Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309-6575 *İlkbahar* 2025 *Cilt: 16-Sayı: 1*
Spring 2025 *Volume: 16-Issue: 1*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

İletişim
e-posta: epodderdergi@gmail.com
Web: https://dergipark.org.tr/tr/pub/epod

Contact
e-mail: epodderdergi@gmail.com
Web: http://dergipark.org.tr/tr/pub/epod

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (EPOD) yılda dört kez yayımlanan hakemli uluslararası bir dergidir. Yayımlanan yazıların tüm sorumluğu ilgili yazarlara aittir.

Journal of Measurement and Evaluation in Education and Psychology (JMEEP) is a international refereed journal that is published four times a year. The responsibility lies with the authors of papers.

Dizinleme / Abstracting & Indexing
Emerging Sources Citation Index (ESCI), DOAJ (Directory of Open Access Journals), SCOPUS, TÜBİTAK TR DIZIN Sosyal ve Beşeri Bilimler Veri Tabanı (ULAKBİM), Tei (Türk Eğitim İndeksi), EBSCO

Hakem Kurulu / Referee Board

Gökhan AKSU (Adnan Menderes Üni.)
Görkem CEYHAN (Muş Alparslan Üni.)
Gözde SIRGANCI (Bozok Üni.)
Gül GÜLER (İstanbul Aydın Üni.)
Gülden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Hacettepe Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan SARIÇAM (Dumlupınar Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil İbrahim SARI (Kilis Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice Çiğdem BULUT (Northern Alberta IT)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hikmet ŞEVGİN (Van Yüzüncü Yıl Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hülya YÜREKLI (Yıldız Teknik Üni.)
İbrahim Alper KÖSE (Bolu Abant İzzet Baysal Üni.)
İbrahim YILDIRIM (Gaziantep Üni.)
İbrahim UYSAL (Bolu Abant İzzet Baysal Üni.)
İlhan KOYUNCU (Adıyaman Üni.)
İlkay AŞKIN TEKKOL (Kastamonu Üni.)
İlker KALENDER (Bilkent Üni.)
İsmail KARAKAYA (Gazi Üni.)
Kadriye Belgin DEMİRUS (Başkent Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent ERTUNA (Sakarya Üni.)
Levent YAKAR (Kahramanmaraş Sütçü İmam Üni.)
Mahmut Sami KOYUNCU (Afyon Üni.)
Mahmut Sami YİĞİTER (Ankara Sosyal B. Üniv.)
Mehmet KAPLAN (MEB)
Mehmet ŞATA (Ağrı İbrahim Çeçen Üni.)
Melek Gülşah ŞAHİN (Gazi Üni.)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Meltem YURTÇU (İnönü Üni.)
Merve ŞAHİN KÜRŞAD (TED Üni.)
Metin BULUŞ (Adıyaman Üni.)
Murat Doğan ŞAHİN (Anadolu Üni.)
Mustafa ASİL (University of Otago)
Mustafa İLHAN (Dicle Üni.)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe Üni.)
Nail YILDIRIM (Kahramanmaraş Sütçü İmam Üni.)
Neşe GÜLER (İzmir Demokrasi Üni.)
Neşe ÖZTÜRK GÜBEŞ (Mehmet Akif Ersoy Üni.)
Nuri DOĞAN (Hacettepe Üni.)
Nükhet DEMİRTAŞLI (Emekli Öğretim Üyesi)
Okan BULUT (University of Alberta)
Onur ÖZMEN (TED Üniversitesi)
Ömer KUTLU (Ankara Üni.)
Ömür Kaya KALKAN (Pamukkale Üni.)
Önder SÜNBÜL (Mersin Üni.)
Özen YILDIRIM (Pamukkale Üni.)
Özge ALTINTAS (Ankara Üni.)
Özge BIKMAZ BİLGEN (Adnan Menderes Üni.)
Özlem ULAŞ (Giresun Üni.)
Recep GÜR (Erzincan Üni.)
Ragıp TERZİ (Harran Üni.)
Sedat ŞEN (Harran Üni.)
Recep Serkan ARIK (Dumlupınar Üni.)
Safiye BİLİCAN DEMİR (Kocaeli Üni.)
Selahattin GELBAL (Hacettepe Üni.)
Seher YALÇIN (Ankara Üni.)
Selen DEMİRTAŞ ZORBAZ (Ordu Üni.)
Selma ŞENEL (Balıkesir Üni.)
Seçil ÖMÜR SÜNBÜL (Mersin Üni.)
Sait Çüm (MEB)
Sakine GÖÇER ŞAHİN (University of Wisconsin Madison)
Sedat ŞEN (Harran Üni.)
Sema SULAK (Bartın Üni.)
Semirhan GÖKÇE (Niğde Ömer Halisdemir Üni.)
Serap BÜYÜKKIDIK (Sinop Üni.)
Serkan ARIKAN (Boğaziçi Üni.)
Seval KIZILDAĞ ŞAHİN (Adıyaman Üni.)
Sevda ÇETİN (Hacettepe Üni.)
Sevilay KİLMEN (Abant İzzet Baysal Üni.)
Sinem DEMİRKOL (Ordu Üni.)
Sinem Evin AKBAY (Mersin Üni.)
Sungur GÜREL (Siirt Üni.)
Süleyman DEMİR (Sakarya Üni.)
Sümeyra SOYSAL (Necmettin Erbakan Üni.)
Şeref TAN (Gazi Üni.)
Şeyma UYAR (Mehmet Akif Ersoy Üni.)
Tahsin Oğuz BAŞOKÇU (Ege Üni.)
Terry A. ACKERMAN (University of Iowa)
Tuğba KARADAVUT (İzmir Demokrasi Üni.)
Tuncay ÖĞRETMEN (Ege Üni.)
Tülin ACAR (Parantez Eğitim)
Türkan DOĞAN (Hacettepe Üni.)
Ufuk AKBAŞ (Hasan Kalyoncu Üni.)
Wenchao MA (University of Alabama)
Yavuz AKPINAR (Boğaziçi Üni.)
Yeşim ÖZER ÖZKAN (Gaziantep Üni.)
Yusuf KARA (Southern Methodist University)
Zekeriya NARTGÜN (Bolu Abant İzzet Baysal Üni.)
Zeynep ŞEN AKÇAY (Hacettepe Üni.)

*Ada göre alfabetik sıralanmıştır. / Names listed in alphabetical order.

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Mart 2025, Sayı: 16-1)

Journal of Measurement and Evaluation in Education and Psychology (March 2025, Issue: 16-1)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2025; 16(1); 1-12

Investigating the Performance of Artificial Neural Networks in Predicting Affective Responses

İzzettin AYDOĞAN * Osman TAT **

Abstract

In this study it is aimed to examine the performance of an artificial neural network trained using items reflecting a latent trait in predicting responses to an item reflecting the same trait. This latent trait is the awareness of being able to communicate with people from different cultures, which is included in the PISA 2018 assessment. Relevant scale items were used as research variables. In addition to determining the extent to which the predicted responses overlap with the actual responses by analyzing the artificial neural network models, it was examined how the predicted responses affect the assumed latent construct and the reliability of the responses. Thus, the performance of artificial neural networks in predicting responses to affective items was evaluated. The responses expected from individuals for the items examined overlap with the responses given by individuals at a relatively moderate. However, it is observed that although the prediction values partially weaken the model fit indices, they still manage to keep them strong. In addition, the predicted values improved the factor loadings and the variance explained for the latent trait. Similarly, it is noticed that the predicted values also positively affect the reliability.

Keywords: Artificial neural networks, machine learning, affective responses, prediction

Introduction

Current advances in machine learning and artificial intelligence are largely driven by artificial neural networks (ANNs) (Goel et al., 2023). The ability of ANNs to analyze complex data, especially those that cannot be simplified by traditional statistical methods, is gradually improving (Tu, 1996). ANNs consist of structures in which neurons are connected to each other by synapses with adjustable weights. Synapses connecting neurons, which are the building blocks of ANNs, function in communication. Information exchange between neurons takes place through synapses. Information flows from the synapse of one neuron to the dendrite of another neuron (Goel et al., 2023). The fact that the weights are adjustable allows the network to be trained by back-propagating the errors throughout the network. The aim of training is to adjust the weights to minimize the error between actual and predicted values (Lillicrap et al., 2020). ANNs have attracted much attention due to their ability to model non-linear relationships between variables (Tu, 1996). Although it is seen as a simple variant, ANNs are biologically similar to the working principles of the human brain (Hasson et al., 2020).

ANNs exhibit successful performances in many important fields such as health, climate, physics, chemistry, biology, engineering, industry, agriculture (Lau et al., 2019; Park et al., 2020). Considering the purpose and frequency of use of ANNs, it is possible to say that they are mostly used for diagnosis, prediction and forecasting. It is widely used in areas such as predicting some features through some predictors with regression logic, missing data assignment, recognition, and classification. Although its application area in educational research is limited (Tu, 1996), studies conducted in relation to education and psychology contents (Aybek & Okur, 2018; Aydoğan & Zırhlıoğlu, 2018; Aydoğdu, 2020; Al-Saleem et al., 2015; Chavez et al, 2023; Flitman, 1997; Guarín et al., 2015; Huang & Fang, 2012; Lau et al., 2019; Rodríguez-Hernández, 2021; Shahiri & Husain, 2015; Umar, 2019; Zacharis, 2016) especially focus on predicting students' academic performance. The basic logic of ANN models

* Assoc. Prof.,Van Yüzüncü Yıl University, Faculty of Education, Van-Türkiye, izettinaydogan@yyu.edu.tr, ORCID ID: 0000-0002-5908-1285
**Assoc. Prof.,Van Yüzüncü Yıl University, Faculty of Education, Van-Türkiye, osmantat@yyu.edu.tr, ORCID ID: 0000-0003-2950-9647

To cite this article:
Aydoğan,İ. ,Tat,O., (2024). Investigating the performance of artificial neural networks in predicting. *Journal of Measurement and Evaluation in Education and Psychology, 16*(1), 1-12. https://doi.org/10.21031/epod.1525454
Received: 31.07.2024
Accepted: 19.11.2024

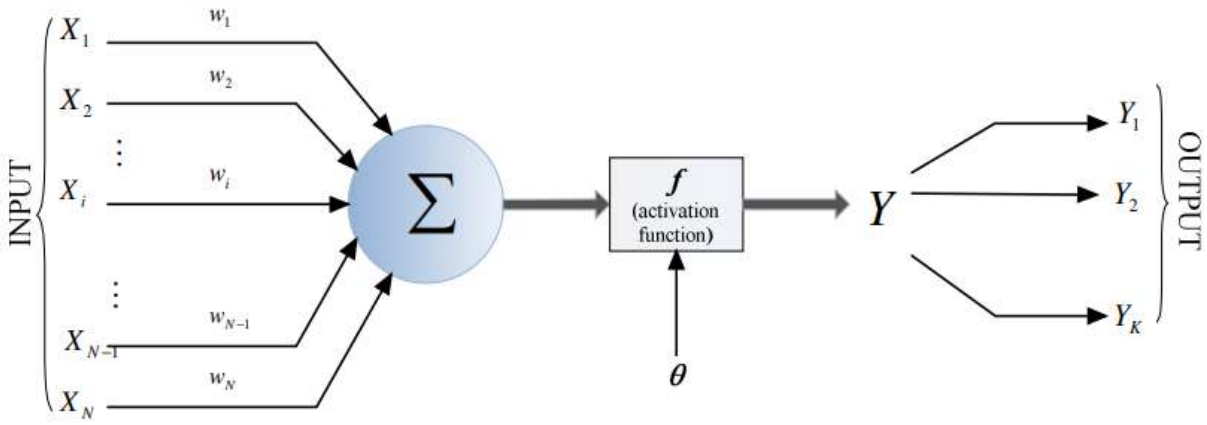
designed in these studies is to make predictions about students' cognitive performance through a number of covariate characteristics such as gender, parent's occupation, socio-economic status, etc. However, in this study, we aim to examine the performance of a network trained for a unidimensional scale, that is, using items that reflect the same latent trait, in predicting responses to another item with the same trait. In other words, using Programme for International Student Assessment (PISA) participants selected from the Lebanese sample, we examine how ANNs trained with responses to items on a scale predict responses to another item that is also part of the same scale. This provides an opportunity to compare the expected and observed values of the responses to a set of items with the same emotional integrity for the ANN trained based on machine learning. Therefore, in addition to determining the extent to which the responses predicted by ANNs overlap with the actual responses, we plan to monitor how the predicted items affect the latent construct and the reliability of the responses. In this context, we aim to evaluate the performance of the ANN method by performing similarity, validity and reliability analyses for actual and predicted responses. The focus here is to answer the question of to what extent we can accurately predict students' responses to another item that is part of the same emotion based on their emotional integrity, or to what extent the expected responses of individuals to a question posed within the same emotional integrity overlap with their responses. The reason for choosing the Lebanese sample is that Lebanon is a society where the emotional state reflected by the implicit feature in the scale we used is strongly experienced. We used the items of the student's intercultural communicative awareness scale administered in PISA 2018 as research data. The latent trait in the scale is the awareness of being able to communicate with people from different cultures. In 2018, Lebanon ranked first among the world countries in terms of the number of refugees per capita (McCarthy, 2019). The number of refugees per thousand inhabitants in Lebanon was 156 in 2018, which is more than twice the number in the second ranked country.

Artificial Neural Networks

ANNs are models that realize the features taught in the training phase through artificial neurons similar to the neuron structure in the human brain based on the principle of continuous improvement (Kose & Arslan, 2017; Vandamme et al., 2007). It has a complex and powerful structure that models non-linear relationships (Kardan et al., 2013). ANNs consist of three layers called input, hidden and output. The information transferred from the input data to the neurons in the input layer is processed through an aggregation function taking into account the weight values and transmitted to the activation function (See Figure 1).

Figure 1.

An Artifical Neuron (Grosan & Abraham, 2011, p.283).



The activation function is the component where calculations are made for the most accurate output. The information coming from the aggregation function is processed here to generate output values and transmitted to the output neurons (Rashid & Ahmad, 2016; Vandamme et al., 2007). In this process, the weight values are constantly adjusted to provide the best output. If the activation values reach the

threshold value during the current iterations, the training phase is terminated and the network has learned. In this phase, new examples are shown to the network to test the learning. After the training phase is completed, the weight values remain constant. In this way, it is ensured that the learned network produces output using the current weight values (Öztemel, 2003).

Methods

Data and Participants

The research data were obtained from the PISA 2018 assessment (https://www.oecd.org/pisa/data/2018database/). PISA is a monitoring and assessment program implemented by the Organisation for Economic Co-operationand Development (OECD) for fifteen-year-old students in many countries around the world. Measures such as demographic information about students and their families, learning environments, information and communication technologies (ICT) and financial competence, as well as affective and cognitive measuresare provided. Thanks to the measurements applied and the results obtained accordingly, it provides the opportunity for countries to evaluate their own education systems and to review the educational outcomes of other countries. In this context, PISA provides important data and results to educators, researchers and administrators in terms of monitoring and evaluating educational processes (OECD, 2018).
The research group consists of Lebanese students who participated in the PISA 2018 assessment. 5614 Lebanese students participated in the PISA 2018 assessment. However, in order to clean the missing data in the data set and to meet the assumptions of the analysis techniques used in the research, some data were deleted and the research was conducted with the remaining 4631 student data.

Variables

The variables of the study consisted of seven items of the student's intercultural communicative awareness (*Awacom*) scale, which PISA officials stated as a part of the global competence domain. *Awacom* includes items that measure individuals' awareness of communicating with people from different cultures (See Table 1).

Tablo 1.
Items of Student's Intercultural Communicative Awareness Scale

Items label	PISA codes	Items
Item1	ST218Q01HA	I carefully observe their reactions.
Item2	ST218Q02HA	I frequently check that we are understanding each other correctly.
Item3	ST218Q03HA	I listen carefully to what they say.
Item4	ST218Q04HA	I choose my words carefully.
Item5	ST218Q05HA	I give concrete examples to explain my ideas.
Item6	ST218Q06HA	I explain things very carefully.
Item7	ST218Q07HA	If there is a problem with communication, I find ways around it (e.g. by using gestures, re-explaining, writing etc.).

The data were obtained through PISA student questionnaires. The scale items are in a four-point Likert response format: strongly disagree-disagree-agree-agree-strongly agree.

Data Preprocessing

The data used in the study were derived from PISA 2018 Lebanese sample data. The Lebanese sample consists of 5614 students; however, since confirmatory factor analysis (CFA), which is a member of structural equation models (SEM), and artificial neural networks (ANN) techniques used in the analysis processes are affected by missing (Ennett et al., 2001; Tabachnick & Fidell, 2019) and extreme (Tabachnick & Fidell, 2019) values, data with these characteristics were deleted by list-based data deletion method. In order to meet the assumptions of the stated techniques, 4631 data suitable for the realization of the research were reached after the deleted data and the research was conducted with this data set.

Data Analysis

Before proceeding with the analysis procedures, CFA analysis was conducted to examine whether the assumed latent construct was provided in terms of the research data. SEM can be categorized in to two categories: measurement and structural models. In measurement models, observed variables and latent variables are associated (Şen, 2020). With the measurement model created, it was revealed whether the seven items in *Awacom* reflect the assumed students' awareness of being able to communicate. Since multivariate normality was not achieved as a result of Mardia's multivariate skewness and kurtosis statistics (skewness and kurtosis <.05) (Tabachnick & Fidell, 2019), the robust standard error maximum likelihood (MLR) estimator was used as a parameter estimator for CFA (Rosseel et al., 2024). The fit indices (RMSEA=.065, CFI=.973, TLI=.96, SRMR=.024) obtained by analyzing the measurement model indicate that *Awcom* can be represented by seven items (Hox et al., 2018). The Cronbach alpha value calculated to determine the reliability of the responses to these seven items was .89, indicating that the reliability of the responses was high (George & Mallery, 2003).

For the aim of the study, responses to two randomly selected items from *Awacom* were predicted by neural networks trained on responses to all items reflecting the same latent trait. The items whose responses were predicted were labeled as Item3 and Item7 (SeeTable 1). The ANN models used in the predictions were created according to different conditions of splitting the data at different rates and the number of layers and the number of neurons in the layers. In machine learning methods, of which ANN is a member, the data are divided in to two parts as training and testing sets in order to avoid the problem of overlearning. The models trained with train data are controlled through other data sets (Brownlee, 2020). For this reason, nine different models were created depending on the split ratio, number of layers and neurons. The models with the most appropriate RMSE values were used as prediction models. Prediction procedures were performed for the relevant items in the test data set. In the analysis processes performed with ANN, the items consisting of four ordinal categories (1-2-3-4) were scaled between 0 and 1 (0-.333-.667-1) according to the model's assumption (Brownlee, 2020). Estimations were made according to the scaled values. After the analysis, these values were converted back to ordinal values by considering close ranges. Accuracy ratio, marginal homogeneity test (Agresti, 2013) and Kappa (Cohen, 1960) statistics were used to reveal the similarity between the estimated values and the actual values for Item3 and Item7. In this way, similarities between predicted and actual values were determined. In addition, sensitivity analyses (Beck, 2018; Lek et al., 1996) were conducted to determine the relationship between responses to actual Item3 and Item7 items and responses to other items. Then, how the subsets containing the predicted Item3 and Item7 items and the subsets containing the actual Item3 and Item7 items represent the assumed *Awacom* latent construct was also examined through CFA analyses. Thus, it was observed how the model fit indices, variance explained by the items and standardized factor loadings changed. In addition, Cronbach's alpha value was used to investigate how the reliability of the responses for the actual and predicted subsets changed. R [caret package (Kuhn, 2023), lavan package (Rosseel et al., 2024), neuralnet package (Fritsch, 2019), neuralnettools package (Beck, 2022), nnet package (Ripley & Venables, 2023)] Mplus and SPSS statistical programs were used for analysis.

Findings

The findings obtained from the prediction of two items labeled Item3 and Item7 by the networks trained with the items of the *Awacom* scale are presented under two separate headings. As a reminder, the rationale for prediction is based on the performance of ANNs that learn from the responses to items on the same scale, i.e. items that measure similar attributes, in predicting the responses to each item on the scale. These findings include the selection method of the networks, the performance of the networks, how the predicted items relate to the other items in the scale, the similarity of the actual and predicted values, the fit metrics in the verification of the assumed latent trait over the test subsets formed by the predicted and actual items, the variance values explained by the items, factor loadings and reliability values.

Predicting Item3 Responses

ANN models with nine different features were developed for predicting the item labeled Item3, where Item3 is used as output data and the other six items are used as input data. The networks have different ratios of test-train data and different numbers of layers and neurons. In evaluating the performance of the networks, the RMSE values produced by the trained network on all, train and test data were taken into account (See Table 2). In selecting the best network, it was preferred that the RMSE value was small and close for all data sets.

Table 2.

Features of ANN Models for Predicting Item3 Responses

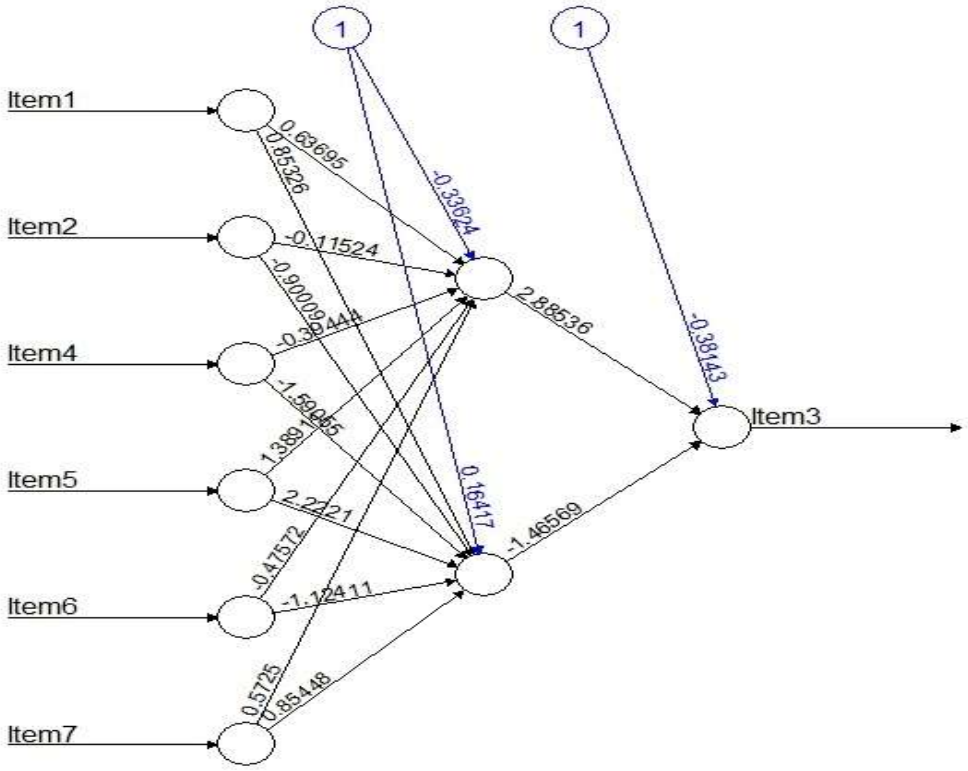
Models	Train/Test Spliting	Hiddens	RMSE		
			All	Train	Test
Model1*		2	.619	.618	.623
Model2	70/30	3	.617	.613	.625
Model3		3:2	.618	.612	.631
Model4		2	.619	.616	.629
Model5	75/25	3	.615	.609	.631
Model6		3:2	.614	.607	.636
Model7		2	.619	.617	.628
Model8	80/20	3	.618	.613	.638
Model9		3:2	.616	.612	.633

* Selected to best model

In this context, the most ideal model for predicting Item3 responses was found to be a single hidden layer network with two neurons in the hidden layer (See Figure 2). It is understood that the selected network performs well with 70-30% of the train and test data.

Figure 2.

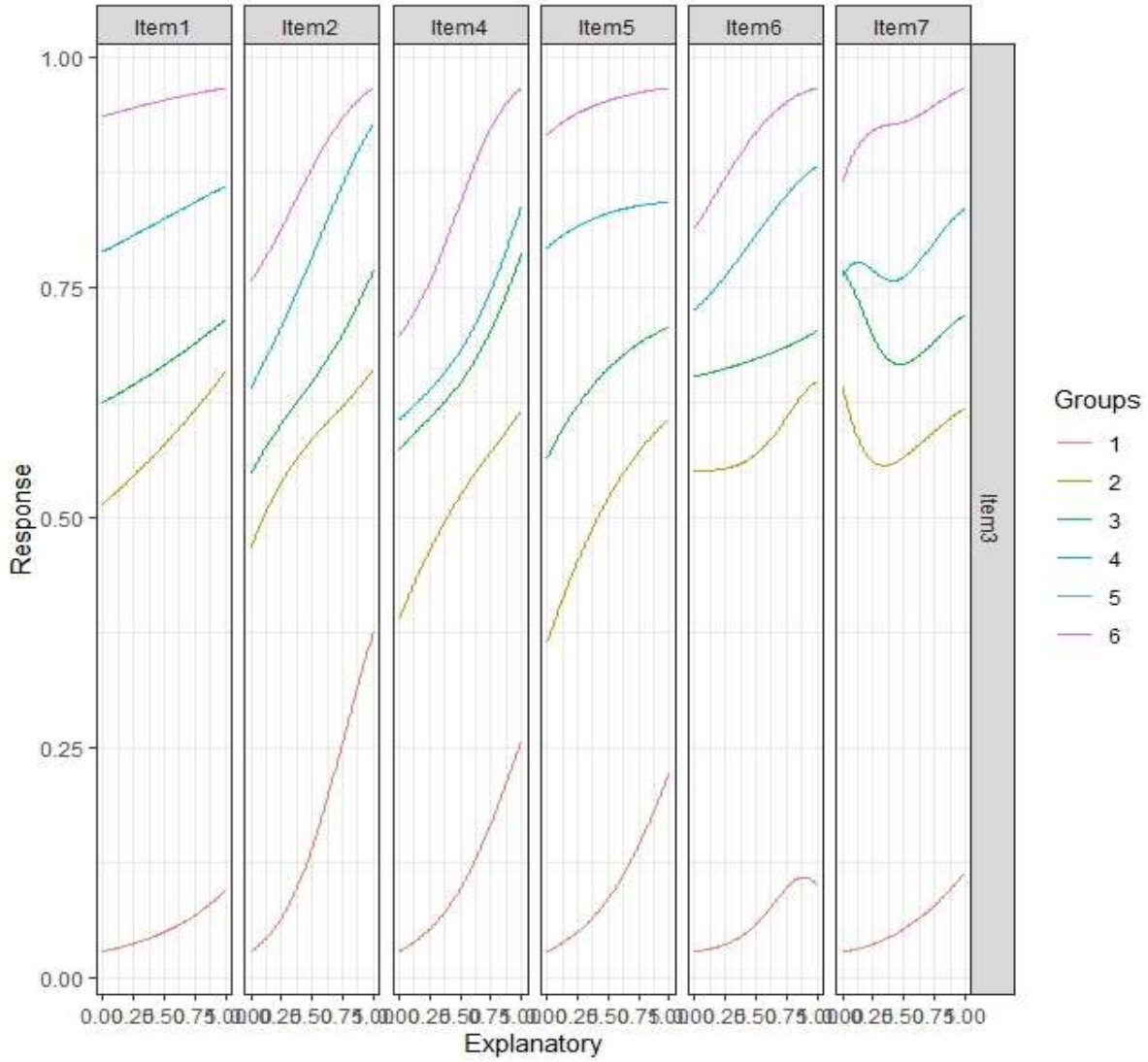
Network Structure of Model1



According to the output of the sensitivity analysis conducted to determine the relationship between the responses to Item3 predicted by Model 1 and the responses to other items, it is understood that the predicted responses to Item3 have a relatively linear relationship with the responses to other items (See Figure 3). It is noteworthy that the responses to the items other than Item7 are positively correlated with the responses to Item3. However, it is not possible to say that the same is the case for Item7. When the judgments expressed by the items are analyzed (See Table 1), the similarity and linearity of the relationship between the responses to the five items other than Item7 and the estimated Item3 values indicate that the predictions support the relevant latent construct (Beck, 2018; Lek et al., 1996).

Figure 3.

Results of Sensivity Analysis for Model1



Predicting Item7 Responses

Similarly, ANN models with nine different features were created where the output variable was Item7 and the input variables were the other six items in the *Awacom* scale. The differentiation in the networks is due to the differences in the ratio considered in the split of the data set and the number of layers and neurons. According to the RMSE values produced by the trained model for all, train and test datasets, the model with the smallest and closest RMSE values was selected as the best model (See Table 3).

Table 3.

Features of ANN Models for Predicting Item7 Responses

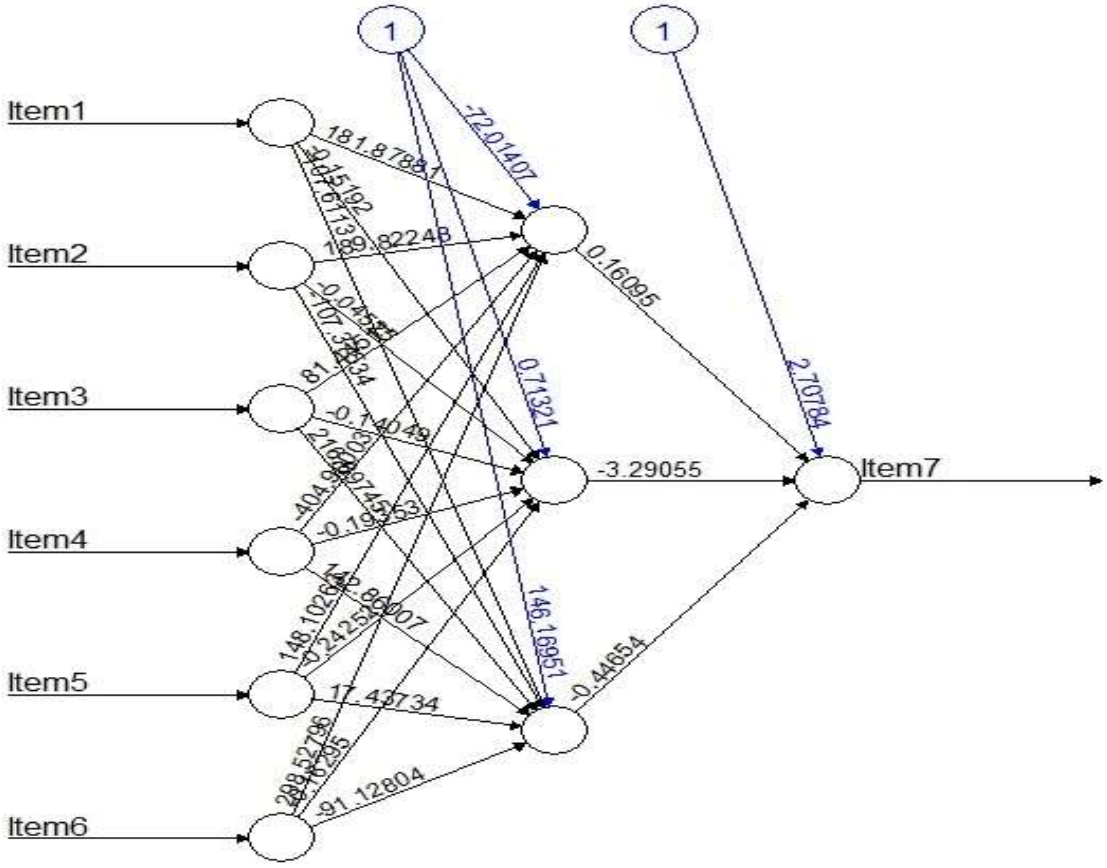
Models	Train/Test Spliting	Hiddens	RMSE		
			All	Train	Test
ModelA		2	.709	.699	.732
ModelB*	70/30	3	.708	.695	.734
ModelC		3:2	.708	.694	.740
ModelD		2	.710	.697	.750
ModelE	75/25	3	.710	.695	.752
ModelF		3:2	.706	.690	.753
ModelG		2	.709	.697	.759
ModelH	80/20	3	.708	.694	.762
ModelI		3:2	.707	.692	.765

* Selected to best model

Based on the RMSE values, it is understood that the lowest and closest values for all data sets are obtained at 70-30% separation of the data set. It is observed that the model in this group, which has ideal values, is a single interlayer network structure with three neurons (See Figure 4). Therefore, ModelB was preferred as the ideal model for predicting the responses to Item7.

Figure 4.

Network Structure of ModelB

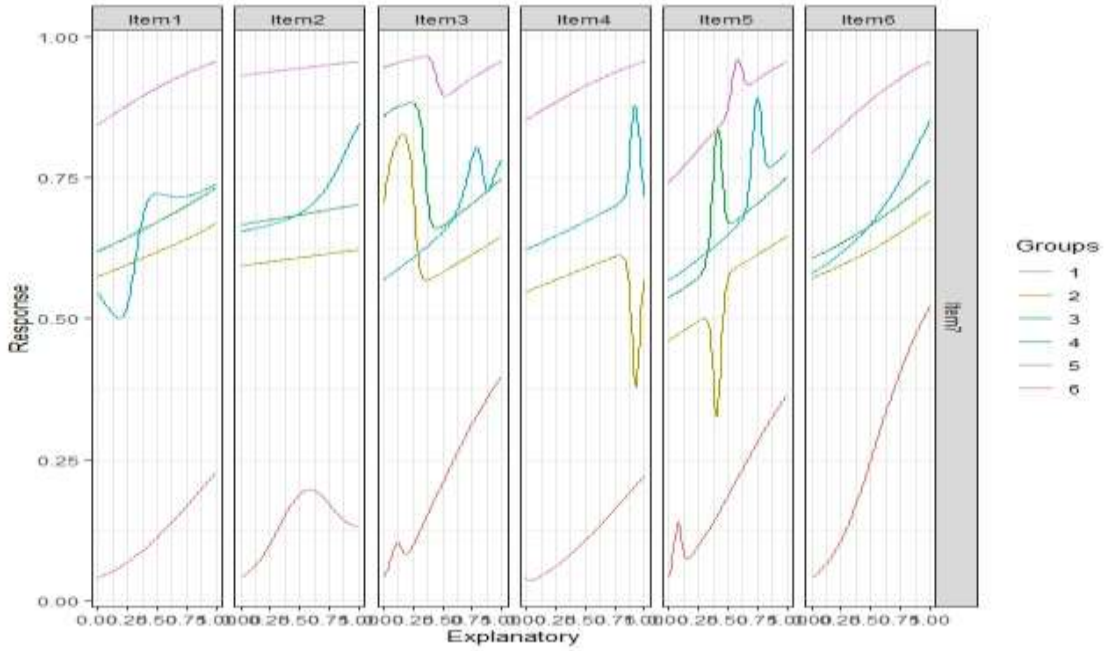


It can be said that the responses to Item7 estimated for ModelB are generally not in a linear relationship with the responses to the other six items used to train the model (See Figure 5). This

finding obtained by sensitivity analysis supports the relationship between Item3 estimated for Model1 and the other items. As observed in Figure 3, the responses to Item3 showed a similar and linear relationship with the responses to the other items except Item7. In this context, it is considered as an expected situation that the responses to Item7 have a non-linear relationship with the responses to other items.

Figure 5.

Results of Sensivity Analysis for ModelB



Performance of ANN Models for Actual and Predicted Values

Accuracy, marginal homogeneity test and Kappa statistics were used to determine how much the Item3 values predicted by Model1 and Item7 values predicted by ModelB corresponded to the actual values (See Table 4). It is understood that the responses to both estimated items are similar to the actual responses at an average rate of .60. There was no statistically significant difference between the mean responses to the two predicted items and the mean responses to the actual items (MH test, p >.05). According to Kappa values, there was a moderate similarity between predicted and actual Item3 values and a low similarity between predicted and actual Item7 values (Landis & Koch, 1977).

Table 4.

Similarity Values of Actual and Predicted Responses

Models	N	Match ratio	MH* test (p)	Kappa
Model1				
Actual Item3	1389	.63	>.05	.42
Predicted Item3				
ModelB				
Actual Item7	1389	.58	>.05	.37
Predicted Item7				

*Marginal homogeneity

CFA analyses were conducted to determine the extent to which the test subsets containing the estimated Item3 and Item7 and the test subsets containing the actual values provided the *Awacom* latent construct, and reliability analyses were conducted to determine how the reliability of the responses to the items estimated for these subsets affected the reliability (See Table 5). At the same time, the variance values and factor loadings explained by the predicted items for the latent construct were examined. The findings revealed that unlike similarity analyses, construct validity and reliability analyses increased the reliability of the responses to the predicted items and supported the latent construct. It shows that the model fit indices were relatively weakened by the two estimated items, but the model fit remained strong. On the other hand, it is understood that the variance values explained by the responses to the estimated items for the latent construct, the standardized factor loadings for the latent construct and the reliability values improved.

Table 5.

Comprassion Model Performance of Actual and Predicted Test Subsets

Subsets	RMSEA	CFI	TLI	R-square	Loading	Alpha
For Item3						
Actual subset	.081	.972	.959	.613	.783	.891
Predicted subset	.106	.965	.948	.983	.992	.903
For Item7						
Actual subset	.081	.972	.959	.489	.700	.891
Predicted subset	.111	.963	.944	.942	.970	.909

Conclusion and Discussion

ANNs, which evolved from the idea of simulating the human brain, provide significant advantages for the realization of many researches with different purposes and content due to its ability to model complex, non-linear relationships, unlike traditional statistical methods, its excellent fault tolerance, and its ability to be a fast and highly scalable machine learning (Zou et al., 2009). This study takes advantage of these important advantages of ANNs and examines how ANNs predict the responses of Lebanese students participating in PISA 2018 to other items that are part of the same emotion based on the integrity of emotion. Using some of the data, the network trained with the items of PISA's student's intercultural communicative awareness scale was able to predict two items of the same scale from another data set.

It is understood that the values predicted by ANN solutions for the responses to two randomly selected items match the actual values at a relatively moderate level. In this context, the values produced by the trained network are expected responses depending on the emotional integrity shaped by the responses to the items. Therefore, the responses expected from individuals for the items examined overlap with the responses given by individuals at a relatively moderate level. However, in the validity and reliability analyses conducted for the latent trait represented by the predicted items together with the other items, it is observed that although the predicted values partially weaken the model fit indices, they still manage to keep them strong. In addition, the estimated values improved the factor loadings and the variance explained for the latent trait. Similarly, when the latent trait aspect is considered, it is noticed that the estimated values also positively affect the reliability.

Especially in recent years, the use of advanced versions of ANNs such as convolutional neural networks (CNN), recurrent neural networks (RNN), emotional neural network (EANN) for predicting individual emotions based on machine learning using images, text, dialog, body movements, etc. in deep emotional fields such as affective state, affective computing, deep learning (Ashwin & Guddeti, 2020; Bakkialakshmi et al., 2022; Carstensen et al., 2016; Chan et al., 2020; Feng, 2022; Jadhav

&Sugandhi, 2018; Jamisola, 2016; Liu et al., 2023; Orozco-del-Castillo et al., 2021; Wang et al., 2022) has gained rapid momentum. Research using these techniques focuses on the relationship between the development of emotional responses by utilizing physiological responses. However, our research is based on a much simpler logic and purpose than these studies. We used ANN to predict the responses to an item in the same emotional state by utilizing emotional integrity. Noting that most of the researches (Aybek & Okur, 2018; Aydoğan & Zırhlıoğlu, 2018; Aydoğdu, 2020; Al-Saleem et al., 2015; Chavez et al., 2023; Flitman, 1997; Guarín et al., 2015; Huang & Fang, 2012; Lau et al., 2019; Rodríguez-Hernández, 2021; Shahiri & Husain, 2015; Umar, 2019; Zacharis, 2016) conducted with ANN in education and psychology are for cognitive prediction by utilizing covarities, we evaluated to what extent this performance of ANNs can be used to predict the responses to any item in scale applications frequently used in education and psychology.

Funding
The research was not financially supported by any institution or organisation.

Declarations

Conflict of Interest: The authors have no relevant financial or non-financial interests to disclose.

Ethical Approval: We declare that all ethical guidelines for the author have been followed. This study does not require any ethics committee approval as it includes open-access data.

The authors of this article declare (Declaration Form #: 2711240128) that Gen-AI tools have NOT been used in any capacity for content creation in this work.

References

Agresti, A. (2013). *Categorical data analysis.* Wiley.

Al-Saleem, M., Al-Kathiry, N., Al-Osimi, S., & Badr, G. (2015). Mining educational data to predict students' academic performance. In *Machine Learning and Data Mining in Pattern Recognition: 11th International Conference, MLDM 2015, Hamburg, Germany, July 20-21, 2015, Proceedings 11* (pp. 403-414). Springer International Publishing.

Ashwin, T. S., & Guddeti, R. M. R. (2020). Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and Information Technologies*, *25*(2), 1387-1415. https://doi.org/10.1007/s10639-019-10004-6

Aybek, H. S. Y., & Okur, M. R. (2018). Predicting achievement with artificial neural networks: The case of Anadolu University open education system. *International Journal of Assessment Tools in Education*, *5*(3), 474-490. https://doi.org/10.21449/ijate.435507

Aydoğan, İ., & Zırhlıoğlu, G. (2018). Öğrenci başarılarının yapay sinir ağları ile kestirilmesi. *Van Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, *15*(1), 577-610. http://dx.doi.org/10.23891/efdyyu.2018.80

Aydoğdu, Ş. (2020). Predicting student final performance using artificial neural networks in online learning environments. *Education and Information Technologies*, *25*(3), 1913-1927. https://doi.org/10.1007/s10639-019-10053-x

Bakkialakshmi, V. S., Sudalaimuthu, T., & Winkler, S. (2022). Effective Prediction System for Affective Computing on Emotional Psychology with Artificial Neural Network. *Easy Chair Preprint*.

Beck, M.W. (2018). NeuralNetTools: Visualization and Analysis Tools for Neural Networks. *Journal of Statistical Software*, *85*(11), 1 .https://doi.org/10.18637/jss.v085.i11

Beck, M.W. (2022). Visualization and analysis tools for neural networks, R package version 1.5.3. Retrieved from https://cran.r-project.org/web/packages/NeuralNetTools/index.html

Brownlee, J. (2020). *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery.

Carstensen, S. L., Madsen, J., & Larsen, J. (2016). Predicting Changes in Affective States using Neural Networks. *arXiv preprint arXiv:1612.00582*. https://doi.org/10.48550/arXiv.1612.00582

Chan, K. Y., Kwong, C. K., Wongthongtham, P., Jiang, H., Fung, C. K., Abu-Salih, B., ... & Jain, P. (2020). Affective design using machine learning: a survey and its prospect of conjoining big data. *International Journal of Computer Integrated Manufacturing*, *33*(7), 645-669. https://doi.org/10.1080/0951192X.2018.1526412

Chavez, H., Chavez-Arias, B., Contreras-Rosas, S., Alvarez-Rodríguez, J. M., & Raymundo, C. (2023, February). Artificial neural network model to predict student performance using nonpersonal information. In *Frontiers in Education* (Vol. 8, p. 1106679). Frontiers Media SA.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Ennett, C. M., Frize, M., & Walker, C. R. (2001). Influence of missing values on artificial neural network performance. In *MEDINFO 2001* (pp. 449-453). Ios Press.

Feng, H. (2022). A Novel Adaptive Affective Cognition Analysis Model for College Students Using a Deep Convolution Neural Network and Deep Features. *Computational Intelligence and Neuroscience*, *2022*(1), 2114114. https://doi.org/10.1155/2022/2114114

Flitman, A. M. (1997). Towards analysing student failures: neural networks compared with regression analysis and multiple discriminant analysis. *Computers & Operations Research*, *24*(4), 367-377. https://doi.org/10.1016/S0305-0548(96)00060-3

Fritsch, S., Guenther, F., Wright, M.N., Suling, M., Mueller, S.M. (2019). Training of neural Networks, R package version 1.44.2. Retrieved from https://cran.r-project.org/web/packages/neuralnet/index.html

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Allyn & Bacon.

Goel, A., Goel, A. K., & Kumar, A. (2023). The role of artificial neural network and machine learning in utilizing spatial information. *Spatial Information Research*, *31*(3), 275-285. https://doi.org/10.1007/s41324-022-00494-x

Grosan, C., & Abraham, A. (2011). Artificial neural networks. *Intelligent Systems: A Modern Approach*, 281-323.

Guarín, C. E. L., Guzmán, E. L., & González, F. A. (2015). A model to predict low academic performance at a specific enrollment using data mining. *IEEE Revista Iberoamericana de tecnologias del Aprendizaje*, *10*(3), 119-125. https://doi.org/10.1109/RITA.2015.2452632

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, *105*(3), 416-434. https://doi.org/10.1016/j.neuron.2019.12.002

Hox, J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge.

Huang, S., & Fang, N. (2012, October). Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques. In *2012 Frontiers in Education Conference Proceedings* (pp. 1-2). IEEE.

Jadhav, N., & Sugandhi, R. (2018, November). Survey on human behavior recognition using affective computing. In *2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN)* (pp. 98-103). IEEE.

Jamisola, R. S. (2016). Conceptualizing a Questionnaire-Based Machine Learning Tool that Determines State of Mind and Emotion. *Lovotics*, *4*(115), 2. http://dx.doi.org/10.4172/2090-9888.1000115

Kardan, A. A., Sadeghi, H., Ghidary, S. S., & Sani, M. R. F. (2013). Prediction of student course selection in online higher education institutes using neural network. *Computers & Education*, *65*, 1-11. https://doi.org/10.1016/j.compedu.2013.01.015

Kose, U., & Arslan, A. (2017). Optimization of self-learning in Computer Engineering courses: An intelligent software system supported by Artificial Neural Network and Vortex Optimization Algorithm. *Computer Applications in Engineering Education*, *25*(1), 142-156. https://doi.org/10.1002/cae.21787

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., … & Hunt, T. (2023). Classification and regression training, R package version 6.0-94. Retrieved from https://cran.r-project.org/web/packages/caret/index.html

Landis, J, R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.https://doi.org/10.2307/2529310

Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences*, *1*(9), 982. https://doi.org/10.1007/s42452-019-0884-7

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., & Aulagnier, S. (1996). Application of neural networks to modeling nonlinear relationships in ecology. *Ecological Modelling*, *90*, 39–52. https://doi.org/10.1016/0304-3800(95)00142-5

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*(6), 335-346. https://doi.org/10.1038/s41583-020-0277-3

Liu, J., Ang, M. C., Chaw, J. K., Kor, A. L., & Ng, K. W. (2023). Emotion assessment and application in human–computer interaction interface based on backpropagation neural network and artificial bee colony algorithm. *Expert Systems with Applications*, *232*, 120857. https://doi.org/10.1016/j.eswa.2023.120857

McCarthy, N. (June19, 2019). Lebanon has by far the most refugees per 1,000 population. Retrieved from https://www.statista.com/chart/8800/lebanon-has-by-far-the-most-refugees-per-capita/

OECD. (2018). PISA 2018 Technical report. OECD Publishing, Retrieved from https://www.oecd.org/pisa/data/pisa2018technicalreport/

Orozco-del-Castillo, M. G., Orozco-del-Castillo, E. C., Brito-Borges, E., Bermejo-Sabbagh, C., & Cuevas-Cuevas, N. (2021, November). An artificial neural network for depression screening and questionnaire refinement in undergraduate students. In *International Congress of Telematics and Computing* (pp. 1-13). Springer International Publishing.

Öztemel, E. (2003).*Yapay sinir ağları.* PapatyaYayıncılık.

Park, C. W., Seo, S. W., Kang, N., Ko, B., Choi, B. W., Park, C. M., ... & Yoon, H. J. (2020). Artificial intelligence in health care: Current applications and issues. *Journal of Korean Medical Science*, *35*(42). https://doi.org/10.3346/jkms.2020.35.e379

Rashid, T. A., & Ahmad, H. A. (2016). Lecturer performance system using neural network with Particle Swarm Optimization. *Computer Applications in Engineering Education*, *24*(4), 629-638. https://doi.org/10.1002/cae.21737

Ripley, B., & Venables, W. (2023). Feed-forward neural networks and multinomial log-linear models, R package version 7.3-18. Retrieved from https://cran.r-project.org/web/packages/nnet/index.html

Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., & Cascallar, E. (2021). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. *Computers and Education: Artificial Intelligence*, *2*, 100018. https://doi.org/10.1016/j.caeai.2021.100018

Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., … & Jorgensen, T. (2024). Latent variable analysis, R package version 0.6-18. Retrieved from https://cran.r-project.org/web/packages/lavaan/lavaan.pdf

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414-422. https://doi.org/10.1016/j.procs.2015.12.157

Şen, S. (2020). *Mplus ile yapısal eşitlik modellemesi uygulamaları.* Nobel Akademik Yayıncılık.

Tabachnick, B. G., & Fidell, L. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, *49*(11), 1225-1231. https://doi.org/10.1016/S0895-4356(96)00002-9

Umar, M. A. (2019). Student academic performance prediction using artificial neural networks: A case study. *International Journal of Computer Applications*, *178*(48), 24-29. http://dx.doi.org/10.5120/ijca2019919387

Vandamme, J. P., Meskens, N., & Superby, J. F. (2007). Predicting academic performance by data mining methods. *Education Economics*, *15*(4), 405. https://doi.org/10.1080/09645290701409939

Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., ... & Zhang, W. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, *83*, 19-52. https://doi.org/10.1016/j.inffus.2022.03.009

Zacharis, N. Z. (2016). Predicting student academic performance in blended learning using artificial neural networks. *International Journal of Artificial Intelligence and Applications*, *7*(5), 17-29. http://dx.doi.org/10.5121/ijaia.2016.7502

Zou, J., Han, Y., & So, S. S. (2009). Overview of artificial neural networks. *Artificial neural networks: methods and applications*, 14-22. https://doi.org/10.1007/978-1-60327-101-1_2

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2025; 16(1); 13-29

Latent Class Analysis and DIF Testing in Mathematics Achievement: A Comparative Study of Korea and Türkiye Using MIMIC Modeling

Hyo Seob SONG* Hee Sun JUNG**

Abstract

This study examines the latent classes of mathematics achievement and investigates differential item functioning (DIF) between Korea and Türkiye. Moreover, it explores the influence of the country on the latent classes of mathematics achievement. To achieve this, data from eighth-grade students in TIMSS 2019 were analyzed using Latent Class MIMIC Modeling. The findings uncovered diverse latent classes of math achievement and detected both uniform and Non-uniform DIF between Korea and Türkiye. Furthermore, the country was found to significantly affect the latent class membership of math achievement. This study highlights the necessity of verifying the measurement invariance of indicator variables in latent class analysis (LCA). It also sheds light on areas where students performed favorably or unfavorably in mathematics achievement tests across these countries by investigating DIF. These findings have important implications for mathematics education in Korea and Türkiye.

Keywords: Mathematics Achievement; Latent Class Analysis (LCA); Multiple Indicator Multiple Cause (MIMIC) Modeling; Measurement Invariance; Differential Item Functioning (DIF); TIMSS 2019

Introduction

Mathematics significantly influences students' academic success and future career prospects (Guhl, 2019; Lubinski et al., 2014). Researchers in mathematics education have utilized international comparative studies (e.g., TIMSS, PISA) to evaluate students' academic achievement (Arıcan et al., 2016; Badri, 2019; Wang et al., 2023; Wiberg, 2019). Since its inception in 1995, the Trends in Mathematics and Science Study (TIMSS) has played a crucial role in assessing national-level mathematics achievement by comparing the relative performance of participating countries over time. Participating countries use assessment results to improve their educational curricula and methods or to enhance achievement (Lee & Stankov, 2018; Şen & Arıcan, 2015). Additionally, TIMSS promotes efforts to advance STEM (Science, Technology, Engineering, Mathematics) education by providing participating countries with data on students' mathematics and science achievement levels (Geesa et al., 2020; Mullis & Martin, 2017). According to the results of TIMSS 2019 conducted by the IEA, there were differences in mathematics achievement among participating countries. Korea achieved a high level of achievement in mathematics, ranking among the top performers, while Türkiye recorded achievement around the international average (Mullis et al., 2020). Such differences in mathematics achievement among countries may arise from students' home resources, attitudes toward mathematics,

* Assistant Professor, U1 University, Institute for Educational Innovation, Chung cheong do - South Korea, songh-s@hanmail.net, ORCID ID: 0000-0001-7554-2849
** Professor, Sungkyunkwan University, Department of Mathematics Education, Seoul - South Korea, hsun90@skku.edu, ORCID ID: 0000-0003-0093-2193

To cite this article:
Song, H. S. & Jung, H. S. (2025). Latent class analysis and dif testing in mathematics achievement: A comparative study of Korea and Türkiye using MIMIC modeling. *Journal of Measurement and Evaluation in Education and Psychology, 16*(1), 13-29. https://doi.org/10.21031/epod.1539828

Received: 28.08.2024
Accepted: 7.03.2025

and cultural differences (Geesa et al., 2019; Klieme & Baumert, 2001), as well as variations in educational curricula across countries (Sohn, 2010). Particularly interesting in the results of TIMSS 2019 between Korea and Türkiye is that while Korea's mathematics achievement was significantly higher than that of Türkiye, Turkish students showed higher mathematics attitudes related to affective achievement compared to Korean students (Mullis et al., 2020). Korea's high mathematics achievement can be attributed to its society's strong emphasis on education, competitive examination, and selection systems (Im & Park, 2010), as well as participation in additional extracurricular education beyond school classes (Dittrich & Neuhaus, 2023; Shin et al., 2019; Woo & Hodges, 2015). Also contributing to the high math achievement of Korean students is the high quality of public education (Im & Park, 2010; Şen & Arıcan, 2015), which includes the implementation of constructivist teaching methods (Hwang & Hwang, 2008) and the competence of math teachers (Ko & Jung, 2020).

Recently, finite mixture models such as Latent Class Analysis (LCA) have been utilized across various research fields, including behavioral science, education, and psychology. Generally, research that applies finite mixture models involves investigating the relationship between predictor and latent class membership (Masyn, 2017; Song et al., 2023; Vermunt, 2010). The integration of predictors and the results of latent class membership has been evolving, and discussions have been held in several studies regarding the timing and method of including predictor variables in mixture models (Masyn, 2017; Nylund-Gibson & Masyn, 2016). Particularly, the 3-step method in latent class modeling is known to produce more robust and accurate results compared to the 1-step method. This is because it excludes covariates in the step of class enumeration, thereby eliminating the risk of class composition varying depending on covariates. However, previous studies have reported that biased estimates of the effects of covariates on latent class variables may occur if the direct effects of covariates on indicator variables are ignored in the 3-step method (Asparouhov & Muthén, 2014; Masyn, 2017). This implies that to estimate the effects of covariates on latent class variables, it is necessary to conduct measurement invariance tests. These tests confirm the direct effects of covariates on each indicator variable within each latent class. This process follows the completion of class enumeration using an unconditional latent class model in the first step of the 3-step method. Based on previous studies that have shown ignoring the direct effects of covariates on indicator variables in LCA can lead to biased estimates of the effects of covariates on latent classes (Clark & Muthén, 2009; Nylund-Gibson & Choi, 2018), Masyn (2017) proposed a method for detecting these direct effects in LCA. Masyn's method combines LCA with the multiple indicator multiple cause (MIMIC) model to confirm the measurement invariance of indicator variables across covariates. This approach enables accurate estimation of the effects of covariates on latent classes and exploration of DIF of indicator variables by covariates. DIF in latent class MIMIC models refers to items where individuals belonging to the same latent class exhibit different expected responses depending on the values of covariates (Masyn, 2017). Uniform DIF is assessed when the difference in expected responses to indicators by covariates is consistent across all classes, while non-uniform DIF is assessed when the difference in expected responses to indicators by covariates varies across one or more classes (Masyn, 2017). Latent classes emerge when not all members exhibit homogeneous response patterns (De Ayala et al., 2002; Samuelsen, 2008). Particularly, results of exploring DIF obtained from the entire population may be biased, thus studies on DIF should be examined across latent classes (Saaatcioglu, 2022). In the studies by Tsaousis, Sideridis, AlGhamdi (2020) and Saaatcioglu (2022), the method proposed by Masyn (2017) was used to explore gender-specific DIF in achievement tests, investigating DIF by gender in the latent class of academic achievement.

To compare academic achievement among countries with different languages and cultures, scale measurement invariance must be secured first (Hambleton, 2001). Recently, a growing body of research has focused on assessing and exploring the causes of measurement invariance across different languages, cultures, and countries in international achievement tests (Demirus & Pektas, 2022; Im & Park, 2010; Sohn, 2010; Yoon & Lee, 2013). Most of these studies apply the technique of DIF to assess the level of equivalence at the item level. For example, Im and Park (2010) compared the mathematics scores of 8th-grade students in Korea and the United States using TIMSS 2003 data, revealing variations in problem reformulation, inference, measurement, and geometry. Demirus and Pektas (2022) examined the presence of DIF in the multiple-choice items of the TIMSS 2015 science achievement test across

various countries, including Türkiye, Australia, New Zealand, Morocco, and Egypt. Their study confirmed that more instances of DIF were observed between countries with diverse cultures and languages, suggesting that language variations contributed to DIF. Sohn (2010) identified DIF between Korean and Finnish students using PISA 2006 mathematics test data. Yoon and Lee (2013) investigated DIF on the TIMSS 2007 mathematics test among students from Korea, the United States, and Singapore. International comparative research using DIF enables the identification of item characteristics that function differentially when compared across countries, even when individuals have similar abilities. This provides insights into the strengths and weaknesses of domestic students and serves as foundational data for improving educational curricula and environments (Sohn, 2010).

This study aims to explore the latent classes of mathematics achievement in the TIMSS 2019 mathematics assessment using the Latent Class MIMIC Modeling proposed by Masyn (2017). It focuses on 8th-grade students in two countries: Korea, the top-performing country on the TIMSS 2019 mathematics test, and Türkiye, which performs around the international average but has been steadily increasing its achievement since joining TIMSS. Additionally, this study explores differential item functioning (DIF) to verify measurement invariance in the mathematics achievement test between Korea and Türkiye. DIF occurs due to violations of measurement invariance across different subgroups (Huang, 2020). Furthermore, it investigates the influence of the country (Korea/Türkiye) on the latent class membership of mathematics achievement. The research questions of this study are as follows:

1. How are latent classes of mathematics achievement identified in combined Korean and Turkish students?

2. Does DIF exist in the mathematics achievement test between Korea and Türkiye?

3. Does the country (Korea/Türkiye) influence the latent class membership of mathematics achievement?

Methods

Data

In this research, data from 8th-grade students in South Korea and Türkiye who participated in TIMSS 2019 were examined. The Trends in International Mathematics and Science Study (TIMSS) is an international assessment of academic performance organized by the International Association for the Evaluation of Educational Achievement (IEA). This assessment measures students' mathematics and science achievements at a global level to evaluate and enhance educational outcomes (Mullis et al., 2020). Initiated in 1995, TIMSS is conducted every four years, targeting 4th-grade and 8th-grade students. The assessment includes mathematics and science achievement tests based on the curricula of the participating countries, along with surveys of schools, teachers, students, and parents about educational contextual factors (Mullis et al., 2020). The TIMSS 2019 8th-grade mathematics assessment comprises 211 items. The framework is divided into two dimensions: the content dimension (Number, Algebra, Geometry, Data and Probability) detailing the subject matter, and the cognitive dimension (Knowing, Applying, and Reasoning) outlining the thinking processes evaluated as students engage with the content (Mullis & Martin, 2017).

A final sample of 553 South Korean students and 582 Turkish students, who participated in Booklet 5 and 6 of the TIMSS 2019 8th-grade mathematics assessment, was selected for this study, as shown in Table 1. The analysis included items from Block 6 of Booklets 5 and 6. Item ME62342, which had missing data for all countries, was excluded from the analysis. Thus, a total of 14 items were analyzed.

Table 1

The number of cases for analysis

Booklet	Excluded Items	Optional Items	Excluded Items	Korea	Türkiye	Total
Booklet 5	Block5	Block6 (ME62150~ ME62123B)		282	290	572
Booklet 6			Block7	271	292	563
Total				553(48.7%)	582(51.3%)	1,135(100%)

Data Analysis

In this research, Latent Class MIMIC Modeling was applied to identify the latent classes of mathematics performance and to examine measurement invariance and DIF of mathematics test items between Korea and Türkiye. Before conducting the analysis, test items were coded with correct answers as 1 and incorrect answers as 0. The country variable was coded as 1 for Korea and 0 for Türkiye. To determine the best number of latent classes for mathematics achievement, various criteria such as information criteria, scree plots, and entropy indices were used, along with considerations for interpretability and discriminant validity between groups (Ram & Grimm, 2009). Likelihood ratio tests were utilized to compare latent class MIMIC models, with effect sizes of identified DIF items evaluated using the Educational Testing Service (ETS) criteria. According to ETS guidelines, a logit value below 0.43 suggests a negligible DIF effect, a value of 0.43 or higher indicates a moderate effect, and a value of 0.64 or higher points to a large effect (Dorans & Holland, 1992). The analysis was performed using Mplus (Version 8.3) and the MplusAutomation package in R (Version 4.2.2), adhering to the method proposed by Masyn (2017), with some modifications detailed as follows:

Step 0: Conduct LCA to identify the optimal number of latent classes. Covariates are included as auxiliary variables to ensure they do not affect the identification of latent classes.

Step 1: Compare a baseline model (M_1.0, No_DIF), where covariates affect latent classes but not indicator variables, with an alternative model (M_1.1, All_DIF), where covariates directly affect both latent classes and all indicator variables. Acceptance of the baseline model (M_1.0) indicates no DIF for individual indicators by covariates, while acceptance of the alternative model (M_1.1) suggests the presence of DIF items for individual indicators by covariates, indicating at least one DIF item in at least one latent class.

Step 2: Conduct an omnibus DIF test to examine DIF for each indicator variable by covariates. This involves comparing model M_2.0.X (covariates affect latent classes but not indicator variables) with model M_2.1.X (covariates have direct effects on both latent classes and indicator variables).

Step 3: Select the optimal model by comparing model M_3.0, where all identified DIF items are treated as non-uniform DIF, with the baseline model (M_1.0, No_DIF) and the alternative model (M_1.1, All_DIF).

Step 4: Determine if the items identified as DIF in Step 2 are uniform DIF items by comparing the fit of model M_4.X (imposes uniform constraints on covariate effects on indicator variables across classes) with model M_3.0 (treats all identified DIF items as non-uniform DIF). If the fit of M_4.X is not significantly worse than that of M_3.0, the item is considered a uniform DIF.

Step 5: Choose the optimal model by comparing the fit of model M_5.0 (covariate effects on indicator variables are equal across latent classes for all identified uniform DIF items) with model M_3.0 (treats all identified DIF items as non-uniform DIF).

Step 6: Select the final model by comparing model M_6.0 (regression coefficients of covariates on latent class membership are constrained to 0) with model M_6.1 (regression coefficients of covariates on latent class membership are freely estimated) in the model chosen from Step 5.

Results

Descriptive Statistics

When examining the item difficulty index for each item in both Korea and Türkiye, it was found that the item difficulty index for all items was higher in Korea compared to Türkiye. Specifically, as shown in Table 2, the item difficulty index for Korean students ranged from 0.41 to 0.92, whereas for Turkish students ranged from 0.09 to 0.60. Particularly, in item 11, the difference in item difficulty between the two countries was 0.57, indicating the largest discrepancy. Additionally, for items 2, 3, 6, 7, 9, and 14, the difference in item difficulty index between the two countries exceeded 0.3, highlighting a notable variation in item difficulty.

Table 2

Math 8th Block6 Item

No	Variable	Domain	Label	Item difficulty index	
				Korea	Türkiye
1	ME62150	Number/Knowing	"DIFFERENCE BETWEEN LOW TEMPERATURE IN CITY X AND Y"	0.79	0.50
2	ME62335	Number/Knowing	"SELECT EQUIVALENT RATIO TO 3:2"	0.92	0.60
3	ME62219	Number/Applying	"KATY ENLARGES A PHOTO - NEW HEIGHT"	0.74	0.38
4	ME62002	Number/Reasoning	"FILL IN BOXES TO MAKE THE SMALLEST PRODUCT"	0.48	0.31
5	ME62149	Algebra/Applying	"IDENTIFY EXPRESSION TO CALCULATE ROBIN'S EARNINGS"	0.48	0.35
6	ME62241	Algebra/Applying	"ROY'S PHONE BUSINESS - EQUATION FOR Y"	0.70	0.27
7	ME62105	Algebra/Reasoning	"AREA OF RECTANGLE WITH SIDES X AND 2X + 1"	0.65	0.27
8	ME62040	Geometry/Applying	"ESTIMATE AREA OF IRREGULAR SHAPE ON 1 CM GRID"	0.60	0.46
9	ME62288A	Geometry/Applying	"FIND VERTICES OF TRAPEZOIDS M AND N"	0.41	0.11
10	ME62288B	Geometry/Applying	"FIND VERTICES OF TRAPEZOIDS M AND N"	0.41	0.09

Table 2 (Continued)

Math 8th Block6 Item

No	Variable	Domain	Label	Item difficulty index	
				Korea	Türkiye
11	ME62173	Geometry/Reasoning	"FIND ANGLE X ON A FOLDED PIECE OF PAPER"	0.76	0.19
12	ME62133	Data and Probability/Applying	"BLACK AND WHITE MARBLES IN A BAG WITH REPLACEMENT"	0.70	0.54
13	ME62123A	Data and Probability/Knowing	"RELAY RACE - MEAN TIME OF RUNNERS "	0.81	0.59
14	ME62123B	Data and Probability/Applying	"RELAY RACE - MEAN TIME WHEN 2 RUNNERS IMPROVE"	0.72	0.36

Measurement Invariance and DIF

Step 0: Before verifying the measurement invariance of the indicator variables and exploring the presence of DIF according to covariates, it is essential to select the optimal number of latent classes. To achieve this, latent class analysis on mathematics achievement was conducted without including covariates, identifying latent classes among the combined Korean and Turkish students. The optimal number of latent classes was determined by comparing the model fit and simplicity indicators as presented in Table 3. As shown in Table 3, the best fit was observed when there were five latent classes. With a large sample size, the values of AIC and BIC tend to decrease as the number of groups increases, and the number of latent classes can be determined using a scree plot (Jedidi et al., 1997).

Examination of the scree plot in Figure 1 reveals that the values of most goodness-of-fit indicators decrease at a slower rate after three latent classes, and AWE shows an increase after three latent classes. Additionally, when there are three latent classes, the entropy index is 0.905, indicating good performance. After considering factors such as goodness-of-fit indices, statistical significance, discriminant between groups, presence of latent classes, and interpretability, the optimal number of latent classes was determined to be three.

Upon examining the composition of classified latent classes in Figure 2, Class 1 (284 participants, 25.0%) exhibited a generally high item difficulty index of over 0.7 for each item, indicating the highest level of mathematics achievement among the three latent classes. Class 2(377 participants, 33.2%) showed moderate levels of mathematics achievement among the three latent classes, with significant differences in item difficulty index for each item. Notably, the item difficulty index for Geometry/Applying items 9 and 10 were below 0.1. Class 3 (474 participants, 41.8%) exhibited an item difficulty index generally below 0.4 across all items, indicating the lowest level of mathematics achievement among the three latent classes.

Consequently, Class 1 to Class 3 were respectively named the high-achievement group, the moderate-achievement group, and the low-achievement group. The item difficulty index by latent class and country is shown in Figure 3, while Figure 4 illustrates the composition of each latent class by country (Korea and Türkiye).

Table 3

LCA Model Fit

Class	Par	LL	BIC	aBIC	CAIC	AWE	BLRT
1-Class	14	-10333	20764	20719	20778	20904	-
2-Class	29	-8364	16932	16840	16961	17223	<0.001
3-Class	44	-8009	16327	16188	16371	16769	<0.001
4-Class	59	-7935	16285	16097	16344	16877	<0.001
5-Class	74	-7878	16277	16042	16351	17020	<0.001

Note. "Par"=parameters, "LL"=log likelihood, "BIC"=bayesian information criterion, "aBIC"=sample size adjusted BIC, "CAIC"=consistent Akaike information criterion, "AWE"=approximate weight of evidence criterion, "BLRT"=bootstrapped likelihood ratio test p-value

Figure 1

Scree Plot

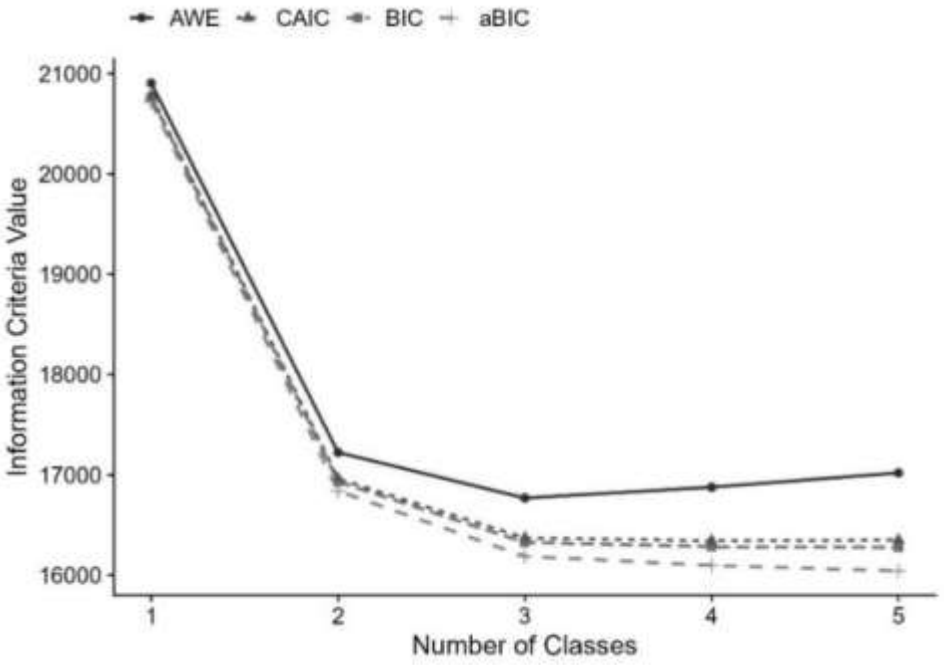


Figure 2

Latent class plots for Math Achievement

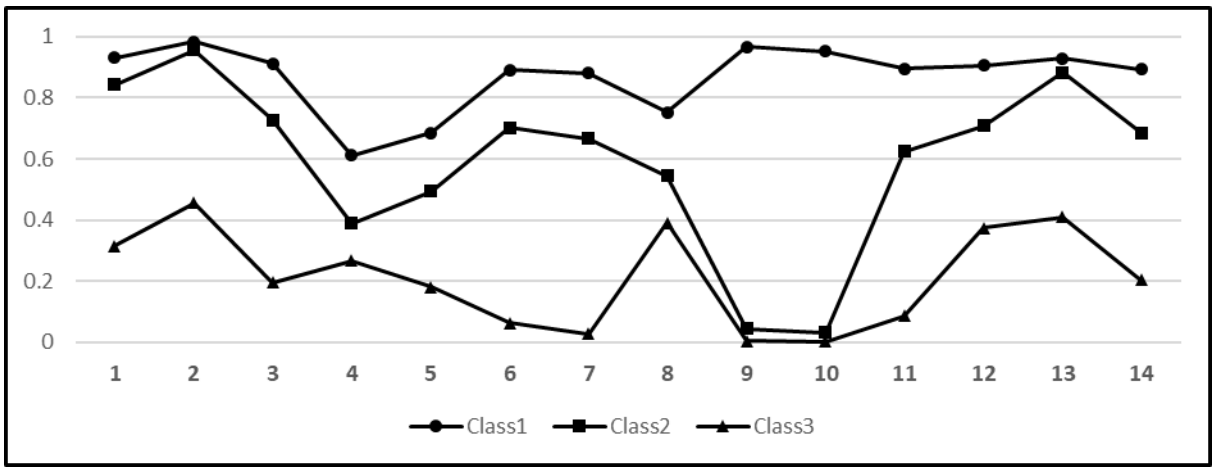


Figure 3

Item difficulty index within latent classes

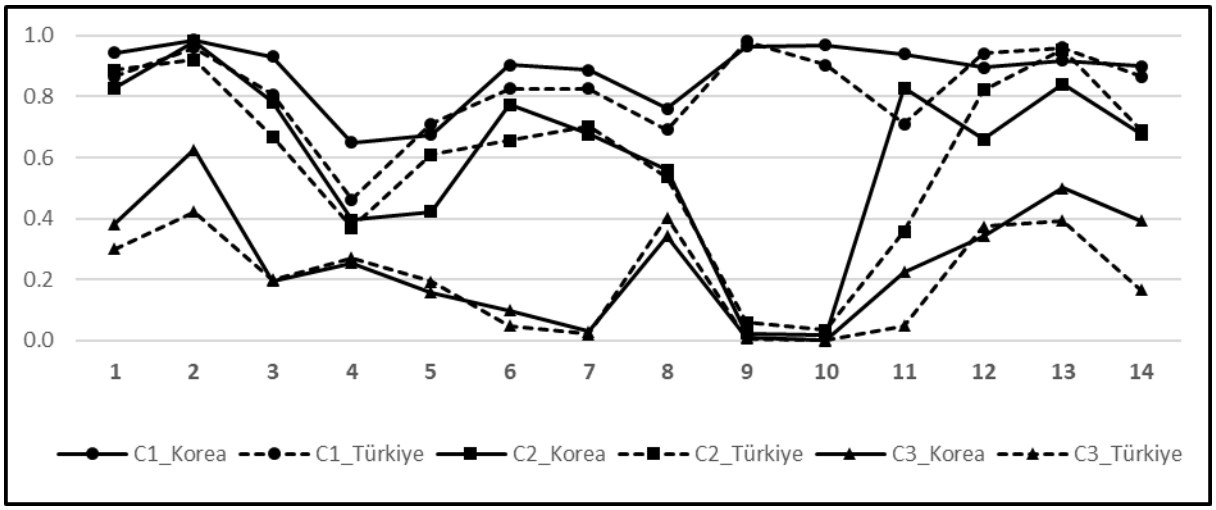
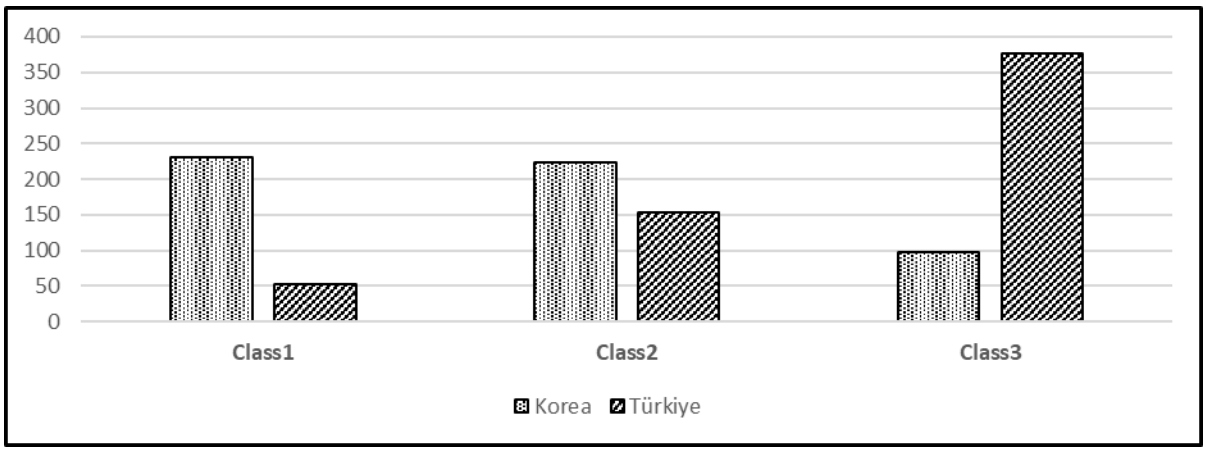


Figure 4

Composition of Korea & Türkiye within latent classes



Step 1: The latent class model selected in Step 0 was augmented with the covariate, the country variable, to compare the baseline model M_1.0 (No_DIF), where the country variable influenced the latent class variable but had no direct effects on the indicators, with the alternative model M_1.1 (All_DIF), where the country variable had direct effects on both the latent class variable and all indicators. As shown in Table 4, the fit of the M_1.1 model was significantly better than that of the M_1.0 model. This indicates that the country variable (Korea/Türkiye) is the source of DIF for at least one of the three latent classes and at least one of the fourteen items.

Step 2: DIF omnibus tests were conducted for each of the fourteen indicator variables by comparing models M_2.0.X, where the country variable (Korea/Türkiye) was set to influence the latent class variable but without direct effects on the indicator variables, and M_2.1.X, where the country variable was set to have direct effects on the indicator variables. As shown in Table 4, it was observed that for items 1, 4, 6, 7, 8, 9, and 10, the fit of the model without direct effects of the covariate on the indicators was not significantly worse than the model with direct effects. Additionally, for items 2, 3, 5, 11, 12, 13, and 14, the fit of the model with direct effects of the covariate on the indicators was significantly better than that without. This indicates that individually, seven items (2, 3, 5, 11, 12, 13, 14) out of the fourteen in the mathematics achievement test exhibit DIF.

Table 4

Model Comparisons for Stepwise DIF Test

Step	Model	Description	LL	npar	Comparison	LRTS	df	p
1	M_1.0	MIMIC: NO DIF	-7844.744	46	M_1.0 vs M_1.1	256.41	42	<0.001
	M_1.1	MIMIC: ALL DIF	-7716.539	88				
2	M_2.0.1	#1: No DIF	-1599.816	7	M_2.0.1 vs M_2.1.1	7.752	3	0.051
	M_2.1.1	#1: Non U DIF	-1595.940	10				
	M_2.0.2	#2: No DIF	-1480.067	7	M_2.0.2 vs M_2.1.2	10.871	3	0.012
	M_2.1.2	#2: Non U DIF	-1480.631	10				
	M_2.0.3	#3: No DIF	-1604.997	7	M_2.0.3 vs M_2.1.3	7.820	3	0.049
	M_2.1.3	#3: Non U DIF	-1601.087	10				
	M_2.0.4	#4: No DIF	-1784.847	7	M_2.0.4 vs M_2.1.4	6.978	3	0.072
	M_2.1.4	#4: Non U DIF	-1781.358	10				
	M_2.0.5	#5: No DIF	-1734.017	7	M_2.0.5 vs M_2.1.5	20.932	3	<0.001
	M_2.1.5	#5: Non U DIF	-1723.551	10				
	M_2.0.6	#6: No DIF	-1496.159	7	M_2.0.6 vs M_2.1.6	3.640	3	0.303
	M_2.1.6	#6: Non U DIF	-1494.339	10				
	M_2.0.7	#7: No DIF	-1464.368	7	M_2.0.7 vs M_2.1.7	4.588	3	0.205
	M_2.1.7	#7: Non U DIF	-1462.074	10				
	M_2.0.8	#8: No DIF	-1806.546	7	M_2.0.8 vs M_2.1.8	1.910	3	0.591
	M_2.1.8	#8: Non U DIF	-1805.591	10				
	M_2.0.9	#9: No DIF	-1184.171	7	M_2.0.9 vs M_2.1.9	5.773	3	0.123
	M_2.1.9	#9: Non U DIF	-1181.284	10				
	M_2.0.10	#10: No DIF	-1170.930	7	M_2.0.10 vs M_2.1.10	4.944	3	0.176
	M_2.1.10	#10: Non U DIF	-1168.466	10				
	M_2.0.11	#11: No DIF	-1517.144	7	M_2.0.11 vs M_2.1.11	91.378	3	<0.001

Table 4 (Continued)

Model Comparisons for Stepwise DIF Test

Step	Model	Description	LL	npar	Comparison	LRTS	df	p
	M_2.1.11	#11: Non U DIF	-1471.455	10				
	M_2.0.12	#12: No DIF	-1701.334	7	M_2.0.12 vs M_2.1.12	15.950	3	0.001
	M_2.1.12	#12: Non U DIF	-1693.359	10				
2	M_2.0.13	#13: No DIF	-1599.589	7	M_2.0.13 vs M_2.1.13	19.798	3	<0.001
	M_2.1.13	#13: Non U DIF	-1589.690	10				
	M_2.0.14	#14: No DIF	-1642.610	7	M_2.0.14 vs M_2.1.14	8.894	3	0.031
	M_2.1.14	#14: Non U DIF	-1638.163	10				
3	M_3.0	all Non U DIF Items			M_1.0 vs M_3.0 M_3.0 vs M_1.0			
	M_4.1	#2 (U DIF) All other (Non U DIF)	-7746.505	65	M_4.1 vs M_3.0	0.366	2	0.416
	M_4.2	#3 (U DIF) All other (Non U DIF)	-7748.457	65	M_4.2 vs M_3.0	4.270	2	0.059
	M_4.3	#5 (U DIF) All other (Non U DIF)	-7747.987	65	M_4.3 vs M_3.0	3.330	2	0.094
4	M_4.4	#11 (U DIF) All other (Non U DIF)	-7746.532	65	M_4.4 vs M_3.0	0.420	2	0.405
	M_4.5	#12 (U DIF) All other (Non U DIF)	-7747.913	65	M_4.5 vs M_3.0	3.182	2	0.102
	M_4.6	#13 (U DIF) All other (Non U DIF)	-7752.585	65	M_4.6 vs M_3.0	12.526	2	0.001
	M_4.7	#14 (U DIF) All other (Non U DIF)	-7750.864	65	M_4.7 vs M_3.0	9.084	2	0.005
5	M_5.0	#13, 14 (Non U DIF) #2, 3 ,5, 11, 12 (U DIF)	-7752.105	57	M_5.0 vs M_3.0	11.566	10	0.072
6	M_6.0	C on Country @ 0	-7787.161	55	M_6.0 vs M_6.1	270.124	2	<0.001
	M_6.1	C on Country (free)	-7752.105	57				

Step 3: To identify the optimal model, the fit of model M_3.0, where the seven identified items with DIF (2, 3, 5, 11, 12, 13, 14) were simultaneously set as non-uniform DIF, was compared with that of the baseline model M_1.0 (No_DIF) and model M_1.1 (All_DIF). The results revealed that the fit of model M_3.0 was significantly better than that of the baseline model M_1.0. Moreover, although the model where all items were set as DIF (M_1.1) exhibited better fit compared to M_3.0, the difference in fit between these two models was not substantial. Considering the improvement in fit from M_1.0 to M_3.0 and the parsimony of the model, model M_3.0 was chosen as the optimal latent class MIMIC model to proceed to the next step.

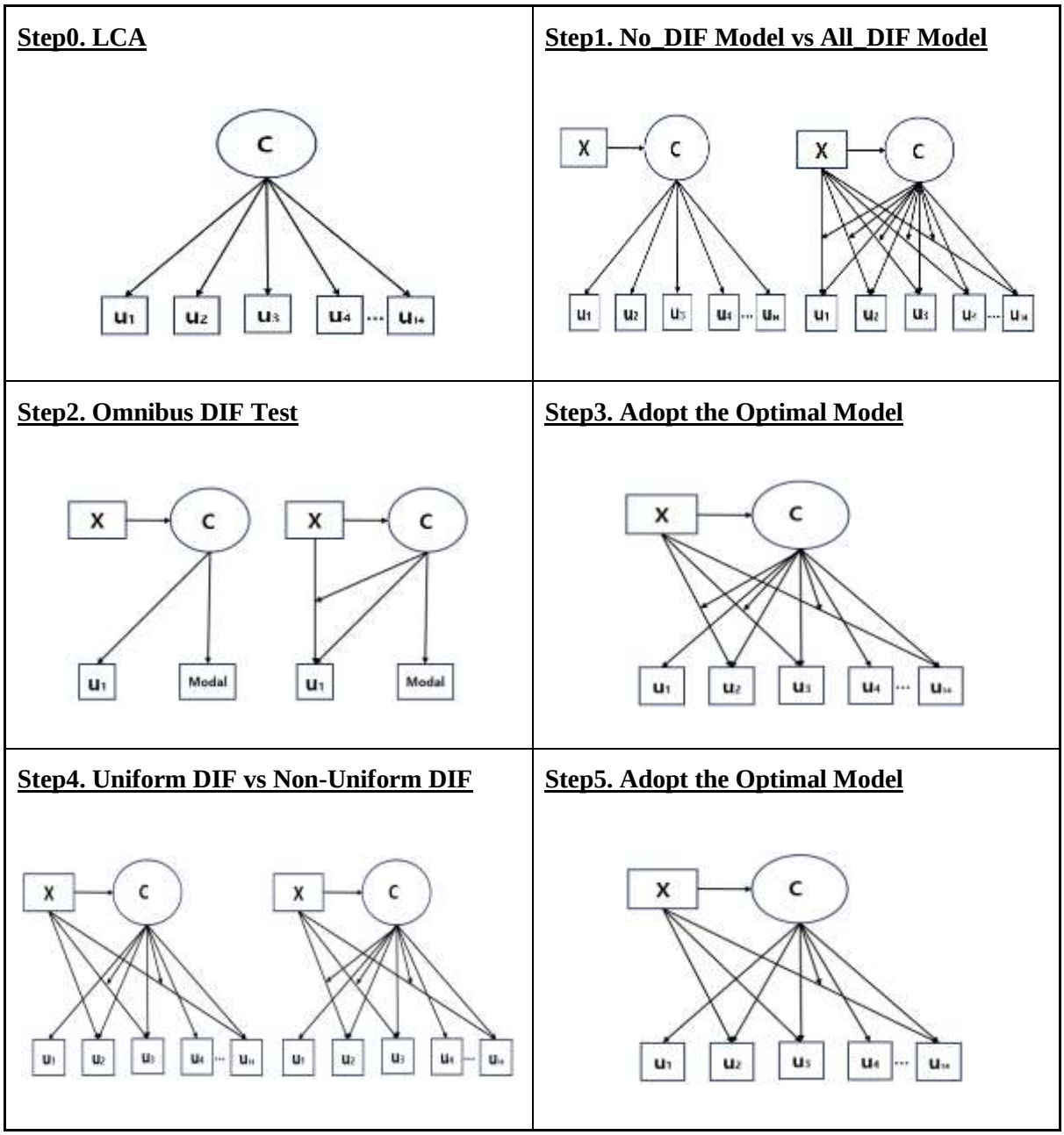
Step 4: To determine whether the seven identified DIF items were uniform DIF, the fit of model M_4.X, where the direct effects of the country variable were constrained to be uniform across classes for each of the seven items, was compared with that of model M_3.0, where all DIF items were treated as non-uniform DIF. As a result, items 2, 3, 5, 11, and 12 were confirmed to be uniform DIF, while items 13 and 14 were confirmed to be non-uniform DIF.

Step 5: The fit of model M_5.0, where the effects of the country on uniform DIF items were constrained to be uniform across classes, was compared with that of model M_3.0. As shown in Table 4, the fit of model M_5.0 was not significantly worse than that of M_3.0, indicating that the imposition of uniform DIF constraints did not significantly deteriorate the fit of the model. Therefore, model M_5.0 was adopted as the optimal model for the next step.

Step 6: Finally, model M_5.0 was re-designated as model M_6.1, and in model M_6.0, the polynomial logistic slope for the effect of the country variable on latent class membership was fixed to 0. In other words, while model M_6.1 allowed the country to freely estimate latent class membership, model M_6.0 did not allow the estimation of latent class membership by the country. Additionally, models M_6.0 and M_6.1 included all uniform and non-uniform DIF effects. The fit of models M_6.0 and M_6.1 was compared, and as shown in Table 4, the fit of model M_6.1, which allowed the country to freely estimate latent class membership, was significantly better than that of M_6.0, indicating the association of the country with latent class membership. Thus, the final adopted latent class MIMIC model, M_6.1, is illustrated in Figure 5, Step 5.

Figure 5

Latent Class MIMIC Modeling



Interpretation of the Final Model

An examination of the composition of Korean and Turkish nationals across the latent classes of mathematics achievement in the ultimately adopted M_6.1 model revealed that the high-achievement group comprised 81.6% Korean and 18.4% Turkish nationals, while the moderate-achievement group consisted of 59.3% Korean and 40.7% Turkish nationals. Furthermore, the low-achievement group comprised 21.2% Korean and 78.8% Turkish nationals.

The seven items of the mathematics achievement test identified as exhibiting DIF effects in the M_6.1 model, along with their respective effect sizes, are presented in Table 5 and Table 6. First, when examining the items identified as exhibiting uniform DIF, item 2 uniformly favored Korea across all classes, with a large DIF effect size. Item 3 similarly favored Korea uniformly across all classes, albeit with a negligible DIF effect size. Conversely, item 5 uniformly favored Türkiye across all classes, with a moderate DIF effect size, while item 11 favored Korea uniformly across all classes, with a large DIF effect size. Additionally, item 12 uniformly favored Türkiye across all classes, with a moderate DIF effect size.

Considering items identified as displaying non-uniform DIF, item 13 exhibited a significant in favored of Türkiye with a large effect size in the moderate-achievement group, while the DIF effects were not significant in the high- and low-achievement groups. On the other hand, item 14 favored Korea significantly with a large effect size in the low-achievement group, while the DIF effects were not significant in the high- and moderate-achievement groups.

Furthermore, Table 7 presents the results of logistic regression analysis on the influence of the country variable on the membership of latent classes of mathematics achievement. In Korea, there was a clear tendency for individuals to belong to either the high-achievement group or moderate-achievement group rather than the low-achievement group. Moreover, individuals in Korea were more likely to be part of the high-achievement group than the moderate-achievement group.

Table 5

Uniform DIF

Uniform DIF					
Item	**Est**	**SE**	**Est/SE**	**p**	**Effect size**
# 2	0.942	-0.220	4.274	<0.001	Large
# 3	0.427	0.176	2.422	0.015	Negligible
# 5	-0.526	0.172	-3.052	0.002	Moderate
# 11	1.980	-0.188	10.517	<0.001	Large
# 12	-0.455	0.177	-2.570	0.010	Moderate

Table 6

Non-Uniform DIF

Non-Uniform DIF									
	High group			**Moderate group**			**Low group**		
Item	**Est**	**p**	**Effect**	**Est**	**p**	**Effect**	**Est**	**p**	**Effect**
# 13	-1.049	0.310	Large	-1.123	0.011	Large	0.408	0.101	Negligible
# 14	0.435	0.406	Moderate	-0.010	0.967	Negligible	1.131	<0.001	Large

Table 7

Logistic regression analysis

	High group	Moderate group	Low group
Country	2.838(17.084)	1.680(5.363)	Ref
Est (odds ratio)	1.159(3.186)	Ref	

Conclusion and Discussion

Mathematics is a subject that significantly impacts students' academic success and future careers. Many countries participate in international academic achievement assessment to compare their performance with other nations and to explore the factors that influence academic achievement. Korea is the top performing country in the TIMSS 2019 math test, while Türkiye, although performing around the international average, has shown a steady increase in its performance since participating in TIMSS. Additionally, Türkiye has a more positive attitude towards mathematics compared to Korea. This study employed the LCA MIMIC method proposed by Masyn (2017) to explore the heterogeneous latent classes of mathematics achievement in the TIMSS 2019 assessment among 8th-grade students in Korea and Türkiye. Subsequently, the DIF of the mathematics assessment was examined according to country (Korea/Türkiye) to explore measurement invariance. The influence of the national variable (Korea/Türkiye) on membership in the latent classes of mathematics achievement was then investigated. The conclusions of this study are as follows:

First, a latent class analysis of mathematics achievement was conducted, identifying three distinct latent classes among the combined group of Korean and Turkish students: high-achievement, moderate-achievement, and low-achievement. The high-achievement group exhibited high item difficulty index of 0.7 or above for most items, with a higher proportion of Korean students in the group compared to Turkish students. The moderate-achievement group showed a wide range of item difficulty index varying from 0.03 to 0.95 across items, and exhibited a difficulty index below 0.1 in some geometry-related items, with a higher proportion of Korean students in the group compared to Turkish students. The low-achievement group demonstrated a consistently low item difficulty index of 0.4 or below for most items, with a higher proportion of Turkish students in the group compared to Korean students.

Secondly, in exploring DIF to verify the measurement invariance of mathematics achievement test items between Korea and Türkiye, a total of 7 out of 14 items were identified as exhibiting DIF. Among these, some items were identified as displaying uniform DIF, while others showed non-uniform DIF. This indicates the presence of direct effects of the country on individual items within the detected latent classes of mathematics achievement, and these direct effects were observed to vary in their application across latent classes, either uniformly or non-uniformly. Notably, while the overall item difficulty index for items indicated higher performance for Korea compared to Türkiye, this study's exploration of heterogeneous latent classes of mathematics achievement and subsequent examination of DIF based on country within these identified classes revealed areas of favorable or unfavorable performance in mathematics between Korea and Türkiye within homogeneous characteristics and ability groups. Furthermore, these results demonstrate that when analyzing the effects of covariates on latent classes, ensuring unbiased results requires conducting measurement invariance tests to confirm the direct effects of covariates on indicator variables.

Third, out of the seven items identified as DIFs, five items were identified as uniform DIFs and two items were identified as non-uniform DIFs. For items 2, 3, 5, 11, and 12, which exhibited uniform DIF, items 2, 3, and 11 favored Korean students in all classes, with large, negligible, and large DIF effect sizes, respectively. Additionally, items 5 and 12 favored Turkish students in all classes, with moderate DIF effect sizes for both items. Next, for items 13 and 14, identified as non-uniform DIF, item 13 favored Turkish students with a large effect size in the moderate-achievement group, while the DIF effect was

not significant in the high- and low-achievement groups. Conversely, item 14 favored Korean students in the low-achievement group with a large effect size, with no significant effect observed in the high- and moderate-achievement groups. Additionally, examining the pattern of uniform/non-uniform DIF based on content areas, it was found that items in the Number, Geometry, and Algebra domains exhibited uniform DIF, whereas items in the Data and Probability domains displayed non-uniform DIF effects. Thus, it was observed that DIF effects varied between uniform and non-uniform across different content areas in mathematics.

Fourth, excluding items with non-significant or negligible DIF effects, examining the mathematics content domain and cognitive domains of items 2, 5, 11, 12, 13, and 14, which exhibit moderate or higher DIF effect sizes, it is found that items 2 and 11 correspond to the Number/Knowing and Geometry/Reasoning domains, respectively, and favor Korea across all latent classes. Item 14 corresponds to the Data and Probability/Applying domain and favors Korea in the low-achievement group. Items 5 and 12 correspond to the Algebra/Applying and Data and Probability/Applying domains, respectively, and favor Türkiye across all latent classes. Additionally, item 13 corresponds to the Data and Probability/Knowing domain and favors Türkiye in the moderate-achievement group. Summarizing the favorable and unfavorable items by country, Korea has one favorable item each in the Number/Knowing, Geometry/Reasoning, and Data and Probability/Applying domains, while Türkiye has one favorable item each in the Algebra/Applying, Data and Probability/Applying, and Data and Probability/Knowing domains. These results differ somewhat from Şen and Arıcan (2015), who reported that Korean students outperformed Turkish students in most math content domains (Number, Algebra, Geometry, Data and Probability). The reason for this partial discrepancy with Şen and Arıcan's (2015) study is that this study classified all students in Korea and Türkiye into heterogeneous latent classes based on their math achievement. It identified areas of favorability or unfavorability in math tests for homogeneous ability groups in Korea and Türkiye by exploring DIF within homogeneous latent classes. In particular, the results of this study showed that in the Data and Probability domain, Korea had a favorable result on one item compared to Türkiye in the low-achievement group. However, Türkiye had a favorable result on one item in each of the latent classes and in the moderate-achievement group compared to Korea. These findings align with Yoon and Lee's (2013) study, which reported that Korean students exhibited unfavorable performance in the Data and Probability domain compared to American students. This was evidenced by the exploration of DIF in the TIMSS 2007 assessment. Thus, it can be inferred that within homogeneous achievement groups, Korean students' performance in the Data and Probability domain is somewhat lower compared to that of Turkish students.

Fifth, examining the distribution of students across math achievement latent classes in each country, 41.8% of Korean students are classified as the high-achievement group, 40.5% as the moderate-achievement group, and 17.7% as the low-achievement group. In contrast, 9.1% of Turkish students are in the high-achievement group, 26.3% are in the middle-achievement group, and 64.6% are in the low-achievement group. The multinomial logistic regression analysis examined the impact of the country (Korea/Türkiye) on latent class membership in math achievement. The results indicated that students from Korea were more likely to be part of the high- and moderate-achievement groups rather than the low-achievement group. Additionally, students in Korea were more likely to be in the high-achievement group than in the moderate-achievement group.

In this study, the relationship between country (Korea/Türkiye) and membership in latent classes of math achievement was examined. To ensure the validity and robustness of the results, measurement invariance tests, including the detection of differential item functioning (DIF), were conducted. These tests were crucial for providing unbiased results in the identification of latent classes and assessing the impact of covariates in the LCA. Through the examination of measurement invariance for indicator variables and DIF in LCA, it was possible to identify areas of favorability or unfavorability across countries for individual items in mathematics achievement tests within homogeneous ability groups. Particularly noteworthy is the utilization of the MIMIC model in LCA for exploring DIF, which differs from previous studies (Kalaycioğlu & Berberoğlu, 2011; Lyons-Thomas et al., 2014; Yildirim, 2006) that applied classical test theory, item response theory, and logistic regression analysis in the exploration of DIF. Subsequent research can identify the causes of favorable or unfavorable areas in math

achievement tests by country through a content-based approach to mathematics education. This can provide insights for enhancing the curriculum and educational methods within each country's mathematics education system.

Declarations

Author Contribution: Hyo Seob SONG: 1st Author, conceptualization, methodology, data analysis, writing & editing. Hee Sun JUNG: Corresponding Author, investigation, data analysis, visualization, supervision, writing - review & editing.

Ethical Approval: All ethical guidelines for authors have been followed. Ethical approval is not required for this study as it utilizes publicly available data.

Conflict of Interest: The authors declare no potential conflicts of interest.

Funding: No funds, grants, or other support were received during the preparation of this manuscript.

References

Arıkan, S., Van de Vijver, F., & Yagmur, K. (2016). Factors contributing to mathematics achievement differences of Turkish and Australian students in TIMSS 2007 and 2011. *EURASIA Journal of Mathematics, Science and Technology Education*, *12*(8), 2039-2059. https://doi.org/10.12973/eurasia.2016.1268a

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using M plus. *Structural equation modeling: A multidisciplinary Journal*, *21*(3), 329-341. https://doi.org/10.1080/10705511.2014.915181

Badri, M. (2019). School Emphasis on Academic Success and TIMSS Science/Math Achievements. *International Journal of Research in Education and Science*, *5*(1), 176-189. https://eric.ed.gov/?id=EJ1197994

Clark, S. L., & Muthén, B. (2009). Relating latent class analysis results to variables not included in the analysis. https://www.statmodel.com/download/relatinglca.pdf

Demirus, K. B., & Pektas, S. (2022). Investigation of Timss 2015 science test items in terms of differential item functioning according to language and culture. *International Journal of Education Technology & Scientific Researches*, *7*(18), 1166-1178. http://dx.doi.org/10.35826/ijetsar.499

De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, *2*(3-4), 243-276. https://doi.org/10.1080/15305058.2002.9669495

Dittrich, K., & Neuhaus, D. A. (2023). Korea's 'education fever' from the late nineteenth to the early twenty-first century. *History of Education*, *52*(4), 539-552. https://doi.org/10.1080/0046760X.2022.2098391

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization 1, 2. *ETS Research Report Series*, *1992*(1), i-40. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Geesa, R. L., Izci, B., Song, H., & Chen, S. (2019). Exploring Factors of Home Resources and Attitudes Towards Mathematics in Mathematics Achievement in South Korea, Turkey, and the United States. *Eurasia Journal of Mathematics, Science and Technology Education*, *15*(9), eM_1751. https://doi.org/10.29333/ejmste/108487

Geesa, R. L., Izci, B., Chen, S., & Song, H. S. (2020). The Role of Gender and Attitudes toward Science in Fourth and Eighth Graders' Science Achievement in South Korea, Turkey, and the United States. *Journal of Research in Education*, *29*(2), 54-87. https://eric.ed.gov/?id=EJ1274027

Guhl, P. (2019). The impact of early math and numeracy skills on academic achievement in elementary school.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European journal of psychological assessment*, *17*(3), 164-172. https://doi.org/10.1027/1015-5759.17.3.164

Huang, J. (2020). *Assessing robustness of the Rasch mixture model to detect differential item functioning: A Monte Carlo simulation study* (Doctoral dissertation, University of Denver). https://digitalcommons.du.edu/etd/1784/

Hwang, H. J., & Hwang, H. K. (2008). An Effect of the Constructivist Discussion on Learning Attitude in Mathematics and Children's Mathematics Achievement. *Education of Primary School Mathematics*, *11*(1), 59-74. https://koreascience.kr/article/JAKO200801440607130.page

Im, S., & Park, H. J. (2010). A comparison of US and Korean students' mathematics skills using a cognitive diagnostic testing method: linkage to instruction. *Educational Research and Evaluation*, *16*(3), 287-301. https://doi.org/10.1080/13803611.2010.523294

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity, *Marketing Science 16*(1), 39-59. https://doi.org/10.1287/mksc.16.1.39

Kalaycioğlu, D. B., & Berberoğlu, G. (2011). Differential item functioning analysis of the science and mathematics items in the university entrance examinations in Turkey. *Journal of Psychoeducational Assessment*, *29*(5), 467-478. https://doi.org/10.1177/0734282910391623

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European journal of psychology of education*, *16*, 385-402. https://link.springer.com/article/10.1007/BF03173189

Ko, D. H., & Jung, H. S. (2020). Analysis on the Effect of Mathematics Class Characteristics and Mathematical Confidence on Mathematical Academic Achievement: Applying Hierarchical Linear Model. *School Mathematics*, *22*(2), 313-332. https://doi.org/10.29275/sm.2020.06.22.2.313

Lee, J., & Stankov, L. (2018). Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA. *Learning and Individual Differences*, *65*, 50-64. https://doi.org/10.1016/j.lindif.2018.05.009

Lubinski, D., Benbow, C. P., & Kell, H. J. (2014). Life paths and accomplishments of mathematically precocious males and females four decades later. *Psychological Science*, *25*(12), 2217-2232. https://doi.org/10.1177/0956797614551371

Lyons-Thomas, J., Sandilands, D., & Ercikan, K. (2014). Gender Differential Item Functioning in Mathematics in Four International Jurisdictions. *Education & Science/Egitim ve Bilim*, *39*, 20-32.

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 180-197. https://doi.org/10.1080/10705511.2016.1254049

Mullis, I. V., & Martin, M. O. (2017). *TIMSS 2019 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement. https://eric.ed.gov/?id=ed596167

Mullis, I. V., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. https://www.skolporten.se/app/uploads/2020/12/timss-2019-highlights-1.pdf

Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, *4*(4), 440-461. https://doi.org/10.1037/tps0000176

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 782-797. https://doi.org/10.1080/10705511.2016.1221313

Ram, N., & Grimm, K. J. (2009). Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International journal of behavioral development*, *33*(6), 565-576. https://doi.org/10.1177/0165025409343765

Saatcioglu, F. M. (2022). Differential item functioning across gender with MIMIC modeling: PISA 2018 financial literacy items. *International Journal of Assessment Tools in Education*, *9*(3), 631-653. https://doi.org/10.21449/ijate.1076464

Samuelsen, K. M. (2008). Examining differential item functioning from a latent mixture perspective. *Advances in latent variable mixture models*, 177-197.

Şen, S., & Arıkan, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2). https://doi.org/10.21031/epod.65266

Shin, K., Jahng, K. E., & Kim, D. (2019). Stories of South Korean mothers' education fever for their children's education. *Asia Pacific Journal of Education*, *39*(3), 338-356. https://doi.org/10.1080/02188791.2019.1607720

Sohn, W. S. (2010). Exploring Potential Sources of DIF for PISA 2006 Mathematics Literacy Items: Application of Logistic Regression Analysis. *Journal of Educational Evaluation*, *23*(2), 371-390. https://scholar-kyobobook-co-kr-ssl.ca.skku.edu/article/detail/4010023071385

Song, H. S., Kim, H. C., & Jung, H. S. (2023). The Effect of Participation in Private Education on Math Affective Attitudes: Measurement Invariance and Differential Item Functioning in Latent Class MIMIC Model. *Journal of Educational Evaluation*, *36*(4), 687-709. http://dx.doi.org/10.31158/JEEV.2023.36.4.687

Tsaousis, I., Sideridis, G. D., & AlGhamdi, H. M. (2020). Measurement invariance and differential item functioning across gender within a latent class analysis framework: Evidence from a high-stakes test for university admission in Saudi Arabia. *Frontiers in Psychology*, *11*, 622. https://doi.org/10.3389/fpsyg.2020.00622

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis*, *18*(4), 450-469. https://doi.org/10.1093/pan/mpq025

Wang, X. S., Perry, L. B., Malpique, A., & Ide, T. (2023). Factors predicting mathematics achievement in PISA: a systematic review. *Large-scale Assessments in Education*, *11*(1), 24. https://doi.org/10.1186/s40536-023-00174-8

Wiberg, M. (2019). The relationship between TIMSS mathematics achievements, grades, and national test scores. *Education Inquiry*, *10*(4), 328-343. https://doi.org/10.1080/20004508.2019.1579626

Woo, H., & Hodges, N. N. (2015). Education fever: Exploring private education consumption motivations among Korean parents of preschool children. *Family and Consumer Sciences Research Journal*, *44*(2), 127-142. https://doi.org/10.1111/fcsr.12131

Yildirim, H. H. (2006). The DIF analysis of mathematics items in the international assessment programs. https://www.proquest.com/openview/37bc178a79e82e2c13ca84983403a39a/1?cbl=2026366&diss=y&pq-origsite=gscholar

Yoon, J. Y., & Lee, Y. S. (2013). A Study of DIF Analyses using TIMSS (2007) Mathematics Test Across South Korea, U. S., and Singapore. *Journal of Educational Evaluation*, *26*(2), 415-439. https://scholar-kyobobook-co-kr-ssl.ca.skku.edu/article/detail/4010023595562

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2025; 16(1); 30-47

Performance of Classification Techniques on Smaller Group Prediction

Cahit POLAT * Kathy GREEN **

Abstract

Classification techniques allow researchers to analyze data based on groups for the purposes of clustering or making predictions about group membership. Since there are many methods for utilizing classification analyses, such as Linear Discriminant Analysis (LDA), Logistic Regression (LR), and Classification and Regression Trees (CART), it is important to know which techniques perform better under which conditions to affect prediction accuracy. In the context of group prediction, it is crucial to consider the impact of group proportional sizes on prediction accuracy, particularly when comparing smaller groups to larger ones. This study evaluated the small group prediction accuracies of LDA, LR, and CART, controlling for the number of groups, correlation, and the number of predictor variables. Results showed that CART performed best for smaller and overall group prediction in most cases. In addition, a notable difference was observed in overall group prediction accuracy compared to small group prediction accuracy, with the overall group prediction accuracy being greater. Data conditions had a greater impact on LR and LDA than CART, and, in certain instances, LR showed superiority over the other two methods. The number of groups was the most influential factor on small group prediction, while the number of predictor variables, correlation, and method were of decreasing influence. In general, overall group prediction accuracy and small group prediction accuracy were negatively related. However, for the categories with an equal number of groups, the two were positively related.

Keywords: method performance evaluation, group membership, classification accuracy, simulation.

Introduction

Classifying cases into groups is widespread in all fields, and statistical or analytical techniques may perform differently depending on the data conditions. The data structure influences the choice of methods of analysis and sets constraints on the study's scope. Classification serves the purpose of identifying group characteristics and predicting group membership and is a valuable statistical approach in various fields such as social sciences, education, health sciences, and other domains. It is further crucial for researchers to assess the significance of predictors in determining the group or class to which observations belong.

Explanatory models are applied to examine relationships between variables, whereas predictive models are utilized to make predictions about categories using a correlational design. Group discrimination and decisions are assessed using these models (Sainani, 2014). Utilizing predictive models, for instance, one may determine the likelihood of contracting an illness based on the findings of diagnostic tests or the mortality rate of a veteran suffering a stroke within a year at a certain severity level (Bates et al., 2014). By applying such models, it is possible to determine, for example, whether certain predictor variables like the student's positive opinion of their teacher, GPA, whether they lived with their biological parents, and the number of days the student missed from school also predict the dropout status of high school students (Suh et al., 2007).

* Asst. Prof. Dr., Harran University, Faculty of Education, Şanlıurfa-Türkiye, cahitpolat@harran.edu.tr, ORCID ID: 0000-0002-1423-5084
** Professor, University of Denver, Faculty of Education, Colorado, US, Kathy.Green@du.edu, ORCID ID: 0000-0002-1676-3139

To cite this article:
Polat, C. & Green, K. (2025). Performance of Classification Techniques on Smaller Group Prediction*, Journal of Measurement and Evaluation in Education and Psychology, 16*(1), 30-47. https://doi.org/10.21031/epod.1598907
Received: 9.12.2024
Accepted: 19.03.2025

There are various techniques for determining group membership, and logistic regression (LR), linear discriminant analysis (LDA), and classification and regression trees (CART), a more recent technique, are widely used ones (Agresti, 2002; Huberty & Olejnik, 2006; Williams et al., 1999). LDA and LR have historically been used extensively in educational and social science research, but CART is a newer technique (Holden et al., 2011). Additionally, in many recent studies, these techniques are applied simultaneously (Castonguay et al., 2022; Hassan et al., 2024; Hoang et al., 2025; Saboor et al., 2022; Selim et al., 2020; Song et al., 2022; Zampogna et al., 2024).

Though they are widely used, limited information exists regarding the effectiveness of these three techniques in predicting categories of observations, especially for relatively smaller groups, and which perform better in certain data scenarios, such as group size ratios, degree of correlation, number of predictor variables, and number of groups in the outcome variable. Therefore, this study aimed to investigate the performance of LDA, LR, and CART for overall and smaller group prediction in addition to whether prediction accuracies are affected by the correlation between predictor variable strength, number of predictor variables, group size ratios, and number of groups in the dependent variable. Finally, this study explored the relationship between overall group prediction accuracy and small group accuracy. We provide a brief overview of each technique below.

Linear Discriminant Analysis (LDA)

LDA procedure calculates the observation score for j[th] group (G_j) as;

$$G_j = c_{j0} + \sum c_{ji} x_i + \ln\left(\frac{n_j}{N}\right) \quad (1)$$

where c_{j0} represents the constant value for the jth group, c_{ji} denotes the coefficient value of the ith variable within the jth group, x_i is the ith variable, n_j indicates the total number of observations in the jth group, and N represents the total number of all observations.

Moreover, the constant value for the j[th] group c_{j0} and the coefficient values c_{ji}s are calculated by the formula;

$$c_{j0} = \frac{1}{2} C_j' M_j \quad (2)$$

where $C_j = W^{-1} M_j$, C_j is the coefficients vector for c_{ji}s, W is the pooled within-group variance-covariance matrix, and M_j is matrix of the means of the variables for group j.

Upon computing the observation scores for each group, the observation is allocated to the group with the highest score. LDA models are exclusively linear functions and assume the absence of multicollinearity and singularity, as well as homogeneity of variance-covariance matrices and multivariate normality (Tabachnick & Fidell, 2013).

Logistic Regression (LR)

LR starts with calculating linear regression model *u* as;

$$u = B_0 + \sum B_j X_{ij} \quad (3)$$

where B_0 represents the linear regression model's intercept., and B_j indicates the j[th] variable's coefficient, X_j.

Then $\widehat{Y_i} = \frac{e^u}{1+e^u}$ is calculated as the probability that the i[th] observation is a member of a group rather than a reference group. It can be seen easily that the natural log of the probability of the odds ratio being in one group versus another reference group is equal to *u* such as;

$$\ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = B_0 + \sum B_j X_{ij}. \quad (4)$$

In many statistical applications, the default threshold for determining observation membership is set at 0.5; hence, if the logit equals or exceeds 0.5, the observation is classified within the group. The cut point may also be established at another value (Soureshani et al., 2013). Logistic Regression (LR) distinguishes itself from many techniques by its flexibility, since it does not relay on certain assumptions such as normality.

Classification and Regression Trees (CART)

CART divides data iteratively to classify objects into more homogeneous groups, which are referred to as nodes. The CART algorithm initiates by locating all subjects in a single node. Subsequently, it assigns them to other nodes by utilizing predictor variables to establish the most homogeneous groups (Breiman et al., 1984). This procedure continues until an ideal group split achieves the desired degree of group membership homogeneity. To mathematically apply this, the node deviances are minimized, and the deviance for ith node (D_i) is computed as;

$$D_i = -2\sum\sum n_{ik} ln(p_{ik}) \quad (5)$$

where n_{ik} denotes the number of subjects from group k in node i, and p_{ik} indicates the proportion of subjects from group k within node i.

The sum $D = \sum D_i$, is used as a measure of homogeneity once the deviances of each group have been calculated; smaller *D*s signify higher homogeneity. The procedure continues until either the requirement for stopping iterations is met or the reduction in *D*s from one step to the next becomes trivial.

Related Research

This section includes a summary of the literature review of related studies. In a comparison of the overall performance of LDA and LR, one study found that LR had a higher prediction accuracy for group membership (Barön, 1991), while others found little or no difference between the two methods (Dey & Astin, 1993; Hess et al., 2001; Meshbane & Morris, 1996). Further, the statistical methods LDA and LR exhibited comparable performance to CART (Dudoit et al., 2002; Ripley, 1994). However, other studies have demonstrated that LDA and LR outperform CART (Preatoni et al., 2005; Williams, 1999) or that CART outperforms LR and LDA (Holden, 2011; Hao et al., 2022). Lastly, while some results indicated that CART performed better than LDA in terms of group membership prediction accuracy (Grassi et al., 2001), others indicated that LR and CART performed similarly (Schumacher et al., 1996). These conflicting findings may be due to different configurations of the data analyzed. In this regard, the overall performance of any method is uncertain in the absence of an assessment of the data's specific characteristics.

While certain studies compared the accuracies of the methods, the comparison results were not generalizable beyond the scope of the research. Hence, some researchers utilized simulated data to compare the performance of techniques rather than utilizing real data from content areas. A substantial advantage of simulated data is the researcher's capacity to manage the data conditions. As a result, numerous studies have compared the performances of methods under controlled conditions. Numerous data factors may have an impact on how well classification techniques perform. For classification accuracy the following conditions have been shown to have an effect: sample size (Bolin & Finch, 2014), group size ratios (Finch & Schneider, 2006; Lei & Koehly, 2003), effect size (Holden et al., 2011), predictor distributions (Pai et al., 2012; Pohar et al., 2004), and homogeneity of variance-covariance matrices (Fan & Wang, 1999; Lei & Koehly, 2003). On the other hand, less researched but important for comparing the methods are correlations between predictor variables (Kiang, 2003), number of variables (Holden & Kelley, 2010), number of groups in the dependent variable (Zavroka & Perret, 2014), model complexity (Holden et al., 2011), dynamic structure of the data, linearity, presence of outliers (Pai et al., 2012), multimodal structure of the data (Kiang, 2003), percent of initial misclassification (Bolin & Finch, 2014), and group separation (Finch et al., 2014).

CART outperforms LDA and LR in various scenarios involving sample size, homogeneity of variance-covariance matrices and effect size, group size ratio, varying model complexities, percentage of initial misclassification, and group separation level (Bolin & Finch, 2014; Finch et al., 2014; Holden et al., 2011); however, it performs less effectively in scenarios involving normal or skewed data (Finch & Schneider, 2006). When the normality and homogeneity of variance-covariance matrices are violated, LR is predicted to outperform LDA (Dattalo, 1994; Ferrer & Wang, 1999; Huberty, 1999). Meanwhile, despite the broad acceptance of the normality assumption for LDA, it may still be resistant to nonnormality (Graf et al., 2023). Under most circumstances, LR and LDA exhibited generally comparable performance, despite some conflicting results (Dey & Austin, 1993; Hess et al., 2011). Kiang (2003) found that when multimodal data and nonlinearity are present, LR performs better than LDA. The dynamic nature of the data and the presence of outliers impact the classification techniques' success (Pai et al., 2012).

The number of groups in the dependent variable (Pohar et al., 2004) and the number of predictor variables (Huberty, 1994; Rausch & Kelley, 2009) had an impact on classification technique performance. The change in the performance of the techniques LDA, LR, and CART were similar when additional groups were included, and the methods' classification accuracies rose as there were more predictor variables. LDA was shown to perform less well under multicollinearity, whereas LR was unaffected by multicollinearity (Pai et al., 2012). Finally, the group size ratio plays an important role in the performance of methods for small and overall group prediction. When proportions are highly unbalanced, small group prediction accuracy tends to be lower while overall group prediction accuracy tends to be larger (Finch & Schneider, 2006). However, the number of studies testing LDA, LR, and CART simultaneously for the effect of data conditions on small prediction accuracy is limited.

Importance of the Study

Although prior research has provided some insight into the parameters influencing the performance of LDA, LR, and CART, further research is necessary to gain a deeper comprehension of the group classification techniques' respective performances. In particular, the number of predictor variables, the number of groups in the dependent variables, and the correlations between predictor variables have not been fully examined. To get more thorough findings, group size ratio should be taken into consideration while evaluating these circumstances. Additionally, classification accuracies of smaller groups should be considered in addition to overall classification accuracy. In cases where data are unbalanced, the prediction of the smallest group may be important. In consideration of this, this study concentrated on the precision of the small group prediction in situations where the sample sizes of the groups were unbalanced. Besides, this study aimed to investigate which of the three methods performs better in terms of smallest group prediction accuracy given varying degrees of correlation between predictor variables, number of groups in the dependent variables, and number of predictor variables. The purpose was to determine whether the number of groups, the level of correlation between predictor variables, the number of predictor variables, and the group size ratios in the dependent variables interact significantly in relation to the classification accuracy of the overall and the smallest group of the three methods. Finally, this study also aimed to investigate the relationship between the accuracy of prediction for small groups and whole groups. Consequently, the research questions for this study are as follows:

1. How do the number of predictor variables, the number of groups, and the correlation between predictor variables affect prediction accuracy for smaller groups?

2. What is the relationship between overall group prediction accuracy and small group prediction accuracy in different data scenarios?

Method

Research Design

Factors associated with data characteristics were controlled in this study. The variables were group size ratio (2 levels: balanced, imbalanced), number of groups (3 levels: 2, 3, 4), correlation between predictor variables (2 levels:.2,.5), and number of predictor variables (3 levels: 2, 5, 10). While the last two conditions are related to the dependent variable, the first two conditions are related to predictor variables. In addition, three distinct analysis techniques (LDA, LR, and CART) were applied. As a result, using each of the three methods, 2x3x2x3 = 36 distinct data conditions were generated and examined. It was considered that all other variables are uncontrollable and random. A fixed sample size of 200 was used, and 1000 simulations were run for each condition. Consequently, the study contained 36x200 = 7,200 simulated observations, each with 1,000 repetitions for each method. For the smallest group prediction, only unbalanced data in terms of group size ratio was applied, while for the overall group prediction, both balanced and unbalanced data were applied. All the predictor variables were simulated as normally distributed with a mean of 0.0 and a standard deviation of 1.0, i.e., a standard normal distribution.

Steps of Data Generation

A Monte Carlo simulation procedure was utilized to produce a dataset with the specified conditions. Monte Carlo techniques apply random sampling to simulate data as it permits the generation of random variables and the management of controlled variables. These techniques involve generating datasets that meet specific criteria using mathematical approximations and probability computations (Paxton et al., 2001).

The function MVRNORM in R software (R Core Team, 2016) was utilized to create data with specific characteristics, ensuring that the predictor variables followed a multivariate normal distribution. Researchers can use the MVNORM package in R to define the correlations among predictor variables and the number of predictor variables. The sample size was set at 200, which is commonly used in simulation studies and a suitable number of observations in quantitative research in the social and educational sciences. Additionally, for LDA, prior probabilities were determined based on the observed group ratios of the respective sample sizes to the total sample size, following the suggestion of Lei and Koehly (2003).

The MVRNORM function generates multivariate normal distribution variables for each group. For example, generating all five predictor variables by MVRNORM yields multivariate normal distributions for each group, but that does not guarantee normality when combining each group for the dependent variable. This function also lets one define predictor variable means and standard deviations for each dependent variable group. However, multivariate normality is not guaranteed for each iteration when creating predictor variables from a multivariate normal distribution for each group and merging them for total datasets.

The groups were designated as group 1, group 2, group 3, and group 4. Groups with lower numerical labels include fewer observations. In unbalanced scenarios, group 1 consistently has the smallest group size. The simulation of a 1000-iteration dataset under appropriate modified and random settings was completed by following the steps outlined below and using the necessary R tools. If data non-convergence occurred during one replication, an additional replication was performed using the R software to compensate, resulting in the completion of 1000 replications. Following the completion of data training, the data were prepared for analysis.

Controlled Variables and Their Patterns

Two degrees of correlation (CORR) were established: 0.2 (indicating low) and 0.5 (indicating medium). Specific values for low correlation (0.2) and medium correlation (0.5) among all predictor variables were entered using the MVRNORM function in R. Adjustments were made to all five predictor variables to achieve a correlation of 0.2 if the correlation coefficient was 0.2. In the same way, in the case where

the correlation coefficient was 0.5, the five variables were adjusted to exhibit a correlation of 0.5. However, when predictor variables were simulated, the average correlation was greater in magnitude when compared to the fixed level. Depending on the data context, when correlations were set to 0.5 in MVRNORM, the simulated data correlations were, for instance, 0.58 or a slightly different value. This was due to arrangements regarding group size ratios and effect sizes. To attain the predetermined correlation conditions, lover-level correlations were introduced to the program and the correlation coefficients were progressively decreased in R throughout the data simulation process until the desired average coefficient values of 0.2 and 0.5 were reached for each of the 36 data scenarios.

The levels of the number of predictor variables (NPV) used in the study were based on generated data with two, five, and ten predictor variables. These levels were set automatically by creating correlation matrices. This study splits the number of groups (GN) in the dependent variable into three levels: two, three, and four, which are the most widely used. To create groups, group size ratios were utilized to count and calculate the number of observations for each group. For example, for three groups with a 10:20:70 group size ratio, 20, 40, and 140 observations were simulated for each group because the total sample size was 200 summed across the group size. Different numbers were assigned to categories. For instance, with three groups in the dependent variable, group 1 had 20 cases, group 2 had 40 cases, and group 3 had 140. After simulating and labeling dependent variable groups (from smaller to larger sizes: group 1, group 2, group 3, and group 4) and predictor variable datasets for each iteration, the outcome variable and predictor variables were randomly matched.

Two different levels of group size ratio (GSR) were controlled in this study: balanced group size ratios and unbalanced group size ratios. A balanced group size ratio exists when the dependent variable's groups have the same number of observations. On the other side, an unbalanced group size ratio exists when the number of instances in the groups is unequal and there is a significant discrepancy in the number of observations between the largest and smallest groups. The group size ratios for balanced groups were set to 50:50, 33:33:33, and 25:25:25:25, respectively, when there were two, three, and four groups. As a result, each group had the same number of instances, with 100 cases per group when there were two groups, 67 cases (1 case omitted from the middle group to set the sample size to 200) when there were three groups, and 50 cases per group when there were four groups. Unbalanced group ratios, on the other hand, were set at 10:90, 10:20:70, and 10:15:20:55 for groups of two, three, and four, respectively. Thus, group sizes were 20 and 180 for the case of two groups, 20, 40, and 140 for the case of three groups, and 20, 30, 40, and 110 for the case of four groups.

Simulating Groups of Dependent Variables

To simulate values for groups for dependent variables, the software was programmed to include the means of predictor variables for each group. The effect size, defined as the standardized difference between consecutive groups, was set at 0.5 using the classification of 0.2, 0.5, and 0.8 as small, medium, and large effect sizes, respectively (Cohen, 1988). The overall group mean was set to zero; to meet this criterion, group means were calculated using their group size ratios. The group means μ_1, μ_2, μ_3 and μ_4 for groups 1, 2, 3, and 4 were calculated using the equations explained below.

Group means were determined based on effect sizes so that consecutive groups' mean difference was 0.5 and the overall mean was 0. Therefore, for balanced two-group case equations $\mu_2 - \mu_1 = 0.5$ and $\mu_1 + \mu_2 = 0$ were solved and, $\mu_1 = -0.25$ and $\mu_2 = 0.25$ were found. For the imbalanced two-group case, equations $\mu_2 - \mu_1 = 0.5$ and, $\mu_1 + 9\mu_2 = 0$ were solved and $\mu_1 = -0.45$ and $\mu_2 = 0.05$ were found. For the three-group balanced case, equations $\mu_2 - \mu_1 = 0.5$, $\mu_3 - \mu_2 = 0.5$, $\mu_1 + \mu_2 + \mu_3 = 0$ were solved and, $\mu_1 = -0.5$, $\mu_2 = 0$ and $\mu_3 = 0.5$ were found. For the three-group imbalanced case, equations $\mu_2 - \mu_1 = 0.5$, $\mu_3 - \mu_2 = 0.5$, $\mu_1 + 2\mu_2 + 7\mu_3 = 0$ were solved and, $\mu_1 = -0.8$, $\mu_2 = -0.3$ and $\mu_3 = 0.20$ were found. For four-group balanced case, equations $\mu_2 - \mu_1 = 0.5$, $\mu_3 - \mu_2 = 0.5$, $\mu_4 - \mu_3 = 0.5$, $\mu_1 + \mu_2 + \mu_3 + \mu_4 = 0$ were solved and, $\mu_1 = -0.75$, $\mu_2 = -0.25$, $\mu_3 = 0.25$ and $\mu_4 = 0.75$ were found. Finally, for four-group imbalanced case, equations $\mu_2 - \mu_1 = 0.5$, $\mu_3 - \mu_2 = 0.5$, $\mu_4 - \mu_3 = 0.5$, $2\mu_1 + 3\mu_2 + 4\mu_3 + 11\mu_4 = 0$ were solved and, $\mu_1 = -1.1$, $\mu_2 = -0.6$, $\mu_3 = -0.1$

and $\mu_4 = 0.4$ were found. Therefore, means of groups were calculated and then introduced to the program.

The observations were generated using R's *c(rep())* function after the predictor variables were assigned their determined values based on correlations between predictor variables, group size ratios, and group sizes. Then all the observations were combined by the *data.frame (,)* function with all predictor variables and the dependent variable.

Analysis of Data

After generating the data with the defined parameters, each analysis method was applied with identical datasets with the same data conditions to predict the outcome variables separately. Therefore, the LDA, LD, and CART analyses were conducted using R's *lda*, *multinom*, and *rpart* functions. Then, an algorithm was created to assess the accuracy of the class predictions obtained from three different methods and to count the number of correct predictions.

To evaluate the performance of the methods, two outcome measures were employed: rate of correct classification for all groups (rccA) and rate of correct classification for the smallest group (rccS) in terms of the group's sample size and number of correct group predictions. The calculation of rccA involved dividing the frequency of all correctly predicted observations by the total number of observations (200). Moreover, rccS was calculated by dividing the frequency of correctly predicted observations for the smallest group by the total number of observations in the smallest group. Hence, this study's analyses were based on proportions, following Edwards' (1985) approach, which used the arcsine transformed value of the proportions as a dependent variable, and the results were the same for the proportions and transformed values.

Calculating the correct prediction rates for all and small groups for each iteration, a second set of data for comparing techniques and data conditions was prepared. A five-way (3x2x3x3x2) factorial analysis of variance (ANOVA) focusing on rccA in connection to Method, Corr, NPV, GN and GSR, and a four-way factorial (3x2x3x3) ANOVA focusing on rccS in connection to Method, Corr, NPV, GN were conducted to evaluate the results of the simulation study. The factorial ANOVA and follow-ups were conducted using SPSS statistical software (IBM Corp., 2025).

Because the statistical significance of interactions and main effects is impacted by sample size, and the sample size of 1000 (number of iterations for each combination of the conditions) is quite large, therefore partial eta squared (η_p^2) was used rather than statistical significance to identify interpretable effects. Partial eta squared is a measure that determines the proportion of total sample variation explained by a specified effect while excluding other main and interaction effects (Pierce et al., 2014; Richardson, 2011). It is calculated using the formula: $\eta_p^2 = \frac{SS_{Effect}}{SS_{Total}+SS_{Error}}$ where SS_{Effect} is sum of squares for the particular effect SS_{total} represents the total sum of squares and, SS_{error} indicates the error sum of squares. Partial eta squares values were utilized to evaluate and compare the importance of main effects and interactions.

The assumptions of factorial ANOVA are independence of data, homogeneity of variance (HOV), and normality of predictor variables. The study's design fulfilled the expectations regarding the independence of observations. On the other hand, according to Levene's test, the assumption of homogeneity of variance was not fulfilled because of large sample sizes (number of iterations), variations in group size ratios, group numbers, and distinctive means. Nearly all the cells met the criteria for normality, except for a few unbalanced situations that included two or five predictor variables and a binary outcome variable (skewnesses were still between -2 and +2). This is based on the general rule that skewness should be between -1 and +1. ANOVA, however, is resistant to HOV and violations of normality, particularly when a sizable dataset with a well-balanced design is present. The consequences of these violations were therefore disregarded.

Following the factorial ANOVA results, further analyses were conducted to explore the main and interaction effects for rccS and rccA. For follow-up analyses in the interactions, the dataset was divided

based on one of the factors in the interaction, and the effects of the other conditions were assessed based on rccS and rccA. To evaluate prediction accuracies of specified data conditions, average rccS ($\bar{X}_{rccS}$), and average rccA ($\bar{X}_{rccA}$) were defined. $\bar{X}_{rccS}$ refers to the mean rate of correct classification for the smallest group and $\bar{X}_{rccA}$ refers to the rate of correct classification for all groups in the specified conditions. Finally, the relationship between rccS and rccA for different cases was analyzed with the Pearson correlation coefficient $r_{\mathrm{rccS-A}}$.

Results

In this section, results for rccA, rccS and the relationship between rccA and rccS are presented separately.

Results for rccA

The overall factorial ANOVA model was statistically significant ($p < .001, \eta_p^2 = .969$) for the outcome variable rccA. All main effects and interactions were significant ($p < .001$). Based on partial eta squared (η_p^2) values GSR ($\eta_p^2 = .914$) was the most influential main effect, while GN ($\eta_p^2 = .907$), Method ($\eta_p^2 = .702$), NPV ($\eta_p^2 = .683$) and Corr ($\eta_p^2 = .502$) had smaller effects. Among all the two-way interactions, Method*GSR ($\eta_p^2 = .485$) was the most influential one while GN*GSR ($\eta_p^2 = .44$), NPV*GN ($\eta_p^2 = .393$), Corr*GN ($\eta_p^2 = .317$), Corr*NPV ($\eta_p^2 = .278$), NPV*GSR ($\eta_p^2 = .268$), Method*GN ($\eta_p^2 = .183$), Corr*GSR ($\eta_p^2 = .16$), Method*Corr ($\eta_p^2 = .156$) and Method*NPV ($\eta_p^2 = .059$) were decreasingly influential. Moreover, Corr*NPV*GN ($\eta_p^2 = .177$) and Method*GN*GSR ($\eta_p^2 = .169$) were the most influential three-way effects, while all the other three-way effects had partial eta squared values less than .1. Finally, all the four-way interactions and the single five-way interaction (Method*Corr*NPV*GN*GSR) had a partial eta squared value less than .1.

Only main effects for rccA are reported here since the focus of this study was prediction of the smallest group. Mean rccA for all the cases was .694 and mean rccA values for levels of main effects are presented at Table 1.

Table 1
rccA Values for Levels of Main Effects: Method, Corr, NPV, GN and GSR

Method	$\bar{X}_{rccA}$	Corr	$\bar{X}_{rccA}$	NPV	$\bar{X}_{rccA}$	GN	$\bar{X}_{rccA}$	GSR	$\bar{X}_{rccA}$
LDA	.655	.2	.718	2	.649	2	.800	Balanced	.613
LR	.681	.5	.669	5	.695	3	.658	Unbalanced	.774
CART	.745			10	.737	4	.622		

Notes. $\bar{X}_{rccA}$*:* Average rccA, Corr: Correlation, NPV: Number of the predictor variables, GN: Number of groups in dependent variable

In rccA, all groups of main effects had prediction accuracy of more than 60%. CART was highest performing method with .745 mean rccA, while LR and LDA had .681 and .655 mean rccA, respectively. As it can be seen from Table 1, higher correlation and higher group numbers resulted in lower mean rccA, while higher NPV resulted in higher rccA. Moreover, unbalanced cases had a greater rccA than balanced cases.

In most cases, CART performed better than LR and LDA. On the other hand, in the case of 10 predictor variables when Corr was .2, GSR was unbalanced, and when Corr was .5, GSR was balanced, LR performed better than CART and LDA. Moreover, when GSR was unbalanced and GN was 4, the cases when Corr was .2 or .5 and NPV was 2, 5 or 10 (6 cases) differences between LR, LDA and CART were trivial.

Results for rccS

Details of the overall factorial ANOVA results for rccS are provided in Table 2.

Table 2
Factorial ANOVA Table for the Effects of Method, Corr, NPV, and GN on rccS

Source	df	F	*p*	η_p^2
Method	2	2319.797	<.001	.079
Corr	1	6434.903	<.001	.107
NPV	2	9079.160	<.001	.252
GN	2	47265.025	<.001	.637
Method * Corr	2	471.969	<.001	.017
Method * NPV	4	72.635	<.001	.005
Method * GN	4	426.726	<.001	.031
Corr * NPV	2	805.290	<.001	.029
Corr * GN	2	654.058	<.001	.024
NPV * GN	4	359.320	<.001	.026
Method * Corr * NPV	4	71.864	<.001	.005
Method * Corr * GN	4	130.099	<.001	.010
Method * NPV * GN	8	189.364	<.001	.027
Corr * NPV * GN	4	56.785	<.001	.004
Method * Corr * NPV * GN	8	29.854	<.001	.004
Error	53946			
Total	53999			

Notes. Corr: Correlation, NPV: Number of the predictor variables, GN: Number of groups in the dependent variable

The overall factorial ANOVA model for rccS was statistically significant and had a meaningful partial eta squared value ($p < .001, \eta_p^2 = .712$). All the interactions and main effects were statistically significant ($p < .001$). Based on partial eta square values, GN ($\eta_p^2 = .637$) was the most influential effect, and NPV ($\eta_p^2 = .252$), Corr ($\eta_p^2 = .107$) and the method ($\eta_p^2 = .079$) were, in order, smaller. Interaction between Method and GN (Method*GN) ($\eta_p^2 = .031$) was the most effective two-way interaction, while interaction between Corr and NPV ($\eta_p^2 = .029$), NPV and GN ($\eta_p^2 = .026$), Corr and GN ($\eta_p^2 = .024$), Method and Corr ($\eta_p^2 = .017$) and, Method and NPV ($\eta_p^2 = .005$) were the two-way effects, in order. In addition, the interaction between Method, NPV, and GN ($\eta_p^2 = .027$) was the strongest three-way interaction, while Method*Corr*GN ($\eta_p^2 = .010$), Method*Corr*NPV ($\eta_p^2 = .005$) and Corr*NPV*GN ($\eta_p^2 = .004$) were smaller. Finally, the only four-way interaction was the interaction between Method, Corr, NPV, and GN had effect size $\eta_p^2 = .004$.

Main Effects in rccS

As stated above GN had a greater effect than the other variables on rccS and Method had the lowest effect. Among all the unbalanced cases the overall mean rccS was .325 and mean rccS values for the method, levels of correlation, NPV, and GN are shown in Table 3.

Table 3
Overall Mean rccS for Levels of Correlation, NPV, GN and Methods

Method	$\bar{X}_{rccS}$	Corr	$\bar{X}_{rccS}$	NPV	$\bar{X}_{rccS}$	GN	$\bar{X}_{rccS}$
LDA	.291	.2	.371	2	.226	2	.099
LR	.302	.5	.278	5	.329	3	.335
CART	.380			10	.418	4	.539

Notes. $\bar{X}_{rccS}$*:* Average rccS, Corr = Correlation, NPV = Number of the predictor variables, GN = Number of groups in dependent variable

Based on the findings, LDA with .291 average rccS demonstrated the lowest overall performance, followed by LR with .302 and CART with .38 average rccS, which was the highest performing method in rccS overall. In terms of correlation, a lower degree of correlation (.2) resulted in better performance than a higher degree of correlation (.5). Moreover, the cases having a larger number of predictor variables had better performance in terms of rccS such as cases of 2 predictor variables had average rccS of .226, while cases of 5 and 10 predictor variables had .329 ad .418 average rccS, respectively. Finally, having more groups resulted in greater average rccS in this setting. Change in method in terms of average rccS from highest accuracy to lowest was .089, while change in Corr was .093, change in NPV was .192 and change in GN was .44. Therefore, it can be observed that data conditions had greater effects than method in terms of prediction accuracy of small groups.

While evaluating mean rccS values for main effects gives an overall idea about prediction accuracy for the smallest groups, it is important to evaluate interactions so that change in prediction accuracy for a main effect when change in other factors occurs may be investigated. Therefore, for the main effect of Method, data were divided into groups, and prediction accuracies were evaluated based on changes in other variables.

Two-way interactions in rccS

All the two-way interactions for rccS were statistically significant but had smaller effect sizes compared to the effect sizes of the main effects. Comparing Method interactions with other variables based on partial eta squared values, it was observed that the interaction of Method with GN ($\eta_p^2 = .031$) had a greater effect than the interaction of Method with Corr ($\eta_p^2 = .017$), and interaction of Method with NPV ($\eta_p^2 = .005$). Mean rccS scores of the methods at the levels of Corr, NPV, and GN are presented in Table 4.

Table 4
Mean rccS Values for Interactions of Method with Corr, NPV and GN

Method	Corr	$\bar{X}_{rccS}$	NPV	$\bar{X}_{rccS}$	GN	$\bar{X}_{rccS}$
LDA	.2	.348	2	.186	2	.048
	.5	.234	5	.295	3	.305
			10	.391	4	.519
LR	.2	.364	2	.193	2	.050
	.5	.240	5	.307	3	.316
			10	.407	4	.540
CART	.2	.402	2	.298	2	.200
	.5	.359	5	.387	3	.384
			10	.456	4	.557

Notes. $\bar{X}_{rccS}$: Average rccS, Corr: Correlation, NPV: Number of the predictor variables, GN: Number of groups in dependent variable.

Increasing Corr resulted in decreases in the mean rccS for all the methods; increasing Corr from .2 to .5 resulted in .114 decrease in LDA, .124 decrease in LR and .043 decrease in CART for mean rccS. Therefore, it can be inferred that CART is the least affected method by the change in Corr, and LR and LDA had similar changes in mean rccS when changing Corr. On the other hand, increasing NPV resulted in increases in the mean rccS for all the methods. Increasing NPV from 2 to 10 resulted in .205 increase in LDA, .214 increase in LR and .158 increase in CART. Thus, CART was the least affected model in the change of NPV and LR and LDA had similar performances in favor of LR. Finally, increasing GN resulted in increases in the mean rccS for all the methods; increasing GN from 2 to 4 resulted in .471 increase in LDA, .490 increase in LR and .357 increase in CART. Thus, the change in GN had a greater impact on LR and LDA than CART. In conclusion, it was observed that LR was most the sensitive method to data conditions, while LDA was the second and CART was the least affected method by data conditions. GN was the most influential data condition on the method's rccS performances, and NPV and Corr were lesser.

Besides two-way interactions, three-way interactions were also analyzed in detail, as the effect size for Method*NPV*GN ($\eta_p^2 = .027$) was close to the effect sizes of two-way interactions. The four-way interaction was not inspected due to the small effect size ($\eta_p^2 = .004$).

Three-way Interactions in rccS

There were four three-way interactions in the design of this study, and the interactions that included Method were evaluated in detail. Mean rccS values for the interaction between Method, NPV, and GN are presented in Table 5.

Table 5

Mean rccS Values of the Methods for the Different Levels of NPV and NG

NPV	GN	$\bar{X}_{rccS}(LDA)$	$\bar{X}_{rccS}(LR)$	$\bar{X}_{rccS}(CART)$
2	2	.013	.013	.102
	3	.174	.181	.297
	4	.372	.386	.493
5	2	.038	.039	.206
	3	.309	.328	.382
	4	.536	.552	.573
10	2	.094	.099	.291
	3	.431	.441	.472
	4	.648	.683	.606

Notes. $\bar{X}_{rccS}(LDA)$: Average rccS in LDA, $\bar{X}_{rccS}(LR)$: Average rccS in LR, $\bar{X}_{rccS}(CART)$: Average rccS in CART, GN: Number of groups in dependent variable, NPV: Number of the predictor variables.

According to the results presented in Table 5, when controlling for Method and NPV, an increase in GN resulted in an increase in mean rccS in all the levels of NPV in the methods. When there were 2 predictor variables, increasing the number of groups from 2 to 4 LDA increased the mean rccS score from .013 to .372 (difference = .359) while LR increased the mean rccS score from .013 to .386 (difference = .373) and CART from .102 to .493 (difference = .391). Thus, CART was the model that was improved most by the change in GN. Moreover, CART was the best performing model for all NPV cases when GN was 2. Similarly, CART was the best performing model in the case when there were 5 predictor variables, and LR was the most improved model in rccS (from .039 to .552, difference = .513). Similarly, in the case when there were 10 predictor variables LR was the most improved model in rccS and it was the best performing model when the number of groups was 4. On the other hand, when the numbers of the groups were 2 and 3, CART was the best performing method. Thus, increasing NPV and GN produce results in favor of LR and LDA rather than CART.

For three-way interaction Method*Corr*GN, average rccS values for the methods at different levels of Corr and GN are presented in Table 6.

Table 6
Mean rccS Values for the Methods at Different Levels of Corr and GN

Corr	GN	$\bar{X}_{rccS}(LDA)$	$\bar{X}_{rccS}(LR)$	$\bar{X}_{rccS}(CART)$
.2	2	.065	.068	.219
	3	.376	.396	.401
	4	.603	.629	.586
.5	2	.032	.033	.181
	3	.234	.237	.366
	4	.435	.451	.529

Notes. $\bar{X}_{rccS}(LDA)$: Average rccS in LDA, $\bar{X}_{rccS}(LR)$: Average rccS in LR, $\bar{X}_{rccS}(CART)$: Average rccS in CART, Corr: Correlation, GN: Number of groups in dependent variable

According to the results in Table 6, at the fixed levels of Corr, an increase in GN resulted in an increase in average rccS for all the methods. When correlations between variables were .2, increasing the number of groups 2 to 4, LDA improved mean rccS from .065 to .603 (difference = .538) while LR improved mean rccS score from .068 to .629 (difference = .561) and CART from .219 to .586 (difference = .367). Hence, LR was the most affected model by the change in GN. Moreover, while CART was the best performing model in cases when there were 2 or 3 groups, LR was the best performing model for the case when there were 4 groups.

Table 7
Mean rccS Values for the Methods at Different Levels of Corr and NPV

Corr	NPV	$\bar{X}_{rccS}(LDA)$	$\bar{X}_{rccS}(LR)$	$\bar{X}_{rccS}(CART)$
.2	2	.207	.214	.310
	5	.356	.371	.402
	10	.480	.508	.494
.5	2	.165	.172	.286
	5	.233	.242	.372
	10	.303	.307	.419

Notes. $\bar{X}_{rccS}(LDA)$: Average rccS in LDA, $\bar{X}_{rccS}(LR)$: Average rccS in LR, $\bar{X}_{rccS}(CART)$: Average rccS in CART, Corr: Correlation, NPV: Number of the predictor variables.

Fixing Corr at .2, an increase in NPV resulted in an increase in mean rccS for all the methods; increasing NPV from 2 to 10 mean rccS in LDA increased from .207 to .48 (difference = .273), in LR from .214 to .508 (difference = .294), and in CART from .31 to .494 (difference = .184). Hence, in the cases when Corr was .2, CART was the least affected method by the change in NPV and it was notable that LR exceeded the CART in terms of mean rccS at the highest level of NPV. On the other hand, for the cases when Corr was .5 change in NPV from 2 to 10 resulted in similar changes in rccS's of LDA (difference = .138), LR (difference = .135), and CART (difference = .133). Furthermore, CART's performance was superior to the other two methods when Corr was .5 at all the different levels of NPV.

Relationship Between rccS and rccA

To analyze the relationship between the smallest group prediction accuracy and prediction accuracy for all groups, the Pearson correlation coefficient was first employed for all the cases together, then for the different levels of main effects, and finally, for different levels of main effects at different levels of GN.

The whole data for rcsS and rccA were normally distributed based on skewness values between -1 and 1. Besides, the rccS and rccA values demonstrated a normal distribution for the main effects and their respective levels within the GN levels, with skewness ranging from -1 to 1. However, exceptions occurred when GN was 2, where skewness values for rccS and rccA ranged from 1 to 2. Specifically, when GN was 2 and NPV was 2, the skewness for rccS reached 2.686, while for rccA it was 2.282. The outcomes of these cases were carefully analyzed and compared with Spearman correlation coefficients. The Spearman and Pearson correlation coefficients were close to each other, and differences between these values did not change the direction of the analyses, so only Pearson correlations are reported.

There was a notable difference between the overall rccS and rccA values: for all the unbalanced cases the overall mean rccS was .323, while the overall mean rccA was .774. Moreover, the correlation between rccS and rccA for all the cases was -.461, which means there was a negative and medium correlation between rccS and rccA. Besides, correlation values for different levels of main effects are presented in Table 8.

Table 8

Correlation between rccS and rccA at Different Levels of GN, Method, NPV and Corr

GN	$r_{\mathrm{rccS-A}}$	Method	$r_{\mathrm{rccS-A}}$	NPV	$r_{\mathrm{rccS-A}}$	Corr	$r_{\mathrm{rccS-A}}$
2	.797	LDA	-.515	2	-.620	.2	-.444
3	.637	LR	-.486	5	-.629	.5	-.568
4	.676	CART	-.477	10	-.463		

Notes. $r_{\mathrm{rccS-A}}$: Pearson Correlation between rccS and rccA, Corr: Correlation, NPV: Number of the predictor variables, GN: Number of groups in dependent variable.

When there were 2 groups, the correlation between rccS and rccA was .797 while it was .637 and .676 for the cases of group number were 3 and 4, respectively. In LDA, $r_{\mathrm{rccS-A}}$ was -.515 while it was -.486 and -.477 in LR and CART, respectively. Moreover, it was -.620, -.629 and -.463 when the number of predictor variables was 2, 5, and 10, respectively. Finally, in the case when the correlation between variables was .2, the correlation between rccA and rccS was -.444, while it was -.568 when the correlation between predictor variables was .5. Since the correlation between rccS and rccA was negative for the groups of method, NPV and Corr and it was positive for GN, a more detailed analysis was conducted by splitting data into GN for further analysis. Correlation values between rccS and rccA at the levels of the method, NPV, and Corr into levels of GN are presented in Table 9.

Table 9

Correlation between rccS and rccA at the Levels of Method, NPV and Corr for Fixed Levels of GN

GN	Method	$r_{\mathrm{rccS-A}}$	NPV	$r_{\mathrm{rccS-A}}$	Corr	$r_{\mathrm{rccS-A}}$
2	LDA	.762	2	.792	.2	.794
	LR	.791	5	.787	.5	.798
	CART	.818	10	.775		
3	LDA	.766	2	.532	.2	.623
	LR	.775	5	.489	.5	.593
	CART	.457	10	.537		
4	LDA	.767	2	.431	.2	.705
	LR	.784	5	.489	.5	.483
	CART	.352	10	.684		

Notes. $r_{\mathrm{rccS-A}}$: Pearson Correlation between rccS and rccA, Corr: Correlation, NPV: Number of the predictor variables, GN: Number of groups in dependent variable.

On splitting data by GN, $r_{\mathrm{rccS-A}}$ was positive for all levels of Method and NPV and Corr even though it was negative before splitting. This demonstrates the impact of GN on the relationship between rccS and rccA. In the case when GN was 2, for the methods, the highest correlation between rccS and rccA was for CART ($r_{\mathrm{rccS-A}} = .818$) and the lowest correlation was for LDA ($r_{\mathrm{rccS-A}} = .762$). For different degrees of NPV and Corr correlations between rccS and rccA were high and there were trivial differences in terms of $r_{\mathrm{rccS-A}}$. In the case when GN was 3, there was no notable difference between LR ($r_{\mathrm{rccS-A}} = .766$), LDA ($r_{\mathrm{rccS-A}} = .775$) and CART ($r_{\mathrm{rccS-A}} = .457$) in terms of $r_{\mathrm{rccS-A}}$. Moreover, for different levels of NPV and Corr when GN was 3, there were not important differences in terms of $r_{\mathrm{rccS-A}}$. Finally, when GN was 4, the difference between LR and CART in terms of $r_{\mathrm{rccS-A}}$ became greater since $r_{\mathrm{rccS-A}}$ was .784 for LR and .352 for CART. Moreover, increasing NPV resulted in increase in $r_{\mathrm{rccS-A}}$ while increasing Corr resulted in a decrease in $r_{\mathrm{rccS-A}}$. Finally, differences in $r_{\mathrm{rccS-A}}$ between cases of .2 Corr and .5 Corr when GN was 2, 3, and 4 were .004, .03, and .222, respectively. Thus, when GN was 4 the difference was notably greater than the cases when GN was 2 and 3.

Discussion

This study delved into the comparative effectiveness of three prevalent classification methods CART, LDA, and LR to evaluate their performance in predicting group membership specifically for proportionally small groups across various controlled conditions. Even though there were certain instances in which LR performed better, one of the primary findings that emerged from this research was that CART consistently displays superior performance across most settings. LR tended to outperform LDA and CART in the cases with a high number of predictor variables, low correlation between variables, and an abundance of groups. Consistent with these results, specifically in the simulation studies, the superiority of CART is supported by existing research (Finch et al., 2014; Holden et al., 2011). In addition to that, for most of the cases, LR and LDA had similar performances, though in almost every case LR showed slightly better accuracy. Hence, even though there are conflicting findings indicating that LDA performs better than LR (Williams, 1999), the finding that LR performs better than LDA (Barön, 1991) or CART, particularly when assumptions for LDA are satisfied, and that there are insignificant differences between LR and LDA (Hestie et al., 2009) are supported by the literature.

This research demonstrates that an important component affecting prediction accuracy is the ratio of group sizes, especially when evaluating smaller groups' predictions. This emphasizes the unequal impact that group size can exert on classification accuracy. Moreover, the number of groups is identified as a significant determinant of accuracy. In agreement with previous studies, an increase in the number of groups resulted in a decrease in overall prediction accuracy (Finch & Schneider, 2007; Pohar et al.,2004). On the other hand, this study also demonstrated that the prediction accuracy of small groups was enhanced as the number of groups increased.

By the design of this study, the number of groups is engaged with degrees of group separation. Since groups were separated by a determined mean difference between consecutive groups, cases with a higher number of groups had greater levels of group separation. For example, for the two group cases, the mean difference between large and small groups was .5 while for the four group cases difference between large and small groups was 1.5. Therefore, differences between large and small groups might affect discrimination and prediction of small groups. When group sizes are unbalanced and group separation is large, small groups can be recognized more accurately. Still, this research highlights that smaller group classification accuracy benefits from an augmentation in the number of groups. Consequently, a larger group separation makes it easier to predict smaller groups and smaller groups are more readily discriminated from larger groups. On the other hand, overall group prediction may be decreased due to the members of larger groups predicted as in the smaller groups. Besides, the performance of methods for overall classification diminishes with an increase in the number of groups, signifying that managing multi-group situations continues to be difficult. It was concluded that all the controlled conditions had a greater impact on small group prediction than on overall prediction accuracy in terms of the percentage

of correctly predicted observations. Finally, the results showed that all the controlled data conditions had a greater impact on the accuracy of small group prediction than on overall group predictions.

In this study, it was found that an increase in the number of predictor variables improved the classification accuracy across all the methods and data conditions. This finding aligned with Finch and Schneider (2007) who stated that the accuracy of group membership prediction is improved with the addition of new predictor factors. This pattern appears stronger in LDA and LR compared to CART, indicating that these two methods might have a superior ability to utilize complicated, high-dimensional data.

In line with earlier studies, correlation influenced classification accuracy (Kiang, 2003). Furthermore, the results of this study align with the observation made by Pai et al. (2012) concerning the ineffectiveness of multicollinear factors, as increased correlation diminishes the contributions of additional variables. The maximum correlation level for this study was .5; so, at higher values, negligible or little contributions may be anticipated. This study revealed that the impact of correlation was diminished for CART compared to LR and LDA regarding overall and small group prediction accuracy. When predictor variables demonstrate minimal correlations, predictive accuracy often increases, benefiting all three techniques, especially CART and LR. This enhancement is particularly important for smaller groups, where precise classifications are critical. Furthermore, CART demonstrates superior robustness in managing imbalanced datasets compared to LDA and LR, which often encounter difficulties in such scenarios. Nonetheless, LR exhibits optimal performance when the data is balanced and evenly distributed among groups.

This study indicates that overall prediction accuracy is remarkably greater than that of small group prediction accuracy, a conclusion corroborated by Chiang (2021). This study also highlights the correlation between the accuracy of predictions for all groups and the accuracy of predictions for the smallest groups. In all the situations, a moderate negative correlation was found; however, for the same number of groups, a significant positive correlation was found. Therefore, the impact of group size and degree of separation on the relationship between small and overall group prediction accuracy was examined. It was concluded that small group and overall group prediction accuracies have parallel characteristics at the same number of groups, while for mixed numbers of groups they tend to have inverse characteristics.

This study makes useful suggestions for practitioners: Less than 10 predictors and smaller groups are best suited for CART, but larger datasets with more groups and predictor variables are better suited for LR. However, unless certain requirements are satisfied, such as equal covariance and normality, LDA is not advised.

While this study offers a thorough evaluation of the performance of CART, LDA, and LR in terms of small group prediction, in addition to the effect of the data conditions on prediction accuracy, it recognizes a few limitations. Since the study uses simulated data, it might not accurately represent actual circumstances. For instance, the data's group separation was maintained at fixed standardized mean differences, which restricts the study's generalizability to situations with non-normal distributions or variable group separation. Further research is encouraged to investigate the consequences of varied sample sizes, non-normal data distributions, and variable levels of group separation. The complex nature of numerous controlled circumstances necessitated the simulation of data under the assumption of multivariate normality for each category, representing an additional restriction of this work. Additionally, factors such as the presence of categorical predictor variables, multimodality, varying sample sizes between groups, and heterogeneity of variance-covariance matrices were not addressed in this work.

It is advised to look at more recent approaches that may provide better results in specific situations, like support vector machines, random forests, and neural networks, as well as investigating more sophisticated classification methods outside of CART, LDA, and LR. The handling of unbalanced datasets and methods for improving the classification of smaller groups are two areas of special interest for further study. This is particularly important because smaller groups frequently have less prediction accuracy, which can produce biased results in practical applications. Furthermore, particular attention

should be paid to how existing techniques might be enhanced to optimize accuracy and judgment in progressively difficult classification tasks, thus promoting the field of predictive modeling.

In summary, the study offers a comprehensive analysis of three widely used classification techniques, highlighting their performance in controlled settings. CART is notable for its adaptability, yet in high-dimensional, multi-group situations, LR proves to be a formidable competitor. For LDA to work effectively, stricter requirements must be met. Practitioners looking to select the best approach for their data classification requirements might benefit from the study's insights. We encourage future developments in classification techniques, especially when handling unbalanced data and smaller groups, indicating the significance of ongoing research and development in the predictive modeling space.

Declarations

This paper was adapted from the first author's doctoral dissertation. A part of this paper was presented at the 2019 annual meeting of the American Educational Research Association (AERA).

Conflict of Interest: The authors declare no conflicts of interest.

Ethical Approval: This study utilized simulated data; therefore, ethical approval was not required. The authors confirm that all ethical guidelines were followed.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

Barön, A. E. (1991). Misclassification among methods used for multiple group discrimination-the effects of distributional properties. *Statistics in Medicine*, *10*(5), 757-766. doi: https://doi.org/10.1002/sim.4780100511

Bates, B. E., Xie, D., Kwong, P. L., Kurichi, J. E., Ripley, D. C., & Stineman, M. G. (2014). One-year all-cause mortality after stroke: A prediction model. *PM&R*, *6*(6), 473-483. doi: https://doi.org/10.1016/j.pmrj.2013.11.006

Bolin, J., & Finch, W. (2014). Supervised classification in the presence of misclassified training data: A Monte Carlo simulation study in the three-group case. *Frontiers in Psychology*, *5* doi:10.3389/fpsyg.2014.00118

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Castonguay, A. C., Zoghi, Z., Zaidat, O. O., Burgess, R. E., Zaidi, S. F., Mueller-Kronast, N., ... & Jumaa, M. A. (2023). Predicting functional outcome using 24-hour post-treatment characteristics: Application of machine learning algorithms in the STRATIS registry. *Annals of Neurology*, *93*(1), 40-49. https://doi.org/10.1002/ana.26528

Chiang, Y. C. (2021). *Evaluating the performance of classification and regression trees, random forests, and K-means clustering under controlled conditions* (Doctoral dissertation, Indiana University).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates.

Dattalo, P. (1995). A comparison of discriminant analysis and logistic regression. *Journal of Social Service Research*, *19*(3-4), 121-144.

Dey, E. L., & Astin, A. W. (1993). Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Research in Higher Education*, *34*, 569-581. doi: https://doi.org/10.1007/BF00991920

Edwards, A. L. (1985). *Experimental design in psychological research* (5th ed.). NY: Harper & Row.

Fan, X., & Wang, L. (1999). Comparing linear discriminant function with logistic regression for the two-group classification problem. *The Journal of Experimental Education, 67*(3), 265-286. doi:10.1080/00220979909598356

Ferrer, A. J. A., & Wang, L. (1999). *Comparing the Classification Accuracy among Nonparametric, Parametric Discriminant Analysis and Logistic Regression Methods* (pp. 1-24, Rep.). Montreal: Paper presented at the Annual Meeting of the American Educational Research Association. https://eric.ed.gov/?id=ED432591

Finch, H. W., Bolin, J. E., & Kelley, K. (2014). Group membership prediction when known groups consist of unknown subgroups: A Monte Carlo comparison of methods. *Frontiers in Psychology*, *5*. doi:10.3389/fpsyg.2014.00337

Finch, H., & Schneider, M. K. (2007). Classification accuracy of neural networks vs. discriminant analysis, logistic regression, and classification and regression trees: Three- and five-group cases. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3*(2), 47-57. doi:10.1027/1614-2241.3.2.47

Finch, W. H., & Schneider, M. K. (2006). Misclassification rates for four methods of group classification. *Educational and Psychological Measurement, 66*(2), 240-257. doi:10.1177/0013164405278579

Graf, R., Zeldovich, M., & Friedrich, S. (2023). Comparing linear discriminant analysis and supervised learning algorithms for binary classification—A method comparison study. *Biometrical Journal*, *66*(1), 2200098. doi: https://doi.org/10.1002/bimj.202200098

Grassi, M., Villani, S., & Marinoni, A. (2001). Classification methods for the identification of "case" in epidemiological diagnosis of asthma. *European Journal of Epidemiology, 17, 19-29.* doi: https://doi.org/10.1023/A:1010987521885

Hassan, H. A., Hemdan, E. E. D., El-Shafai, W., Shokair, M., & Abd El-Samie, F. E. (2024). Detection of attacks on software defined networks using machine learning techniques and imbalanced data handling methods. *Security and Privacy*, *7*(2), e350. doi: https://doi.org/10.1002/spy2.350

Hastie, T., Tibshirani R., Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hess, B., Olejnik, S., & Huberty, C. J. (2001). The efficacy of two improvement-over-chance effect sizes for two-group univariate comparisons under variance heterogeneity and nonnormality. *Educational and Psychological Measurement*, *61*(6), 909-936. doi: https://doi.org/10.1177/00131640121971572

Hao, Z., Yang, B., Ruggiano, N., Ma, Y., Guo, Y., & Pan, X. (2022). Depression prediction amongst Chinese older adults with neurodegenerative diseases: A performance comparison between decision tree model and logistic regression analysis. *The British Journal of Social Work*, *52*(1), 274-290. doi: https://doi.org/10.1093/bjsw/bcaa237

Hoang, M. L., Matrella, G., & Ciampolini, P. (2025). Metrological evaluation of contactless sleep position recognition using an accelerometric smart bed and machine learning. *Sensors and Actuators A: Physical*, *385*, 116309. doi: https://doi.org/10.1016/j.sna.2025.116309

Holden, J. E., Finch, W. H., & Kelley, K. (2011). A comparison of two- group classification methods. *Educational and Psychological Measurement, 71*(5), 870-901. doi:10.1177/0013164411398357

Holden, J. E., & Kelley, K. (2010). The effects of initially misclassified data on the effectiveness of discriminant function analysis and finite mixture modeling. *Educational and Psychological Measurement, 70*(1), 36-55. doi:10.1177/0013164409344533

Huberty, C. J. (1994). *Applied discriminant analysis*. John Wiley & Sons.

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (Vol. 498). John Wiley & Sons.

IBM Corp. (2025). IBM SPSS Statistics for Windows, Version 30.0.0. Armonk, NY: IBM Corp. Retrieved from https://www.ibm.com/us-en/marketplace/statistical-analysis-and-reporting

Kiang, M. Y. (2003). A comparative assessment of classification methods. *Decision Support Systems*, *35*(4), 441-454. doi: https://doi.org/10.1016/S0167-9236(02)00110-0

Lei, P., & Koehly, L. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case. *Journal of Experimental Education, 72*(1), 25-49. doi: https://doi.org/10.1080/00220970309600878

Meshbane, A., & Morris, J. D. (1995). A method for selecting between linear and quadratic classification models in discriminant analysis. *Journal of Experimental Education*, *63*(1), 263-273. doi: https://doi.org/10.1080/00220973.1995.9943813

Pai, D. R., Lawrence, K. D., Klimberg, R. K., & Lawrence, S. M. (2012). Analyzing the balancing of error rates for multi-group classification. *Expert Systems with Applications*, *39*(17), 12869-12875. doi: https://doi.org/10.1016/j.eswa.2012.05.006

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo Experiments: Design and implementation. *Structural Equation Modeling*, 8(2), 287-312. doi: https://doi.org/10.1207/S15328007SEM0802_7

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement*, *64*(6), 916-924. doi: https://doi.org/10.1177/0013164404264848

Pohar, M., Blas, M., & Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: A simulation study. *Metodoloski Zvezki*, *1*(1), 143-161. http://mrvar.fdv.uni-lj.si/pub/mz/mz1.1/pohar.pdf

Preatoni, D. G., Nodari, M., Chirchella, R., Tosi, G., Wauters, L. A., & Martinoli, A. (2005). Identifying bats from time-expanded recordings of search calls: Comparing classification methods. *Journal of Wildlife Management*, *69*(1), 1601-1614. doi: https://doi.org/10.2193/0022-541X(2005)69[1601:IBFTRO]2.0.CO;2

R Core Team (2016). R: A Language and Environment for Statistical Computing. R foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.r-project.org/

Rausch, J. R., & Kelley, K. (2009). A comparison of linear and mixture models for discriminant analysis under nonnormality. *Behavior Research Methods*, *41*(1), 85-98. doi: https://doi.org/10.3758/BRM.41.1.85

Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135-147. doi: https://doi.org/10.1016/j.edurev.2010.12.001

Ripley, B. D. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society: Series B (Methodological)*, *3*(1), 409-456. doi: https://doi.org/10.1111/j.2517-6161.1994.tb01990.x

Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., & Ullah, N. (2022). A method for improving prediction of human heart disease using machine learning algorithms. *Mobile Information Systems*, *2022*(1), 1410169. doi: https://doi.org/10.1155/2022/1410169

Sainani, K. L. (2014). Explanatory versus predictive modeling. *PM&R*, *6*(9), 841-844. doi: https://doi.org/10.1016/j.pmrj.2014.08.941

Schumacher, M., Rossner, R., & Vach, W. (1996). Neural networks and logistic regression: Part I. *Computational Statistics: Data Analysis*, *21*(1), 661-682. doi: https://doi.org/10.1016/0167-9473(95)00032-1

Selim, G. E. I., Hemdan, E. E. D., Shehata, A. M., & El-Fishawy, N. A. (2021). Anomaly events classification and detection system in critical industrial internet of things infrastructure using machine learning algorithms. *Multimedia Tools and Applications*, *80*(8), 12619-12640. doi: https://doi.org/10.1007/s11042-020-10354-1

Soureshjani, M. H., & Kimiagari, A. M. (2013). Calculating the best cut off point using logistic regression and neural network on credit scoring problem-A case study of a commercial bank. *African Journal of Business Management*, *7*(16), 1414. doi: 10.5897/AJBM11.394

Song, G., Ai, Z., Zhang, G., Peng, Y., Wang, W., & Yan, Y. (2022). Using machine learning algorithms to multidimensional analysis of subjective thermal comfort in a library. *Building and Environment*, *212*, 108790. doi: https://doi.org/10.1016/j.buildenv.2022.108790

Suh, S., Suh, J., & Houston, I. (2007). Predictors of categorical at-risk high school dropouts. *Journal of Counseling & Development*, *85*(2), 196-203. doi: https://doi.org/10.1002/j.1556-6678.2007.tb00463.x

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Boston: Pearson Education.

Williams, C. J., Lee, S. S., Fisher, R. A., & Dickerman, L. H. (1999). A comparison of statistical methods for prenatal screening for Down syndrome. *Applied Stochastic Models in Business and Industry*, *15*(2), 89-101. doi: https://doi.org/10.1002/(SICI)1526-4025(199904/06)15:2<89::AID-ASMB366>3.0.CO;2-K

Zampogna, B., Torre, G., Zampoli, A., Parisi, F., Ferrini, A., Shanmugasundaram, S., ... & Papalia, R. (2024). Can machine learning predict the accuracy of preoperative planning for total hip arthroplasty, basing on patient-related factors? An explorative investigation on Supervised machine learning classification models. *Journal of Clinical Orthopaedics and Trauma*, *53*, 102470. doi: https://doi.org/10.1016/j.jcot.2024.102470

Zavorka, S., & Perrett, J. J. (2014). Minimum sample size considerations for two-group linear and quadratic discriminant analysis with rare populations. *Communications in Statistics-Simulation and Computation*, *43*(7), 1726-1739. doi: https://doi.org/10.1080/03610918.2012.744041

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2025; 16(1); 48-58

Investigation of Activities For Reading Comprehension Skills: A G-Theory Analysis

Gülden KAYA UYANIK* Serap ATAOĞLU**

Abstract

This study aimed to investigate the effectiveness of activities prepared to improve reading comprehension skills based on the variables of number of raters, evaluation criteria, and number of activities. Twelve raters evaluated five reading comprehension activities created by the researcher. A descriptive survey method grounded in a quantitative approach was employed. The study utilized five reading comprehension activities, commonly used in high school textbooks, and a rubric developed by the researcher, consisting of sixteen criteria based on relevant literature. After performing reliability and validity analyses on the rubric, the experts assessed the activities using this tool. The data collected from their evaluations were analyzed through generalizability theory. The EduG program was used to estimate variance values for both main and interaction effects according to generalizability theory, calculate the scores' reliability using G and Φ (Phi) coefficients, and conduct decision (D) studies. The findings revealed that each of the reading comprehension activities used to improve students' comprehension skills is different from each other. Additionally, it was concluded that increasing the number of criteria included in the rubric and increasing the number of expert raters would lead to a more accurate and effective evaluation of the activities.

Keywords: reading comprehension, activity, rater, rubric, generalizability theory

Introduction

Reading, one of the four basic language skills, plays an important role in language teaching and is defined as a receptive skill. The necessity of reading skills is not only crucial for the content of language learning, but also for other courses. Despite the use of various technological tools in today's education systems and the continuous development of these tools, reading maintains its place and importance in education and training practices, and education and training activities are widely based on reading skills (Smith, Snow, Serry, & Hammond, 2021). Meanwhile, research findings outlined in the literature (Floris & Divina, 2015; Hunt & Beglar, 2005) emphasize that reading alone is not enough. This skill is fully utilized and serves its purpose when the content of the text is understood. Reading comprehension skills are very important in educating individuals capable of thinking, questioning, producing, and inferring. In instances where students are unable to read fluently and comprehend the text adequately, it cannot be asserted that the act of reading has fully achieved its intended purpose. Reading comprehension includes readers' ability to recognize and perceive symbols in the text, thinking skills, and lifelong knowledge and experiences. The interest and desire for reading, the intended goals of reading, one's opinions about reading, and the location where the act of reading takes place all influence the process of reading comprehension (Akyol, 2005). Individuals who can comprehend what they read can be successful in various fields, including social, scientific, political, economic, and so on. The healthy execution of comprehension and expression skills in mother tongue lessons also affects students' success in other courses. Understanding the problem is very important in order to solve the problems encountered in the lessons (Güvendir, 2014).

* Assoc. Prof. Dr., Sakarya University, Faculty of Education, Sakarya-Türkiye, guldenk@sakarya.edu.tr, ORCID ID: 0000-0002-8100-6994
** Milli Eğitim Bakanlığı Sakarya, Sakarya-Türkiye, imra5425@gmail.com, ORCID ID: 0000-0002-3849-0493

To cite this article:
Kaya Uyanık, G., & Ataoğlu, S. (2025). Investigation of activities for reading comprehension skills: A G-Theory analysis. *Journal of Measurement and Evaluation in Education and Psychology, 16*(1), 48-58. https://doi.org/10.21031/epod.1548738

Received: 12.09.2024
Accepted: 18.02.2025

Each country conducts its own national exam (for Turkey - ABIDE) to identify and improve students' reading comprehension skills, and there are international practices such as PIRLS, TIMSS, and PISA. One of the purposes of these large-scale exams is to evaluate the effectiveness of teaching methods and materials that help students acquire reading skills and comprehension skills and the impact of these skills on other academic achievements (Mullis et al., 2009).

Several factors can help students develop reading comprehension skills. Some of these factors include the development of organs that assist reading, the use of strategies for reading comprehension, and enjoying the act of reading (Kim et al. 2021). Furthermore, factors such as the content and structural features of the text, its attractiveness to the student, its sufficiency regarding vocabulary and grammar rules, the student's level of knowledge, desire to read, and internalization of the text are also influential in reading comprehension (Baştuğ et al. 2019). Considering all these factors under one concept, it is indispensable to carry out activities within education and training activities for the improvement of reading comprehension skills, which is crucial for every age period. The activities designed for this purpose aimed at reading comprehension skills are among the instruments constantly used in education and training activities. Considering the reports of national and international large-scale exams, it is seen that the basis of all problems is the ability to understand and interpret what is read. To facilitate the development of these skills, activities targeting different age groups are designed within the educational processes, and these prepared activities are frequently employed in lessons.

The dictionary definition of "activity," which we frequently hear in the teaching field via the constructivist approach, is "the state of being active." Activities are significant in developing individuals' language skills, ensuring permanent learning, and helping them develop the habit of reading (Clarke et al., 2010). Activities are used to teach students both specific and general skills. The specific aim of the activities is to transform the learning outcome into behavior. The general aim is to equip students with skills such as creative thinking, critical thinking, etc. In other words, while the activities ensure the acquisition of the determined outcomes, they also play an important role in differentiating students' perspectives and making what they learn permanent. Students can improve their reading skills and gain creative thinking skills with the help of activities that they can relate to their lives and include problems they may encounter (Başpınar, 2013). Considering their effects on the acquisition and development of desired skills, it is possible to say that the activities have an important contribution to the teaching process. Activities allow students to develop their language and thinking skills by providing them with relevant acquisitions in a suitable period based on a predetermined plan, thus enabling them to learn easily, quickly, permanently, and systematically (Güneş, 2017).

When it comes to the activities prepared for reading comprehension in educational processes, the potential obstacles in the evaluation of these activities should also be considered. It is seen from the studies that one of the factors affecting the literature is the issue of who and how the activities prepared for reading comprehension skills will be evaluated. (Long & Pang, 2015; Myford & Wolfe, 2003; Snyder, Caccamise & Wise, 2005; Şata & Karakaya, 2021; Wiseman, 2012). In this regard, studies that reveal the effect of the rater in reading comprehension activities and the characteristics of the rubric used in scoring are required. One of the theories that helps to reveal the effect of these statistical properties is the Generalizability Theory (G-Theory).

Generalizability (G) Theory is based on the analysis of variance and is similar to Classical Test Theory (CTT). However, unlike the CTT, in G-theory, the sources of the error rate in the observed scores can be obtained in detail. In G-theory, error rates can be determined separately for each error source and for the interaction of these sources (Shavelson and Webb, 1991). G theory uses the concepts of the "facet, object of measurement, condition, and design." The concept of facet is the definition used for each of the sources of variability in the universe (Brennan, 2001). The source of variance in the universe whose effect is examined for the research purpose is the "object of measurement." In the theory, variance due to the object of measurement is desirable, while large variance due to facets is undesirable. The different levels that facets have are called conditions (Guler, Kaya Uyanik, & Tasdelen Teker, 2012). For instance, consider a scenario in which five raters evaluate a 10-question examination administered to a class of 50 students. In this context, the students' exams represent the objects of measurement, while

raters and questions serve as sources of error, which are examined and treated as facets. The condition for the rater facet in the study was five, whereas the condition for the question facet was ten.

Another concept that needs to be addressed in G-theory is the designs of facets. Crossed or nested are the types of designs that are considered in this theory (Shavelson & Webb, 1991). A crossed design is when all the conditions of one facet are associated with all the conditions of another facet. A nested design is a type of design in which a condition of one facet is associated with some conditions of another facet. These designs also differ in terms of notation, with the crossed design represented by "x" and the nested design represented by ":" (Shavelson & Webb, 1991).

In G-Theory, there are two different coefficients, G and Phi, as reliability coefficients. The main difference between these coefficients is that the sources of variance of the object of measurement considered in the assessments are examined in relative and absolute terms. The G coefficient is used for relative assessment, and the Phi coefficient is a usable absolute assessment (Brennan, 2001).

In G-theory, reliability can be obtained for two different cases called Generalizability (G) and Decision (D) studies. The G study is concerned with generalizing to the universe based on the universe in which the measurements are made, thus aiming to provide information about the sources of variability in the sample. In the D study, scenarios are created for a specific purpose by using the information obtained in the G study, and decision-making is aimed at these scenarios (Brennan, 2001; Guler, Kaya Uyanik, & Tasdelen Teker, 2012; Nalbantoglu & Gelbal, 2011).

This study, which emphasizes the importance of reading and reading comprehension skills in educational activities, aimed to question the effectiveness of the activities prepared for reading comprehension by evaluating them by different raters and increasing their efficiency by identifying their deficiencies. For this purpose, a completely crossed randomized design of a (activity) x r (rater) x c (criterion) was created. With the design created, answers to the questions of variance values for the main and interaction effects and reliability of the test were sought. In addition, Decision (D) studies were conducted, and scenarios suitable for the features of the facets were created. In this regard, the main problem of the research is as follows:

What are the variance values for the main and interaction effects of the a (activity) x r (rater) x c (criterion) completely crossed randomized design and the reliability of the test as a result of the examination of the activities prepared for reading comprehension skills by different raters with the specified criteria? In the study, answers were sought to three sub-problems.

1. What are the variance values for the main and interaction effects in a (activity) x r (rater) x c (criterion) completely crossed randomized design?

2. What are the G and Φ (Phi) coefficients calculated for the reliability of scores in a (activity) x r (rater) x c (criterion) completely crossed randomized design?

3. What are the reliability values obtained from scenarios created with different numbers of raters and criteria in a (activity) x r (rater) x c (criterion) completely crossed randomized design?

Method

Research Design

In this study, which was conducted to examine the activities prepared to measure reading comprehension skills by different raters and to make suggestions for increasing the efficiency and effectiveness of these activities, the "descriptive survey" design, one of the quantitative research methods, was used. The descriptive survey model aims to describe a past or ongoing situation as it exists. The individual, object, or event to be researched is defined within its own conditions, as it is. The researcher does not attempt to intervene, influence, or change shape (Karasar, 2010). The main purpose of this model is to describe and explain the situation in detail (Çepni, 2010).

Study Group

The study group consisted of 12 raters who are teachers working in measurement and evaluation centers located in different provinces of Turkey and are experts in their fields. Demographic information of the raters is given in Table 1.

Table 1
Demographic Information of the Study Group

Variables	Category	Frequency (f)	Percent (%)
Gender	Female	8	66,7
	Male	4	33,3
Profession	Turkish language and literature	4	33,3
	Curriculum development	4	33,3
	Measurement and Evaluation	4	33,3
Educational Status	Bachelor	2	16,7
	Master	6	50
	Doctorate	4	33,3
Professional seniority	6-10 years	1	8,3
	11-15 years	6	50
	16-20 years	3	25
	21-25 years	2	16,7
	Total	12	100

Table 1 shows that, of the participants, 8 were female, 4 were male, 4 were experts in Turkish language and literature, 4 were in curriculum development, and 4 were in measurement and evaluation. Two participants had bachelor's degrees, six had master's degrees, and four had doctorate degrees. It was observed that the least experienced participant had 6 years of seniority and 92% had more than 10 years of experience. As can be understood from these data, the study was conducted with experienced and expert evaluators.

Data Collection Tools

In this study, five activities to measure reading comprehension skills in the 10th and 12th grade Skill-Based Turkish Language and Literature Books written by the researcher within the General Directorate of Secondary Education (GDSE) of the Ministry of National Education (MoNE) and a rubric consisting of 16 items created by the researcher to determine the suitability of these activities were used as data collection tools.

Activities

The activities used as one of the data collection tools of this study were selected from the Skill-Based Activity Books written by Turkish Language and Literature and Turkish teachers employed by the GDSE in the 2020-2021 academic year. Three of the selected activities are at 10th-grade level, and the other two are at 12th-grade level. One of the researchers who conducted this study wrote the activities for the GDSE. The written activities aim to identify and improve students' reading and reading comprehension skills. The activities used in the research aimed to develop reading skills among domain skills and critical thinking skills among general skills. The learning outcome-based activities were sent to field experts, curriculum development experts, measurement and evaluation experts, language experts, and guidance experts employed under the GDSE. The opinions of these experts were taken, the activities revised in line with the feedback received were finalized, and the activities collected in an interactive book were uploaded to the GDSE and made available to students and teachers.

Rubric

The rubric was developed by the researchers. First, a question pool of 26 items was prepared based on the literature to determine the effect of the activities used in the study on improving reading comprehension skills. Upon examination, repetitive items, items with little relationship with the content, and items with a broad scope were removed, and a trial form consisting of 16 items was prepared. The prepared form was presented to expert opinion. Opinions were received from two faculty members who were experts in the field of Turkish education and two faculty members who were experts in the field of measurement and evaluation. The items were arranged according to the feedback from the experts and the experts reached a consensus on the final version of the form. At the end of these processes, the final form consisting of 16 items was created.

Data Analysis

The data obtained from the activities examined through the rubric and the raters were analyzed according to the Generalizability Theory. In the study, the variance values for the main and interaction effects of the a (activity) x r (rater) x c (criterion) completely crossed randomized design, which was formed by analyzing the activities prepared for reading comprehension skills (object of measurement) by different raters with the specified criteria, were examined. G and Φ (Phi) coefficients were calculated for the reliability of the test scores of the design used in the study. In addition, decision (D) studies were conducted, and future scenarios were created. The EduG program was utilized to estimate the main and interaction effect's variance values according to the generalizability theory, to calculate the reliability of the scores, and to carry out D studies.

Results

The variance values for the main and interaction effects of the a (activity) x r (rater) x c (criterion) completely crossed randomized design, which was formed by evaluating the activities prepared for reading comprehension by different raters using specified criteria, were investigated, and the results are given in Table 2.

Table 2

Variance Values and Total Variance Explanation Rates Estimated by the G Study Regarding the axrxc Design

Source of Variance	df	Sum of Squares	Mean Squares	Variance	%
a	4	99.09792	24.77448	7.56847	63.2
r	11	10.56146	0,96013	0.29312	2.8
c	15	5.05521	0.33701	0.00209	0.2
ar	44	43.02708	0.97788	1.63254	14.3
ac	60	21.96875	0.36615	1.08145	9.8
rc	165	12.05729	0.07307	0.01005	1.1
Arc,e	660	13.10625	0.01985	0.96780	8.6
Total	959	204.87396			100%

An analysis of the variance estimated and total variance explained ratios of the axrxc fully crossed randomized design in Table 2 shows that the variance component estimated for the main effect of activity (a) explains 63.2% of the total variance. In generalizability studies, the main effect taken as the object of measurement is evaluated as the variance of the universe score and refers to the differentiation between activities in this study in terms of the measured feature (Shavelson & Webb, 1991; Brennan, 2001; Guler, Kaya Uyanik, & Tasdelen-Teker, 2012; Kaya Uyanik & Guler, 2016). The ratio of the variance estimated for activities to the total variance should be large. This indicates that differences between activities can be revealed in the dimension obtained by measurement (Brennan, 2001; Kaya Uyanik & Guler, 2016). According to the results obtained in this study, it can be said that the evaluation of activities based on criteria can reveal the differences between activities. The variance component estimated for the rater main effect (0.29) explains 2.8% of the total variance. This value is the third smallest value. The rater's main effect is due to inconsistency between raters' ratings. Therefore, it is desirable that this effect is low. The variance component estimated from the G study for the main effect

of criterion (c) explains 0.02% of the total variance. The criterion's main effect shows the degree of differentiation of the difficulty level of each measurement unit (item) in the rubric. According to the results obtained, it can be interpreted that the attainability levels of the criteria used to measure reading comprehension skills were similar to each other.

The activity x rater (ar) interaction effect explains 14.3% of the total variance, and this value is the second highest value. The activity and rater interaction refers to the inconsistency of the raters in terms of generosity-severity in scoring for some activities. In this case, it was concluded that although the raters gave generally consistent results, there were differences between their scoring in some activities. The activity x criterion (ac) interaction effect explains 9.8% of the total variance and is the third largest variance obtained. This shows that the relative status of certain activities differs from one criterion to another. In other words, it can be interpreted that the scores given to the criteria differ from activity to activity. Rater x criterion (rc) interaction effect explains 1.1% of the total variance. This value is the second smallest variance obtained. For this result, it can be interpreted that there is no difference between the raters according to the criteria. The variance component of the activity x rater x criterion (residual) interaction effect variance explains 8.6% of the total variance. A large residual variance is an indication of a large interaction of activity, rater, criterion, unmeasured sources of variability, and/or random errors. When the values obtained are examined, it is observed that the rate of random errors is low for the study.

G and Φ (Phi) coefficients were calculated for the reliability of scores in the axrxc completely crossed randomized design. In the rubric containing the criteria in the study, there are sixteen criteria in total and these criteria were scored by twelve raters. In this case, the G coefficient was 0.885, and the Phi coefficient was 0.863. It can be said that the measurements obtained from the measurement tool used are reliable.

Decision (D) studies are conducted using the variance values calculated over the data used in the generalizability study. D study allows the estimation of the coefficients G and Phi for the reliability values by decreasing and increasing the conditions of the facets in the universe G, respectively. Table 4 shows the values of G and Phi coefficients calculated by keeping the criterion facet constant and decreasing and increasing the number of raters, and the values of G and Phi coefficients calculated by keeping the rater facet constant and decreasing and increasing the number of criteria in the D study.

Table 3

axrxc Fully Crossed Randomized Design D Study Results

	Number of Rater	G coefficient	Φ coefficient
Number of Criteria: 16	5	0.782	0.743
	10	0.850	0.845
	15	0.886	0.864
	20	0.887	0.875
	25	0.889	0.877
	Number of Criteria		
Number of Rater: 12	5	0.740	0.726
	10	0.810	0.799
	15	0.882	0.861
	20	0.906	0.887
	25	0.914	0.909

In Table 3, G and Phi coefficients were calculated for the two different cases. In the first case, the number of criteria was kept fixed at 16, and the number of raters varied from 5 to 25. Also, in the second case, the number of raters was constant at 12, and the number of criteria varied from 5 to 25. When the number of criteria was kept fixed and the number of raters was changed, it was observed that the reliability value increased as the number of raters increased. However, it was observed that after 15 raters, the increase in reliability was significantly low for every 5-rater increase. Similarly, G and Phi coefficients were calculated when the number of raters was kept constant at 12, and the number of criteria was 5, 10, 15, 20, and 25. When the number of raters was kept constant and the number of criteria was changed, the

highest reliability value was obtained from the scenario where the number of criteria was 25. It was observed that the reliability value increased as the number of criteria increased.

Discussion and Conclusion

The present study examined the design obtained by evaluating the activities for reading comprehension skills by using rubrics with the expert raters. The analysis revealed the criteria for more effective and efficient development of activities prepared to measure reading comprehension skills in terms of structure and content. An axrxc design was used in the study. An analysis of the axrxc design examined in the study regarding sources of variance shows that the largest source of variance is due to activity. Considering that the study involved activities designed for reading comprehension, these activities are diverse rather than being of a single type. Furthermore, the high value of this source of variance indicates that the effect of each activity to be used for reading comprehension is different. It can be concluded that students can better develop this skill if they interact more with reading comprehension activities. In addition, considering that the situation emphasized by each activity will be different, it is an important conclusion that the scoring keys should be arranged accordingly. Supporting this result, the study also revealed that for the axrxc design, the activity criterion interaction effect was also a significant source of variance. The interaction of activity and criterion indicates that the criteria are met in some activities and not in others. This result shows that not every activity met every criterion, so it can be stated that the criteria including the features that should be present in reading comprehension activities cannot be provided with only one activity alone, and it would be more accurate to use more than one activity. Considering the information obtained from these two findings, a small number of activities aimed at measuring reading comprehension skills will create a deficiency in terms of meeting the criteria. Therefore, a large number of activities will increase reading comprehension skills. Recent studies have emphasized the importance of using activities that serve this purpose in order to increase reading comprehension skills (Akyol & Ketenogluarter Kayabasi, Topuz 2018; Collins, et al. 2020; Siti & Mumu, 2022; Brillianandа & Wibowo, 2023).

The necessity of increasing the number of activities and using a rubric in evaluating these activities has been emphasized in many studies in the literature because it reveals the learning objectives clearly and understandably, reduces the errors involved in the evaluation, and provides an opportunity to complete missing learning (Arter, 2002; Dunbar, Brooks, & Miller, 2006; Hall & Salmon, 2003; Oaklef, 2009; Wolf and Steven, 2007). Another question that comes to mind is the number of criteria in the rubrics used in the evaluation. In the decision studies conducted in the study, it was observed that the reliability of the study increased as the number of criteria increased. In the study, the maximum value for the number of criteria was 25, and the highest reliability was obtained from this value. Similarly, when the number of criteria was reduced to five, a value around 0.70 was obtained. In this respect, it can be concluded that the number of criteria should be at least five and that there is no upper limit. One of the most valid ways to ensure objectivity and inter-rater reliability in multi-rater measurements is to use rubrics (Jonsson,& Svingby,2007). The reliability and validity of rubrics have been examined from various perspectives. While some researchers have focused on the objectivity of rubrics (Rezaei, & Lovorn,2010 ; Spandel, 2006; Wolfe, 1997), others have critiqued them as being overly reductive (Kohn, 2006; Mabry, 1999). However, the interaction result in all the studies mentioned is that using rubrics is a more reliable way than not using them. In this case, what to consider when using rubrics is another important issue that increases reliability. In the literature, it has been emphasized that the number of items should be increased as well as different factors (Henson, R, & Thompson, 2002; Hellman, Fuqua & Worley, 2006). At this point, when the studies in the literature and the results obtained from this study are interpreted together, using rubrics increases reliability and it can be said that for a more reliable measurement, the criteria in the rubrics should be at least 5 and reliability will increase as the number of criteria increases.

Another important source of variance for the axrxc design is the event rater interaction effect. The activity and rater interaction refer to the inconsistency of the raters in terms of generosity-rigor in scoring for some activities. In this case, it was concluded that although the raters gave generally consistent results, there were differences between their scoring in some activities. This result revealed that the

activities may have different effects on different raters, leading to a scoring disadvantage. Within this perspective, it can be claimed that scoring the activities to measure reading comprehension skills by a single rater would create deficiencies, whereas evaluating the activities by more than one rater would increase the efficiency of the activities. This result is consistent with the literature in terms of both rater reliability and reducing the error rate of the process (Alkan & Doğan, 2023; Kim, 2020; Kim et al. 2021). This result obtained from the present study and other studies in the literature raises the question of the required number of raters. The finding obtained through the decision study conducted in the present study has been a relevant answer in this context. Based on the decision studies conducted to answer the question of how much the number of raters should be increased, it was concluded that the reliability of the study increased as the number of raters increased, but the increase in the reliability value was not very high if the number of raters was above 15, so it would not be practical to increase the number of raters above 15.

The findings of the study necessitate separate discussions for classroom assessments and large-scale assessments. According to the results, when the number of raters is five, the reliability coefficient exceeds the interactionally accepted threshold of 0.70 for Cronbach's Alpha. Considering that for classroom assessments with multiple raters, an acceptable reliability coefficient can be as low as 0.60 (DeVellis & Thorpe, 2021), it can be stated that even with fewer than five raters, acceptable reliability can still be achieved. Thus, in classroom assessments, having multiple raters invariably yields more reliable results compared to assessments conducted with a single rater. On the other hand, Cizek (2009) highlights that there should be procedural distinctions between classroom and large-scale assessments. For large-scale examinations, the acceptable threshold for reliability is higher than that for classroom assessments. When evaluated in the context of large-scale examinations, the finding that 15 raters represent an upper limit is both significant and practical. At both national and international levels, large-scale examinations often involve open-ended questions that require multiple raters. The number of raters required for evaluating these exams becomes a critical factor in managing the assessment process. In Turkey, for instance, the pilot implementation and the first official administration of the "four-skill Turkish language exams"—which consist of both open-ended and multiple-choice questions—were conducted in 2024. These exams were administered to approximately 10,000 students across 4th, 7th, and 11th grades. For the writing and speaking skills components, open-ended assessments were used, and multiple raters were involved in the evaluation process for each grade level. In the pilot study, which involved approximately 2,000 participants, it was reported that five raters were sufficient for reliable scoring (MoNE, 2020). However, the significant discrepancy between the number of participants in the pilot study and the actual implementation (approximately 10,000) indicated the need for an increased number of raters. Based on the results of this study, it can be concluded that 15 raters are sufficient for the four-skill language examinations. On the other hand, it is worth noting that working with 15 raters is not easy. In this context, it is recommended that the selection of raters and the harmony processes between the obtained scores should be carried out with scientific steps.

The most significant finding of this study, which involved the scoring of reading comprehension activities by different raters based on specific criteria, is that these activities exhibit a high level of variance. Accordingly, it can be stated that the activities differentiate in terms of assessing students' reading comprehension skills. This suggests, indirectly, that students need to encounter a wide variety of activities in order to develop their reading comprehension skills. In this regard, it is recommended that teachers, school administrators, and educational policymakers emphasize the importance of numerous reading activities to enhance students' reading comprehension abilities.

In scoring using rubrics, the difference between the raters decreases and compliance increases. Therefore, it is recommended that the activities for reading comprehension be scored using a rubric to determine the reliability of the rater and obtain more reliable results. A 16-item rubric was used in this study. The findings indicate that as the number of criteria increases, the reliability of the raters also improves. Therefore, it is recommended to increase the number of criteria in the scoring rubric used to assess reading comprehension skills as much as possible. On the other hand, according to the results of this study, it is considered important for the reliability of the rubric that the number of criteria should not be fewer than five. Additionally, it was observed that after 20 criteria, increasing the number to 25

did not result in a sharp improvement. In this context, it is suggested that the rubric should include at least five criteria, and considering usability and practicality, there is no need to exceed 25 criteria.

Similar to the present study, in which it was found that a high number of raters increased reliability, different raters could be used in scoring, and the number of raters could be increased up to 15. It is recommended to keep this number around 15, especially in large-scale exams, as increasing the number of raters above 15 will not make a big difference in the results.

The present study, intended to determine the effectiveness level of reading skills activities and their deficiencies concerning structure and content, can also be applied to writing, speaking, and listening skills, which are among the basic language skills, and their rater reliability can be examined. Most of the raters who contributed to this study are experts and experienced in their fields. It could be taken into consideration that experienced raters make more accurate interpretations and judgments than less experienced raters (Jorgenson, 1975), and similar studies could be conducted by grouping raters according to their experience. This study examined the activities for reading comprehension skills prepared by the researcher and used in the MoNE. Similar studies can be conducted by utilizing different types, content, and grade-level activities to determine reading comprehension skills.

Declarations

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

Ethical Approval: We declare that all ethical guidelines for authors have been followed by all authors. Ethical approval for the study was received from Sakarya University, Educational Sciences Ethics Committee dated 15.02.2023 numbered E-61923333-050.99-222299

References

Akyol, H. (2005). *Turkish primary reading and writing teaching*, Ankara: PegemA.

Akyol, H., & Ketenoğlu Kayabaşı, Z. E. (2018). Improving the Reading Skills of a Students with Reading Difficulties: An Action Research. *Education and Science, 43*(193). https://doi.org/10.15390/EB.2018.7240

Alkan, M., & Doğan, N. (2023). A Comparison of Different Designs in Scoring of PISA 2009 Reading Open Ended Items According to Generalizability Theory. *Journal of Measurement and Evaluation in Education and Psychology, 14*(2), 106-117. https://doi.org/10.21031/epod.1210917

Arter, J. (2002). Rubrics, scoring guides, and performance criteria. In C. Boston (Ed.), *Understanding Scoring Rubrics a Guide for Teachers* (p. 21-31). Office of Educational Research and Improvement.

Başpınar Yörük, N. (2013). *An investigation on the use of creativity development methods in 6th grade Turkish course reading activities.* (Master thesis), University of Necmettin Erbakan Üniversitesi, Konya. Accessed from YOK Thesis Center database (Dissertation No: 348744).

Baştuğ, M., Hiğde, A., Çam, E., Örs, E., & Efe, P. (2019). *Strategies, techniques, practices to improve reading comprehension skills.* Ankara: PegemA.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer Verlag.

Brilliananda, C., & Wibowo, S. E. (2023). Reading Strategies for Post-Pandemic Students' Reading Comprehension Skills. *International Journal of Elementary Education, 7*(2).

Cizek, G. J. (2009). Reliability and validity of information about student achievement: Comparing large-scale and classroom testing contexts. *Theory into practice, 48*(1), 63-71.

Clarke, P. J., Snowling, M. J., Truelove, E. & Hulme, C. (2010). Ameliorating Children's Reading-Comprehension Difficulties: A Randomized Controlled Trial. Psychological Science, 21(8), 1106–1116. https://doi.org/10.1177/0956797610375449

Collins, A. A., Compton, D. L., Lindström, E. R., & Gilbert, J. K. (2020). Performance variations across reading comprehension assessments: Examining the unique contributions of text, activity, and reader. *Reading and Writing, 33*(3), 605-634.

Çepni, S. (2010). *Introduction to research and project work.* Trabzon: Celepler.

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications.* Sage publications.

Dunbar, N. E., Brooks, C. F. & Miller, T. K. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education, 31*(2), 2006, 115-128.

Floris, F. D., & Divina, M. (2015). A study on the reading skills of EFL university students. *Teflin Journal, 20*(1), 37–47.

Güler, N., Kaya Uyanık, G., & Taşdelen Teker, G. (2012). *Generalizability theory.* Ankara: PegemA.

Güneş, F. (2017). Activity approach in teaching Turkish, *Journal of Native Language Education, 5*(1), 48-64. https://doi.org/10.16916/aded.286415

Güvendir, M. A. (2014). Öğrenci başarılarının belirlenmesi sınavında öğrenci ve okul özelliklerinin Türkçe başarısı ile ilişkisi. *Eğitim ve Bilim, 39*(172).

Hall, E. K. & Salmon, S. J. (2003). Chocolate chip cookies and rubrics helping students understand rubrics in inclusive settings. *Teaching Exceptional Children, 35*(4), 8-11.

Hellman, C. M., Fuqua, D. R., & Worley, J. (2006). A reliability generalization study on the survey of perceived organizational support: The effects of mean age and number of items on score reliability. *Educational and psychological measurement, 66*(4), 631-642.

Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies. *Measurement and Evaluation in Counseling and Development,* 35, 113-126.

Hunt, A., & Beglar, D. (2005). A framework for developing EFL reading vocabulary. *Reading in a Foreign Language, 17*(1), 23–59.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review, 2*(2), 130-144.

Jorgenson, G. W. (1975). An analysis of teacher judgments of reading level. American Educational Research Journal, 12 (1), 67-75. https://doi.org/10.2307/1162581

Karasar, N. (2010). *Scientific research method.* Ankara: Nobel.

Kaya Uyanık, G., & Güler, N. (2016). Examining the reliability of concept map scores: An example of a crossover mixed design in generalizability theory. *Hacettepe University Faculty of Education Journal 31*(1). 97-11. http://doi.org/10.16986/HUJE.2015014136

Kim, J. S., Relyea, J. E., Burkhauser, M. A., Scherer, E., & Rich, P. (2021). Improving elementary grade students' science and social studies vocabulary knowledge depth, reading comprehension, and argumentative writing: A conceptual replication. *Educational Psychology Review*, 1-30.

Kim, Y. S. G. (2020). Hierarchical and dynamic relations of language and cognitive skills to reading comprehension: Testing the direct and indirect effects model of reading (DIER). *Journal of Educational Psychology, 112*(4), 667.

Kohn, A. (2006). The trouble with rubrics. *English Journal,* 95(4), 12–15.

Long, H., & Pang, W. (2015). Rater effects in creativity assessment: A mixed methods investigation. Thinking Skills and Creativity, 15, 13-25.

Mabry, L. (1999). Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappan,* 80(9), 673–679.

MoNE. (2020). *Turkish Language Exam in Four Skills: Pilot Study Results.* https://www.meb.gov.tr/meb_iys_dosyalar/2020_01/20094146_Dort_Beceride_Turkce_Dil_Sinavi_Ocak_2020.pdf

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement, 4*(4), 386-422.

Nalbantoğlu F., & Gelbal S. (2011). Comparison of different designs with generalizability theory at the communication skills station scale, *Hacettepe University Faculty of Education Journal, 41*, 509-518. https://dergipark.org.tr/tr/download/article-file/87423

Oaklef, M. (2009). Using rubrics to assess information literacy: An examination of methodology and ınterrater reliability. *Journal of the American Society for Information Science and Technology, 60*(5), 969-983. https://doi.org/10.1002/asi.21030

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing writing,* 15(1), 18-39. https://doi.org/10.1016/j.asw.2010.01.003

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage. http://doi.org/10.1002/9781118445112.stat00068

Siti, M., & Mumu, M. (2022). The effect of critical multiliteracy learning model on students' reading comprehension. *International Journal of Educational Qualitative Quantitative Research (IJE-QQR), 1*(1), 28-33.

Smith, R., Snow, P., Serry, T., & Hammond, L. (2021). The role of background knowledge in reading comprehension: A critical review. *Reading Psychology, 42*(3), 214-240.

Snyder, L., Caccamise, D., & Wise, B. (2005). The assessment of reading comprehension: Considerations and cautions. *Topics in Language Disorders*, *25*(1), 33-50.

Spandel, V. (2006). In defense of rubrics. *English Journal*, 96 (1), 19–22. https://doi.org/10.58680/ej20065683

Şata, M., & Karakaya, İ. (2021). Investigating the Effect of Rater Training on Differential Rater Function in Assessing Academic Writing Skills of Higher Education Students. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(2), 163-181. https://doi.org/10.21031/epod.842094

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*(3), 150-173.

Wolf, K. & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *The Journal of Effective Teaching, 7*(1), 3-14.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, 4(1), 83–106.