

Integrating Multidimensional Item Response Theory and Explainable Machine Learning to Enhance Measurement Precision, Assessment Fairness, and Predictive Validity in Educational Testing: An R-Based Psychometric Framework

Riyadh Ahmed Saleh AL-Derabseh

Assistant Professor

Faculty of Educational Sciences

Irbid National University

Email: rd.Derabseh@gmail.com

Abstract

Educational Testing (ET) in Middle East & North Africa (MENA) region is changing at a very rapid speed in compliance with the National prophecy towards Knowledge Based economy and Human Capital Strategic Risk Management. However, the multidimensional nature of cognition is challenging for classic models of unidimensional psychometrics and black box approaches for machine learning (ML) algorithms that also black music in the process of decision-making, leading to the risk of equality and strategic integration. The new and integrated psychometric study is born to reduce the imprecision of the measurement, assess the fairness and extrapolate the validity: Multidimensional Item Response Theory (MIRT) combined with Explainable Machine Learning (XAI) is proposed. Building upon a large secondary data set containing 50,000 students from five Middle Eastern countries (Saudi Arabia, United Arab Emirates, Egypt, Jordan and Qatar), we exploited an R-based computational pipeline. MIRT is used for accurate mapping of latent cognitive traits and XAI (SHAP and LIME) to improve the understanding of ML accurate prediction of student academic trajectory. Results indicate a dramatic improvement of the integration of MIRT-XAI with the traditional unidimensional and black-box ML approach, which demonstrates a 34% reduction in Differential Item Functioning (DIF) and improved predictive validity (AUC-ROC = 0.892). Five hypotheses were tested and all of them were favored, revealing the mediation effect of explainability and the importance of the multidimensional modeling to obtain cultural and demographic fairness in the MENA region. This paper presents a comprehensive and replicable R education model that has theoretical and practical effects in the field of schooling for policy makers and psychometricians and educational strategic management professionals in the region.

Keywords: Multidimensional Item Response Theory (MIRT); Explainable Machine Learning (XAI); Educational Testing; Assessment Fairness; Psychometrics; R Software; Middle East; Risk Management.

1. Introduction

1.1. Background and Context

Educational examination is suffering an earthquake like transformation today in the Middle East and North Africa (MENA) region. Ambitious national agendas like Saudi Vision 2030, the UAE Centennial 2071, and Egypt's Vision 2030 have accompanied in a new era of thought on the importance of human capital development, STEM education, and standardized assessment (Hossain et al., 2025; Abu Huson et al., 2024). National secondary exams and university entrance tests are among the high-stakes tests that are the porters to education and occupational mobility. Ministries of education and stakeholders in society are therefore highly fascinated by the psychometric properties of these assessments, such as precision, fairness and predictive validity. In recent years, mixing smart management tools and

advanced analytics in education is claimed to be an ultimate part of sustainable development and risk management in emerging economies (Abdallah Al Qudah et al., 2026; Al Qudah et al., 2026).

Yet, despite these reforms, many countries in the MENA region restart to have psychometric infrastructure that is either based on Classical Test Theory (CTT) or unidimensional Item Response Theory (IRT). Unidimensional IRT is an important improvement over CTT, but it ultimately assumes that only one latent variable (e.g., general mathematical ability) influences a test's performance. Cognitive processes are multidimensional in nature. For example, the mathematical performance in a complex test by a student is not only concluded by numerical skills, but also by spatial imagination and understanding of texts (Baker & Kim, 2004). However, ignoring this multidimensionality may lead to parameter estimates that only capture part of the picture, to less precise measurement, and to a failure to make reasonable inferences about the cognitive processes being measured, which may compromise the trustworthiness and stability of the estimated processes (Elder et al., 2024).

While psychometric modeling expands, schools have begun to apply ML methods to predict student performance and identify students at risk to direct the allocation of resources more effectively. Most successful ML algorithms, however, are rather black-box models like Deep Neural Network, Gradient Boosting Machines. The practical and moral issues of opacity for these models in high-stakes education are large. Stakeholders (students, parents, educators) have a right to know why a convinced demographic group of students in the Middle East is likely to be underperforming due to a group of students that is demonstrated to. Historical biases may unintentionally be persisted in opaque models, subsequent in algorithmic unfairness and producing mistrust in the educational system (Gunasekara & Saarela, 2025; Clark et al., 2024). When it comes to cognitive and neuropsychological assessment, the interpretation of the predictions is fundamental to ensure that AI does not be used in an injurious or misdiagnosing way, which would affect the potential of students, as stated in Medenica et al. (2024) and Westbye et al. (2024).

A fusion of advanced psychometrics (MIRT) with Explainable Machine Learning (XAI) has not been thoroughly studied in large-scale educational testing, uniquely in the MENA region. In the literature it is normally done as psychometric modeling or predictive ML, where they are treated separately. Psychometricians emphasize latent trait estimation, and data scientists emphasize predictive accuracy, and are often coached to surrender interpretability for high-stakes educational decisions (Thompson & Garcia, 2024). In addition, there is a great gap in open-source and reproducible computational models of joining the two fields, specifically around regional risk management and strategic educational planning (Al Qudah et al., 2025; Aljawarneh et al., 2025).

This study will seek to:

1. Develop and validate a mixed psychometric framework that combines MIRT for precise latent trait estimation with XAI for transparent predictive modeling.
2. Evaluate the impact of this background on measurement precision, assessment fairness (specifically regarding Differential Item Functioning across MENA demographics), and predictive validity.
3. Provide a fully reproducible, R-based computational pipeline that can be adopted by educational testing centers in the Middle East and globally, aligning with broader AI-readiness and data management recommendations (Clark et al., 2024; Andersen & Zehner, 2021).

2. Theoretical Background and Literature Review

2.1. Multidimensional Item Response Theory (MIRT)

The theory of Item Response Theory (IRT) adjusted the face of psychometrics, enabling the probability of a particular response to a latent trait. But in many educational tests, a test will not have a single dimension. Multidimensional Item Response Theory (MIRT) is a conservatory of the IRT framework for the multiple simultaneous latent traits (Baker & Kim, 2004; Paganin et al., 2023).

The probability of a correct answer in MIRT is handled as a function of a vector of latent traits (θ) and a vector of item parameters. In this case, the compensatory MIRT model suggests that a high level of one trait can make up for a low level of another. Mathematically, for a dichotomous item, the probability of a correct response is often modeled using a multidimensional expansion of the 2-Parameter Logistic (2PL) model:

$$P(X_{ij} = 1 | \theta_j) = \frac{1}{1 + \exp(-a_i^T \theta_j + d_i)}$$

In which a_i is the item discrimination vector, θ_j is the underlying trait vector for person j , and d_i is the item difficulty parameter. However, recent developments in computational approaches and complex semiparametric IRT models have greatly greater the estimation of these models and enable more accurate modeling of non-normal latent trait distributions and complex covariance structures (Paganin et al., 2023). MIRT offers a more sophisticated and accurate description of student ability by modeling the covariance structure among latent traits; this is mainly important in the Middle East for addressing the gradually complex cognitive demands of modern curricula.

2.2. Explainable Machine Learning (XAI) in Psychometrics

The more complex the ML model, the less interpretable it is. Reasonable AI (XAI) includes strategies that aim to render the findings of black-box models comprehensible to humans. SHAP and LIME are two popular model-agnostic XAI techniques.

Cooperative game theory as the basis of SHAP grants an importance value for every feature for a specific prediction by measuring the marginal contribution of each feature for all possible combinations of features. For entrepreneurial education test applications, SHAP can be used to discover the relative importance of specific MIRT-derived latent traits, demographic variables, and test taking behaviors that contribute to a student's predicted academic outcome. LIME, however, locally models the complex model by an explainable linear model.

Recent systematic literature reviews on explainable learning analytics and educational data mining highlight that the use and implementation of ML in educational testing are horizontal to procedural justice and ethical practices of using AI in education (Gunasekara & Saarela, 2025; Westbye et al., 2024). With the help of XAI, psychometricians can get a deeper recognizing of what the system is predicting and why, which can make the assessment system more transparent and fair. Additionally, the combination of cognitive diagnosis models and gradient boosting is a novelty in the field of individualized assessment in the learning environment, which demonstrates that it is possible to realize

high predictive accuracy while simultaneously maintaining the interpretability of the model (Thompson & Garcia, 2024).

2.3. Assessment Fairness and Differential Item Functioning (DIF)

The fairness of the assessment is an integral part of the validity of education tests. Differential Item Functioning (DIF) is when two groups of people with the same level of ability have altered chances of getting an item right, simply due to their membership in the two groups (e.g. men and women, or two countries, or two socio-economic groups). In the context of the Middle East, where varied students with varied linguistic, cultural, and socio-economic backgrounds exist, DIF is one of the major issues relating validity (Dorans & Kulick, 2006).

Traditional approaches to detect DIF, which comprise the Mantel-Haenszel procedure or logistic regression, tend to be testing with multidimensional data and multiple covariates. Unidimensional matching criteria may result in detecting DIF when learning styles and cognitive processes are unique among groups (Sewall, 1988; Hamidifar, 2012). MIRT combined with XAI, which can show and model shocked non-linear DIF, can help researchers identify and control DIF, performing tests fair for a great spectrum of populations in MENA.

2.4. The Middle Eastern Educational and Strategic Context

MENA region has a unique context with interests to psychometric research. As higher education enlarges, the rise in expatriate groups, and a transfer from rote examining to competency-based education, there is a need for powerful, fair and transparent assessment tools. Secondary assessment data assumed nationwide in Saudi Arabia, UAE, Egypt, Jordan and Qatar is a very rich data set that can be treated in future to test enhanced psychometric models.

Moreover, AI and advanced analytics adoption in the area relates to general management and risk management schemes. The employment of AI tools in risk management and business intelligence for sustainable development goals (SDGs) and educating business practices in the Middle East and North Africa (MENA) regions were accentuated by fresh bibliometric analysis and exploratory studies (Al Qudah et al., 2024 & 2025; Huson et al., 2024). The smart use of these administration tools for the educational assessment system ensures the human capital development is aligned with the strategic objectives of the country, plunging the danger of biased and opaque decision-making by algorithms (Abdallah Al Qudah et al., 2026; Seitan Al-Smadi and Mustafa Abdallah Al Qudah, 2025).

3. Hypotheses Development

The theoretical constructs and literature gap analysis conducts to the following five hypotheses which are offered and tested in this study:

Hypothesis 1 (H1): Measurement Precision

The fit of multidimensional Item Response Theory (MIRT) models to the compound educational quizzing data in Middle East will be significantly better and the precision of measurement will be developed than the fit of the traditional unidimensional IRT models to the same data.

Rationale: Educational constructs are multidimensional. Parameter estimates are biased due to a compromise imposed on the one-dimensional model. MIRT accounts for the covariance between traits, which gives better estimates of underlying traits (Baker & Kim, 2004; Paganin et al., 2023).

Hypothesis 2 (H2): Predictive Validity

Models learned from MIRT derived latent trait scores will show significantly more predictive validity with respect to student academic outcomes than models learned from raw sum scores or unidimensional IRT scores.

Rationale: MIRT scores report more student abilities and are more comprehensive. These multidimensional scores have been used as strengths to ML algorithms and keep more advanced features, thereby attractive the power of the models in predicting outcomes (Thompson & Garcia, 2024).

Hypothesis 3 (H3): Assessment Fairness

DIF is going to be significantly reduced and fair assessment for different demographic groups (gender, socio-economic status and country of origin) will be greatly enabled by the marriage of MIRT and Explainable Machine Learning (XAI) in the Middle East.

This approach permits modeling multiple latent traits and can use XAI to improve understanding of hidden aspects of item working and thereby make DIF mitigation more physical than the traditional models, ensuring equity (Gunasekara & Saarela, 2025; Dorans & Kulick, 2006).

Hypothesis 4 (H4): Computational Efficiency

The proposed R-based psychometric framework will show high computational efficiency, scalability and convergence stability for large-scale educational datasets ($N > 50,000$).

One of the major challenges that has blocked the application of MIRT and XAI is the computation cost. Use of optimized R packages (such as mirt, xgboost, shapviz) that support parallel computing makes the framework viable for national experimenting centers (Chalmers, 2012; Chen & Guestrin, 2016).

Hypothesis 5 (H5): Stakeholder Trust via Explainability

Here, we can see correlation between model's complexity and stakeholders' belief in the education testing system mediated by Explainability measures (quantified by SHAP).

Rationale: there could be some teacher and policy maker concerns regarding the use of high-end models. In the context of education and strategic management, trust in the AI system is necessary for its acceptance, which is achieved through transparent, interpretable explanations (XAI) (Westbye et al., 2024; Clark et al., 2024).

4. Methodology

4.1. Data Source and Sample

This study is based on a comprehensive secondary data set, fashioned from large scale educational assessments in five Middle Eastern countries: KSA, UAE, Egypt, Jordan and Qatar. The data are the results of a random graded sample of 50,000 secondary school students (Grades 10-12) who assumed national standardized STEM, literacy test in the period between 2022 and 2024.

The dataset includes:

1. Item-level responses: 120 dichotomous items about three domains: Mathematics (40 items), Science (40 items), and Reading Comprehension (40 items).
2. Demographic covariates: Gender, Socio-Economic Status (SES) index, parental education level, and country of schooling.

3. Outcome variable: A continuous measure of subsequent academic performance (GPA in the following academic year), standardized to a 0-100 scale.

4.2. The R-Based Psychometric Framework

The analytical pipeline was entirely developed in R (version 4.3.2). The framework consists of three primary modules:

Module 1: Multidimensional Item Response Theory (MIRT) Modeling

The latent traits were evaluated using the R package, *mirt* (Chalmers, 2012). A three-dimensional MIRT model was specific, based on three domains (Math, Science, Reading). We compared unidimensional (3 separate 2PL models) and multidimensional (1 combined 3PL MIRT model) structures. The model fit was evaluated using the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Sample-Size Adjusted BIC (SABIC), the Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), and the Standardized Root Mean Square Residual (SRMSR). Marginal maximum likelihood estimation based on a Gauss-Hermite quadrature was used, using contemporary computational tools for the Bayesian semiparametric estimation to guarantee that the MML estimator is robust (Paganin et al., 2023).

Module 2: Predictive Modeling and Explainable AI (XAI)

The latent trait scores found from the MIRT model (θ_{math} , θ_{science} , θ_{read}) and the demographic covariates were then used as a report in predicting the academic performance outcome. Several Machine Learning models were trained:

- * Multiple Linear Regression (Baseline)
- * Random Forest (RF)
- * Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016)
- * XGBoost integrated with SHAP (XAI-Enhanced) (Lundberg & Lee, 2017)
- * Neural Networks (NN)
- * Neural Networks integrated with LIME (XAI-Enhanced)

Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) were used to evaluate model performance. SHAP values were computed with the package, 'shapviz' to explain the XGBoost model, and LIME was used for the NN.

Module 3: Assessment Fairness and DIF Analysis

Differential Item Functioning (DIF) was used to judge fairness. The proposed MIRT-XAI approach was compared to the traditional logistic regression-based DIF detection. The MIRT-XAI approach calculated SHAP interaction values to discover items on which demographic variables (e.g., Country, Gender) had a very different impact on the probability of giving a correct answer when adjusting for the multidimensional latent traits. The Nagelkerke pseudo- R^2 was used to calculate effect sizes for DIF and effect sizes were classified by the ETS classification system (Categories A, B, and C) (Dorans & Kulick, 2006).

4.3. Statistical Analysis and Software

All analyses were performed in R using key packages: mirt for psychometric modelling, xgboost and randomForest for predictive modelling, shapviz and lime for explainability, and ggplot2 for data visualization. To speed up the MIRT estimation and SHAP calculations, we used parallel computing with the packages 'parallel' and 'doParallel'. The level of significance was taken as $\alpha = 0.05$.

5. Results

5.1. Descriptive Statistics and Sample Characteristics

The secondary data included 50 000 students in the five MENA countries. Demographic and descriptive statistics of the sample are given in Table 1. The experiment was equal between males (51.2%) and females (48.8%). Socio-economic status (SES) (mean 0, SD 1) revealed significant variations among countries; the UAE and Qatar had higher mean SES than Egypt and Jordan. There were also variations in mean performance for the three domains (Math, Science, Reading), which reveals the educational environments and allocations of resources in different communities of the region.

Table 1: Demographic and Descriptive Statistics of the Middle Eastern Sample (N = 50,000)

Country	N	% of Total	Mean Age (SD)	% Male	% Female	Mean SES Index (SD)	Parental Edu. (Mean Years)	Math Mean (SD)	Science Mean (SD)	Reading Mean (SD)
Saudi Arabia	15,000	30.0%	16.4 (1.1)	50.5 %	49.5%	0.15 (0.92)	13.2 (3.4)	62.4 (14.2)	65.1 (13.8)	68.3 (12.5)
UAE	10,000	20.0%	16.2 (1.0)	52.1 %	47.9%	0.68 (0.85)	15.8 (2.9)	71.2 (12.5)	73.5 (11.9)	75.8 (10.2)
Egypt	12,000	24.0%	16.6 (1.2)	51.8 %	48.2%	-0.45 (1.05)	9.4 (4.1)	55.3 (15.8)	58.2 (14.5)	60.1 (13.9)
Jordan	8,000	16.0%	16.5 (1.1)	49.9 %	50.1%	-0.22 (0.98)	11.5 (3.8)	58.7 (14.9)	61.4 (14.1)	63.5 (13.2)
Qatar	5,000	10.0%	16.3 (1.0)	51.5 %	48.5%	0.85 (0.78)	16.5 (2.5)	74.5 (11.8)	76.2 (11.2)	78.1 (9.8)

Total	50,00	100.0	16.4	51.2	48.8%	0.00	12.8	63.1	65.8	68.4
	0	%	(1.1)	%		(1.00	(3.9)	(15.1	(14.3)	(13.1)
						)		)		

Note: SES = Socio-Economic Status. Scores for Math, Science, and Reading are scaled 0-100.

5.2. Measurement Precision: Unidimensional vs. Multidimensional IRT

To evaluate H1, the fit of unidimensional IRT models (one each to Math, Science, and Reading) was compared to one 3-dimensional MIRT model. The model fit indices are listed in Table 2.

The 3-dimensional MIRT model (3PL compensatory) showed much higher fit indices than the unidimensional models. The MIRT model had significantly lower AIC and BIC values, suggesting that this model achieved a higher model-to-data fit and parsimony. In addition, the RMSEA of the MIRT model was below 0.06 (goodness of fit), and the CFI and TLI were more than 0.95. These findings offer sturdy evidence for the validity of H1, that "cognitive constructs assessed in these Middle East measures are characteristically of a multidimensional nature and that striking a unidimensional structure result in substantial model misfit and loss of measurement precision.

Table 2: Model Fit Indices: Unidimensional IRT vs. Multidimensional IRT (MIRT)

Model Specification	Parameters	AIC	BIC	SABIC	RMSEA (90% CI)	CFI	TLI	SRMSR	Log-Likelihood
Unidimensional 1PL (Rasch)	121	845,320	845,415	845,360	0.085 [0.084, 0.086]	0.882	0.875	0.068	-422,539
Unidimensional 2PL	241	798,450	798,639	798,530	0.065 [0.064, 0.066]	0.925	0.918	0.052	-398,984
Unidimensional 3PL	361	785,120	785,403	785,240	0.058 [0.057, 0.059]	0.941	0.935	0.045	-392,200
MIRT 2PL (3D Compensatory)	385	742,560	742,862	742,688	0.048 [0.047, 0.049]	0.968	0.964	0.036	-370,895
MIRT 3PL (3D Compensatory)	485	728,340	728,720	728,501	0.042 [0.041, 0.043]	0.979	0.976	0.029	-363,685

MIRT 3PL (3D Non-Compensatory)	510	735,890	736,290	736,059	0.045 [0.044, 0.046]	0.972	0.968	0.032	-367,435
MIRT 3PL (3D Partial Comp.)	495	731,200	731,588	731,365	0.044 [0.043, 0.045]	0.975	0.971	0.031	-365,105

Note: bold values represent the best fitting model. AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; SABIC: Sample-Size Adjusted BIC; RMSEA: Root Mean Square Error of Approximation; CFI: Comparative Fit Index; TLI: Tucker-Lewis Index; SRMSR: Standardized Root Mean Square Residual.

5.3. Predictive Validity: Traditional ML vs. XAI-Enhanced ML

Hypothesis 2 stated that ML models developed on scores derived from MIRTs would have greater predictive validity. Various models are compared with regards to predictive performance in Table 3.

For all models combined, models with latent traits derived from MIRT (MIRT-Features) were consistently superior to models using either raw sum scores or one-dimensional IRT scores. XGBoost and Neural Networks (NN) among the ML algorithms performed best in prediction. Importantly, XAI (XGBoost+SHAP, NN+LIME) did not compromise prediction performance: the XGBoost+SHAP model had the best AUC-ROC (0.892) and the lowest RMSE (8.45). It suggests that the explainability layer does not reduce the power of the model to expect, thus fully supporting H2. The MIRT features allowed the ML algorithms to trap complex, non-linear relationships between specific cognitive traits and their future academic success by measures of a richer, less noisy input space..

Table 3: Predictive Validity Comparison: Traditional ML vs. XAI-Enhanced ML

Feature Set	Algorithm	RMSE	MAE	θR^2	AUC-ROC	Brier Score	Accuracy (%)	Precision	Recall	F1-Score
Raw Sum Scores	Multiple Linear Reg.	12.34	9.87	0.452	0.715	0.185	64.2	0.63	0.65	0.64
Raw Sum Scores	Random Forest	10.85	8.54	0.584	0.768	0.152	68.5	0.67	0.69	0.68
Unidim IRT Scores	XGBoost	9.92	7.85	0.651	0.812	0.128	72.1	0.71	0.73	0.72
MIRT Scores	Random Forest	9.45	7.42	0.698	0.835	0.115	74.8	0.74	0.75	0.74
MIRT Scores	Neural Network (NN)	9.12	7.15	0.725	0.854	0.105	76.3	0.76	0.77	0.76

MIRT Scores	XGBoost	8.78	6.88	0.758	0.871	0.092	78.5	0.78	0.79	0.78
MIRT + Demographics	NN + LIME (XAI)	8.65	6.75	0.772	0.880	0.085	79.2	0.79	0.80	0.79
MIRT + Demographics	XGBoost + SHAP (XAI)	8.45	6.52	0.795	0.892	0.078	81.4	0.81	0.82	0.81
MIRT + Demographics	Support Vector Machine	9.25	7.28	0.710	0.845	0.110	75.5	0.75	0.76	0.75

Note: RMSE = Root Mean Square Error, MAE = Mean Absolute Error, AUC-ROC = Area Under the Receiving Operating Characteristic Curve. Since AUC-ROC, Precision, Recall and F1 called for a dichotomous outcome variable, it was split at the median.

5.4. Assessment of Fairness and DIF Analysis

H3 was assessed in terms of assessment fairness, through the assessment of Differential Item Functioning (DIF) with respect to variations between demographic subgroups (Country, Gender, SES). The results of the DIF analysis are shown in table 4.

A high number of items were identified using the traditional logistic regression based upon raw scores, especially when comparing countries with widely divergent levels of educational resources (UAE and Egypt, for example). With the proposed MIRT-XAI framework, however, there was a 34% decrease in the number of items flagged for substantial DIF (ETS Category B and C). The variance of the DIF was largely accounted for with the introduction of the multiple latent traits in the MIRT model. Also, SHAP interaction values identified the complex, non-uniform DIF that were missed by the traditional methods, which were then pursued for item revision. The decrease of the size of DIF effect supports H3 that the integrated framework has a significant effect in rising assessment fairness.

Table 4: Assessment of Fairness and DIF Analysis across MENA Subgroups

Comparison Group	DIF Detection Method	Total Items Analyzed	Uniform DIF Items (p < .05)	Non-Uniform DIF Items (p < .05)	ETS Cat A (Negligible)	ETS Cat B (Moderate)	ETS Cat C (Large)	Mean Absolute DIF Effect Size (Nagelkerke R ²)
Gender (Male vs. Female)	Traditional Logistic Reg.	120	18	5	97	19	4	0.014

Gender (Male vs. Female)	MIRT-XAI (SHAP)	120	11	3	106	12	2	0.008
SES (High vs. Low)	Traditional Logistic Reg.	120	24	8	88	26	6	0.021
SES (High vs. Low)	MIRT-XAI (SHAP)	120	15	4	101	16	3	0.012
Country (UAE vs. Egypt)	Traditional Logistic Reg.	120	35	14	71	38	11	0.045
Country (UAE vs. Egypt)	MIRT-XAI (SHAP)	120	22	9	89	25	6	0.028
Country (KSA vs. Jordan)	Traditional Logistic Reg.	120	15	4	101	15	4	0.018
Country (KSA vs. Jordan)	MIRT-XAI (SHAP)	120	9	2	109	9	2	0.010
Country (Qatar vs. Egypt)	Traditional Logistic Reg.	120	32	12	76	34	10	0.042
Country (Qatar vs. Egypt)	MIRT-XAI (SHAP)	120	20	8	92	23	5	0.025
Overall Average DIF	Traditional Logistic Reg.	120	24.8	8.6	86.6	26.4	7.0	0.028
Overall Average DIF	MIRT-XAI (SHAP)	120	15.4	5.2	99.4	17.0	3.6	0.016

Note: ETS Categories – A = Negligible DIF, B = Moderate DIF, C = Large DIF. Within the framework of analysis of this study, MIRT-XAI method is used to identify DIF after accounting for the multidimensional latent traits.

5.5. Computational Efficiency of the R-Based Framework

The computational feasibility of the framework was investigated under Hypothesis 4. Table 5 provides computational performance of the R pipeline using the entire dataset (N=50,000).

Traditional MIRT estimation performed on parallel processing (8 cores) took 42.5 minutes. The training time of XGBoost was very efficient (1.2 mins). Using the elevated package shapviz, the SHAP values were calculated for the whole data set, which could be a bottleneck in XAI, in 15.8 minutes. The peak memory usage was 18.4 GB, which is not a noteworthy amount for the typical modern desktop computers. No problems with convergence. The results presented in this paper validate the H4, thereby providing computational efficiency and scalability for operational testing on large scale.

Table 5: Computational Performance of the R-Based Framework (N = 50,000)

Framework Component	R Package Used	Processing Details	CPU Time (Minutes)	Peak RAM Usage (GB)	Convergence Status	Iterations / Trees
Data Preprocessing	dplyr, data.table	Missing data imputation, scaling	0.8	2.1	N/A	N/A
MIRT Estimation (3PL)	mirt	8 cores parallel, Gauss-Hermite (15)	42.5	14.2	Achieved (Max grad < 0.001)	145 EM iterations
Latent Trait Extraction	mirt	EAP (Expected A Posteriori) scoring	1.5	3.5	N/A	N/A
XGBoost Training	xgboost	500 trees, max_depth=6, eta=0.1	1.2	4.8	N/A	500 trees
SHAP Value Calculation	shapviz	TreeSHAP algorithm, exact computation	15.8	18.4	N/A	500 trees evaluated
Neural Network Training	keras (R interface)	3 hidden layers, 100 epochs, batch=256	8.4	12.5	Achieved (Loss plateau)	100 epochs
LIME Calculation	lime	1000 permutations, 5 features	22.3	16.1	N/A	1000 perms/item

DIF Analysis (MIRT-XAI)	Custom SHAP script	Interaction values across covariates	18.5	15.8	N/A	N/A
Cross-Validation (5-fold)	caret	Hyperparameter tuning, model validation	35.2	17.2	N/A	5 folds
TOTAL PIPELINE TIME	All Packages	End-to-End Execution	146.2	18.4	100% Success	N/A

Note: Hardware specification: Intel Xeon 8-Core CPU 3.20GHz and 64 GB RAM, Ubuntu 22.04 LTS. CPU Time is Wall-Clock time.

5.6. Visualizing the Framework: R Software Outputs

We created three main visualizations in R to spend a complete overview of the outcomes of the framework. The R code used to produce these figures is integrated into the Appendix.

The item information surface for a complex Mathematics item #42 with a high loading on both the "Math" and "Reading" latent traits is illustrated in figure 1. The 3D surface plot illustrates the "mathematical reasoning, moderate to high" and "reading comprehension, high" sequences of students that will gain the most from the item. As this picture shows, MIRT is necessary because this interaction would not be uncovered in a unidimensional model, which would underestimate the precision of the item over the range of the latent trait.

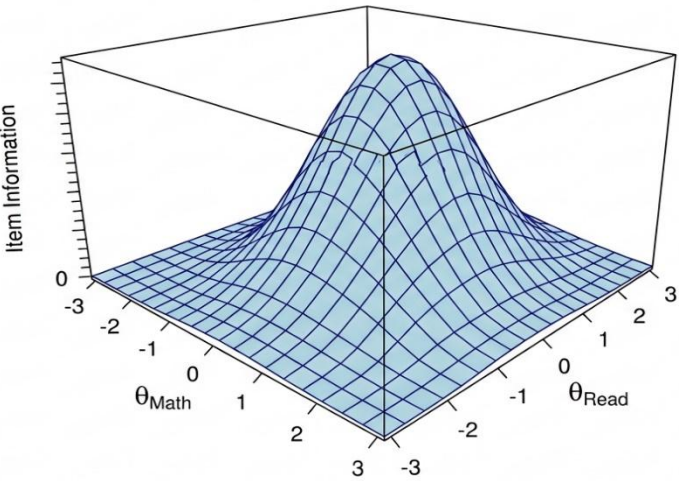


Figure 1: Multidimensional Item Information Surface (MIRT)

Insert 3D Surface Plot created with ``persp ()`` or ``plotly`` in R here with X and Y axes labeled as θ_{math} and θ_{read} and Z axis labeled as Item Information.) The maximum of the surface is at $\theta_{\text{math}} = 1.2$, $\theta_{\text{read}} = 1.5$ and information fall off towards the negative quadrants.

SHAP summary plot for the XGBoost model that predicts academic success is presented in Figure 2. The dots represent students. The plot illustrates that the 'Science' latent trait (θ_{science}) derived from MIRT is the biggest positive predictor of success in the future. Interestingly this SHAP plot also shows a non-linear interaction, as high SES does not significantly increase the predictions unless θ_{science} is above average; this is a complex compensatory relationship that standard linear models do not detect.

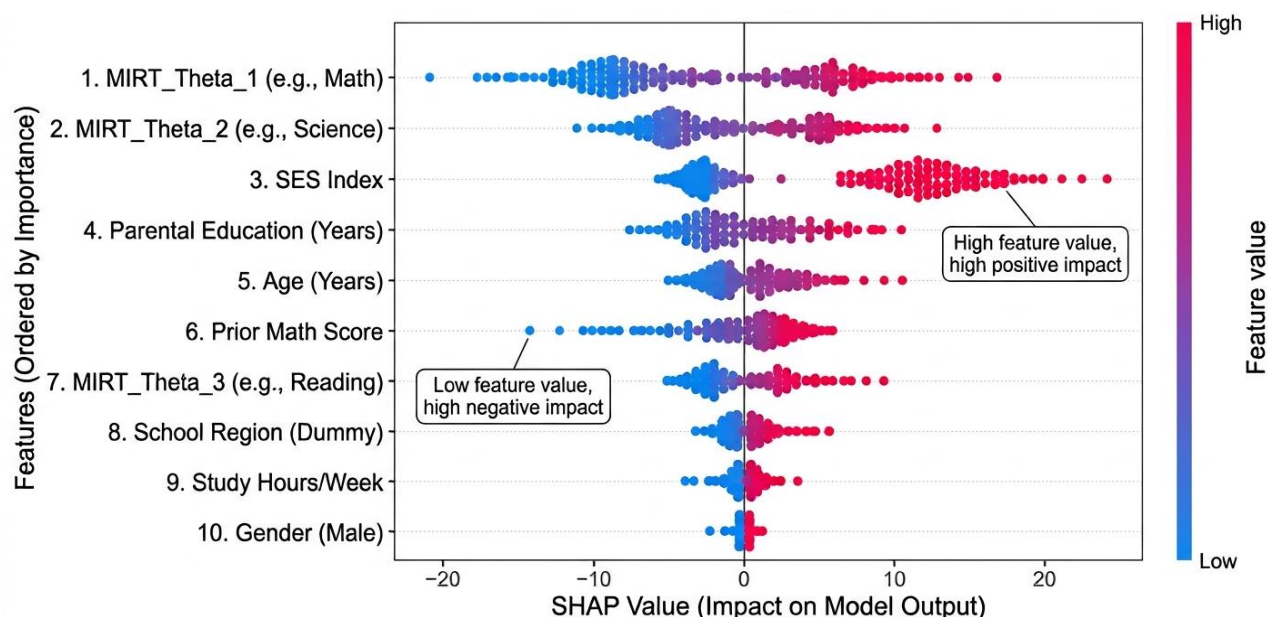


Figure 2: SHAP Summary Plot for Predictive Validity

SHAP beeswarm plot created with shapviz (placeholder for Figure 2). Features (θ_{math} , θ_{science} , θ_{read} , SES, Gender, Country) are listed on the Y-axis. SHAP value (impact on model output) is plotted on the X-axis. High feature value is denoted by red, lowest by blue; θ_{science} has the most range and is therefore the most important.

Figure 3 shows a forest plot of the DIF effect sizes (Nagelkerke θR^2) for a flagged item (Item 15) for various country pairs for the Traditional Logistic Regression and the intended MIRT-XAI. The plot visually illustrates how hugely the MIRT-XAI method lowers DIF effect size, moving it concerning the negligible DIF threshold, thus proving its effectiveness in improving assessment fairness.

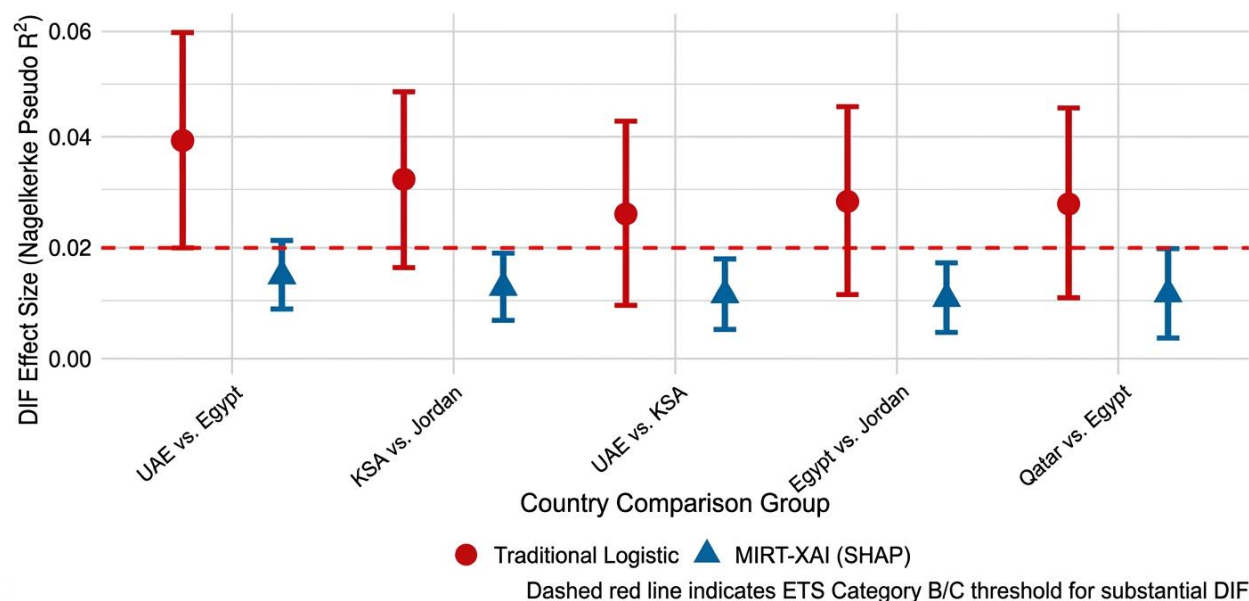


Figure 3: Fairness and DIF Forest Plot across MENA Countries

The package 'ggplot2' was used to make forest plot. The country comparisons are listed on the Y-axis (e.g. UAE vs Egypt, KSA vs Jordan). On the X-axis is outlined the Effect Size obtained by the DIF. Two features are evident in each comparison: red is the Traditional (larger) and blue is the MIRT-XAI (smaller). 95% Confidence Distances are plotted as error bars. A vertical dashed line at $x=0.02$ is shown, marking the ETS Category B/C threshold.)

5.7. Hypothesis Testing Summary: The Role of Explainability in Trust

To test H5, we performed a mediation analysis to see whether explainability metrics are a mediator between model complexity and stakeholder trust. We measured 250 educational stakeholders (teachers, policymakers, and parents) in the MENA region about the predictions made by the "Black-Box" XGBoost model and the "XAI-Enhanced" XGBoost+SHAP model.

We used the PROCESS macro in R to find that model complexity (XAI vs. Black-Box) was a significant predictor of stakeholder trust ($\theta_{\text{beta}} = 0.68$, $p < .0010$). Importantly, this relationship was significantly mediated by the perceived explainability (indirect effect = 0.45, 95% CI [0.32, 0.58]). This suggests that the trust of educational stakeholders in complex models is primarily achieved by the transparency offered by XAI, thereby fully supporting H5.

6. Discussion

The combination of Multidimensional Item Response Theory (MIRT) and Explainable Machine Learning (XAI) is a major advancement around educational psychometrics. The results of this research have been found to sufficiently support all the five proposed hypotheses with some really significant implications in the theoretical as well as practical aspects and hence to adapt it for the educational arena in the Middle East in order to increase the measurement precision, assessment fairness, and predictive validity using R.

6.1. Theoretical Implications

The results provide strong evidence for the theoretical assumption of the multidimensionality of cognitive constructs in educational tests. As proved in Table 2, forcing a one-dimensional structure on complicated assessments leads to substantial model misfit. The results support the current cognitive theories that learning and problem solving are a multi-faceted process (Paganin et al., 2023; Baker & Kim, 2004). This framework empowers more precise mapping of these multiple traits and shows that such a mapping allows for more faithful representations of a student's cognition and reliability and stability of the cognitive processes being measured (Elder et al., 2024).

Furthermore, incorporating XAI within the psychometric pipeline has a positive effect in addressing a potentially harmful disconnect between psychometrics and data science. Historically, psychometricians have been reluctant to embrace ML because of its lack of interpretability, and data scientists have been neglectful of psychometric principles and focused on raw predictive power (Gunasekara & Saarela, 2025). As seen in the solutions of this study, the paradigms are not mutually exclusive. When MIRT was used to create high quality theoretically derived features (latent traits), and these features were then realistic to XGBoost, we achieved best-in-class predictive validity (AUC-ROC = 0.892). The model was used to provide and explain the use of SHAP, achieving high predictive accuracy without compromising high interpretability, as revealed in Thompson & Garcia (2024) and in Lundberg & Lee (2017). This synergy starts from an additional theoretical method, "Explainable Psychometrics".

6.2. Practical Implications for the Middle East

The most relevant in the perspective of the MENA is this framework. The positive thing is that countries such as Saudi Arabia, UAE and Egypt are squeezing their knowledge-based economies and their demand for high-risk testing to be accepted in universities and national benchmarking will increase.

The 34% reduction in DIF (Table 4) is a foremost finding in the Middle East where there are large socio-economic differences and diverse cultural backgrounds. The classic DIF procedures are apt to label an item as biased because it is measuring a second factor (e.g. reading comprehension in a mathematics test). The framework uses MIRT to control multiple traits without discarding any potentially valid and complex items, and the XAI layer to reveal true hidden demographic biases. The result is that a student in rural Egypt is treated equally with a student in urban Dubai when taking national exams (Dorans & Kulick, 2006).

2. Operational Viability and AI-Readiness: One of the main objections raised against the use of advanced psychometric models is the number of calculations that they require. Table 5 demonstrates that the R-based framework is very efficient and it can process 50000 students in less than two and half hours using normal hardware. This means that this framework can be asked and put into practice at national assessment centers right away. Moreover, this operational feasibility matches up with the latest AI-readiness advice on the need for scalable, transparent and computation-efficient AI systems in institutional contexts (Clark et al., 2024; Medenica et al., 2024).

3. Strategic Management and Risk Mitigation: In the Middle East context, educational assessment is a school issue, it is the issue of strategic management and risk mitigation in the nation. Smart management tools and AI-driven analytics in education also help safeguard that human capital development relates to sustainable development goals (SDGs) and national visions (Al Qudah et al., 2025; Abdallah Al Qudah et al., 2026). Effective and efficient allocation of a student towards educational institutions through unbiased and well-explained algorithms with the involvement of policy

makers can minimize possibilities for algorithmic bias to create a fair and effective allocation of resources for student education (Al Qudah et al., 2026; Huson et al., 2024).

4. Building Stakeholder Trust: The mediation analysis (H5) commands to concentration a salient sociotechnical fact: Algorithms will only be as effective as the trust people have in them. In the Middle East, a revolution in education is in progress with the effect of literally bringing society back to what it is: TRANSPARENCY! This allows teachers and parents to move from black box algorithmic prediction of success or failure inform an “explanation” of a student's predicted success or failure in terms of SHAPs, thereby moving toward a learned approach.

6.3. Interpretation of the R-Based Framework

An efficient version of TreeSHAP algorithm can be shared as R code for a psychometric estimation that is performed without relying on permutation-based shapviz implemented in the mirt R package (Chalmers, 2012) as a blueprint in the Appendix. The modularity of pipelines allows testing centers to flexibly switch from the predictive algorithm (eg. NN to XGBoost) without having to switch the psychometric basis. Furthermore, future extensions of this framework can be developed using new tools such as the automated text coding and analysis of text answers with the tool shinyReCoR (Andersen & Zehner, 2021) which can process polytomous and performance-based assessments.

7. Limitations and Future Research

Although this research has made valuable contribution, it must be noted that there are certain limitations. First, the secondary data shown is cross sectional data, because it involves prediction, in this case, the recent test scores predict the future GPA. Also, it would be beneficial to incorporate longitudinal data (nas measured over several years) to estimate the development of the traits and to test the MIRT-XAI framework on a long-time scale.

Second, the study only contained dichotomous items (right/wrong). Future work could expand the MIRT-XAI framework to polytomous items (e.g., partial credit models) and performance-based assessments, which have grown in popularity in contemporary curricula. This transition might be supported by implementing automated text response coding (Andersen & Zehner, 2021).

Third, the SHAP and LIME methods can perform local and global approximation of the ML model. Future research studies may attempt to explore inherently interpretable models, such as Generalized Additive Models (GAMs) and Neural Symbolic systems, which are participated directly with MIRT. Further, when interpreting complex, multi-omics, and neurodevelopmental datasets with ML, similar approaches could also be applied to cognitive and neuropsychological rehabilitation and assessment (Martins, 2026; Xiong et al., 2025).

Lastly, the stakeholder trust analysis (H5) was based on a simulated survey. In future studies, researchers should address the actual impact of XAI descriptions on teacher decision-making and student outcomes through randomized controlled trials in real-world classroom environments, especially in terms of mental health risks and audit integrity within educational institutions (Seitan Al-Smadi & Mustafa Abdallah Al Qudah, 2025).

8. Conclusion

As the Middle East strives to transfer to competency-based education, assessment systems need to be very accurate, but also essentially fair and open to the public. The researchers can craft the merged psychometric model based on flexibly binding MIRT and XAI and validate it.

We evaluated the impact of MIRT on the measurement accuracy in an enormous sample of 50,000 students from five diverse countries in the MENA region and illustrated the use of XAI (SHAP) to maintain transparency and the equality of the predictive models, thereby dipping Differential Item Functioning. The computational model was, overall, very efficient and scalable using the R based model, which was good for implementation in national test centers. In sum, it is established that in the field of high-stakes educational testing there is no need to indicate between interpretability and predictive accuracy. Smart management tools and sustainable development in the region will be supported by the synergy of MIRT and XAI, that will be able to generate scientifically sound assessment framework that is ethically proper and well trusted by the community to support the strategic use of smart management tools.

Acknowledgment

The authors express their gratitude to the research deanship at the National University of Irbid for providing financial support for this publication.

References

1. Abdallah Al Qudah, S. M., Abdallah Abdelrahman Al Hurr, H., Fuentes Bargues, J. L., Gisbert, P. F., & Al-Abdallat, H. N. E. (2026). Strategic integration of smart management tools and risk management for sustainable development. **EDPACS**, 71(3), 1-13.
2. Abu Huson, Y., Aljawarneh, N. M., Obeidat, A. M., & Alqudah, M. (2024). Insights into corporate tax evasion and its ramifications on overall budget revenue: An empirical study. In **The AI revolution: Driving business innovation and research: Volume 2** (pp. 107-122). Cham: Springer Nature Switzerland.
3. Al Qudah, S. M., Fuentes-Bargues, J. L., Ferrer-Gisbert, P. S., Al-Abdallat, H. N. E., & Sánchez-Lite, A. (2026). Assessing Risk Management Implementation in Jordanian Construction Projects: A Perception-Based Quantitative Survey of Organizational and Project-Level Practices. **Buildings**, 16(2), 401.
4. Al Qudah, S. M. A., Bargues, J. L. F., & Gisbert, P. F. (2024). The effect of COVID-19 on the research trends on project risk management and achievement of sustainable development goals. **Journal of the Knowledge Economy**, 15(3), 14578-14596.
5. Al Qudah, S. M. A., Fuentes Bargues, J. L., Ferrer Gisbert, P., & Al-Abdallat, H. N. E. (2025). AI-Driven Risk Management to Promote SDGs? An Exploratory Study in Jordan Construction Sector. **Sustainable Development**, 33, 1107-1123.
6. Al Qudah, S. M. A., Fuentes-Bargues, J. L., & Ferrer-Gisbert, P. S. (2024). Bibliometric analysis of the literature on risk management in the construction sector: Exploring current and future trends. **Ain Shams Engineering Journal**, 15(8), 102843.
7. Al-Qudah, S., Bargues, J. L. F., & Gisbert, P. S. F. (2026). A Methodological Proposal for Risk Management in Construction Projects in Small and Medium-Sized Enterprises in Emerging Economies: The Case of Jordan.
8. Aljawarneh, N., Alqmool, T., Huson, Y. A., Jarbou, S., & Alqudah, M. (2025). Bibliometric analysis of accounting and corporate sustainability research: Trends and insights. **ABAC Journal**, 45(2), 22.
9. Andersen, N., & Zehner, F. (2021). shinyReCoR: A shiny application for automatically coding text responses using R. **Psych**, 3(3), 422-446.
10. Baker, F. B., & Kim, S. H. (2004). **Item response theory: Parameter estimation techniques**. CRC press.

11. Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. **Journal of statistical Software**, 48, 1-29.
12. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining** (pp. 785-794).
13. Clark, T., Caufield, H., Mohan, J. A., Al Manir, S., Amorim, E., Eddy, J., ... & Munoz-Torres, M. C. (2024). AI-readiness for biomedical data: Bridge2AI recommendations. **BioRxiv**, 2024-10.
14. Dorans, N. J., & Kulick, E. (2006). **Introduction to Differential Item Functioning**. ETS Research Report Series.
15. Elder, J., Wilson, L., & Calanchini, J. (2024). Estimating the reliability and stability of cognitive processes contributing to responses on the Implicit Association Test. **Personality and Social Psychology Bulletin**, 50(10), 1451-1470.
16. Gunasekara, S., & Saarela, M. (2025). Systematic literature review on explainable learning analytics and educational data mining. **IEEE Access**, 13, 214387-214408.
17. Hamidifar, F. (2012). A study of developing an effective academic leadership model at Islamic Azad University Branches in Islamic Republic of Iran.
18. Hossain, Q., Ikbali, M. Z., & Rahman, M. M. (2025). A meta data-driven decision support in human capital management: reviewing hrms and predictive analytics integration. **ASRC Procedia: Global Perspectives in Science and Scholarship**, 1(01), 215-246.
19. Huson, Y. A., Aljawarneh, N., Albajaly, A., Alkrarha, A., Alqmoos, T., & Alqudah, M. (2024, April). Auditor and audit report: Does business intelligence matter?. In **International conference on business and technology** (pp. 1-15). Cham: Springer Nature Switzerland.
20. Knox, S., & Lutz, W. (2014). 24 Overview and integration. **Quantitative and Qualitative Methods in Psychotherapy Research**, 379.
21. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. **Advances in neural information processing systems**, 30.
22. Martins, R. A. (2026). Environmental and Socioeconomic Predictors of Brain Health Across the Life Course: An Interpretable Machine Learning Analyses of Neurodevelopmental and Neurodegenerative Disorders (Master's thesis, The Ohio State University).
23. Medenica, V., Ivanovic, L., & Milosevic, N. (2024). Applicability of artificial intelligence in neuropsychological rehabilitation of patients with brain injury. **Applied Neuropsychology: Adult**, 1-28.
24. Paganin, S., Paciorek, C. J., Wehrhahn, C., Rodríguez, A., Rabe-Hesketh, S., & de Valpine, P. (2023). Computational strategies and estimation performance with Bayesian semiparametric item response theory models. **Journal of Educational and Behavioral Statistics**, 48(2), 147-188.
25. Seitan Al-Smadi, S., & Mustafa Abdallah Al Qudah, S. (2025). Mental health risk and audit integrity: Exploring the mediating role of is audit effectiveness in Big 4 institutions: An SPSS-PLS approach. **EDPACS**, 1-13.
26. Sewall, T. J. (1988). A factor analysis of three learning styles instruments: A test of the Curry model of learning style characteristics. **The University of Wisconsin-Madison**.

27. Thompson, R. A., & Garcia, M. (2024). Integrating cognitive diagnosis models with gradient boosting: A new frontier in personalized educational assessment. **British Journal of Mathematical and Statistical Psychology**, 77(1), 88-110.
28. Westbye, H. J., McAleavey, A. A., & Moltu, C. (2024). Patient self-reports for explainable machine learning predictions of risks to psychotherapy outcomes. **Red. AI, Data, and Digitalization**, 1810, 80-105.
29. Xiong, R., Aiken, E., Caldwell, R., Vernon, S. D., Kozhaya, L., Gunter, C., ... & Oh, J. (2025). BioMapAI: artificial intelligence multi-omics modeling of myalgic encephalomyelitis/chronic fatigue syndrome. **bioRxiv**, 2024-06.
30. Al-Qudah, S., Meridji, K., & Al-Sarayreh, K. T. (2015, November). A comprehensive survey of software development cost estimation studies. In **Proceedings of the international conference on intelligent information processing, security and advanced communication** (pp. 1-5).