRT models. The fourth subtest, however, stood out with lower RMSE and higher reliability compared to the other three. When the proportion of polytomously scored items reached 50%, subtests 3 and 4, which are both composed entirely of polytomous items, demonstrated the lowest RMSE and highest reliability values. A similar pattern was observed in the Bifactor model; however, its performance across all subtests was still inferior to that of the other two models (see Table 6).

Similar patterns were identified in the classification accuracy analyses conducted at the subtest level. As the correlation level increased, classification accuracy values for the subtests improved under the MIRT and HO-IRT models, while these values tended to decline for the Bifactor model. Although the difference between MIRT and HO-IRT was statistically significant at each correlation level, the difference in mean classification accuracy values remained quite small. Across all correlation levels, the Bifactor model yielded the lowest classification accuracy. Regarding test difficulty, all three models produced higher classification accuracy when the test was easy or moderately difficult. Subtests 3 and 4, particularly under the condition where 50% of items were polytomously scored, yielded higher classification accuracy values across all models.

In conclusion, the subscore-level analyses largely mirrored the findings obtained from the overall score analyses. The MIRT model generally produced the lowest RMSE and the highest reliability and classification accuracy values, followed by the HO-IRT model, while the Bifactor model showed the weakest performance, particularly under high correlation conditions. The prominence of Subtest 4 under conditions with a high proportion of polytomous items is meaningful, given that it consists entirely of polytomously scored items. This highlights the potential for notable differences in estimation performance depending on the composition of the subtest.

Results of Real Data Analysis

In the real data analyses, the ability estimation performance of the MIRT, HO-IRT, and Bifactor models was evaluated at both the overall and subtest levels in terms of error, reliability, and classification accuracy.

Data from the English proficiency test, which includes Listening, Reading, Vocabulary, and Writing subtests, were analyzed using the MIRT, HO-IRT, and Bifactor models. For dichotomously scored items ($n = 82$) and polytomously scored items ($n = 3$), item parameters were estimated. Across all three models, the average difficulty and guessing parameters were comparable. However, the HO-IRT model yielded the highest average discrimination values, while the Bifactor model produced relatively lower discrimination estimates. The average item difficulty was around -0.5 across models, suggesting that the test was of moderate difficulty but slightly skewed toward the easier side. Guessing parameters for dichotomous items were also consistent across models, averaging approximately 0.19. The means and standard deviations of the ability parameters estimated using the MIRT and HO-IRT models are close to 0 and 1, respectively, whereas the standard deviations of the estimates obtained with the Bifactor model are lower.

The means and standard deviations of the conditional standard errors of the ability estimates obtained for each subtest and the overall test are presented in Table 7. The HO-IRT model produced the lowest error values for the subtests, followed by the MIRT model and then the Bifactor model. At the overall test level, the lowest error value was obtained from the MIRT model, while the HO-IRT model yielded the highest error value.

Table 7.

Mean and Standard Deviation of Conditional Standard Error of Ability Estimates for Overall Test and Subtests in Real Data Analysis

Model	Subtests				Total test (85 items)
	Listening (30 items)	Reading (33 items)	Vocabulary (20 items)	Writing (2 items)	
	Mean (sd)	Mean (sd)	Mean (sd)	Mean (sd)	
MIRT	0.432 (0.050)	0.404 (0.073)	0.445 (0.085)	0.522 (0.047)	0.252 (0.029)
HO-IRT	0.320 (0.038)	0.290 (0.042)	0.352 (0.060)	0.452 (0.050)	0.295 (0.035)
Bifactor	0.825 (0.042)	0.811 (0.066)	0.739 (0.075)	0.752 (0.037)	0.288 (0.035)

Whether the differences in errors were statistically significant was tested using repeated measures ANOVA. According to the results, the conditional standard errors of ability estimates significantly differed across models at both the subtest and overall test levels ($p < .001$). The effect sizes were particularly large for the subtests, with partial eta-squared (η_p^2) values ranging from .978 (Writing) to .992 (Listening), indicating that the choice of model had a substantial impact on estimation precision. At the overall test level, the effect size was lower ($\eta_p^2 = .759$), though still large. Pairwise comparisons showed that all three models produced significantly different error values from one another. Specifically, in the subtests, the MIRT and HO-IRT models yielded considerably lower standard errors than the Bifactor model, suggesting superior estimation accuracy. For the overall test, the MIRT model produced the lowest standard error, indicating it provided the most precise ability estimates among the three.

The reliability and classification accuracy values of the ability estimates were presented in Table 8. At the overall test level, the highest reliability coefficient was obtained with the MIRT model, and the

values of all three models were found to be quite similar. At the subtest level, the MIRT and HO-IRT models produced high reliability, whereas the reliability coefficients of the Bifactor model were considerably lower. In addition, the scores obtained from the writing subtest, which consists of only two items, have lower reliability compared to the others. The highest classification accuracy values were generally obtained with the HO-IRT model. However, the values obtained with the HO-IRT and MIRT models were quite close to each other. The Bifactor model, on the other hand, yielded significantly lower classification accuracy both at the overall test level and in the subtests.

Table 8.

Marginal Reliability Coefficients and Classification Accuracy Values for the Real Data Analysis

	Model	Subtests				Total test (85 items)
		Listening (30 items)	Reading (33 items)	Vocabulary (20 items)	Writing (2 items)	
Reliability	MIRT	0.856	0.874	0.841	0.778	0.934
	HO-IRT	0.896	0.914	0.873	0.794	0.912
	Bifactor	0.312	0.332	0.448	0.432	0.916
Classification Accuracy	MIRT	0.898	0.913	0.882	0.865	0.931
	HO-IRT	0.913	0.924	0.895	0.870	0.920
	Bifactor	0.860	0.703	0.726	0.745	0.808

The analyses conducted with real data largely align with the simulation findings. At the subtest level, the Bifactor model yielded higher estimation errors and lower reliability and classification accuracy. The MIRT model provided the lowest estimation error and the highest reliability coefficient for the overall test, and produced results quite similar to the HO-IRT model for the subtests. The HO-IRT model, on the other hand, demonstrated particularly strong performance in terms of classification accuracy. These findings indicate that the choice of scoring model can significantly affect measurement accuracy and decision-making processes.

Discussion

This study compared the performance of MIRT, HO-IRT, and Bifactor models using both simulation and real data to support the joint reporting of overall and subscale scores in multidimensional tests. Accuracy, reliability, and classification accuracy of ability estimates were evaluated. In the simulation study, test difficulty, inter-dimensional correlation, and proportion of polytomous items were manipulated.

The goal was to identify models that reliably estimate both overall and diagnostic subscale scores. Subscores are valuable for tracking progress and identifying strengths and weaknesses (Haladyna & Kramer, 2004), but both overall and subscale scores must be valid and reliable for meaningful interpretation (Yao, 2012). All three models consider multidimensionality beyond unidimensional approaches. When subdomains are highly correlated, modeling this relationship enhances estimation accuracy (de la Torre & Patz, 2005), as supported by prior research (de la Torre et al., 2011; DeMars, 2005; Shin, 2007; Wang et al., 2004).

The real test used in this study, an English proficiency exam, measures general proficiency and includes listening, reading, vocabulary, and writing subdomains. While a unidimensional analysis is possible, multidimensional modeling explains item responses better (Reckase & Xu, 2015), and MIRT particularly aids in diagnosing subskills. Modeling correlated subscores separately may contradict language acquisition theories (Longabach & Peyton, 2018); thus, models accounting for subdomain relationships are preferable.

Simulation results showed that MIRT and HO-IRT yielded low RMSE and high reliability/classification accuracy across conditions, with comparable performance (de la Torre et al., 2011; Yao, 2010; Soysal, 2017). The Bifactor model performed weakest, with higher errors and lower reliability and classification accuracy. Liu et al. (2019) discouraged its use for reporting subscores, recommending alternative scoring approaches. Real data findings confirmed these patterns: HO-IRT and MIRT produced the most accurate

subscore estimates, while Bifactor lagged. For the overall score, MIRT achieved the lowest RMSE and highest reliability/classification accuracy. Both MIRT and HO-IRT effectively estimated general ability. Yao (2010) found MIRT yielded the lowest RMSE and highest reliability using the maximum information method. Similarly, de la Torre and Song (2009) showed that HO-IRT accurately estimated general ability when modeling total and subscores jointly.

Model differences in overall score estimation were more pronounced at low correlations but diminished as the correlation increased, consistent with Yao (2010). Conversely, for subscores, Bifactor's performance declined with increasing correlation (Soysal, 2017). In real data where correlation was high (0.75), HO-IRT and MIRT produced similar, accurate, and reliable subscore estimates (de la Torre et al., 2011). Bifactor generally produced more error and lower reliability/classification accuracy. This was especially evident under low correlation for total scores and high correlation for subscores. MIRT and HO-IRT were more suitable for reporting general and sub-abilities together. Alternative scoring approaches within the Bifactor framework may improve performance (DeMars, 2013; Liu et al., 2019), but the classical Bifactor model used here showed the weakest results.

One key factor was inter-dimensional correlation. In simulations, defined as 0.2, 0.5, and 0.8, and high in real data, increasing correlation improved HO-IRT and Bifactor model performance in total scores. MIRT showed smaller changes. At the subscore level, MIRT and HO-IRT improved with higher correlation, while Bifactor's performance worsened. These results are relevant for tests with related subdomains, such as language skills (Alderson, 2007; Chapelle, 2011; Solano-Flores & Trumbull, 2003). Previous studies also supported MIRT's suitability for highly correlated structures (de la Torre, 2008, 2009; de la Torre & Patz, 2005; Hong et al., 2010; Sheng & Wickle, 2008; Yao & Boughton, 2007). Under HO-IRT, accuracy improves significantly with higher correlation (de la Torre & Hong, 2010). For Bifactor, increased correlation led to higher RMSE and poorer subscore performance (Liu et al., 2019; Soysal, 2017; Yao, 2010).

Another variable was the proportion of polytomous items. Simulations showed that higher proportions, especially under MIRT, reduced RMSE and improved reliability/classification accuracy. While the pattern varied for HO-IRT and Bifactor depending on other conditions, all models performed better at the subscore level with more polytomous items. Subtests with more such items had higher reliability and lower error. Similar findings were reported by Longabach & Peyton (2018). Polytomous items provide richer information and enhance reliability (Ercikan et al., 1998; Shin, 2007). Item format can also increase multidimensionality, which should be considered when interpreting test scores (Traub, 1993; Yao & Schwarz, 2007; Lee et al., 2020). In the real data analysis, the limited number of polytomous items likely led to lower reliability and classification accuracy for the subtest in which they appeared, consistent with prior findings on item count effects (Bulut, 2013; Yao, 2010; Reckase & Xu, 2015).

Test difficulty also influenced estimation. Simulations showed that all models performed best (lowest RMSE and highest reliability) under moderate difficulty. Easy or difficult tests reduced performance, especially under high difficulty. Classification accuracy was higher in easier conditions. The same pattern held for subscores. As correlation increased, differences between difficulty levels diminished. These results align with the literature. Paek (2015) found higher reliability under moderate or easy difficulty. Prior studies also reported better estimation when item difficulty matched ability (Embretson, 2007; Morse et al., 2012; Yao, 2012). For example, Yao (2012) found that subdomains with difficulty averages near the ability mean had higher reliability. In this study, the best results occurred when mean ability was near zero, and test difficulty was moderate, where item difficulty also averaged near zero. Real data confirmed this, with a mean ability close to zero and moderate test difficulty.

In summary, this study offered a detailed comparison of models used to estimate general and subdomain scores in multidimensional tests. Results showed that MIRT and HO-IRT models are particularly effective for reporting subscores in highly correlated tests, while MIRT consistently outperformed others for the overall score. Selecting models sensitive to test features and dimensional relationships is critical for producing accurate and diagnostic results for informed decision-making.

Declarations

Ethical Approval: Ethical approval was obtained from the Applied Ethics Research Center of the public university from which the real data used in the study were obtained.

Author Contribution: This study was derived from the first author's doctoral dissertation, conducted under the supervision of the second author.

Funding: The authors received no financial support for the research, authorship, and/or publication of this study.

Consent to Participate: This study used secondary data. For the achievement test from which the real data were obtained, participants were provided with an option prior to the examinations to give consent for their exam data to be used for scientific research purposes.

Consent to Publish: The institution from which the data were obtained did not grant permission for its name to be disclosed. Accordingly, the institution is referred to as a "public university" throughout the manuscript.

Competing Interests: The authors declare that they have no competing interests.

References

- Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, 13(2), 113-127. <https://doi.org/10.1177/014662168901300201>
- Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, 7(4), 255-278. https://doi.org/10.1207/s15324818ame0704_1
- Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22(3), 37-51. <https://doi.org/10.1111/j.1745-3992.2003.tb00136.x>
- Alderson, J. (2007). The challenge of (diagnostic) testing: Do we know what we are measuring? In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 21-40). University of Ottawa Press.
- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Brooks/Cole.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. AERA.
- Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9(1), 37-48. <https://doi.org/10.1177/014662168500900104>
- Brennan, R. L. (2012). *Utility indexes for decisions about subscores*. Center for Advanced Studies in Measurement and Assessment (CASMA). Research Report, 33.
- Bulut, O. (2013). *Between-person and within-person subscore reliability: Comparison of unidimensional and multidimensional IRT models* (Doctoral Dissertation). University Of Minnesota.
- Chalmers, R. P. (2012). A Multidimensional item response theory package for the R Environment. *Journal of Statistical Software*, 48(6), 1-29. <https://doi.org/10.18637/jss.v048.i06>
- Chang, Y.-W., & Davison, M. L. (1992, April). *A comparison of unidimensional and multidimensional IRT approaches to test information in a test battery*. Paper presented at the annual meeting of the American Educational Research Association, San-Francisco, CA, United States.
- Chapelle, C. (2011). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 21-33). Routledge.
- Chin, T. Y. (2011). *Accuracy and robustness of diagnostic methods: Comparing performance across domain score, multidimensional item response, and diagnostic categorization models* (Doctoral Dissertation). University Of Nebraska.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
- de la Torre, J. (2008). Multidimensional scoring of abilities: The ordered polytomous response case. *Applied Psychological Measurement*, 32(5), 355-370. <https://doi.org/10.1177/0146621607303784>

- de la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement*, 33(6), 465-485. <https://doi.org/10.1177/0146621608329890>
- de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement*, 34(4), 267-285. <https://doi.org/10.1177/0146621608329501>
- de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, 30(3), 295-311. <https://doi.org/10.3102/10769986030003295>
- de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, 33(8), 620-639. <http://doi.org/10.1177/0146621608326423>
- de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscore. *Applied Psychological Measurement*, 35(4), 296-316. <http://doi.org/10.1177/0146621610378653>
- DeMars, C. E. (2005, August). *Scoring subscales using multidimensional item response theory models*. Poster presented at the Annual Meeting of the American Psychological Association, Washington, DC, United States.
- DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, 13(4), 354-378. <https://doi.org/10.1080/15305058.2013.799067>
- Embretson, S. E. (2007). Impact of measurement scale in modeling development processes and ecological factors. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 63-87). Lawrence Erlbaum Associates Publishers.
- Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, 35(2), 137-154. <https://doi.org/10.1111/j.1745-3984.1998.tb00531.x>
- Fan, F. (2016). *Subscore reliability and classification consistency: a comparison of five methods* (Doctoral Dissertation). University of Massachusetts Amherst.
- Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33(2), 204-229. <https://doi.org/10.3102/1076998607302636>
- Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, 75(2), 209-227. <https://doi.org/10.1007/S11336-010-9158-4>
- Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation & the Health Professions*, 27(4), 349-368. <https://doi.org/10.1177/0163278704270010>
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
- Hong, Y., Lam, T. C., & de la Torre, J. (2010). *Ancillary variables and multidimensional scoring of polytomous responses*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO, United States.
- Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, 44(1), 1-21. <https://doi.org/10.1111/j.1745-3984.2007.00024.x>
- Kahraman, N., & Kamata, A. (2004). Increasing the precision of subscale scores by using out-of-scale information. *Applied Psychological Measurement*, 28(6), 407-426. <https://doi.org/10.1177/0146621604268736>
- Lane, S. (2005, April). *Status and future directions for performance assessments in education*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.
- Lathrop, Q. (2015). Practical Issues in Estimating Classification Accuracy and Consistency with R Package cacIRT. *Practical Assessment, Research, and Evaluation*, 20(1), Article 18. <https://doi.org/10.7275/43vm-p442>
- Lee, W. C., Kim, S. Y., Choi, J., & Kang, Y. (2020). IRT approaches to modeling scores on mixed-format tests. *Journal of Educational Measurement*, 57(2), 230-254. <https://doi.org/10.1111/jedm.12248>
- Li, C. (2018). *Subscore reporting for double-coded innovative items embedded in multiple contexts* (Doctoral Dissertation). Graduate School of the University of Maryland, College Park.
- Liu, Y., Li, Z., & Liu, H. (2019). Reporting valid and reliable overall scores and domain scores using bi-factor model. *Applied psychological measurement*, 43(7), 562-576. <https://doi.org/10.1177/0146621618813093>
- Longabach, T. (2015). *A comparison of subscore reporting methods for a state assessment of English language proficiency* (Doctoral Dissertation). University of Kansas.

- Longabach, T., & Peyton, V. (2017). A comparison of reliability and precision of subscore reporting methods for a state English language proficiency assessment. *Language Testing*, 35(2), 297-317. <http://doi.org/10.1177/0265532217689949>
- Morse, B. J., Johanson, G. A., & Griffeth, R. W. (2012). Using the graded response model to control spurious interactions in moderated multiple regression. *Applied Psychological Measurement*, 36(2), 122-146. <https://doi.org/10.1177/0146621612438725>
- Nandakumar, R. (1994). Assessing dimensionality of a set of item responses-comparison of different approaches. *Journal of Educational Measurement*, 31(1), 17-35. <https://doi.org/10.1111/j.1745-3984.1994.tb00432.x>
- Paek, I. (2015). An investigation of the impact of guessing on coefficient α and reliability. *Applied Psychological Measurement*, 39(4), 264-277. <https://doi.org/10.1177/0146621614559516>
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9(4), 401-412. <https://doi.org/10.1177/014662168500900409>
- Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21(1), 25-36. <https://doi.org/10.1177/0146621697211002>
- Reckase, M. D. (2009). *Multidimensional item response theory (statistics for social and behavioral sciences)*. Springer.
- Reckase, M. D., & Xu, J. R. (2015). The evidence for a subscore structure in a test of English language competency for English language learners. *Educational and Psychological Measurement*, 75(5), 805-825. <https://doi.org/10.1177/0013164414554416>
- Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, 92(6), 544-559. <https://doi.org/10.1080/00223891.2010.496477>
- Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16(1), 19-31. <https://doi.org/10.1007/s11136-007-9183-7>
- Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment Research & Evaluation*, 7(1). <https://doi.org/10.7275/an9m-2035>
- Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment Research & Evaluation*, 10(1). <https://doi.org/10.7275/56a5-6b14>
- Sass, D. A., Schmitt, D. A., & Walker, C. M. (2008) Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education*, 21(1), 65-88. <https://doi.org/10.1080/08957340701796415>
- Sheng Y. & Wikle C. K. (2007). Comparing multi-unidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, 68(6), 413-430. <https://doi.org/10.1177/0013164406296977>
- Shin, D. (2007). *A comparison of methods of estimating subscale scores for mixed-format tests* (Pearson Educational Measurement Research Report).
- Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, 47(2), 150-174. <https://doi.org/10.1111/j.1745-3984.2010.00106.x>
- Sinharay, S., Haberman, S. J., & Wainer, H. (2011). Do adjusted subscores lack validity? Don't blame the messenger. *Educational and Psychological Measurement*, 71(5), 789-797. <http://doi.org/10.1177/0013164410391782>
- Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, 32(2), 3-13. <https://doi.org/10.3102/0013189X032002003>
- Soysal, S. (2017). A comparison of the estimation of total test scores and subtest scores using hierarchical item response theory models [Toplam Test Puanı ve Alt Test Puanlarının Kestiriminin Hiyerarşik Madde Tepki Kuramı Modelleri ile Karşılaştırılması] (Doctoral Dissertation). Hacettepe University
- Soysal, S., & Kelecioğlu, H. (2018). A comparison of the estimation of total test scores and subtest scores using hierarchical item response theory models [Toplam test ve alt test puanlarının kestiriminin hiyerarşik madde tepki kuramı modelleri ile karşılaştırılması]. *Journal of Measurement and Evaluation in Education and Psychology*, 9(2), 178-201. <https://doi.org/10.21031/epod.404089>
- Traub, R. E. (1993). On the equivalence of traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 29-44). Lawrence Erlbaum.
- Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement*, 37(2), 113-140. <https://doi.org/10.1111/j.1745-3984.2000.tb01079.x>

- Wang, H. (2012). *A Monte Carlo comparison of polytomous item estimation based on higher-order item response theory models versus higher-order confirmatory factor analysis models* (Doctoral Dissertation). University of Pittsburgh.
- Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, 9(1), 116–136. <http://doi.org/10.1037/1082-989X.9.1.116>
- Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and non-compensatory two-dimensional data on unidimensional IRT estimation. *Applied Psychological Measurement*, 12(3), 239–252. <https://doi.org/10.1177/014662168801200303>
- Wedman, J., & Lyrén, P. E. (2015). Methods for examining the psychometric quality of subscores: A review and application. *Practical Assessment, Research, and Evaluation*, 20(1), 21. <https://doi.org/10.7275/ng3q-0d19>
- Yao, L. (2003). *BMIRT: Bayesian multivariate item response theory* [Computer Software]. Defense Manpower Data Center.
- Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement*, 47(3), 339–360. <http://doi.org/10.1111/j.1745-3984.2010.00117>
- Yao, L. (2012). Multidimensional item response theory for score reporting. In Y. Cheng, & H.- H. Chang (Eds.) *Advances in modern international testing: Transition from summative to formative assessment*. Information Age.
- Yao, L. (2016). *The BMIRT toolkit*.
- Yao, L., & Boughton, K. A. (2007). A Multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, 31(2), 83–105. <http://doi.org/10.1177/0146621606291559>
- Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, 30(6), 469–492. <https://doi.org/10.1177/0146621605284537>
- Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, 27(1), 119–140. <https://doi.org/10.1177/0265532209347363>
- Zhang, J. (2007). Conditional covariance theory and detect for polytomous items. *Psychometrika*, 72(1), 69–91. <http://doi.org/10.1007/s11336-004-1257-7>