

I

J

A

T

E

Volume 7

Issue 2

2020

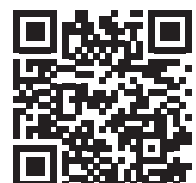
*International Journal of
Assessment Tools in Education*

<https://dergipark.org.tr/en/pub/ijate>

<http://www.ijate.net>

e-ISSN: 2148-7456

© IJATE 2020





e-ISSN 2148-7456

<https://dergipark.org.tr/en/pub/ijate>
<http://www.ijate.net>

Volume 7

Issue 2

2020

Dr. İzzet KARA

Publisher

International Journal of Assessment Tools in Education

&

Pamukkale University,

Education Faculty,

Department of Mathematic and Science Education,

20070, Denizli, Turkey

Phone : +90 258 296 1036

Fax : +90 258 296 1200

E-mail : ijate.editor@gmail.com

Frequency : 4 issues per year starting from June 2018 (March, June, September, December)

Online ISSN: 2148-7456

Website : <http://www.ijate.net/index.php/ijate>

<http://dergipark.org.tr/en/pub/ijate>

Design & Graphic: IJATE

Support Contact

Dr. İzzet KARA

Journal Manager & Founding Editor

Phone : +90 258 296 1036

Fax : +90 258 296 1200

E-mail : ikara@pau.edu.tr

International Journal of Assessment Tools in Education (IJATE) is a peer-reviewed and academic online journal.

The scientific and legal responsibility for manuscripts published in our journal belongs to the authors(s).



International Journal of Assessment Tools in Education

International Journal of Assessment Tools in Education (IJATE) is an international, peer-reviewed online journal. IJATE is aimed to receive manuscripts focusing on evaluation and assessment in education. It is expected that submitted manuscripts could direct national and international argumentations in the area. Both qualitative and quantitative studies can be accepted, however, it should be considered that all manuscripts need to focus on assessment and evaluation in education.

IJATE as an online journal is sponsored and hosted by **TUBITAK-ULAKBIM** (The Scientific and Technological Research Council of Turkey).

There is no submission or publication process charges for articles in IJATE.

IJATE is indexed in:

- Emerging Sources Citation Index (ESCI),
- Education Resources Information Center (ERIC),
- TR Index (ULAKBIM),
- European Reference Index for the Humanities and Social Sciences (ERIH PLUS),
- Directory of Open Access Journals (DOAJ),
- Index Copernicus International
- SOBIAD,
- JournalTOCs,
- MIAR 2015 (Information Matrix for Analysis of the Journals),
- idealonline,
- CrossRef,
- ResearchBib,

Editors

Dr. Eren Can Aybek, *Pamukkale University*, Turkey

Dr. Özen Yıldırım, *Pamukkale University*, Turkey

Editorial Board

Dr. Beyza Aksu Dünya, *Bartın University*, Turkey

Dr. R. Şahin Arslan, *Pamukkale University*, Turkey

Dr. Stanislav Avsec, *University of Ljubljana*, Slovenia

Dr. Murat Balkıs, *Pamukkale University*, Turkey

Dr. Gülşah Başol, *Gaziosmanpaşa University*, Turkey

Dr. Bengü Börkan, *Boğaziçi University*, Turkey

Dr. Kelly D. Bradley, *University of Kentucky*, United States

Dr. Okan Bulut, *University of Alberta*, Canada

Dr. Javier Fombona Cadavieco, *University of Oviedo*, Spain

Dr. William W. Cobern, *Western Michigan University*, United States

Dr. R. Nükhet Çıkrıkçı, *İstanbul Aydın University*, Turkey

Dr. Safiye Bilican Demir, *Kocaeli University*, Turkey

Dr. Nuri Doğan, *Hacettepe University*, Turkey

Dr. Selahattin Gelbal, *Hacettepe University*, Turkey

Dr. Anne Corinne Huggins-Manley, *University of Florida*, United States

Dr. Violeta Janusheva, *"St. Kliment Ohridski" University*, Republic of Macedonia

Dr. Francisco Andres Jimenez, *Shadow Health, Inc.*, United States

Dr. Nicole Kaminski-Öztürk, *The University of Illinois at Chicago*, United States

Dr. Orhan Karamustafaoglu, *Amasya University*, Turkey

Dr. Yasemin Kaya, *Atatürk University*, Turkey

Dr. Hulya Kelecioğlu, *Hacettepe University*, Turkey

Dr. Hakan Koğar, *Akdeniz University*, Turkey

Dr. Omer Kutlu, *Ankara University*, Turkey

Dr. Sunbok Lee, *University of Houston*, United States

Dr. Froilan D. Mobo, *Ama University*, Philippines

Dr. Hamzeh Moradi, *Sun Yat-sen University*, China

Dr. Nesrin Ozturk, *Ege University*, Turkey

Dr. Turan Paker, *Pamukkale University*, Turkey

Dr. Abdurrahman Sahin, *Pamukkale University*, Turkey

Dr. Murat Dogan Sahin, *Anadolu University*, Turkey

Dr. Ragıp Terzi, *Harran University*, Turkey

Dr. Hakan Türkmen, *Ege University*, Turkey

Dr. Hossein Salarian, *University of Tehran*, Iran

Dr. Kelly Feifei Ye, *University of Pittsburgh*, United States

English Language Editors

Dr. Hatice Altun, *Pamukkale University*, Turkey

Dr. Çağla Atmaca, *Pamukkale University*, Turkey

Dr. Sibel Kahraman, *Pamukkale University*, Turkey

Arzu Kanat Mutluoğlu - *Pamukkale University*, Turkey

Copy & Language Editor

Anıl Kandemir, *Middle East Technical University*, Turkey

Table of Contents

Research Article

1. [Wald Test Formulations in DIF Detection of CDM Data with the Proportional Reasoning Test / Pages: 145-158](#)
Likun HOU, Ragıp TERZİ, Jimmy DE LA TORRE
2. [The Psychometric Properties of School Belonging Scale for Middle School Students / Pages: 159-176](#)
Bekir DIREKCI, Mehmet CANBULAT, Ibrahim TEZCİ, Serdar AKBULUT
3. [Adaptation of the STEM Value-Expectancy Assessment Scale to Turkish Culture / Pages: 177-190](#)
Arif ACIKSOZ, Yakup Özkan OZKAN, Ilbilge DOKME
4. [Analyzing Different Module Characteristics in Computer Adaptive Multistage Testing / Pages: 191-206](#)
Melek Gülşah ŞAHİN
5. [Investigation of Measurement Invariance of Science Motivation and Self-Efficacy Model: PISA 2015 Turkey Sample / Pages: 207-222](#)
Metehan GÜNGÖR, Kübra ATALAY KABASAKAL
6. [The Development of a Scale to Evaluate Foreign Language Skills at Preparatory Schools / Pages: 223-235](#)
Recep Şahin ARSLAN
7. [The Development of Teachers' Knowledge of the Nature of Mathematical Modeling Scale / Pages: 236-254](#)
Reuben ASEMPAPA
8. [Parametric or Non-parametric: Skewness to Test Normality for Mean Comparison / Pages: 255-265](#)
Fatih ORCAN
9. [Factors Affecting Academic Self-efficacy of Syrian Refugee Students: A Path Analysis Model / Pages: 266-279](#)
Hasibe YAHSİ SARI, Selahattin GELBAL, Halil SARI
10. [Psychometric Characteristics of Written Response Instruments Used in Postgraduate Theses Completed in Special Education / Pages: 280-304](#)
Gamze SARIKAŞ, Safiye BİLİCAN DEMİR

Wald Test Formulations in DIF Detection of CDM Data with the Proportional Reasoning Test

Likun Hou ¹, Ragip Terzi ^{2,*}, Jimmy de la Torre ³

¹ Educational Testing Service, NJ, USA

² Department of Educational Measurement and Evaluation, Harran University, Sanliurfa, Turkey

³ Division of Learning, Development and Diversity, The University of Hong Kong, Pokfulam, Hong Kong

ARTICLE HISTORY

Received: Feb 16, 2020

Revised: Apr 7, 2020

Accepted: Apr 17, 2020

KEYWORDS

Cognitive Diagnosis Model,
Proportional Reasoning,
Wald Test,
DIF

Abstract: This study aims to conduct differential item functioning analyses in the context of cognitive diagnosis assessments using various formulations of the Wald test. In implementing the Wald test, two scenarios are considered: one where the underlying reduced model can be assumed; and another where a saturated CDM is used. Illustration of the different Wald test to detect DIF in CDM data was based on the items' performance of the Proportional Reasoning test among low- and high-performing school students. A benchmark simulation study was included to compare the performance of the Wald test in each scenario. The agreement of the latent attribute classification based on different cognitive diagnosis models was also discussed.

1. INTRODUCTION

Cognitive diagnosis models (CDMs) are a family of multidimensional latent class models that are used to obtain finer grained information on students' learning progress. CDMs classify examinees based on attribute mastery profiles that determine students' membership in latent groups. Each latent group is denoted by a binary vector with 1s and 0s, indicating mastery and nonmastery of each of the attributes being measured, respectively.

To date, despite the benefits of cognitive diagnosis assessments (CDAs), the application of CDMs has been limited. Some researchers (Tatsuoka, 1984; Tjoe & de la Torre, 2014) have created some tests based on CDA through an intensive study. In these studies, specific latent attributes were constructed as finer-grained and interrelated, but separable skills within a domain of interest. However, many psychometric questions about the CDM framework still remain. One such question is about differential item functioning (DIF) in CDMs. DIF analyses are regularly carried out for the purpose of test fairness and validity (Camilli, 2006). In the context of CDMs, DIF occurs if students with the same attribute mastery profile but from distinct observed groups have different probabilities of correctly answering an item (Hou, de la

CONTACT: Ragip Terzi ✉ terziragip@harran.edu.tr 📧 Harran University, School of Education, Sanliurfa, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

Torre, & Nandakumar, 2014; Li, 2008). DIF analysis is necessary to examine parameter or construct invariance (Zumbo, 2007). Invariance pertains to the item responses that should be independent conditioned on attribute profiles. Therefore, DIF analysis is important to investigate the invariance of attribute-item interactions across groups (Hou et al., 2014).

Currently, there exist a few studies for DIF detection purposes in CDMs (e.g., Hou et al., 2014; Li, 2008; Milewski & Baron, 2002; Zhang, 2006). Milewski and Baron (2002) examined group differences in skill mastery profiles controlling for overall ability where skill strengths and weaknesses were analyzed. However, they did not investigate whether an item was biased due to a specific skill. Furthermore, the Mantel-Haenszel (MH; Holland & Thayer, 1988) and SIBTEST methods (Shealy & Stout, 1993) were applied by Zhang (2006) to examine DIF for the deterministic inputs, noisy “and” gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model based on total test scores and attribute profile scores. However, the two methods were limited to detect only uniform DIF. Moreover, the estimates of the item parameters and attribute mastery profiles were contaminated because of including potential DIF items in the procedures. The study of Milewski and Baron (2002) was extended by Li (2008) to a modified higher-order DINA model (HO-DINA; de la Torre & Douglas, 2004), where DIF and differential attribute functioning (DAF) were simultaneously investigated. In addition to DIF described previously, DAF occurs if students with the same attribute mastery profile but from different observed groups have different probabilities of mastering an attribute. The higher-order (HO) structure in this procedure explains the relationship among items, attributes, and general ability; however, it was also limited to uniform DIF detections. Given these limitations, Hou et al. (2014) introduced the Wald test for DIF detection purposes in the DINA model. This procedure has two major advantages. First, separate calibrations were performed for the reference (R) and focal (F) groups so as not to require test purification for DIF contaminations. Second, the procedure can effectively detect both uniform and nonuniform DIF. The Wald test also outperformed the MH and SIBTEST procedures in detecting uniform DIF.

This study aims to carry out DIF analyses in the context of CDMs using various formulations of the Wald test. In implementing the Wald test, two scenarios were considered: one where the underlying reduced model (i.e., DINA model) was assumed; another scenario where a saturated CDM was used. The purpose of this study is to illustrate the performance of the different Wald tests in detecting DIF in CDM data; thus, the Proportional Reasoning test data (Tjoe & de la Torre, 2014) for schools with different proficiency levels were used. In particular, DIF items are detected when the groups are defined as high-performing school versus low-performing school.

1.1. Theoretical Framework

1.1.1. G-DINA Model

In the last two decades, the DINA model has been a very commonly used reduced CDM. This model classifies examinees into two groups, those who do have and who do not have all the required attributes. In other words, missing any one of the required attributes is the same as missing all of them. However, this restriction may be too strict under certain situations. de la Torre (2011) proposed the generalized DINA (G-DINA) model where examinees are classified into 2^{K_j} latent groups, and K_j is the number of the required attributes for item j (i.e., $K_j = \sum_{k=1}^K q_{jk}$). Therefore, examinees who have mastered different attributes can have different probabilities of correctly answering an item.

Let item j require the first $1, \dots, K_j$ attributes. The reduced attribute vector can be denoted by α_{lj}^* , which represents the columns of the required attributes (i.e., $l = 1, \dots, 2^{K_j}$). $P(X_j = 1 | \alpha_{lj}^*) = P(\alpha_{lj}^*)$ can represent the probability of correctly answering an item j by examinees

with attribute pattern α_{ij}^* . The item response function of the G-DINA model for the identity link is given by

$$P(\alpha_{ij}^*) = \delta_{j0} + \sum_{k=1}^{K_j} \delta_{jk} \alpha_{lk} + \sum_{k'=k+1}^{K_j} \sum_{k=1}^{K_j-1} \delta_{jkk'} \alpha_{lk} \alpha_{lk'} + \dots + \delta_{j12\dots K_j} \prod_{k=1}^{K_j} \alpha_{lk}, \quad (1)$$

where δ_{j0} is the intercept for item j ; δ_{jk} is the main effect of α_k ; $\delta_{jkk'}$ is the interaction effect of α_k and $\alpha_{k'}$; and $\delta_{j12\dots K_j}$ is the interaction effect of $\alpha_1, \dots, \alpha_{K_j}$.

The G-DINA model is a commonly used saturated model that subsumes several reduced CDMs such as the DINA model, the DINO model, the A -CDM, the LLM, and the R-RUM. These reduced models can be obtained from the G-DINA model by applying appropriate parameterization (de la Torre, 2011). For example, after setting all the parameters in Equation (1) to zero, except for δ_{j0} and $\delta_{j12\dots K_j}$, the DINA model can be formulated as,

$$P(\alpha_{ij}^*) = \delta_{j0} + \delta_{j12\dots K_j} \prod_{k=1}^{K_j} \alpha_{lk}. \quad (2)$$

In this present study, the DINA and G-DINA models were employed as reduced and saturated models, respectively. The former model assumes a specific underlying process, whereas, the latter does not.

1.1.2. The Wald Test

The Wald test (Morrison, 1967) has been used in various statistical analyses for decades. In particular, the Wald test in the context of CDMs has been applied to a number of studies (de la Torre, 2011; de la Torre & Lee, 2013; Hou et al., 2014; Ma, Iaconangelo, & de la Torre, 2016; Terzi, 2017). The Wald test for CDM applications was first introduced by de la Torre (2011) to investigate whether the G-DINA model can be replaced by one of the reduced models (i.e., DINA, DINO, or A -CDM). The null hypothesis to test the fit of a reduced model with $p < 2^{K_j}$ parameters can be written as $\mathbf{R}_{jp} \times \mathbf{P}_j = 0$, where $\mathbf{P}_j = \{P(\alpha_{ij}^*)\}$, and \mathbf{R}_{jp} is the $(2^{K_j} - p) \times 2^{K_j}$ restriction matrix. The Wald statistic W_j to test the null hypothesis for item j is computed as

$$W_j = [\mathbf{R}_{jp} \times \mathbf{P}_j]' [\mathbf{R}_{jp} \times \text{Var}(\mathbf{P}_j) \times \mathbf{R}'_{jp}]^{-1} [\mathbf{R}_{jp} \times \mathbf{P}_j], \quad (3)$$

where $\text{Var}(\mathbf{P}_j)$ is the variance-covariance matrix of the item parameters for the saturated model computed from the inverse of the information matrix. Under the null hypothesis for the DINA model (i.e., $p = 2$), the Wald statistic is assumed to be asymptotically χ^2 distributed with $2^{K_j} - p$ degrees of freedom.

Moreover, the Wald test has also been applied at the item level by comparing the fit of a saturated model to the fits of reduced models to come up with the most appropriate CDM (de la Torre & Lee, 2013). They found that the Wald test had excellent power to determine the true underlying model even for small sample sizes, while controlling the Type-I error for large sample sizes with a small number of attributes. The Wald test application in the study of de la Torre and Lee (2013) was extended by Ma et al. (2016), in that the Wald test was evaluated across several popular additive models and was shown that it can identify correct reduced models and improve attribute classifications. Hou et al. (2014) further carried out the Wald test for DIF detection in the context of CDMs, where the Wald test was able to detect both uniform and nonuniform DIF in the DINA model.

1.2. DIF in Cognitively Diagnostic Assessments

In contrast to IRT, DIF for CDMs needs to be redefined because the examinees are provided with the mastery profile of latent discrete attributes instead of locating examinees on the latent continuum. DIF in CDMs can be represented as $\Delta_{j\alpha_l} = P(X_j = 1|\alpha_l)_F - P(X_j = 1|\alpha_l)_R$, where $\Delta_{j\alpha_l}$ denotes DIF in item j for examinees with the attribute mastery profile α_l ; $P(X_j = 1|\alpha_l)_F$ is the success probability on item j for examinees with the attribute mastery profile α_l in the F group; and similarly $P(X_j = 1|\alpha_l)_R$ in the R group. There is no DIF if $\Delta_{j\alpha_l} = 0$ for all attribute mastery profiles.

Because there are two parameters (the slip and guessing parameters) in the DINA model, DIF can be investigated by examining the differences in the slip and guessing parameters between the F and R groups. Item j exhibits DIF if:

$$\Delta_{s_j} = s_{R_j} - s_{F_j} \neq 0, \quad (4)$$

and/or

$$\Delta_{g_j} = g_{R_j} - g_{F_j} \neq 0. \quad (5)$$

For the G-DINA model, each item parameter corresponds to the probability of success on item j for examinees with the reduced attribute vector α_{lj}^* . Thus, DIF in the G-DINA model is the difference in the item parameters between the F and R groups, represented by $\Delta_{j\alpha_{lj}^*} = P(X_j = 1|\alpha_{lj}^*)_F - P(X_j = 1|\alpha_{lj}^*)_R$, where $\Delta_{j\alpha_{lj}^*} \neq 0$ denotes DIF in item j for examinees with the attribute mastery profile α_{lj}^* .

1.2.1. The Wald Test for DIF Analysis

The Wald test detects DIF in the CDM through multivariate hypothesis testing. To detect DIF in the DINA model, the null hypothesis is written as:

$$H_0: \begin{cases} s_{Fj} - s_{Rj} = 0 \\ g_{Fj} - g_{Rj} = 0 \end{cases} \quad (6)$$

The alternative hypothesis is that at least one of the item parameters is different between the F and R groups. There are two steps to implement the Wald test. In the first step, item parameters are calibrated for the F and R groups separately. The first step translates into applying an unconstrained model to the data, where no constraints in the item parameters across the F and R groups are used. The parameter estimates for item j across the two groups are represented as

$$\hat{\beta}_j^* = (\hat{\beta}_{Rj}, \hat{\beta}_{Fj}) = (\hat{g}_{Rj}, \hat{s}_{Rj}, \hat{g}_{Fj}, \hat{s}_{Fj})'. \quad (7)$$

In the second step, the null hypothesis of the equality of item parameters of the F and R groups is tested. The null hypothesis given in Equation (6) can be expressed in terms of the constrained model as follows:

$$H_0: \mathbf{R}_j \cdot \hat{\beta}_j^* = \mathbf{0}, \quad (8)$$

where \mathbf{R}_j is a 2×4 matrix of restrictions, given as follows:

$$\mathbf{R}_j = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}. \tag{9}$$

The Wald statistic W_j to test the null hypothesis is computed as:

$$W_j = [\mathbf{R}_j \times \widehat{\boldsymbol{\beta}}_j^*]' [\mathbf{R}_j \times \text{Var}(\widehat{\boldsymbol{\beta}}_j^*) \times \mathbf{R}_j']^{-1} [\mathbf{R}_j \times \widehat{\boldsymbol{\beta}}_j^*], \tag{10}$$

where $\text{Var}(\widehat{\boldsymbol{\beta}}_j^*)$ is the variance-covariance matrix of the item parameters, written as:

$$\text{Var}(\widehat{\boldsymbol{\beta}}_j^*) = \begin{pmatrix} \text{Var}(\widehat{\boldsymbol{\beta}}_{Rj}) & 0 \\ 0 & \text{Var}(\widehat{\boldsymbol{\beta}}_{Fj}) \end{pmatrix}, \tag{11}$$

and under the null hypothesis $H_0: \mathbf{R}_j \cdot \boldsymbol{\beta}_j^* = \mathbf{0}$, and W_j is asymptotically χ^2 distributed with two degrees of freedom under the DINA model.

Similarly, in the G-DINA model, the first step of the Wald test is to estimate the item parameters separately for the F and R groups in the form of the vector written as follows:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_j^* &= (\widehat{\boldsymbol{\beta}}_{Rj}, \widehat{\boldsymbol{\beta}}_{Fj})' \\ &= (\widehat{P}(\alpha_{0j}^*)_R, \dots, \widehat{P}(\alpha_{1j}^*)_R, \dots, \widehat{P}(\alpha_{1j}^*)_F, \dots, \widehat{P}(\alpha_{1j}^*)_F, \dots, \widehat{P}(\alpha_{1j}^*)_F)'. \end{aligned} \tag{12}$$

In the second step, the null hypothesis of the equality of item parameters of the F and R groups is tested, as in $H_0: \mathbf{R}_j \cdot \boldsymbol{\beta}_j^* = \mathbf{0}$. Since there are 2^{K_j} parameters to be estimated for each group, there are 2^{K_j} constraints and the dimension of the restriction matrix \mathbf{R}_j is $2^{K_j} \times 2^{K_j+1}$. For example, for an item requiring two attributes for a correct response ($K_j = 2$), \mathbf{R}_j is given as:

$$\mathbf{R}_j = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}. \tag{13}$$

Under $H_0: \mathbf{R}_j \cdot \boldsymbol{\beta}_j^* = \mathbf{0}$, the Wald statistic W_j in this example is assumed to be asymptotically χ^2 distributed with four degrees of freedom. Similar to the use of the Wald test for DIF detection in the DINA model, it only requires the estimation of the unconstrained model, that is, the item parameters are calibrated for the F and R groups separately.

In the G-DINA model, the Wald test can also be used to detect DIF when the underlying restricted model is specified (e.g. DINA model). It is carried out the same way as it is in the G-DINA model, but the restriction matrix \mathbf{R}_j is structured differently, depending on which restricted model is assumed. For example, when the DINA model is assumed as the underlying restricted model, there are $2^{K_j+1} - 2$ constraints and the dimension of the restriction matrix \mathbf{R}_j is $(2^{K_j+1} - 2) \times 2^{K_j+1}$. For an item requiring two attributes for a correct response ($K_j = 2$), \mathbf{R}_j is given as:

$$\mathbf{R}_j = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}. \tag{14}$$

Under the null hypothesis $H_0: \mathbf{R}_j \cdot \boldsymbol{\beta}_j^* = \mathbf{0}$, W_j is assumed to be asymptotically χ^2 distributed with $2^{K_j+1} - 2 = 6$ degrees of freedom. It should be noted that the Wald test for comparing the reduced and saturated models only requires estimation of the saturated model. That is, finding $\hat{\boldsymbol{\beta}}_{Rj}$, $\hat{\boldsymbol{\beta}}_{Fj}$, variance-covariance matrices of $\hat{\boldsymbol{\beta}}_{Rj}$, $\hat{\boldsymbol{\beta}}_{Fj}$, and \mathbf{R}_j is sufficient to implement the Wald test. The implementation of the Wald test for DIF analysis rests on an important property of the chosen CDM that its item parameters are absolutely invariant. When the model reasonably fits the data, one can expect the chosen CDM to yield relatively invariant item parameter estimates.

2. METHOD

2.1. Real Data Application: Proportional Reasoning Data

The Proportional Reasoning (PR) data consist of responses of 301 students from the reference (R) group and 506 students from the focal (F) group. The Q-matrix for the PR test is given in Table 1. In estimating both the DINA model and G-DINA model parameters, the MMLE algorithm written in Ox (Doornik, 2002) was implemented with a convergence criterion of 0.001. DIF analyses were conducted using the Wald test in conjunction with the DINA model, with the G-DINA model where the underlying restricted model was not specified, and with the G-DINA model where the underlying DINA model was assumed.

Table 1. Q-matrix for the PR Data

Item	α_1	α_2	α_3	α_4	α_5	α_6	Item	α_1	α_2	α_3	α_4	α_5	α_6
1	1	0	1	0	0	1	17	1	1	0	0	0	0
2	0	1	1	0	0	0	18	1	0	1	0	1	1
3	1	0	1	1	1	0	19	1	0	1	1	1	0
4	1	1	1	0	0	0	20	0	0	1	1	1	0
5	1	1	1	0	0	0	21	1	0	1	0	0	0
6	0	0	1	0	0	0	22	1	1	0	0	0	0
7	1	0	0	0	0	1	23	1	0	1	0	0	0
8	0	1	1	0	0	0	24	1	0	1	0	1	1
9	0	0	0	1	0	0	25	1	1	1	0	0	0
10	0	1	0	0	0	0	26	0	1	1	0	0	0
11	1	0	0	0	0	0	27	1	0	0	1	1	0
12	0	0	0	0	1	0	28	1	0	1	0	1	1
13	1	0	0	0	1	0	29	1	0	1	0	1	1
14	1	1	1	0	0	0	30	0	0	0	1	0	0
15	0	0	1	0	0	0	31	1	1	0	0	0	0
16	0	0	1	0	0	0							

2.2. Benchmark Simulation for PR Data Analyses

To fully understand the performance of the Wald test for DIF detection in the three procedures and to generate the empirical distribution of the Wald statistics based on the PR data, a benchmark simulation study was performed that mimicked the PR data. In the benchmark study, Type-I error and power of the Wald test were assessed by generating 500 datasets using the estimated values of the item parameters for the R and F groups combined; and the sample sizes for the R and F groups matched those in the real data. The item parameter estimates under the DINA model were used to generate the datasets, provided in Table 2 for the R, F, and the combined groups. It is known that the theoretical power rate of the Wald tests calculated was inflated; that is why, the empirical distributions of the Wald statistic were obtained.

Table 2. Item Parameter Estimates of the DINA Model for R, F, and Combined Groups

Item	$N_R = 301$		$N_F = 506$		$N_T = 807$	
	$s (SE)$	$g (SE)$	$s (SE)$	$g (SE)$	$s (SE)$	$g (SE)$
1	.024 (.012)	.531 (.049)	.190 (.029)	.451 (.031)	.125 (.018)	.469 (.027)
2	.126 (.025)	.670 (.049)	.220 (.036)	.484 (.028)	.169 (.021)	.514 (.025)
3	.015 (.011)	.844 (.034)	.028 (.015)	.745 (.024)	.026 (.009)	.757 (.021)
4	.282 (.034)	.366 (.051)	.400 (.042)	.240 (.024)	.339 (.026)	.253 (.022)
5	.089 (.022)	.568 (.052)	.117 (.029)	.416 (.027)	.127 (.018)	.451 (.025)
6	.064 (.018)	.708 (.054)	.124 (.025)	.384 (.033)	.086 (.015)	.445 (.029)
7	.009 (.017)	.005 (.088)	.201 (.031)	.040 (.044)	.141 (.020)	.054 (.038)
8	.228 (.032)	.447 (.053)	.365 (.041)	.238 (.024)	.291 (.025)	.269 (.022)
9	.246 (.034)	.390 (.077)	.248 (.031)	.229 (.039)	.257 (.022)	.215 (.038)
10	.045 (.014)	.851 (.045)	.017 (.017)	.643 (.028)	.037 (.011)	.695 (.024)
11	.011 (.007)	.972 (.030)	.036 (.011)	.877 (.030)	.024 (.007)	.892 (.025)
12	.518 (.035)	.001 (.107)	.600 (.035)	.222 (.043)	.584 (.024)	.228 (.037)
13	.601 (.034)	.473 (.073)	.504 (.038)	.275 (.031)	.554 (.025)	.311 (.028)
14	.730 (.033)	.345 (.050)	.656 (.040)	.324 (.026)	.709 (.024)	.341 (.024)
15	.053 (.017)	.656 (.057)	.126 (.026)	.454 (.033)	.084 (.015)	.489 (.029)
16	.075 (.021)	.435 (.062)	.229 (.031)	.397 (.033)	.164 (.019)	.412 (.029)
17	.122 (.024)	.507 (.064)	.218 (.037)	.278 (.025)	.161 (.020)	.313 (.024)
18	.014 (.010)	.908 (.027)	.025 (.013)	.594 (.028)	.016 (.007)	.662 (.023)
19	.144 (.030)	.576 (.049)	.271 (.036)	.231 (.025)	.204 (.022)	.278 (.023)
20	.178 (.034)	.286 (.047)	.459 (.040)	.182 (.023)	.346 (.025)	.195 (.021)
21	.253 (.032)	.363 (.056)	.433 (.036)	.158 (.024)	.333 (.024)	.184 (.022)
22	.168 (.028)	.220 (.062)	.481 (.042)	.081 (.015)	.284 (.025)	.086 (.015)
23	.269 (.032)	.380 (.057)	.546 (.036)	.220 (.027)	.405 (.024)	.245 (.024)
24	.326 (.037)	.490 (.047)	.509 (.039)	.277 (.026)	.401 (.026)	.307 (.023)
25	.142 (.027)	.447 (.053)	.320 (.040)	.353 (.026)	.241 (.023)	.372 (.024)
26	.271 (.033)	.305 (.050)	.507 (.042)	.227 (.024)	.378 (.026)	.230 (.021)
27	.106 (.026)	.567 (.056)	.148 (.030)	.276 (.027)	.130 (.019)	.305 (.026)
28	.114 (.025)	.567 (.046)	.142 (.029)	.298 (.027)	.136 (.019)	.353 (.024)
29	.012 (.009)	.800 (.036)	.108 (.027)	.483 (.029)	.040 (.011)	.533 (.025)
30	.427 (.038)	.179 (.068)	.510 (.033)	.104 (.029)	.494 (.023)	.101 (.029)
31	.201 (.029)	.566 (.062)	.274 (.039)	.263 (.025)	.224 (.023)	.306 (.024)

3. RESULT

Results are reported in this section of the paper. In the first part, preliminary results based on PR data were discussed. In the next part, results of a benchmark simulation study to mimick the PR data were presented.

3.1. Preliminary Results: PR Data Analyses

The first Wald test was conducted with the item parameters calibrated along with the restriction matrix formulated in the DINA model. The second Wald test was conducted with the item parameters calibrated along with the restriction matrix formulated in the G-DINA model where no underlying constrained model was specified. The last Wald test was also conducted with the item parameters calibrated by the G-DINA model, but the restriction matrix was formulated in the G-DINA model framework where underlying DINA model was specified.

Table 3. Preliminary DIF Results for PR Data

Item	DINA			G-DINA (No Model Assumed)			G-DINA (DINA Model Assumed)		
	Wald Statistic	p-value	DIF	Wald Statistic	p-value	DIF	Wald Statistic	p-value	DIF
1	30.6	0.000	√	–	–	–	–	–	–
2	16.7	0.000	√	22.5	0.000	√	51.3	0.000	√
3	6.6	0.038	–	10.2	0.856	–	604.2	0.000	√
4	10.7	0.005	–	21.2	0.007	–	40.8	0.000	√
5	7.7	0.022	–	26	0.001	–	169.6	0.000	√
6	32.9	0.000	√	33.8	0.000	√	33.8	0.000	√
7	32.1	0.000	√	–	–	–	–	–	–
8	21.8	0.000	√	20.6	0.000	–	25	0.000	–
9	3.9	0.143	–	11.9	0.003	–	11.9	0.003	–
10	15.9	0.000	–	36.4	0.000	√	36.4	0.000	√
11	9.9	0.007	–	7.7	0.022	–	7.7	0.022	–
12	4.7	0.095	–	20.8	0.000	√	20.8	0.000	√
13	8.4	0.015	–	9.9	0.042	–	42.6	0.000	√
14	2.1	0.349	–	28.6	0.000	–	44.3	0.000	√
15	17.3	0.000	√	16	0.000	–	16	0.000	–
16	18.7	0.000	√	29.7	0.000	√	29.7	0.000	√
17	17.8	0.000	√	19.9	0.001	–	73.4	0.000	√
18	67.6	0.000	√	–	–	–	–	–	–
19	52.4	0.000	√	51	0.000	–	314.5	0.000	√
20	36.4	0.000	√	24	0.002	–	90	0.000	√
21	28.2	0.000	√	21.1	0.000	√	44.4	0.000	√
22	46.2	0.000	√	44.6	0.000	√	62.8	0.000	√
23	43.3	0.000	√	32.1	0.000	√	41.3	0.000	√
24	29.8	0.000	√	–	–	–	–	–	–
25	17.2	0.000	√	27.5	0.001	–	81.1	0.000	√
26	22.8	0.000	√	12.7	0.013	–	18.9	0.004	–
27	25.2	0.000	√	64.5	0.000	√	155.5	0.000	√
28	26.5	0.000	√	–	–	–	–	–	–
29	61.8	0.000	√	–	–	–	–	–	–
30	5.1	0.077	–	2.1	0.343	–	2.1	0.343	–
31	24.3	0.000	√	40.7	0.000	√	49.5	0.000	√

Notes:

1. $\alpha = 0.01=31$ was used as the critical value because the theoretical χ^2 distribution can lead to inated Type-I error.
2. For some of the items, the inverse of the weighted variance-covariance matrix cannot be computed.

Results given in [Table 3](#) showed that the Wald test in the G-DINA model where no underlying constrained model was assumed detected the lowest number of DIF items ($n = 11$), while the Wald test in the DINA model detected the highest number of DIF items ($n = 21$). The Wald test in the G-DINA model with the DINA model in the restriction matrix detected 19 DIF items. The agreement among the three Wald tests calculated based on the kappa coefficient was 0.18.

3.2. Benchmark Simulation Study

For the Wald test to adhere well to the nominal significance level ($\alpha = 0.05$), the observed Type-I error should be within the range of (0.04, 0.06) based on the exact binomial distribution where the standard error of p was computed as $[p(1 - p)/n]^{1/2}$. Additionally, the critical values of the empirical distributions of the Wald statistics were used to calculate the empirical power of the Wald tests in the benchmark power study and to determine the significance of DIF detection in this dataset. A cutoff of 0.80 indicates excellent power and moderate power between 0.70 and 0.80 (Cohen, 1992).

[Table 4](#) summarizes the results of the benchmark simulation. The Wald test to detect DIF in the DINA model adhered well to the nominal significance level for six items (3, 5, 11, 14, 18, and 29). The observed Type-I error were slightly inflated (within the range of [0.06, 0.10]) for eight items (1, 2, 4, 19, 23, 26, 28, and 31). For the most of the other items, the observed Type-I error were largely inflated. For the most of the items, the Wald test had moderate to excellent power. However, for items 1, 7, 9, 12, 14, 16, 20, 25, and 30, empirical power was inadequate. The observed Type-I error were largely inflated to detect DIF in the G-DINA model. For some of the items, the inverse of the weighted variance-covariance matrix cannot be computed therefore the Wald statistic cannot be acquired, noted as “N/A” in the table. For 10 items (2, 6, 8, 11, 19, 21, 22, 23, 27, and 31), the Wald test had moderate to excellent power when it was used to detect DIF in the G-DINA model. While for the Wald test to detect DIF in the G-DINA model with the DINA model assumed as the underlying restricted model, it had moderate to excellent power only for four items (6, 11, 22, and 31). Because of the highly inflated Type-I error and low power in the G-DINA model, the Wald test in the DINA model was selected to detect DIF in the PR data.

[Table 5](#) presents empirical DIF analysis results on the PR data. Critical values of the empirical distributions were used to determine if an item has DIF. As can be seen in [Table 5](#), most of the items showed DIF except for items 9, 12, 13, 14, and 30 in the DINA model. Most of the DIF items in the PR data were also identified as displaying moderate to excellent power, except for items 1, 7, 16, 20, and 25. Hence, one can be sure that these items are DIF items. Among the five non-DIF items in the PR data, only one item 13 displayed excellent power, therefore this item is a non-DIF item. For those nine items displaying poor power, one has to be cautious in interpreting DIF in these items. It is possible that some of these items are DIF items but are not identified as such because the Wald test for DIF detection in the DINA model is not sensitive enough given the characteristics of the data. One of the reasons could be the small sample size. The other reasons including the items with low discriminating power and small DIF sizes also contribute to the low power.

Table 4. Benchmark Simulation Study Results

Item	DINA		G-DINA (No Model Assumed)		G-DINA (DINA Model Assumed)	
	Type-I Error	Empirical Power	Type-I Error	Empirical Power	Type-I Error	Empirical Power
1	0.07	0.54	N/A	N/A	N/A	N/A
2	0.10	0.97	0.40	0.71	0.54	0.61
3	0.02	0.90	0.73	0.06	0.96	0.12
4	0.09	0.84	0.64	0.28	0.87	0.28
5	0.05	0.93	0.72	0.21	0.90	0.15
6	0.18	0.98	0.29	0.90	0.29	0.90
7	0.27	0.04	N/A	N/A	N/A	N/A
8	0.11	0.98	0.45	0.71	0.61	0.55
9	0.53	0.15	0.61	0.10	0.61	0.10
10	0.15	0.74	0.25	0.67	0.25	0.67
11	0.06	0.89	0.11	0.79	0.11	0.79
12	0.30	0.16	0.47	0.16	0.47	0.16
13	0.17	0.85	0.57	0.35	0.75	0.21
14	0.06	0.14	0.72	0.08	0.88	0.11
15	0.17	0.83	0.29	0.60	0.29	0.60
16	0.13	0.68	0.22	0.36	0.22	0.36
17	0.11	0.98	0.53	0.64	0.71	0.47
18	0.02	1.00	N/A	N/A	N/A	N/A
19	0.10	1.00	0.81	0.94	0.99	0.12
20	0.21	0.65	0.74	0.32	0.94	0.21
21	0.14	0.99	0.43	0.75	0.60	0.60
22	0.16	0.95	0.41	0.82	0.58	0.79
23	0.10	0.99	0.38	0.79	0.54	0.57
24	0.11	1.00	N/A	N/A	N/A	N/A
25	0.11	0.67	0.74	0.14	0.91	0.12
26	0.10	0.75	0.45	0.29	0.58	0.20
27	0.15	1.00	0.85	0.74	0.97	0.66
28	0.07	1.00	N/A	N/A	N/A	N/A
29	0.04	1.00	N/A	N/A	N/A	N/A
30	0.47	0.15	0.55	0.15	0.55	0.15
31	0.10	1.00	0.52	0.89	0.68	0.76

Table 5. Empirical DIF Results for PR Data

Item	DINA			G-DINA (No Model Assumed)			G-DINA (DINA Model Assumed)		
	Wald Statistic	DIF	Power	Wald Statistic	DIF	Power	Wald Statistic	DIF	Power
1	30.60	√	–	–	–	–	–	–	–
2	16.70	√	√	22.50	–	√	51.30	√	–
3	6.60	√	√	10.20	–	–	604.20	√	–
4	10.70	√	√	21.20	–	–	40.80	–	–
5	7.70	√	√	26.00	–	–	169.60	√	–
6	32.90	√	√	33.80	√	√	33.80	√	√
7	32.10	√	–	–	–	–	–	–	–
8	21.80	√	√	20.60	–	√	25.00	–	–
9	3.90	–	–	11.90	–	–	11.90	–	–
10	15.90	√	√	36.40	√	–	36.40	√	–
11	9.90	√	√	7.70	–	√	7.70	–	√
12	4.70	–	–	20.80	–	–	20.80	–	–
13	8.40	–	√	9.90	–	–	42.60	–	–
14	2.10	–	–	28.60	–	–	44.30	–	–
15	17.30	√	√	16.00	√	–	16.00	√	–
16	18.70	√	–	29.70	√	–	29.70	√	–
17	17.80	√	√	19.90	–	–	73.40	√	–
18	67.60	√	√	–	–	–	–	–	–
19	52.40	√	√	51.00	–	√	314.50	√	–
20	36.40	√	–	24.00	–	–	90.00	√	–
21	28.20	√	√	21.10	–	√	44.40	√	–
22	46.20	√	√	44.60	√	√	62.80	√	√
23	43.30	√	√	32.10	√	√	41.30	√	–
24	29.80	√	√	–	–	–	–	–	–
25	17.20	√	–	27.50	–	–	81.10	–	–
26	22.80	√	√	12.70	–	–	18.90	–	–
27	25.20	√	√	64.50	–	√	155.50	√	–
28	26.50	√	√	–	–	–	–	–	–
29	61.80	√	√	–	–	–	–	–	–
30	5.10	–	–	2.10	–	–	2.10	–	–
31	24.30	√	√	40.70	√	√	49.50	–	√

Notes:

1. Power with √ indicates moderate to excellent power, above 0.70.
2. For some of the items, the inverse of the weighted variance-covariance matrix cannot be computed.

There were six attributes in the model. Table 6 lists the estimates of the attribute prevalence for the R and F groups. Among the six listed attributes, Attribute 1 was the easiest one to master for the R group and Attribute 6 was the easiest one for the F group. Attribute 3 was the most difficult one to master for the R group and Attribute 2 was the most difficult one for the F group. Overall, the R group has a higher prevalence of mastering each attribute.

Table 6. *Attribute Prevalence Estimates for the Comparison Groups*

Item	Posterior Probability	
	R	F
1	0.889	0.710
2	0.765	0.368
3	0.725	0.476
4	0.744	0.571
5	0.841	0.596
6	0.802	0.755

4. DISCUSSION and CONCLUSION

Designing assessments in CDMs for diagnostic purposes depends on assurance that the methodological advancement is needed for their analysis and commonly use. The invariance of item parameters for various groups of interest should be checked to assure the appropriate use of CDMs. In this sense, DIF analysis is critical for test validation to investigate whether the groups identified ahead of time influence test inference. This study presents the Wald test to detect DIF in different CDM contexts, including the Wald test in the DINA model, in the G-DINA model where the underlying restricted model was not specified, and in the G-DINA model where the underlying DINA model was assumed. For these purposes, low- versus high-performing school districts based on the Proportional Reasoning test were examined for DIF analyses.

From the preliminary DIF detection results, 11 items were identified as DIF items when the Wald test was used in the G-DINA model; 21 items were identified as DIF items in the DINA model; and 19 items were identified as DIF items when the Wald test was used with the saturated G-DINA model but with the DINA model in the restriction matrix. The kappa coefficient of 0.18 indicated a low agreement among the three Wald tests in determining which items were flagged as DIF items.

In addition to the preliminary DIF analyses, a simulation study was implemented to serve as the benchmark to assess the Type-I error and power of the three Wald tests. The Wald test in the DINA model showed a better performance of detecting DIF than the other two tests in terms of the lower Type-I error and more adequate power overall. Based on the empirical DIF results, the Wald test in the DINA model had moderate to excellent power on 22 items. However, the Wald test in the G-DINA model had moderate to excellent power on 10 items; and the Wald test in the G-DINA model where the DINA model was assumed in the restriction matrix had acceptable power only on four items. Because the proposed Wald tests are based on item parameter estimation, the poor performance of the Wald test in the G-DINA model may relate to the small sample size of the real data in application.

Adding to previous studies of using the Wald test to detect DIF in the DINA model, this study explored different ways of constructing the Wald tests in various CDM context and compared the performance of the Wald tests to detect DIF in each of the three scenarios described above. It also discussed how to implement a benchmark simulation study to assess the Type-I error and power of the Wald test applied to real data. Although the proposed Wald tests in the G-

DINA model framework is not as good as the one in the DINA model given the small sample size of the real data, it provides a different way of constructing the test for DIF detection in a more general theoretical framework and can be used to different data application in the future.

Acknowledgements

An early draft of this paper was presented at the annual meeting of National Council on Measurement in Education, San Antonio, TX, USA. (2017, April).


Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Likun Hou  <https://orcid.org/0000-0002-1381-8907>

Ragip Terzi  <https://orcid.org/0000-0003-3976-5054>

Jimmy de la Torre  <https://orcid.org/0000-0002-0893-3863>

5. REFERENCES

- Camilli, G. (2006). Test fairness. *Educational Measurement*, 4, 221-256.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179-199.
- de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.
- de la Torre, J., & Lee, Y. S. (2013). Evaluating the wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, 50, 355-373.
- Doornik, J. A. (2002). *An object-oriented matrix programming using Ox (Version 3.1) [Computer software]*. London, UK: Timberlake Consultants Press.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 301-321.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (p. 129–145). Hilldale, NJ: Lawrence Earlbaum Associates.
- Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, 51, 98-125.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258-272.
- Li, F. (2008). *A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning* (Doctoral dissertation). University of Georgia, Athens, GA.
- Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, 40, 200-217.
- Milewski, G. B., & Baron, P. A. (2002). *Extending DIF methods to inform aggregate reports on cognitive skills*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Morrison, D. F. (1967). *Multivariate statistical methods*. New York, NY: McGraw-Hill.

- Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/dif from group ability differences and detects test bias/dif as well as item bias/dif. *Psychometrika*, 58, 159-194.
- Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems*. Computer-based Education Research Laboratory, University of Illinois.
- Terzi, R. (2017). *New Q-matrix validation procedures* (Doctoral dissertation). Rutgers, The State University of New Jersey, New Brunswick, NJ.
- Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, 26, 237-255.
- Zhang, W. (2006). *Detecting differential item functioning using the DINA model* (Doctoral dissertation). University of North Carolina at Greensboro, Greensboro, NC.
- Zumbo, B. D. (2007). Three generations of dif analyses: Considering where it has been where it is now, and where it is going. *Language Assessment Quarterly*, 4, 223-233.

The Psychometric Properties of School Belonging Scale for Middle School Students

Bekir Direkci^{1,*}, Mehmet Canbulat¹, Ibrahim Hakki Tezci¹, Serdar Akbulut¹

¹Akdeniz University, Education Faculty, Antalya / Turkey

ARTICLE HISTORY

Received: Nov 04 2019

Revised: Mar 03 2020

Accepted: Apr 15 2020

KEYWORDS

Scale development,
School belonging,
Middle school students,
Factor analysis,

Abstract: This study aimed at developing a valid and reliable scale to determine middle school students' sense of school belonging. In this respect, the relevant literature on the concept of belonging was reviewed, interviews were conducted with field experts and middle school students to determine items to be included in the scale. An item pool was created based on the findings of these processes. Later, a pilot form was prepared by taking the opinions of 2 field and 2 measurement and evaluation experts so as to ensure that the scale items represent the structure measured. This form was administered to 287 middle school students studying in the 2018-2019 academic year, and the final scale obtained as a result of exploratory factor analysis was applied to 568 middle school students in a different school. For validity evidence, exploratory factor analysis (EFA), confirmatory factor analysis (CFA) and hypothesis test findings; for reliability, findings of Cronbach Alpha and composite reliability coefficients were used. According to the exploratory factor analysis, the scale consisted of 4 factors with 23 items, and the total variance explained was 63.88%. As a result of the second-order confirmatory factor analysis of the obtained structure, the fit indices of the unidimensional model showed that the model was verified. The internal consistency coefficient of the developed model was $\alpha = .92$ and the composite reliability coefficient was .97. These findings showed that the scale had psychometric properties that could be used in future research.

1. INTRODUCTION

The education system, one of the most significant indicators of the countries' development levels, comprises a systematic process. Many sub-elements such as students, teachers, parents, school administrations, course materials, school infrastructure and so on are administered properly and regularly by the Ministry of National Education and its affiliated institutions so that children as the future of the countries can involve in a contemporary educational process. Although there are partial differences between the educational timetables across countries, this process continues in a similar manner in each country. As in many other industrialized countries, children start school to carry out educational activities from an early age and spend at least 30 weeks of a year in formal education institutions starting from early childhood in

CONTACT: Bekir Direkci ✉ bdirekci@akdeniz.edu.tr 📍 Akdeniz University, Education Faculty, Turkish Language Teaching Dept., Antalya/Turkey

ISSN-e: 2148-7456 /© IJATE 2020

Turkey. During the 12-year compulsory education process designed as 4 + 4 + 4, children attending school receive training in many different disciplines according to their age and developmental characteristics. For the Turkish context, this process is carried out in two different ways as normal (full-day) and half-day (dual) education and it is planned to start normal (full-day) education in all schools by the end of 2019. The full-day education is applied in all OECD countries, with an average of 7-8 hours of schooling. However, children in Turkey attend additional support and training courses at school right after their compulsory courses are over. Therefore, the time spent by some children with schoolmates and teachers on an ordinary school day may even exceed that of their parents (Cemalcılar, 2010). When we consider the length of duration taken into consideration, we may state that it is important for children to feel happy during their time in school and to see school as a second home for their social, academic and cognitive development. The fact that the children love the school, feel the sense of belonging and have a positive attitude towards it have a positive effect on achieving the anticipated objectives of the curriculum and increasing academic success. Now that the students' sense of school belonging is boosted, the potential negative perspectives and attitudes towards the school are thought to disappear since the concept of belonging is a multidimensional structure that we encounter in every aspect of our lives with a different form. It is sometimes attributed to an institution such as a family and school, an individual or a community, and an area or place within the scope of the need for a common structure or origin such as religious or ethnic identity (Sarı, 2013). Since school is naturally perceived as a form of society, it is important to discuss and examine the concepts of society within the school. Just as an individual's sense of belonging to social groups and society brings out the feeling of protecting and improving this structure, it is very important for a student to feel himself as a part of the school so as to protect and promote it (Akar Vural, Özelçi, Çengel, & Gömleksiz, 2013). Therefore, the concept of belonging, which is a sociological and psychological term and has an important place in Maslow's hierarchy of needs, is a critical not only for the society but also for the school to achieve curriculum objectives. In this study, researchers discussed the concept of belonging within the context of student-school relationship. When the relevant literature is examined, it is seen that different researchers investigated and presented findings related to its various aspects. Considering the large time period spent by students in the school, the studies were conducted to measure or increase the sense of school belonging. These studies mainly revealed the positive effects of belonging on various psychological, social and academic outputs (Ireson & Hallam, 2005; Osborne & Walker, 2006; Roeser, Midgley & Urdan, 1996). In general, students with higher sense of school belonging were found to be less anxious and isolated, more autonomous and prosocial, more successful and more intrinsically motivated in classes (Cemalcılar, 2010; Finn, 1989; Goodenow & Grady, 1993; Sarı, 2013; Van Ryzin, Gravely & Roseth, 2009; Voelkl, 1997). In addition, it was revealed that these students placed more emphasis on education, participated in-and-out-of-class activities more, had higher self-esteem and higher attendance rates and better relations with teachers and peers, and they were more satisfied with their current situation (Cemalcılar, 2010). On the other hand, lack of sense of belonging was associated with feelings of alienation and loneliness, low academic achievement, negative attitudes towards school, behavioural problems, low school attendance rate, social rejection, isolation and dropout (Edwards & Mullis, 2001; Voelkl, 1997). Moreover, the lack of sense of school belonging was reported to be a strong predictor of loneliness (Hagerty, Williams, Coyne & Early, 1996; Pretty, Andrewes & Collett, 1994).

Especially in the international literature, the importance of the school belonging for students, its development and relationship with other outcomes of education have been the subject of many studies. As noted above, the most prominent finding in these studies was the effect of school belonging on students' academic achievement. The researchers such as Booker (2006); Cemalcılar (2010); Finn (1989); Goodenow (1992) and Osterman (2000) revealed that the sense

of school belonging was positively correlated with high achievement, academic motivation and academic self-efficacy, and showed a high negative relationship with drop-out rate.

Anderman (2002); Hagborg (1994); Isakson and Jarvis (1999) also reached similar findings. They characterised the sense of high school belonging to high academic achievement and supported the positive relationship between school belonging and academic achievement. Bond et al. (2007), who carried out studies with middle school students, found that students' school belonging level promoted their academic achievement and the rate of continuing to further educational stages. Adelabu (2007) and Israelashvili (1997) associated students' school belonging level with their future expectations. In their studies, they revealed that there was a positive relationship between students' sense of school belonging and their future expectations. In other words, they enounced that the students with a high level of school belonging had a more positive perspective towards the future.

Pehlivan (2006), who examined the reasons for the absenteeism of middle school students, specified the reasons as boredom at school, disliking school and lessons, lack of friends' encouragement and expectations about education. He suggested that these reasons were closely related to the sense of school belonging. Booker (2006) stated that school belonging was an important part of a whole because it affected students' school attendance, academic achievement and educational outcomes related to psychological well-being. Goodenow (1992) stated that the inadequacy or low level of school belonging would have to be considered as a decrease in participation in school and lessons and as a result of this, it would be possible to face with low academic achievement and even drop out.

The OECD report (Willms, 2003) of the PISA study, published in 2000 and conducted in 43 countries to examine the 15-year-old student group, states that there is a direct link between the sense of school belonging and students' participation in school activities. In this report, another factor related to school belonging is expressed as "dropping out". As it is underlined in the report, it is suggested that the students who do not develop a sense of school belonging try to create a different channel for belonging necessity, which leads to the emergence of antisocial behaviour models or the outbreak of violence-prone student groups such as school gangs.

As stated in the relevant literature, the sense of school belonging plays a crucial role in educational life of the students. For this reason, several scale development studies were carried out to determine the level of students' sense of school belonging. In the literature, the first study we come across is the study of Goodenow (1993) who developed the Psychological Sense of School Membership (PSSM) Scale through the data of 755 students studying in middle (N=454) and high (N=301) schools. As a result of analyses, a final 18-item scale having .80 internal consistency value. The scale items consisted of the statements which measured subjective and individual perspectives of the students towards the school rather than an objective evaluation. This scale is one of the most frequently used data gathering instruments regarding the school belonging. It was used by many researchers such as Isakson and Jarvis (1999), McMahon et al. (2008) and Sarı (2013) in the relevant field. Based on the findings of Goodenow (1993), the scale had the required psychometric features both in English and Spanish versions. The relevant scale was adapted by Alkan (2015) who confirmed the construct validity and the efficiency of internal consistency.

Like Goodenow (1993), Aslan and Duru (2017) developed a scale to measure students' sense of school belonging, as well. They collected their data from middle school and high school students in order to obtain a practical scale that can be used in the studies carried out both school levels. In the end of the study, they developed a 10-item scale consisting of two sub factors, satisfaction and loneliness. The scale acquired the required psychometric features both in EFA and CFA, and it was brought into use of the researchers.

Malone, Pillow and Osman (2012), on the other hand, focused on general belongingness and developed a scale to measure this phenomenon. The data were collected through online computer-administered surveys and the analyses of EFA and CFA attested the usability of a 12-item scale consisting of two sub-factors, acceptance/Inclusion and lack of rejection/exclusion. The psychometric properties of this scale were also examined by Duru (2015) who confirmed the two-factor structure and highlighted that the scale can be used to measure general belongingness levels of the university students.

When the literature related to the concept of school belonging is examined, it is observed that this concept is of great importance both in students' current educational life and in shaping the future road map. Measuring the students' school belonging with a valid and reliable measurement tool, identifying the ones having low levels of belonging and carrying out studies to increase their belonging to school will make an important contribution on behalf of countries. The school belonging scale, which is intended to be developed within the scope of this research, aims to fill this gap in the literature and to provide a valid and reliable measurement tool for future research. The scales presented above either focused on the concept of general belonging or were developed by collecting data from middle-school and high school students. This scale is, on the other hand, merely focused on the concept of school belonging and collected the data from middle school students. Therefore, it is thought that it will reflect the school belonging levels of the middle school students more precisely.

2. METHOD

2.1. Study Group

The study group consisted of 855 middle school students who were divided into three groups. The first form of the scale was administered to the Group I. After analysing and performing exploratory factor analysis, the final form was administered to Group II. The data obtained from this application were used in second-order confirmatory factor analysis. Later, the scale was applied to Group III to compute the hypothesis test. Descriptive statistics on research groups are presented in [Table 1](#).

Table 1. Descriptive statistics for research groups

Total Data	Raw Data	Analysis Data		<i>n</i>	%	
855	287	218 (Group I)	Gender	Female	113	51.83
				Male	105	48.17
			Class	5	47	21.5
				6	40	18.3
				7	76	34.8
				8	55	25.4
	312	276 (Group II)	Gender	Female	139	50.36
				Male	137	49.64
			Class	5	61	22.10
				6	74	27.89
				7	86	31.88
				8	55	19.92
256	212 (Group III)	Gender	Female	113	53.30	
			Male	99	46.70	
		Class	5	42	19.80	
			6	33	15.60	
			7	78	36.80	
			8	59	27.80	

The groups were determined according to some principles. First of all, it was decided to carry out the study in schools located in Antalya province in order to provide ease of access to data and economic principle. All the data obtained were collected from public schools. Secondly, it was aimed at increasing the generalizability of the study to the middle school students by including the students from all middle school stages. Therefore, EFA and CFA analyses were performed with the data collected from different groups considering the criteria that EFA and CFA cannot be performed with the same groups (Fabrigar, Wegener, Strahan, & MacCallum, 1999). Another validity method was hypothesis testing. The analysis and study groups included in the research are summarized in [Table 2](#).

Table 2. *Study groups and statistical analyses performed in the research*

Study Group	Statistical Analysis	Evidence
Group I	Exploratory Factor Analysis (EFA)	Construct Validity
Group II	Confirmatory Factor Analysis (CFA)	Construct Validity
Group I+II	Item Analysis, Cronbach Alpha, Composite Reliability	Reliability and Item Discrimination
Group III	Hypothesis Test (<i>t</i> -test)	Construct Validity

2.2. Procedure

While forming the items to be included in the school belonging, the literature on the concept of belonging was reviewed and basic knowledge and theories related to this concept were analysed. However, due to the lack of the number of studies merely focusing the concept of school belonging in this field, an item pool was formed only after in-depth interviews were conducted with field experts and focus group interviews were made with middle school students. During the construction of the item pool, it was asked to 2 field and 2 measurement and evaluation experts to reflect the construct to be measured. The field experts were the academicians who worked the concept of belonging and carried out scale development studies.

A five-point Likert-type rating was used for the statements in the scale: Strongly Disagree (1), Disagree (2), Partly Agree (3), Agree (4), and Strongly Agree (5). After the feedback received from expert opinions, necessary changes and arrangements were made in the scale items. The items having clarity and understandability problems were corrected and those of which both experts had a consensus on removing were eliminated. In the last stage, two language experts checked the scale to ensure the suitability of the scale in terms of language. The items in the scale were reviewed and arranged in line with the opinions and ideas of the experts regarding the use of punctuation marks and spelling. The scale was applied to the pre-trial group of 15 students before being applied to the study groups. During the implementation, students' reactions were monitored and it was concluded that the instructions and items prepared for the scale were clearly understood. The data obtained from the preliminary application were not included in the data of the main research groups.

2.3. Data Analysis

Several analyses were performed to reveal the psychometric properties of the measurements after administering the 37-item pilot form to three research groups. First, exploratory factor analysis (EFA) was computed to obtain evidence about the construct of the measurements. Before applying the EFA, it is necessary to examine whether data meet the assumptions of the factor analysis. The sample size is the first step of this analysis (İlhan & Çetin, 2014) There are different views concerning the number of participants that should be included in factor analysis studies. Cattell (1978) suggests that the number of people in the study group should be 3 to 6

times greater than the number of items in the scale for factor analysis and 200 participants are acceptable for factor analysis and 500 individuals are considered as an optimum number. Hair, Anderson and Grablowsky (1979) stated that the number of the study group should be 20 times as many as the number of scale items in the factor analysis. Comrey and Lee (1992) defined the criteria for factor analysis for the number of participants as 100 inadequate, 200 as average, 300 as good, 500 as satisfactory and 1000 as excellent. Ferguson and Cox (1993) stated that 100 participants should be the minimum for the factor analysis. Kline (1994), on the other hand, stated that the number of data could be reduced up to 100 in cases where the number of factors is low and significant, but a sample of 200 people is required for the reliability of the results in more complex structures. When considering various opinions in determining the number of sample for factor analysis, it is seen that there are different criteria. In this respect, it is suggested that each researcher should be able to meet at least two of the mentioned criteria in accordance with the characteristics of the research (Çokluk, Şekercioğlu & Büyüköztürk, 2012). In this study, it can be said that the data of 218 people in the first group is sufficient for factor analysis since it meets multiple criteria stated above. Kaiser-Meyer-Olkin (KMO) Test for Sampling Adequacy and the Bartlett test where normality is tested are other statistical ways to test the assumptions of the factor analysis. Factor analysis can be performed when the KMO value is higher than .60, which is an indication of sufficient sample size, and when the Bartlett test is statistically significant, which indicates that multivariate normality is achieved (Büyüköztürk, 2016).

EFA encompasses various techniques such as maximum likelihood factor analysis, principal component factor analysis and unweighted least-squares analysis. (Tabachnick & Fidell, 2007). Stevens (1996) stated that compared to other techniques, principal component analysis is a factorization technique that should be preferred primarily because it is psychometrically more powerful and mathematically easier to perform, and has more positive effects in dealing with factor uncertainty problems. Due to its features, the principal component analysis was found suitable for use as a factorization technique in this study. It is suggested that oblique rotation methods should be preferred in case of inter-factor relationship (Tabachnick & Fidell, 2007). Since the factors in the relevant model are related ($r > .30$), Promax rotation method, one of the oblique rotation methods, was used. When interpreting the results obtained from EFA, the factor loading of .50 was taken as the cut-off point because it was considered as satisfactory to include an item in the theoretically predicted factor (Awang, 2015). Items below this value were removed from the scale. When interpreting the findings obtained from EFA; the common variance values (h^2) shown in all the factors were also taken into consideration (İlhan & Çetin, 2014). It was stated that if the item h^2 value, the expression of the sum of the squares of the factor loadings that an item showed in all factors, is low, the item should be removed from the scale in the factor analysis (Kalaycı, 2010). In general, when the studies in the literature are examined, it is recommended that the .50 value for the common variance should be taken as a criterion (Thompson, 2004). However, it is often not possible to obtain high common variance values because the field of study is in the social sciences and human behaviour represents various latent structures. Costello and Osborne (2005) stated that taking the .40 value as a criterion for the common variance would be more meaningful and accurate for the social sciences. Tabachnick and Fidell (2007) stated that if a common variance of an item is lower than .20, it indicates that the items measure different situations. When this view is taken into consideration, the criterion for the common factor variance should be taken as .20 at least (Şencan, 2005).

Confirmatory Factor Analysis was used to obtain information about the accuracy of EFA results and to test the data-fit measurement model which was formed as a result of a theoretical basis. If χ^2 value obtained from the CFA findings of the model is significant, it is considered as evidence that the model is not confirmed by the collected data. However, it is important to note

that; The value of χ^2 is sensitive to the increase in the number of sampling and tends to be significant as the number of data in the study increases. In this case, χ^2 value which is not meaningful in practice might be significant in the analysis results due to the sample size (Byrne, 2010; Kline, 2011). For this reason, it is necessary to examine the standardized value obtained by dividing χ^2 to the degree of freedom and the other fit indexes in the literature when deciding whether the model is validated in the study (Hu & Bentler, 1999). Several fit indices are used to demonstrate the adequacy of the model tested in the CFA. In this study, the following indices were examined for CFA: chi-square goodness of fit test, goodness of fit index (GFI), adjusted goodness of fit (AGFI), normed fit index (NFI), non-normed fit index (NNFI), incremental fit index (IFI), comparative fit index (CFI), root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), parsimony normed fit index (PNFI) and parsimony goodness of fit index (PGFI).

Hypothesis testing was applied to provide different validity evidence for the study. According to the literature (Cemalcilar, 2010; Edwards & Mullis, 2001; Voelkl, 1997), it is expected that students with high school attendance will have higher school belonging and students with low school attendance will have low scale scores on the scale. On this basis, it is expected that there will be a difference in school belonging scores in low and high absenteeism groups. For this purpose, the individuals of the third research group were divided into two different groups (low and high absenteeism) and the school belonging level was compared by t-test.

The reliability of the scores obtained from the school belonging scale was calculated by using composite reliability and Cronbach Alpha methods. In order to determine the level of discrimination of the items in the scale, 27% upper-lower group comparisons and item-total correlation were checked. Statistical package programs were used to compute Cronbach Alpha reliability and item analysis.

3. FINDINGS

3.1. Construct validity

EFA, CFA and hypothesis testing were used to test the construct validity of the items in the scale.

3.1.1. Exploratory Factor Analysis (EFA)

KMO value calculated for the adequacy of the sample size was found to be .916. Besides, the Bartlett test, which was computed to check the multivariate normality assumption, was significant ($\chi^2 = 3103.889$, $df = 253$). According to these results, it can be concluded that the data are suitable for factor analysis. As a result of the principal components factor analysis and varimax vertical rotation method in EFA, the four-factor structure explaining 63.88% of the total variance was found to be appropriate for the theoretical basis. The scree plot obtained for determining the number of factors is shown in [Figure 1](#).

The scree plot is a suggested auxiliary graph for determining the number of factors. Compared to determining the factors through eigenvalue, this graph generally provides a more clear-cut picture of the factors and makes it easier to read the structure graphically. When interpreting the obtained graph, the point at which linearity starts is taken as a cut-off point in determining the number of factors. As seen in the graph, the line gains linearity after 4 bars which is interpreted as an indication of the 4-factor structure in the data set.

According to the findings obtained from the EFA, 5 items were removed since their factor loadings were below the acceptable level. 9 items were excluded from the scale since they showed high factor loadings in more than one sub-factor. Accordingly, all of the items on the scale had a factor loading above the pre-determined cut-off point (.50). In addition, it was found that the common factor variances of all items in the scale were .30 and above and met the

required criteria. When the items in the factors and the theoretical basis were taken into consideration, the first factor was named as School Engagement (SE), the second factor was Teacher Support (TS), the third factor was Friend Support (FS) and the fourth factor was Alienation to School (AS). Descriptive statistics related to the factors are presented in Table 3.

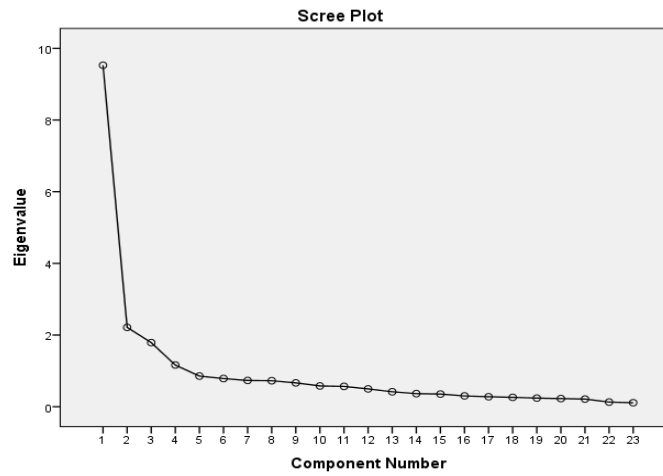


Figure 1. Scree Plot of School Belonging Scale

Table 3. School Belonging Scale Factor Structure and Factor Loadings

Factor	Item	Factor Loadings				Communalities
		Factor 1	Factor 2	Factor 3	Factor 4	
School Engagement	M1	.75	.08	-.03	.04	.57
	M6	.82	-.08	.06	.00	.68
	M7	.87	-.03	-.09	-.16	.79
	M8	.60	.02	.08	.10	.38
	M11	.53	.09	.01	.21	.33
	M12	.74	.04	.01	.06	.56
	M21	.85	.10	-.11	-.09	.75
	M24	.88	-.11	-.01	-.17	.82
	M26	.68	.02	.16	.10	.49
M28	.50	.12	.13	.12	.30	
<i>Variance Explained 41.43 %</i>						
Teacher Support	M17	-.06	.87	.14	-.10	.78
	M31	-.10	.91	.06	-.01	.85
	M33	-.04	.97	-.05	.03	.94
	M34	.14	.64	-.03	.05	.43
	M35	.02	.93	-.03	-.03	.87
	M37	.15	.84	-.13	-.01	.75
<i>Variance Explained 9.62 %</i>						
Friend Support	M4	.02	-.02	.74	-.10	.56
	M25	-.12	.01	.75	.17	.61
	M29	.16	.07	.68	-.09	.50
	M32	.02	-.05	.78	-.05	.61
<i>Variance Explained 7.78 %</i>						
Alienation to School	M10	.16	-.10	-.06	.79	.66
	M19	-.12	-.01	-.04	.74	.57
	M20	-.10	.05	.03	.82	.69
<i>Variance Explained 5.06 %</i>						
TOTAL VARIANCE EXPLAINED						
63.89 %						

3.1.2. Confirmatory Factor Analysis (CFA)

The second-order CFA was applied to test whether the data of the second study group confirm that the structure consisting of 23 items and four factors obtained as a result of EFA is basically a model measuring a single dimension. This analysis provides evidence of whether the structure is unidimensional. The fact that the first structure composed of the items that are compatible with their own sub-factors has sufficient fit indexes is a prerequisite for performing second-order CFA. In the first stage of the study, the second-order CFA was applied as a result of the agreement between the absolute and acceptable level of the CFA goodness of fit indices from the first stream. The fit indices related to the unidimensional model obtained are as follows; $\chi^2/df=2.18$, GFI=.91, AGFI=.85, CFI=.97, NFI=.95, NNFI=.97, IFI=.97, RMSEA=.065, SRMR=.061, PNFI=.85 ve PGFI=.71. In order to reveal the model-data relationship of the structure, the absolute and acceptable values of the fit indices and the fit index values obtained are shown in Table 4. As can be seen in the table, the fitness level of the unidimensional model obtained from the CFA is sufficient and that the model is validated. The obtained model is presented in Figure 2 below.

Table 4. Fit Index Values Obtained in CFA

Fit Indices Examined	Criteria for Absolute Fit	Criteria for Acceptable fit	Fit Indices Obtained	Result
χ^2/df *	$0 \leq \chi^2/df \leq 2,5$	$2,5 \leq \chi^2/df \leq 5$	2.18	Acceptable fit
GFI**	$.95 \leq GFI \leq 1.00$	$.90 \leq GFI \leq .95$.91	Acceptable fit
RMSEA***	$.00 \leq RMSEA \leq .05$	$.05 \leq RMSEA \leq .08$.065	Acceptable fit
AGFI****	$.90 \leq AGFI \leq 1.00$	$.85 \leq AGFI \leq .90$.85	Acceptable fit
IFI**	$.95 \leq IFI \leq 1.00$	$.90 \leq IFI \leq .95$.97	Absolute fit
NFI**	$.95 \leq NFI \leq 1.00$	$.90 \leq NFI \leq .95$.95	Absolute fit
NNFI**	$.95 \leq NNFI \leq 1.00$	$.90 \leq NNFI \leq .95$.97	Absolute fit
SRMR*	$.00 \leq SRMR \leq .05$	$.05 \leq SRMR \leq .10$.061	Acceptable fit
CFI**	$.95 \leq CFI \leq 1.00$	$.90 \leq CFI \leq .95$.97	Absolute fit
PNFI****	$.95 \leq PNFI \leq 1.00$	$.50 \leq PNFI \leq .95$.85	Acceptable fit
PGFI****	$.95 \leq PGFI \leq 1.00$	$.50 \leq PGFI \leq .95$.71	Acceptable fit

*(Kline, 2011) **(Bentler, 1980; Marsh, Hau, Artelt, Baumert & Peschar, 2006) *** (Byrne & Campbell, 1999) ****(Meydan & Şeşen, 2011) ***** (Schermelleh-Engel & Moosbrugger, 2003)

Table 5 shows t-values for the unidimensional model obtained from the second-order CFA. As can be seen in the table, the t-test values were found to be between 10.38 and 13.44 for SE factor, between 10.37 and 14.24 for TS factor, between 5.72 and 6.91 for FS factor, and between 6.03 and 6.20 for AS factor. T value which is greater than 1.96 is an indication of significance at .05 level; but if it is higher than 2.56, it indicates that it is significant at .01 level (Jöreskog & Sörbom, 2000; Kline, 2011). Accordingly, all the t values obtained in the CFA were found to be significant at .01 level. Finally, the t-values obtained from the CFA indicated that the number of the data was sufficient for factor analysis and the model-data fit was validated, so there were no items to be removed from the model.

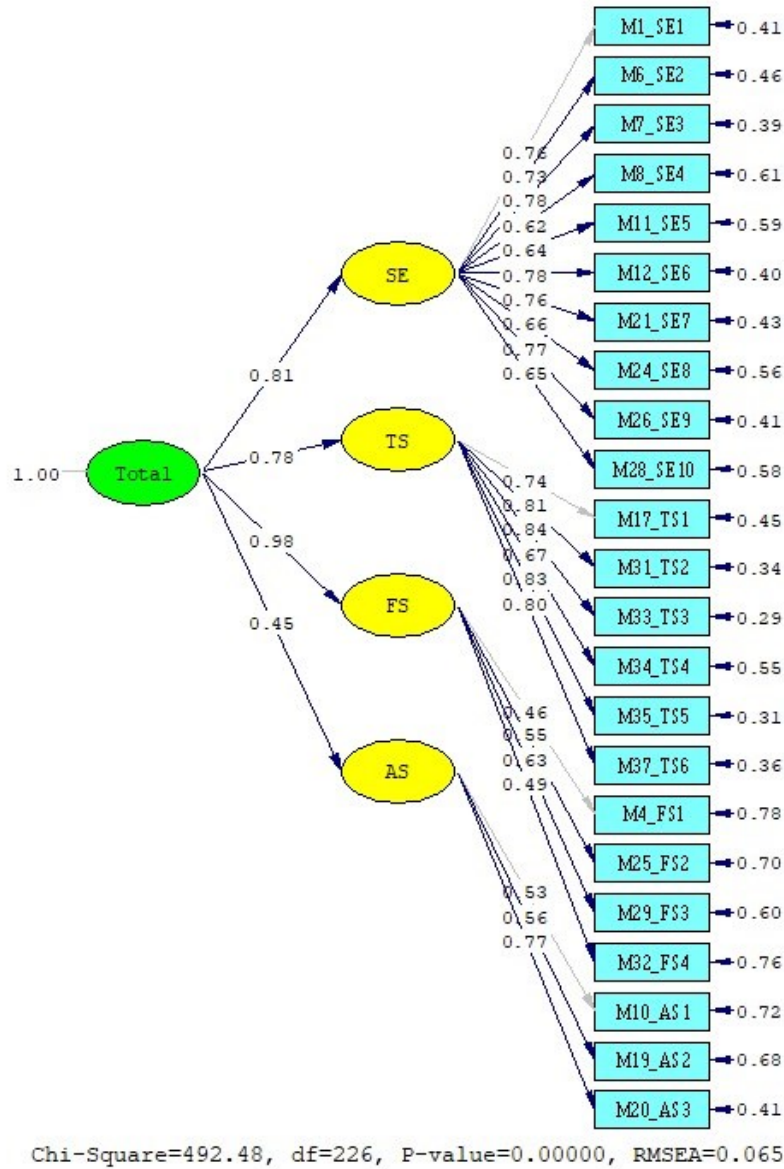


Figure 2. Path Diagram regarding the model

Table 5. t-test Values obtained from CFA for School Belonging Scale for Middle School Students

Item	t value	Item	t value	Item	t value
M1	13.00*	M26	10.90*	M4	5.72*
M6	11.14*	M28	10.75*	M25	6.08*
M7	12.60*	M17	14.24*	M29	6.91*
M8	13.44*	M31	14.00*	M32	6.51*
M11	10.38*	M33	13.65*	M10	6.03*
M12	13.25*	M34	13.36*	M19	6.20*
M21	13.23*	M35	10.37*	M20	6.11*
M24	10.67*	M37	11.11*		

* significant at the .01 level.

3.1.3. Hypothesis Test

When the studies conducted in the literature are examined, there is evidence that the absentee rate is related to the feelings of school belonging. It is expected that students with less absentee rate will have more school belonging than students with more absentee rate. A hypothesis test is an analysis to perform whether the scale reflects such a situation in the literature. In the hypothesis test, the group was divided into two groups as low and high absentee rate. The *t*-test results of the scale scores according to absentee rate are shown in Table 6. The null and alternative hypotheses mentioned in the research are as follows:

H_0 = School belonging scale scores of the students do not differ significantly according to the absentee rate.

H_1 = School belonging scale scores of the students differ significantly according to the absentee rate.

Table 6. Independent Samples *t*-test Results by Gender

Absence	N	\bar{X}	SD	df	<i>t</i> value	<i>p</i>	η^2
Low	106	84.58	18.44	210	3.43	.001	.053
High	106	75.88	18.46				

The scores of the students on the school belonging scale show a significant difference according to the absentee rate, $t(210) = 3.43, p < .05$. Based on the findings, the students with a low absentee rate (84.58) were more positive than the students with the high absentee rate (75.88). The effect size calculated for the difference according to attendance is .053. The significant difference is close to the medium effect size.

3.2. Reliability

Reliability of the scores obtained from school belonging scale was calculated through Cronbach Alpha and Composite Reliability methods. The Cronbach's alpha reliability coefficients of the measures were found to be .91 for school engagement, .92 for teacher support, .72 for friend support and .71 for alienation to school, and .92 for the entire scale. Accordingly, the entire test can be said to be reliable in terms of internal consistency. Composite reliability coefficients of measurements were as follows; .92 for SE factor; .94 for TS factor; .83 for FS factor and .72 for AS factor. The overall reliability of the scale was .97. Since the reliability coefficient of .70 and above is accepted as reliable (Domino & Domino, 2006), it can be said that the reliability coefficients of the scale are sufficient. The results for the reliability analysis are presented in Table 7.

Table 7. Reliability Coefficients for Belonging to School Scale and its Sub-factors

Scale	Cronbach Alpha	Composite Reliability
School Engagement	.91	.92
Teacher Support	.92	.94
Friend Support	.72	.83
Alienation to School	.71	.72
Scale (total)	.93	.97

3.3. Item Analysis

Item analyzes were used for additional evidence of the validity and reliability of the scale. When the findings in Table 8 are examined. It is seen that the *t*-test values of the 27% lower and upper group scores of the school belonging scale items ranged between 3.09 and 12.86 ($p < .01$). In addition, the results of item-total correlations were between .302 and .757. Item-total

correlation provides information about the level of the item discrimination. According to the literature, the items having .30 and above are considered as sufficient for discrimination. Therefore, it can be said that all items in the scale have a value above the cut-off point and therefore all of the items in the scale are distinctive items.

Table 8. Analysis of School Belonging Scale Items

New Item Number	Old Item Number	Reliability if item deleted	Item-Total Correlation	\bar{X}	SD	t value
SE1	M1	.927	.718	3.61	1.21	12.30*
SE2	M6	.927	.683	3.23	1.30	12.01*
SE3	M7	.929	.580	2.79	1.34	9.16*
SE4	M8	.928	.632	3.57	1.20	8.79*
SE5	M11	.928	.628	3.46	1.37	8.57*
SE6	M12	.926	.710	3.35	1.46	11.51*
SE7	M21	.926	.712	2.96	1.50	12.61*
SE8	M24	.929	.575	2.56	1.48	8.93*
SE9	M26	.926	.757	3.51	1.26	12.86*
SE10	M28	.927	.660	3.17	1.40	10.06*
TS1	M17	.927	.657	3.55	1.40	11.13*
TS2	M31	.927	.664	3.65	1.35	10.72*
TS3	M33	.926	.713	3.58	1.35	11.32*
TS4	M34	.928	.625	3.18	1.28	9.65*
TS5	M35	.926	.721	3.53	1.41	12.52*
TS6	M37	.926	.716	3.75	1.33	11.20*
FS1	M4	.932	.401	4.66	.563	4.24*
FS2	M25	.931	.424	3.91	1.16	5.42*
FS3	M29	.929	.552	4.13	1.13	8.13*
FS4	M32	.932	.404	3.75	1.24	5.64*
AS1	M10	.932	.376	4.10	1.13	4.61*
AS2	M19	.935	.302	3.95	1.22	3.09*
AS3	M20	.932	.333	4.26	1.09	3.50*

*significant at the .01 level.

3.4. Interpretation of School Belonging Scale Scores

The school belonging scale consists of 23 items and has a 5-point Likert-type rating. As a result of exploratory factor analysis (EFA), the 23-item scale having 4 sub-factors were obtained. In order to confirm the structure obtained in EFA, confirmatory factor analysis (CFA) was applied and satisfying goodness of fit indices were obtained. Therefore, the second-order CFA analysis were computed whether the scale had unidimensional. The analysis results attested that it was “unidimensional”. This finding can be interpreted that the researchers and educators using this scale can make interpretations based on both sub-factors and total score of the scale. The score range can be between 23-115. The increase in the scores obtained from the school belonging scale can be interpreted as an indication of a higher level of school belonging.

4. DISCUSSION and CONCLUSION

In this study, it was aimed to develop a valid and reliable scale which can be used to measure school belonging level of the middle school students. When the different studies in the national and international literature were examined, it was found that the sense of school belonging had a significant impact on the affective, cognitive and social development characteristics of the students inside and outside the school. According to the studies in the literature, the students with a high sense of school belonging were academically more successful, more willing to study and learn, pro-social, had better teacher-student relationships, felt less lonely and anxious, participated in-and-out-of-class activities more, were more satisfied with their current situation

and highly motivated (Cemalcilar, 2010; Finn, 1989; Goodenow & Grady, 1993; Sari, 2013; Voelkl, 1997). On the other hand, students with a low sense of belonging are characterized by negative attitudes towards the school, behavioural problems, low academic achievement, alienation, and high emotional and low attendance rates (Edwards & Mullis, 2001; Voelkl, 1997). In line with these findings, a valid and reliable measurement tool is required to accurately measure school belonging level. Various measurement tools (Goodenow, 1993; Malone, Pillow & Osman, 2012) are available in the international literature developed for this purpose. However, since they are in different languages, efficient and effective results cannot be obtained due to the linguistic competence factor when they are administered to students. In addition to this, some measurement tools either consider the concept of belonging as a general belonging (Malone, Pillow & Osman, 2012) and do not concentrate on the phenomenon of school belonging (Keskin & Pakdemirli, 2016) or are based on the data of the students studying in two different levels, middle school and high school (Aslan & Duru, 2017). Therefore, the scale entitled with "School Belonging Scale" developed within the scope of this study is considered to be important in filling the gap in national and international literature as a tool which possesses valid and reliable psychometric properties in determining the school belonging level of the middle school students.

In this study, the research data were collected from 855 middle school students who were divided into three groups (Group I=287, Group II=312, Group III=256). Missing values and outliers were eliminated from the total data and the analyses were performed with the rest (Group I=218, Group II=276, Group III=212). The data of the first group were used in the exploratory factor analysis. According to the EFA findings, 5 items were excluded from the scale since their factor loadings were below the pre-determined value of .50 and 9 items were overlapping, so they were removed. As a result, a 23-item scale consisting of four sub-factors were obtained from the 37-item initial version. Considering the items in the factors and the theoretical basis, the first factor was named as School Engagement (SE), the second factor was Teacher Support (TS), the third factor was Friend Support (FS) and the fourth factor was Alienation to School (AS). The first factor (SE) consisted of 10 items whose factor loadings ranged from .50 to .88. The second one (TS) comprised of 6 items whose factor loadings were between .64 and .97. The third one (FS) composed of 4 items having factor loadings between .68 and .78. The last one (AS) consisted of 3 items having factor loadings between .79 and .82. The total variance explained was 63.89 % and it was the first factor to contribute it in the highest amount (41.34%). The others' degree of contribution were as follows: 9.62%, 7.78% and 5.06%.

The results of the EFA were tested with the second-order confirmatory factor analysis (CFA) to confirm whether the structure consisting of 23 items and 4 factors was essentially verified as a model measuring a dimension. The data of the second group were used to perform this analysis. As a result of the second-order CFA analysis, 11 goodness of fit indices (including χ^2/df value) were examined. Since the fit indices obtained from the analysis were between absolute (IFI, NFI, NNFI and CFI) and acceptable (χ^2/df , GFI, RMSEA, AGFI, SRMR, PNFI and PGFI) values, the model was confirmed. This finding was also observed in the path diagram of the model.

In addition to EFA and CFA, hypothesis testing was also performed. When the researches in the related literature were examined, it was noteworthy that the absentee rate at school was related to students' sense of belonging to the school (Cemalcilar, 2010; Edward & Mullis, 2001; Voelkl, 1997). It was expected that students with low absentee rate would have higher levels of belonging to students than students with high absentee rate. Accordingly, the data were divided into two groups as low and high absenteeism. As a result of the t-test performed to these groups, it was found that the students' school belonging scale scores showed a significant difference

according to their attendance status. This finding showed that the evidence for absentee rate and sense of belonging to the literature was put forward by the developed scale and confirmed the hypothesis of this relationship.

In addition to the evidence of construct validity, Cronbach's alpha and composite reliability coefficients were calculated for the reliability of the school belonging scale. The Cronbach's alpha reliability coefficients of the measures were found to be .91 for school engagement, .92 for teacher support, 0.72 for friend support and 0.71 for alienation to school, and .92 for the entire scale. Accordingly, the entire scale can be said to be reliable in terms of internal consistency. Composite reliability coefficients of measurements were; .92 for SE factor; .94 for TS factor; .83 for FS factor and .72 for AS factor. The overall reliability of the scale was .97. In the literature, the value of the reliability coefficient of .70 and above is accepted as an indication that the measurements are reliable (Domino & Domino, 2006). When this information is considered, it can be said that both Cronbach Alpha and composite reliability coefficients of the scale are sufficient.

It was found that the item-total correlation values of the items were between .302 and .757. As stated before, these values provide information about the level of discrimination of the items in the scale. Considering that the items above .30 have sufficient value in terms of discrimination in the literature, it can be said that each of the items in the scale is discriminative.

The findings of the second-order CFA analysis showed that the 23-item school belonging scale, which had a five-point Likert-type rating, was a unidimensional model although it had a four sub-factor structure. In this respect, a total score can be obtained from the scale and interpretations can be made on this score. The range of scores from the scale varies between 23 and 115. The increase in the scores obtained from the school belonging scale means that the students' level of belonging to the school is high (Appendix, Table A1).

In addition to the strengths listed above of the research, the research has some limitations. These limitations bring some suggestions for future research and researchers. First of all, the data collected within the scope of this research is limited to the students attending middle school (5, 6, 7 and 8th grade). As Bademci (2013) states, reliability findings are considered to be characteristics related to measurements, whereas interpretations made as a result of measurements are accepted as validity characteristics. In this respect, it is necessary to renew the validity and reliability analyses for the data to be collected from different study groups. Another suggestion is for researchers who will conduct research using the school belonging scale. When the literature related to school belonging is examined, it is found that school belonging is related to variables such as academic achievement, number of attendance, student-teacher relationship, attitudes towards school, the participation rate in classroom-outside activities. It is thought that when the researchers collect data by taking these variables into consideration, they may easily make descriptive definitions of their study groups.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Bekir Direkci  <https://orcid.org/0000-0002-6951-8567>

Mehmet Canbulat  <https://orcid.org/0000-0002-1781-2684>

İbrahim Hakkı Tezci  <https://orcid.org/0000-0003-0273-8853>

Serdar Akbulut  <https://orcid.org/0000-0002-5809-1481>

5. REFERENCES

- Adelabu, D.D. (2007). Time perspective and school membership as correlates to academic achievement among African American adolescents. *Adolescence*, 42(167), 525-538.
- Akar-Vural, R., Yılmaz-Özelçi, S., Çengel, M., & Gömleksiz, M. (2013). The development of the “Sense of Belonging to School” Scale. *Eurasian Journal of Educational Research*, 53, 215-230.
- Alkan, N. (2015). Psychological sense of university membership: An adaptation study of the PSSM scale for Turkish university students. *The Journal of Psychology*, 150(4), 431-449.
- Anderman, E.M. (2002). School effects on psychological outcomes during adolescence. *Journal of Educational Psychology*, 94(4), 795–809.
- Awang, Z. (2015). *SEM made simple: a gentle approach to learning Structural Equation Modelling*. Bandar Baru Bangi: MPWS Rich Publication.
- Aslan, G., & Duru, E. (2017). Initial development and validation of the school belongingness scale. *Child Indicators Research*, 10, 1043–1058.
- Bademci, V. (2013). Değerbiçiciler arası (interrater) ölçüm güvenirliğinin Cronbach’ın alfası ile kestirilmesi [Estimation of Interrater Score Reliability by the Cronbach’s Alpha]. *Journal of Industrial Arts Education Faculty*, 30, 55-62.
- Bentler, P.M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, 31, 419-456.
- Bond, L., Butler, H., Thomas, L., Carlin, J., Glover, S., Bowes, G., & Patton, G. (2007). Social and school connectedness in early secondary school as predictor of late teenage substance use, mental health, and academic outcomes. *Journal of Adolescent Health*, 40 (4), 357-366.
- Booker, K.C. (2006). School belonging and the African American adolescent: What do we know and where should we go? *The High School Journal*, 89(4), 1-7.
- Byrne, B.M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications and programming*. New York, NY: Taylor and Francis Group.
- Byrne, B.M., & Campbell, T.L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-Cultural Psychology*, 30, 557–576.
- Büyüköztürk, Ş. (2016). *Sosyal bilimler için veri analizi el kitabı- istatistik, araştırma deseni, SPSS uygulamaları ve yorum* (22. Baskı). Ankara: Pegem Akademi
- Cattell, R.B. (1978) *The scientific use of factor analysis in behavioral and life sciences*. Plenum, New York.
- Cemalcılar, Z. (2010). Schools as socialization contexts: Understanding school factors’ impact on students’ sense of school belonging. *Applied Psychology: An International Review*, 59(2), 243-272.
- Comrey, A.L., & Lee, H.B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Costello, A.B., & Osborne, J.W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, 10(7), 1-9.
- Çokluk, Ö., Şekercioglu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları* [Multivariate Statistics for Social Sciences: SPSS and LISREL Applications]. Ankara: Pegem Academy Publishing.
- Domino, G., & Domino, M.L. (2006). *Psychological testing: An introduction*. Cambridge: Cambridge University Press.
- Duru, E. (2015). Genel aidiyet ölçeğinin psikometrik özellikleri: Geçerlik ve güvenirlik çalışması. *Türk Psikolojik Danışma ve Rehberlik Dergisi*, 5(44), 37-47.

- Edwards, D., & Mullis, F. (2001). Creating a sense of belonging to build safe schools. *Journal of Individual Psychology*, 57(2), 196–203.
- Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272-299.
- Ferguson, E., & Cox, T. (1993). Exploratory factor analysis: A users' guide. *International Journal of Selection and Assessment*, 1(2), 84-94.
- Finn, J. (1989). Withdrawing from school. *Review of Educational Research*, 59, 117-142.
- Goodenow, C. (1992). Strengthening the links between educational psychology and the study of social contexts. *Educational Psychologist*, 27, 177-196.
- Goodenow, C. (1993). The psychological sense of school membership among adolescents: Scale development and educational correlates. *Psychology in the Schools*, 30, 79-90.
- Goodenow, C., & Grady, K. (1993). The relationship of school belonging and friends' values to academic motivation among urban adolescent students, *The Journal of Experimental Education*, 62, 60-71.
- Hagborg, W.J. (1994). An exploration of school membership among middle and high school students. *Journal of Psychological Assessment*, 12, 312-323.
- Hagerty, B.M., Williams, R.A., Coyne, J.C., & Early, M.R. (1996). Sense of belonging and indicators of social and psychological functioning. *Archives of Psychiatric Nursing*, 10(4), 235–244. [http://dx.doi.org/10.1016/S0883-9417\(96\)80029-X](http://dx.doi.org/10.1016/S0883-9417(96)80029-X)
- Hair, J.F., Anderson, R.E., Tatham, R.L., & Grablowsky, B.J. (1979). *Multivariate data analysis*. Tulsa, OK: Pipe Books.
- Hu, L.-t., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55.
- Ireson, J., & Hallam, S. (2005). Pupils' liking for school: Ability grouping, self-concept, and perceptions of teaching. *British Journal of Educational Psychology*, 75(2), 297–311.
- Isakson, K., & Jarvis, P. (1999). The adjustment of adolescents during the transition into high school: A short term longitudinal study. *Journal of Youth and Adolescence*, 28(1), 1-26.
- Israelashvili, M. (1997). School adjustment, school membership, and adolescents' future expectations. *Journal of Adolescence*, 20, 525–535.
- İlhan, M., & Çetin, B. (2014) Sınıf değerlendirme atmosferi ölçeğinin (SDAÖ) geliştirilmesi: geçerlilik ve güvenilirlik çalışması [Development of Classroom Assessment Environment Scale (CAES): Validity and Reliability Study]. *Education and Science*, 39(176),31-50.
- Jöreskog, K., & Sörbom, D. (2000). *LISREL [Computer Software]*. Lincolnwood, IL: Scientific Software, Inc.
- Kalaycı, Ş. (2010). Faktör analizi [Factor Analysis]. Ş. Kalaycı, (Ed.), *SPSS uygulamalı çok değişkenli istatistik teknikleri* [Multivariate Statistical Techniques with SPSS Applications]. Ankara: Asil Publishing. Assessment, 1(2), 84-94.
- Keskin, R., & Pakdemirli, M.N. (2016). Mesleki aidiyet ölçeği: Bir ölçek geliştirme, geçerlilik ve güvenilirlik çalışması. *Uluslararası Sosyal Araştırmalar Dergisi*, 9(43), 2580-2587.
- Kline, R.B. (1994). *An easy guide to factor analysis*. New York: Routledge.
- Kline, R.B. (2011). *Principles and practice of structural equation modeling*. New York: The Guilford Press.
- Malone, G.P., Pillow, D.R., & Osman, A. (2012). The General belongingness scale (GBS): Assessing achieved belongingness. *Personality and Individual Differences*, 52, 311–316.
- Marsh, H.W., Hau, K.T., Artelt, C., Baumert, J., & Peschar, J.L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, 6(4), 311-360.

- McMahon, S.D., Parnes, A.L., Keys, C.B., & Viola, J.J. (2008). School belonging among low-income urban youth with disabilities: Testing a theoretical model. *Psychology in the Schools*, 45(5), 387-401.
- Meydan, C.H., & Şeşen, H. (2011). *Yapısal Eşitlik Modellemesi – AMOS Uygulamaları*, Ankara: Detay Yayıncılık.
- Osborne, J.W., & Walker, C. (2006). Stereotype threat, identification with academics, and withdrawal from school: Why the most successful students of color might be most likely to withdraw. *Educational Psychology*, 26(4), 563–577.
- Osterman, F. K. (2000). Students' need for belonging in the school community. *Review of Educational Research*, 70(3), 323-367.
- Pehlivan, Z. (2006). *Resmi genel liselerde öğrenci devamsızlığı ve buna dönük okul yönetimi politikaları (Ankara ili örneği)* [The Absenteeism at State Secondary Schools and Related School management Policies-ANKARA Case]. Unpublished Dissertation. Ankara: Ankara University Graduate School of Educational Sciences.
- Pretty, G. M., Andrewes, L., & Collett, C. (1994). Exploring adolescents' sense of community and its relationship to loneliness. *Journal of Community Psychology*, 22(4), 346–358
- Roeser, R.W., Midgley, C., & Urdan, T. (1996). Perceptions of the psychological environment and early adolescents' psychological and behavioral functioning in school: The mediating role of goals and belonging. *Journal of Educational Psychology*, 88, 408–422.
- Sarı, M. (2013). Lise Öğrencilerinde Okula Aidiyet Duygusu [Sense of School Belonging Among High School Students]. *Anadolu University Journal of Social Sciences*, 10(1), 147-160.
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research-Online*, 8(2), 23-74.
- Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenilirlik ve geçerlilik* [Reliability and Validity in Social and Behavioral Measurements]. Ankara: Seçkin Publishing.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston, Pearson Education, Inc.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington DC: American Psychological Association
- Van Ryzin, M. J., Gravely, A. A., & Roseth, C. J. (2009). Autonomy, belongingness, and engagement in school as contributors to adolescent psychological well-being. *Journal of Youth and Adolescence*, 38(1), 1–12. <http://dx.doi.org/10.1007/s10964-007-9257-4>
- Voelkl, K.E. (1997). Identification with school. *American Journal of Education*, 105, 294–317.
- Willms, J. (2003). *Student engagement at school: A sense of Belonging and Participation*, Results from PISA 2000, OECD.

6. APPENDIX

Table A1. *School Belonging Scale for Middle School Students*

Previous Item Number	Turkish Version	English Version (Suggested)*
1	Okulumu Severim.	I like my school.
6	Okulda kendimi huzurlu hissederim.	I feel peaceful at school.
7	Okula gelmek için can atarım.	I long to come to school.
8	Okulda kendimi güvende hissederim.	I feel safe at school.
11	Kendimi bu okulun bir parçası olarak görürüm.	I see myself as a part of this school.
12	Bu okulun bir öğrencisi olduğum için kendimi şanslı hissederim.	I feel lucky to be a student of this school.
21	Okul benim ikinci evimdir.	The school is my second home.
24	Okulda daha fazla zaman geçirmek isterim.	I would like to spend more time at school.
26	Okulda kendimi mutlu hissederim.	I feel happy at school.
28	Okula kendimi ait hissederim.	I feel like I belong to school.
17	Öğretmenlerim duygularıma önem verir.	My teachers care about my feelings.
31	Öğretmenlerim düşüncelerimi söylemem konusunda beni destekler.	My teachers support me in expressing my thoughts.
33	Öğretmenlerim düşüncelerimi dinler.	My teachers listen to my thoughts.
34	Okulda öğretmenlerim beni her etkinliğe dâhil eder.	My teachers involve me in every activity at school.
35	Okuldaki öğretmenler fikirlerimize saygı gösterir.	The teachers at school respect our ideas.
37	Okuldaki öğretmenler bize hoşgörülü davranır.	The teachers at school treat us with tolerance.
4	Okulda arkadaşlarımla zaman geçirmekten hoşlanırım.	I like spending time with my friends at school.
25	Okulda arkadaşlarımla arasında kendimi değerli hissederim.	I feel valuable among my friends at school.
29	Okulda arkadaşlarımla etkinlik yapmaktan mutlu olurum.	I feel happy to do activities with my friends at school.
32	Planlarıma okul arkadaşlarımı dâhil ederim.	I involve my friends in my plans.
10	Okulda kendimi dışlanmış hissederim.	I feel left out at school.
19	Okuldaki diğer öğrencilerle birlikteyken kendimi yabancı gibi hissederim.	I feel like a stranger when I am with other students at school.
20	Okulda kendimi yalnız hissederim.	I feel lonely at school.

* The scale was translated into English by two experts in the Department of English Language Teaching. It was then translated back into the original language (Turkish) by different experts. Therefore, the researchers planning to use English version are required to conduct factor analysis and recheck reliability of the scale.

Adaptation of the STEM Value-Expectancy Assessment Scale to Turkish Culture

Arif Aciksoz^{1,*}, Yakup Ozkan², Ilbilge Dokme³

¹ Ministry of National Education, 42060, Konya, Turkey

² Ministry of National Education, 06620, Ankara, Turkey

³ Department of Science Education, Faculty of Education, Gazi University, Ankara, Turkey

ARTICLE HISTORY

Received: Oct 25 2019

Revised: Mar 21 2020

Accepted: Apr 08 2020

KEYWORDS

STEM,
STEM Education
Expectancy-Value
Theory,
Scale Adaptation

Abstract: This study aimed to obtain a measurement tool in Turkish culture to determine the motivation of university students (pre-service teachers) toward STEM based on the expectancy-value theory. For this purpose, the validity and reliability studies of the Turkish version of the STEM Value-Expectancy Assessment Scale developed by Appianing and Van Eck (2018) were conducted. A confirmatory factor analysis (CFA) was undertaken to check the validity of the scale administered to 196 pre-service science teachers selected by purposeful sampling and Cronbach's alpha internal consistency coefficients were examined for the reliability evaluation. One item that showed a tendency to be loaded on two factors in CFA was removed, and the repeated CFA confirmed a good fit for the two-factor structure as in the original scale. In the reliability analysis, the internal consistency coefficients were calculated as .87 for the whole scale, .82 for the perceived value component, and .82 for the expectations of success in STEM careers component. When the validity and reliability results were evaluated together, it was concluded that the adaptation of the scale to Turkish culture was measurement tool that has high validity and reliability that could be administered to prospective teachers.

1. INTRODUCTION

The developments in society and economy increase the need for individuals who research, question, create solutions for problems they encounter, associate the information with daily life and participate in production (Altunel, 2018; Ananiadou & Claro, 2009; Morrison, 2006; Saavedra & Opfer, 2012). These abilities, required by the 21st century, are related to what individuals can do with the information they have and how they apply what they have learned in authentic contexts, they should be considered as an education which is supposed to be integrated to into curriculum, not as "another thing to be taught" (Larson & Miller, 2011). Thus, education systems should be equipped in a way to reveal the interests and abilities of individuals, to provide them to benefit from the new forms of socializing, contribute to

CONTACT: Arif Açıksoz ✉ arifaciksoz@gmail.com 📍 Ministry of National Education, 42060, Konya, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

economic growth actively and to ask questions and conduct researches (Ananiadou & Claro, 2009; MEB, 2017).

Today, STEM (Science, Technology, Engineering, and Mathematics) education, which is considered to be interdisciplinary, has come to the forefront as an innovative approach to education that aims to raise individuals who can meet the challenges of the twenty-first century. In the STEM approach, the education on science, technology, engineering and mathematics is presented in an integrated manner and associated with daily life (Yıldırım, 2018). This approach enables individuals to overcome the challenges of the century. In this process, students first learn about science and mathematics, then acquire an understanding of how these two disciplines work in the fields of technology and engineering, and develop deep technical and personal skills (Bybee, 2010). Additionally, in this day when the economic success gradually depends on the creation and application of knowledge, students possess valuable skills such as the skeptical and delicate analysis of evidences and theories, evidence-based thinking, development of logical arguments and problem solving (West, 2012). Therefore, STEM education is vital in not only supporting the necessary participation in labor force but also succeeding in global competition in the rapidly developing world (Breiner, Harkness, Johnson, & Koehler, 2012).

The reform movements, initiated in the United States to create workforce in STEM fields, have been being implemented with the participation of many educational communities (American Association for the Advancement of Science, National Research Council, National Sanitation Foundation etc.) for more than 20 years (MEB, 2016; Sanders, 2009). However, despite such reforms, serious difficulties are still encountered in STEM education. One of the most important problems is that STEM education cannot motivate students to acquire sufficient knowledge and skills to meet the challenging economic and leadership needs of the century (Hossain & Robinson, 2012). In particular, the inadequacy of teachers in applying STEM causes students to develop a negative attitude toward related fields, and at the beginning of their education, they become convinced that STEM subjects are very difficult to learn or uninteresting. This results in a shift in students' occupational choices from STEM to other fields (National Science Board, 2007; PCAST, 2010).

Increasing and continuing interest in STEM is of paramount importance as the need for monitoring STEM career paths increases day by day (Romine & Sadler, 2016). This requires students being encouraged to move toward careers in STEM fields (Akgündüz et al., 2015). The relationship between students' motivation and STEM career choices is an important issue in promoting STEM careers (Rosenzweig & Wigfield, 2016). Because the motivation of students directly affects their decision whether to enter an education path that will provide access to a career in a STEM field (Chen & Dede, 2011; Wang, 2013).

1.1. Theoretical Framework

Motivation, generally considered as a way of mobilizing individuals, is defined by social scientists as a psychological process that stimulates, guides, and sustains a behavior in a more technical manner (Mitchell, 1982). In other words, it is a great source of power that affects the direction, amount, and continuity of students' behaviors toward their goals (Akbaba, 2006). Academically motivated students set goals for themselves, make plans, and endeavor to realize these goals (Ekeh & Njoku 2014). Believed to be a vital determinant of academic performance and success, motivation is an important factor to take into consideration in education (Joseph, Anikelechi & Marumo, 2019).

Motivation is a meta-concept that includes a number of related concepts, such as participation, persistence, interest, self-efficacy, and self-concept. As a meta-concept, it also involves a wide range of theoretical constructs, including expectancy-value or internal-external and many

related theories concerning self-efficacy, goal, intelligence, choice, and self-determination (Irvine, 2018). Among these examples, the expectancy-value theory is considered to be one of the most important theories related to the nature of achievement motivation (Wigfield, 1994). Consisting of two main components, expectancy beliefs and subjective task value, this theory suggests that individuals' preferences, persistence, and performance can be explained by their beliefs about how well they will perform and how they value the activity (Eccles & Wigfield, 2002; Gråstén, 2016; Wigfield, 1994). In this respect, motivation for the achievement of a task to be performed in a context should be considered as the sum of the value given to this task and the reward expectations related to this task (Tünkler, 2018; Sarisepetçi, 2018).

The expectancy beliefs component of the expectancy-value theory refers to the beliefs of students concerning how well they will perform an activity in the short and long term (Appianing & Van Eck, 2018). Vroom (1964), one of the pioneers of the expectancy theory, defined expectation as a temporary belief that a certain action would result in a specific purpose and emphasized that beliefs might change over time (Onaran, 1981:73). Researchers also argue that this component overlaps with an individual's self-efficacy perception (Appianing & Van Eck, 2018; Irvine, 2018; Wigfield & Eccles 2000). Self-efficacy is the judgment of individual concerning his/her belief in being able to organize and conduct actions necessary to manage possible situations (Bandura, 1995:2). An individual with high self-efficacy is more willing to make greater efforts and work harder in the face of failure and difficulties than a person who doubts his/her abilities (Titrek, Çetin, Kaymak & Kaşıkçı, 2018). An individual's belief in self-efficacy is influenced by indirect experience (observing the experiences of others), verbal persuasion (being verbally motivated by others), mastery experience (achievements and failures), and physiological and affective situations (Tschannen-Moran & McMaster, 2009).

Subjective task value, the second component of the expectancy-value theory, expresses the importance or meaning an individual attribute to a certain task, and in a way, the incentives for performing that task (Gråstén, 2016; Putwain, Nicholson, Pekrun, Becker, & Symes, 2019). Eccles (2005a) defined subjective task value as the quality of a task that contributes to an increase or decrease in the possibility of an individual's choice and suggested that this component was composed of four sub-components: attainment/importance value, intrinsic value, utility value, and relative cost value (Eccles 2005a; Eccles, 2005b; Eccles & Wigfield, 2002; Wigfield & Eccles, 2000) which are explained in detail below.

Attainment value refers to the importance of performing a task or activity or completing a given job for a student (Eccles & Wigfield, 2002; Irvine, 2018; Patridge, Brustad, & Stellino, 2013; Wigfield, 1994). This value is related to the suitability of the given task or activity to the self-identity of the person (Eccles, 2005b). To clarify, an individual who encounters a task or activity that suits his/her identity will tend to perform it in the best way possible by attaching greater personal attention to it. The high importance value held by a student supports their performance despite possible low expectations of success and provides greater participation in course tasks and activities (Putwain et al., 2019).

Intrinsic value is related to the immediate and naturally occurring pleasure (amount of satisfaction) that a person receives or hopes to receive by performing a task or activity (Eccles, 2005b, Patridge et al., 2013). This component, also described as an individual's subjective interest in a subject, is similar to the concept of intrinsic motivation in some respects, but these two concepts are not identical (Eccles, 2005a; Nagy, Trautwein, Baumert, Köller, & Garrett, 2006; Wigfield, 1994). As explained in the self-determination theory, in cases with high intrinsic value, the intrinsic value component may be seen as similar to intrinsic motivation because positive psychological results present as a reward (Meyer, Fleckenstein, & Köller, 2019).

Utility value refers to how much a task is related to an individual's current or future goals, including career goals (Patridge et al., 2013; Wigfield, 1994). If an individual believes that a task is important for his/her life, such as "I need to take extra courses to attend medical school", utility value increases (Harackiewicz, Rozek, Hulleman, & Hyde, 2012). In this component, the status of engaging in a task is not related to the individual's inner desire but his/her willingness to reach the desired final state (Wigfield & Eccles, 2000). Therefore, it is similar to the structure of external motivation. Simple external interventions on utility value, such as parental encouragement of their offspring's academic efforts can influence this value (Harackiewicz et al., 2012).

Relative cost value, the final component of subjective task value, refers to what a person should compromise on (e.g., doing biology homework instead of watching movies) or sacrifice (e.g., effort, time, and pleasure) to complete a task (Appianing ve Van Eck, 2018; Irvine, 2018). This value is a negative component for the motivation of an individual and decreases the value of the task (Tünkler, 2018). Relative cost value is affected by many psychological states related to the performance of a task, such as anxiety or fear of failure, rejection or discrimination by peers, or anger/disappointment of parents (Eccles, 2005a; Patridge et al., 2013).

1.2. Significance of the Study

The students' orientation toward STEM fields and their sustained efforts in these fields are reflected as a whole of their expectations of success and perception of value. STEM, having an important position in today's world, is influenced as much by the expectations and values of teachers in the education field as those of students. Thus, teacher motivation in the field is one of the important factors that affect student motivation because a motivated teacher both encourages students in his/her class to have high expectations and values and promotes the implementation of educational reforms at an advanced level (Yazıcı, 2009). For this reason, it is very important to measure not only students' but also teachers' expectations and perceptions of success (motivation) to provide high-quality STEM education, which is considered a new reform in the educational field.

The literature contains several studies based on the expectancy-value theory. For example, Burak (2014) examined the motivation of students in the musical instrument learning process using a questionnaire while Tünkler (2018) investigated students' expectations and value perceptions of the social studies course using an inventory developed by Eccles and Wigfield (1995) that the author adapted to the social studies context. Sarısepetçi (2018) developed an achievement motivation scale based on Eccles' theory of achievement motivation and adapted it for middle school students. In another study, Barutcu (2017) examined how the workplaces prepared according to the principles of expectation-value theory affect students' writing skills and motivation with action research, which is one of the qualitative research methods. However, to the best of our knowledge, the Turkey literature contains limited studies and no scales taking the expectancy-value theory as a basis to determine university students' expectations and perceived value of STEM education. Thus, it is hoped that this adaptation study will make an important contribution to the literature by introducing a tool to measure the values and expectations of prospective teachers related to STEM.

1.3. Purpose of the Study

This study aimed to implement the adaptation of The Value-Expectancy STEM Assessment Scale (VESAS) to Turkish culture developed by Appianing and Van Eck (2018) to determine the motivation of pre-service science teachers about STEM in higher education.

2. METHOD

This section presents information on the sample, original scale, and process of adaptation of the scale to Turkish culture, administration of the adapted version to the sample, and analysis of the data obtained.

2.1. Sample

In this study, the research population consists of 12435 females and 3851 males, in total 16286 preservice teachers, from the Science Education Department of Faculty of Education at 66 universities as of 2019. Scale has been applied at two universities (Gazi and Çanakkale Onsekiz Mart Universities) chosen randomly in this population. The original population of the scale was consisted of female students who continue to STEM programs or who left the STEM programs after participating at least one semester. For this reason, criterion sampling which is one of the purposeful samplings is used in order for the scale items to work correctly. It is determined as a criterion that preservice teachers who will answer the adapted scale take at least one semester STEM lecture or course. According to this criterion, 196 preservice teachers, including 166 females and 30 males, filled the scale. The scale filling rate of the chosen universities has been calculated as 29.2%. The range of the preservice teachers is shown in [Table 1](#).

Table 1. Frequency table showing the gender distribution of the sample

Gender	Grade	Frequency	Total Frequency	Percentage
Male	1	6	30	15.3
	2	12		
	3	5		
	4	7		
Female	1	22	166	84.7
	2	65		
	3	39		
	4	40		
TOTAL			196	100

As it is seen in [Table 1](#), the reason that the sample largely consists of female students is that the preservice teacher population consists of 76% female and 24% male. In this regard, a rate close to the population has been achieved.

Although there is no definite opinion on the sample size in the confirmatory factor analysis, Kline has stated that it should be 10 times of the number of items take place in the scale. Since the original scale consisted of 15 items, 166 female students formed a sufficient size for analysis. However, since the increase in the sample size affects the fit indexes, it is examined whether there is a significant difference between male and female preservice teachers who filled the scale. As a result of the independent sample t-test, no significant difference has been found ([Table 2](#)).

Table 2. Independent sample t-test results according to the average of scale scores

Group	N	\bar{x}	SD	t	df	p
Famale	166	4.0111	.52291	-1.260	194	.209
Male	30	4.1389	.50901			

Thus, although the original scale was only applied to female students, based on the t-test results no gender discrimination has been observed in this study, and the sample consists of 196 preservice teachers.

2.2. Original Value-Expectancy STEM Assessment Scale

The fact that the students who enroll in STEM programs in the USA have a high rate of leaving these programs and that more than half of the students who leave are the females, create need for the academicians to understand why they leave the programs. Appianing and Van Eck (2018) aimed to develop a valid and reliable scale in order to measure the values and expectations of females regarding their participation, attendance or renounce decisions for STEM programs.

The original scale based on the expectancy-value theory of Eccles, a motivation theory, was developed in 2015 by the same researchers to measure motivations for information and communications technology (ICT) and was adapted from 22-item Likert type VIES - Value Interest Expectancy Scale. The scale, which was adapted to STEM, is applied to two groups of females in universities located in the middle west of the USA. The first group consists of female students who have completed at least one semester in the STEM program or who have stayed in the STEM program, and the second group consists of female students who have been enrolled in the STEM program for at least one semester but left. 356 students (297 students first group, 59 students second group), who complete the online scale delivered through e-mail to randomly chosen 2055 students, form the sample.

The researchers have conducted Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity to determine whether the 22-item Likert type scale applied to 356 students is suitable for factor analysis or not. The KMO value is found to be .96 and the Bartlett test of sphericity is found to be $p < 0.05$, based on these results it is seen that the sample is sufficient for factor analysis test. As a result of factor analysis, a two-factor structure, which consists of 14 items of which factor loads range between .41 and .97 and 8 items of which factor loads range between .48 and .86, has been found. These two factors have explained 62.29% of the variance. Researchers who see the Cronbach Alpha values for the internal consistency of the scale have found that the reliability coefficient of the first factor to be .95 and the second factor to be .90. Since the reliability coefficient of the first factor is over .90, the correlations of the items under the factor have been examined and 7 similar items have been excluded from the scale. It has been observed that the two-factor structure formed as a result of the renewed factor analysis explains 61.49% of the total variance, the factor loads in the first factor range between .51 and .91 and the factor loads in the second factor range between .65 and .88. The Cronbach Alpha value of the "Perceived Value" component made on 15 items, 7 of which are inverted, is found to be .90, while The Cronbach Alpha value of the "Expectation of Success in STEM career" component is found to be .89.

2.2. Adaptation of VESAS to Turkish Culture

Motivation of teachers for a field is a factor that affects the motivation of their students for this field. In this respect, since it is thought that it will contribute to the development of STEM education in our country, the adaption of STEM Value-Expectancy Assessment Scale of Appianing and Van Eck (2018), which is based on expectancy-value theory of Eccles, a motivation theory, has been decided to be adapted. In the adaptation study, the adaptation stages specified by Hambleton and Patsula (1999) have been taken into consideration. First, the researchers have been contacted by e-mail and their permissions have been asked in ethical aspect to in relation to adapt the scale to Turkish culture. After obtaining the permission of the researchers, the adaptation process has started.

Because the basic structure desired to be measured on the scale developed by Appianing and Van Eck (2018) is STEM motivation, the scale was studied with a group of three people with language competencies, two of them have taken STEM education and one of them is a STEM specialist. It has been agreed that this structure is in our culture and that it is perceived jointly

in different cultures. Therewith, the scale items have been then translated from English to Turkish one by one by the same group, considering their cultural characteristics, and the translated materials have been discussed comparatively especially in terms of cultural compatibility. For instance, the statement of “I dislike STEM courses” has been adapted as “I dislike the courses in STEM field” since there is no STEM course directly in our culture.

The final translations of the items were obtained based on consensus and written in both English and Turkish on the scale adaptation form. Then, 10 English teachers were asked to rate the translations from 0 to 5. In addition, below each item was added the question, “How would you translate it?” to gain the contributions of teachers to the adaptation process. The forms were collected, and the mean scores given by the teachers to the translation of the items were calculated (Table 3).

Table 3. *Mean scores of the translated items*

Item number	Mean score	Item number	Mean score
1	5.0	9	5.0
2	4.9	10	4.7
3	4.7	11	4.9
4	4.9	12	4.9
5	4.8	13	4.8
6	4.5	14	4.8
7	4.9	15	4.2
8	4.9		

In scale adaptation studies, a rate of agreement of more than 80% consistency in the translation of items made by different individuals is considered appropriate (Crocker & Algina, 1986, as cited in Hacıömeroğlu & Bulut, 2016). The evaluation revealed that the translation score ranged from 4.2 to 5.0, indicating that the items had been accurately translated considering the original version and there was a high agreement (over 80%) between the translators and reviewers. The Turkish version of the scale, confirmed to be equivalent to the original English scale in terms of language, was proofread by a Turkish language teacher, and its comprehensibility was confirmed through the examination of five doctoral students. According to their feedback, the necessary revision was undertaken, and the final version of the scale was obtained.

2.3. Administration of the Adapted Scale

The final Turkish version of VESAS was administered to pre-service teachers in one lesson, in which there was high student participation and motivation during the academic year according to the academicians that gave the courses. Prior to the administration of the scale, the academicians were informed about STEM and the scale. The students were encouraged to be sincere in their responses to the scale items, and sufficient time was given for them to complete the scale.

2.4. Data Analysis

The data obtained during the adaptation process of the scale were entered into IBM SPSS Statistics program v. 22.0, the reverse items were corrected, and values were assigned to missing data. Then, a confirmatory factor analysis (CFA) was performed to test the construct validity of the scale. In order to verify the two-factor structure determined by the researchers who developed the original scale in the analysis, a path diagram was constructed and its suitability to the factor structure was checked by examining the standardized loads and *t*-values, as well as the fit indices. In addition, recommendations for modification were taken into consideration to increase the fit of the model. In order to determine the reliability of the scale, Cronbach’s alpha internal consistency coefficients of the whole scale and the factors were

analyzed. $p < 0.05$ level was used to evaluate the statistical significance of all analyzes performed in this study.

3. FINDINGS

3.1. Validity of the Adapted Scale

The results obtained from CFA revealed that the standardized loads varied between .42 and .85 and the t-values were significant at the .01 level. Although some of the fit indices obtained from the first analysis showed a good fit (χ^2/df -2.92, SRMR-.080, AGFI-.82, NFI-.91, IFI-.94), others only had values indicating an acceptable fit (RMSEA-.09, NNFI-.92, Since CFI-.93, GFI-.87); therefore, modification recommendations were sought. According to this, high modification was recommended for item 12 under the expectations of success component. The tendency of this item, which was theoretically aimed to measure the perceived value latent variable, to load on the expectations of success component led to a decrease in the fit of the model. Thus, this item was removed from the scale after obtaining expert opinion. In the second CFA analysis conducted following the exclusion of item 12 from the scale, the recommended modification between items 8 and 10 under the expectations of success component was re-checked. This modification was also performed in order to further increase the model fit. After applying the necessary modifications, the final CFA was conducted, and the results of standardized loads and t-values are presented in Table 4.

Table 4. Standardized loads and t-values

Factors		Standardized loads	t-value	p value
Perceived value	M1	.47	6.59	$p < .01$
	M2	.61	8.95	
	M3	.85	14.17	
	M4	.42	5.79	
	M5	.81	13.12	
	M6	.84	13.86	
	M7	.57	8.31	
Expectation of success in a STEM career	M8	.56	14.74	
	M9	.65	9.53	
	M10	.60	8.65	
	M11	.46	6.37	
	M13	.80	12.69	
	M14	.74	11.34	
	M15	.73	11.11	

The standardized factor loads in CFA show how much the latent variable is represented by the observed variable. As shown in Table 3, the standardized factor loads varied between .42 and .85. for the adapted scale. This means that the perceived value component was least represented by item 4 with a factor load of .42 and most represented by item 3 with a factor load of .85. The second component (expectations of success in STEM careers) was represented least by item 11 and most by item 13 with factor loads of .46 and .80, respectively. In addition, non-significant t-values should be excluded from the analysis (Çokluk et al., 2012), but as shown in Table 3, all t-values were significant at the .01 level. The fit indices obtained from the analyses and their critical values are given in Table 5 according to the measures used.

An examination of the data in Table 4 shows that the IFI and CFI values indicated a perfect fit, while the value obtained by dividing the square value (χ^2) by the degree of freedom (df) and the SRMR, NFI and NNFI values were close to a perfect fit. Since the values of all the

remaining fit measures (RMSEA, GFI, and AGFI) were above the critical level, the model was considered to have a good fit.

Table 5. Fit indices and their critical values

Fit measure	Perfect fit	Acceptable fit	Fit values of the research	Conclusion
X ² /df	0 ≤ χ ² /df ≤ 2	2 < χ ² /df ≤ 3	2.1	Acceptable
RMSEA	0 ≤ RMSEA ≤ .05	.05 < RMSEA ≤ .08	.075	Acceptable
CFI	.97 ≤ CFI ≤ 1.00	.95 ≤ CFI < .97	.97	Perfect fit
IFI	.95 ≤ IFI ≤ 1.00	.90 ≤ IFI ≤ .95	.97	Perfect fit
GFI	.95 ≤ GFI ≤ 1.00	.90 ≤ GFI < .95	.90	Acceptable
NFI	.95 ≤ NFI ≤ 1.00	.90 ≤ NFI < .95	.94	Acceptable
AGFI	.90 ≤ AGFI ≤ 1.00	.85 ≤ AGFI < .90	.85	Acceptable
NNFI	.97 ≤ NNFI ≤ 1.00	.95 ≤ NNFI < .97	.96	Acceptable
SRMR	0 ≤ SRMR ≤ .05	.05 < SRMR ≤ .10	.058	Acceptable

(Çapık, 2014; Kline, 2011; Schermelleh-Engel, Moosbrugger, Müller, 2003; Sümer, 2000)

3.2. Reliability of the Adapted Scale

Cronbach’s alpha internal consistency coefficient and reliability analyses of the adapted scale were performed. The internal consistency coefficients, corrected item-total relationship, and the alpha values after correction were analyzed for each factor and the whole scale, and the findings are shown in Table 6.

Table 6. Cronbach’s alpha coefficients of the factors and scale

Item	Cronbach’s Alpha	\bar{X}	S	Corrected Item-Total Correlation	Cronbach’s Alpha Value After Correction	Cronbach’s Alpha of the Scale
Perceived value						
M1	.82	4.61	.68	.462	.81	.87
M2		4.17	.86	.559	.80	
M3		4.19	.75	.735	.77	
M4		4.22	1.04	.409	.83	
M5		4.40	.63	.700	.78	
M6		4.19	.78	.683	.78	
M7		4.35	.76	.554	.80	
Expectation of success in a STEM career						
M8	.82	3.25	.98	.552	.80	
M9		4.21	.78	.582	.80	
M10		3.89	.78	.572	.80	
M11		3.87	1.09	.432	.83	
M13		4.13	.85	.639	.79	
M14		4.24	.73	.617	.79	
M15		4.25	.74	.705	.78	

The internal consistency coefficients were calculated as .82 for both perceived value and expectations of success in STEM careers components (Table 6). The internal consistency coefficient of the whole scale was found to be .87. The values obtained for the two components and the whole scale were greater than .70, would be indicated that the adapted scale was reliable.

Table 6 shows that the reliability coefficients would slightly increase by removing item 4 from the perceived value component and item 11 from the expectations of success in STEM careers component. However, considering that these items did not result in significant changes in the fit indices obtained from CFA and their exclusion would not have led to a significant increase in reliability, it was decided to retain both items in order to maintain as much consistency with the original scale as possible. The results also showed that the corrected-item correlations ranged from .41 to .74 for the perceived value component and from .43 to .71 for the expectations of success in STEM careers. Since the threshold value for the corrected-item total correlations is .30, it can be stated (Büyüköztürk, 2007) that the items under each component adequately measured the desired construct.

4. DISCUSSION and CONCLUSION

In the educational context, the expectancy-value theory stipulates that students' motivation for success and behaviors (preferences) are a function of their beliefs (expectations) about their abilities and perceived importance (value) of a particular task (Wigfield, Tonks, & Klauda, 2009). Thus, the development of interest in a field, including that in a future career, is only possible by increasing the values and expectations of students (Hidi & Renninger, 2006). Therefore, students' expectations and values concerning STEM are important when examining their STEM orientation and choices (Svoboda, Rozek, Hyde, Harackiewicz, & Destin, 2016). When a student's expectations regarding their success in and value of STEM fields are high, it is more likely that he/she would make further efforts in STEM fields and graduate from the related education programs. Otherwise, the opposite can be seen (Appianing & Van Eck, 2018). Therefore this study aimed to adapt VESAS developed by Appianing and Van Eck (2018) to determine individuals' motivation for STEM.

The analysis of the data obtained from the administration of the scale was conducted by CFA. Considering that CFA is a method that enables the validation of a previously formed structure with the available data from a theoretical basis, the factor structure of the adapted scale was found to be adequate for this analysis (Çapık, 2014; Çokluk, Şekercioğlu, & Büyüköztürk, 2012). According to the results of CFA, some items provided a good fit (χ^2/df -2.92, SRMR-.080, NFI-.91, IFI-.94, RMSEA-.09, NNFI-.92, CFI-.93, GFI-.87, AGFI-.82) while others only indicated an acceptable fit; thus, possible modifications were explored. After obtaining expert opinion, item 12 (I feel that I will have something to be proud of as a STEM expert), which tended to load on both components at the same time, was removed from the scale, and CFA was repeated for the remaining 14 items. The results of the second CFA revealed a good fit for all index values (χ^2/df -2.1, RMSEA-.75, CFI-.97, GFI-.90, AGFI-.85, SRMR-.058, IFI-.97, NFI-.94, NNFI-.96) and confirmed that the data obtained from the Turkish version of the scale complied with the theoretical structure of the original tool.

After verifying the construct validity of the scale, Cronbach's alpha internal consistency coefficient was examined for the reliability of each component and the whole scale. The internal consistency coefficients were calculated as .82 for both perceived value and expectations of success in STEM careers components, and .87 for the whole scale. Cronbach's alpha value varies between different disciplines or fields of study in the social sciences, the .70 threshold offered by Nunally (1978) is accepted. Considering that this coefficient exceeded the threshold value of .70 in all calculations, it was concluded that the factors of the scale and the scale itself were reliable as a measurement instrument.

In conclusion, this study may be successfully implemented the adaptation of VESAS to Turkish culture, which aims to determine the individuals' motivations related to STEM, and confirmed the validity and reliability of the adapted version through relevant analyses. The importance of STEM fields is increasing day by day, and considering that students' career choices mostly

depend on their persistence, performance, and motivation in the related fields, it is necessary to measure students' motivation toward STEM (Appianing & Van Eck, 2018). When the studies in Turkey were examined, no scale was found based on the expectancy-value theory to measure STEM motivation. Thus, the adapted scale has an important place as it fills a gap in the literature by acting as a guide for future research about expectancy-value.

STEM is a new educational approach in Turkey; therefore, the shortage of pre-service teachers receiving effective STEM education created a limitation for this study. It is considered that repeating the study with a larger number of pre-service teachers will contribute to the validity and reliability of the this scale. In addition, it is highly recommended that investigation on the motivation of STEM education based on the expectation-value theory should be diversified at different levels (primary and secondary schools).

Acknowledgements

This scale was adapted from the STEM Value-Expectancy Assessment Scale (VESAS) developed by Appianing and Van Eck (2018). The authors would like to thank Appianing and Van Eck for allowing their scale to be adapted.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Arif Açıksöz  <https://orcid.org/0000-0002-6770-3777>

Yakup Özkan  <https://orcid.org/0000-0002-2757-8123>

İlbilge Dökme  <https://orcid.org/0000-0003-0227-6193>

5. REFERENCES

- Akbaba, S. (2006). Eğitimde motivasyon [Motivation in education]. *Kazım Karabekir Eğitim Fakültesi Dergisi*, 13, 343-361.
- Akgündüz, D., Aydeniz, M., Çakmakçı, G., Çavaş, B., Çorlu, M.S., Öner, T., & Özdemir, S. (2015). *STEM eğitimi Türkiye raporu: Günün modası mı yoksa gereksinim mi?* [STEM education Turkey Report: Fashion of the day or need?]. İstanbul Aydın Üniversitesi
- Altunel, M. (2018). STEM eğitimi ve Türkiye: Fırsatlar ve riskler [STEM education and Turkey: Opportunities and risks]. *SETA Perspektif*, 207, 1-7.
- Ananiadou, K., & Claro, M. (2009). 21st Century Skills and Competences for New Millennium Learners in OECD Countries. *OECD Education Working Papers*, No. 41, OECD Publishing.
- Appianing, J., & Van Eck, R.N (2018). Development and validation of the Value-Expectancy STEM Assessment Scale for students in higher education. *International Journal of STEM Education*, 5(24), 1-16.
- Bandura, A. (1995). *Self-Efficacy in Changing Societies*. Cambridge University Press. Retrieved July 19, 2019, from <https://www.researchgate.net/>
- Barutçu, T. (2017). *Beklenti-değer temelli öğretimde yazma becerileri ve motivasyon ilişkisi [The relation between writing skills and motivation in teaching based upon the expectancy-value]* (Doctoral thesis). Available from YÖK National Thesis Center database. (Thesis No: 485941).
- Breiner, J., Harkness, S., Johnson, C., & Koehler, C. (2012). What is STEM? A discussion about conceptions of STEM in education and partnerships. *School Science and Mathematics*, 112(1), 3–11.

- Burak, S. (2014). Motivation for instrument education: A Study with the perspective of expectancy-value and flow theories. *Eurasian Journal of Educational Research*, 55, 123-136.
- Büyüköztürk, Ş. (2007). *Sosyal bilimler için veri analizi el kitabı [Handbook of data analysis for social sciences]*. Ankara: Pegem A Yayıncılık
- Bybee, R. W. (2010). What is STEM education?. *Science*, 329(5995), 996.
- Chen, J.A., & Dede, C.J. (2011). Youth STEM motivation: Immersive Technologies to engage and empower underrepresented students. STEM Learning and Research Center: Retrieved June 27, 2019, from <http://stelar.edc.org/>
- Çapık, C. (2014). Geçerlik ve güvenirlik çalışmalarında doğrulayıcı faktör analizinin kullanımı [Use of confirmatory factor analysis in validity and reliability studies]. *Anadolu Hemşirelik ve Sağlık Bilimleri Dergisi*, 17(3), 196-205.
- Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları [Multivariate statistics for social sciences: SPSS and LISREL applications]*. Ankara: Pegem Akademi.
- Eccles, J.S. (2005). Subjective task value and eccles et al. model of achievement-related choices. In A.J. Elliot & C.S. Dweck (Eds.). *Handbook of competence and motivation*, 105-121.
- Eccles, J.S. (2005). Studying gender and ethnic differences in participation in math, physical science and information technology. *New Directions for Child and Adolescent Development*, 110, 7-14.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109-132.
- Ekeh, P.U., & Njoku, C. (2014). Academic optimism, students' academic motivation and emotional competence in an inclusive school setting. *European Scientific Journal*. 10(19), 127-141.
- Grástén, A. (2016) Children's expectancy beliefs and subjective task values through two years of school-based program and associated links to physical education enjoyment and physical activity. *Journal of Sport and Health Science*, 5(4), 500-509.
- Hacıömeroğlu, G., & Bulut, A.S. (2016). Integrative STEM Teaching Intention Questionnaire: A validity and reliability study of the Turkish form. *Eğitimde Kuram ve Uygulama*, 12(3), 654-669.
- Harackiewicz, J.M., Rozek, C.S., Hulleman, C.S., & Hyde J.S. (2012). Helping Parents to Motivate Adolescents in Mathematics and Science: An Experimental Test of a Utility-Value Intervention. *Psychological Science*, 23(8), 899-906.
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, 41, 111-127.
- Hossain M., & Robinson, M.G. (2012). How to motivate US students to pursue STEM (science, technology, engineering and mathematics) careers. *US-China Educ Rev A*, 4, 442-451.
- Hambleton, R.K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, 1(1), 1-30.
- Irvine, J. (2018). A framework for comparing theories related to motivation in education. *Research in Higher Education Journal*, 35, 1-30.
- Joseph, C.H., Anikelechi, I.G., & Marumo, P. (2019). Academic motivation of school going adolescents: Gender and age difference. *Gender and Behaviour*, 17(1), 12306-12315.
- Kline R.B. (2005). *Principles and Practice of Structural Equation Modeling*. New York: Guilford Press, 154-186.
- Lotta C.L., & Teresa N.M. (2011). 21st Century skills: Prepare students for the future. *Kappa Delta Pi Record*., 47(3), 121-123.

- MEB (2016). *STEM Eğitimi Raporu [STEM education report]*. Ankara
- MEB (2017). *STEM Eğitimi Öğretmen El Kitabı [STEM education teacher handbook]*. Ankara
- Meyer, J., Fleckenstein, J., & Köller, O. (2019). Expectancy value interactions and academic achievement: Differential relationships with achievement measures. *Contemporary Educational Psychology*, 58, 58–74.
- Mitchell T.R. (1982). Motivation: New directions for theory, research, and practice. *Academy of Management Review*, 7(1), 80-88.
- Morrison, J. (2006). *TIES STEM education monograph series, attributes of STEM education*. Baltimore, MD: Teaching Institute for Excellence in STEM. Partner for Public Education: Retrieved July 15, 2019, from <https://www.partnersforpubliced.org/>
- Nagy, G., Trautwein, U., Baumert, J., Köller, O., & Garrett, J. (2006). Gender and Course Selection in Upper Secondary Education: Effects of academic self-concept and intrinsic value. *Educational Research and Evaluation*, 12(4), 323-345.
- National Science Board. (2007). A National action plan for addressing the critical needs of the U.S. science, technology, engineering, and mathematics education system. National Science Foundation. <https://www.nsf.gov/pubs/2007/nsb07114/nsb07114.pdf> (accessed on 16/07/2019)
- Nunnally, J.C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Onaran, O. (1981). *Çalışma yaşamında güdülenme kuramları [Motivation theories in working life]*. Ankara: Ankara Üniversitesi Siyasal Bilgiler Fakültesi Yayınları
- PCAST (2010). *Prepare and Inspire: K-12 Education in Science, Technology, Engineering, and Math (STEM) Education for America's Future Executive Report*. Washington National Science Foundation: Retrieved July 13, 2019, <https://nsf.gov/>
- Partridge, J., Brustad, R., & Stellino, M.B. (2013). Theoretical perspectives: Eccles' expectancy-value theory. *Advances in Sport Psychology*, 3, 269–292.
- Putwain, D.W., Nicholson, L.J., Pekrun, R., Becker, S., & Symes, W. (2019). Expectancy of success, attainment value, engagement, and Achievement: A moderated mediation analysis. *Learning and Instruction*, 60, 117-125.
- Romine, W.L., & Sadler, T.D. (2016). Measuring changes in interest in science and technology at the college level in response to two instructional interventions. *Reserch in Science Education*, 46(3), 309-327.
- Rosenzweig, E.Q., & Wigfield, A. (2016). STEM motivation interventions for adolescents: A promising start but further to go. *Educational Psychologist*, 51(2), 146-163.
- Saavedra, A. R., & Opfer, D. (2012). Learning 21st-century skills requires 21st century teaching. *Phi Delta Kappan*, 94(2), 8-13.
- Sanders, M. (2009). Stem, STem Education, STEMmania. *Technology Teacher*, 68(4), 20-26.
- Sarisepetçi, M. (2018). An adaptation of the success motivation scale based on the expectation-value theory. *International Journal of Education Science and Technology*, 4(1), 28-40.
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive Goodness-Of-Fit Measures. *Methods Of Psychological Research Online*, 8(2), 23-74.
- Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar [Structural equation models: Basic concepts and sample applications]. *Türk Psikoloji Yazıları*, 3(6), 49-74.
- Svoboda, R.C., Rozek C.S., Hyde, J.S., Harackiewicz, J.M., & Destin, M. (2016). Understanding the relationship between parental education and STEM course taking through identity-based and expectancy-value theories of motivation. *AERA Open*, 2(3), 1-13.

- Tschannen-Moran, M., & McMaster, P. (2009). Sources of self-efficacy: Four professional development formats and their relationship to self-efficacy and implementation of a new teaching strategy. *The Elementary School Journal*, 110(2), 228-245.
- Titrek, O., Çetin, C., Kaymak, E., & Kaşıkçı, M. M. (2018). Academic motivation and academic self-efficacy of prospective teachers. *Journal of Education and Training Studies*, 6(11a), 77-87.
- Tünkler, V. (2018). İlköğretim öğrencilerinin sosyal bilgiler dersine yönelik yeterlik beklentileri ve değer algılarının incelenmesi [Examining the adequacy expectations and value perceptions of primary school students towards social studies course]. *Hacettepe University Journal of Education*, 34(4), 1107-1120.
- Wang, X. (2013). Why students choose STEM majors: Motivation, high school learning, and postsecondary context of support. *American Educational Research Journal*, 50(5), 1081–1121.
- West, M. (2012). STEM education and the workplace. Office of the Chief Scientist, Occasional Paper Series, Issue 4. Canberra: Australian Government. Retrieved July 20, 2019, <https://www.chiefscientist.gov.au/>
- Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, 6(1), 49-78.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68–81.
- Wigfield, A., Tonks, S., & Klauda, S. L. (2009). Expectancy-value theory. In K. Wentzel & A. Wigfield (Eds.). *Handbook of motivation at school*, 55–75. New York, NY: Routledge.
- Yazıcı, H. (2009). Öğretmenlik mesleği, motivasyon kaynakları ve temel tutumlar: Kuramsal bir bakış [Teaching profession, sources of motivation and basic attitudes: A theoretical perspective]. *Kastamonu Eğitim Dergisi*, 17(1), 33-46.
- Yıldırım, B. (2018). *Türkiye'nin 2023, 2053 ve 2071 hedefleri için STEM eğitim raporu*. [Turkey's 2023, 2053 and 2071 targets for STEM education report]. Muş Alparslan Üniversitesi. Retrieved June 20, 2019, <https://www.researchgate.net/>

Analyzing Different Module Characteristics in Computer Adaptive Multistage Testing

Melek Gulsah Sahin ^{1,*}

¹Gazi University, Gazi Education Faculty, Department of Educational Sciences, Turkey

ARTICLE HISTORY

Received: Jan 19, 2020

Revised: Mar 6, 2020

Accepted: Apr 19 2020

KEYWORDS

ca-MST,
Panel design,
Module length,
Item discrimination
sequence,
Difficulty difference
condition

Abstract: Computer Adaptive Multistage Testing (ca-MST), which take the advantage of computer technology and adaptive test form, are widely used, and are now a popular issue of assessment and evaluation. This study aims at analyzing the effect of different panel designs, module lengths, and different sequence of a parameter value across stages and change in b parameter range on measurement precision in ca-MST implementations. The study has been carried out as a simulation. MSTGen simulation software tool was used for that purpose. 5000 simulees derived from normal distribution ($N(0,1)$) were simulated. 60 different conditions (two panel designs (1-3-3; 1-2-2), three module lengths (10-15-20), 5 different a parameter sequences (“0.8; 0.8; 0.8” - “1.4; 0.8; 0.8”-“0.8;1.4; 0.8” - “0.8; 0.8;1.4” - “1.4; 1.4; 1.4”) and two b parameter difference (small; large) conditions) were taken into consideration during analysis. Correlation, RMSE and AAD values of conditions were calculated. Conditional RMSE values corresponding to each ability level are given in a graph. Dissimilar to other studies in the literature, this study examines b parameter difference condition in three-stage tests and its interaction with a parameter sequence. Study results show that measurement precision increases as the number and length of the modules increase. Errors in measurement decrease as item discrimination values increase in all stages. Including items with a high value of item discrimination in the second or last stage contributes to measurement precision. In extreme ability levels, large difficulty difference condition produces lower error values when compared to small difficulty difference condition.

1. INTRODUCTION

In line with the fact that computer technology has led to various differences in all domains of life, it has changed the way cognitive/affective tests are carried out. Traditional paper-and-pencil tests have gradually been replaced by computer based testing (CBT) in time. When the qualities of tests that are administered on the basis of CBT are considered, it can clearly be observed that there is a version which prescribes all individuals to take the same form as well as the other version which assigns the use of an adaptive form (computerized adaptive testing-CAT) that makes it possible to determine the test items in accordance with the abilities of individuals who take the test. In the background, the adaptive form performs some operations

CONTACT: Melek Gülşah Şahin ✉ mgulsahsahin@gazi.edu.tr 📍 Gazi University, Gazi Education Faculty, Department of Educational Sciences, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

that are based on Item Response Theory (IRT), which is a testing theory that increases the measurement precision about the estimation of item and ability parameters and that brings about a great number of advantages in terms of implementation. Computer Adaptive Testing (CAT) has been frequently preferred for use thanks to the numerous advantages it offers and it has been discussed in many studies up until now. CAT has item level adaptation as its basis, because the algorithms that are created for item selection build up each test form while the test is being taken by the examinee via controlling an item, estimating a tentative score, and later choosing the next item to be used from the active item bank through making use of some specific statistical optimization criteria (Luecht & Nungester, 1998). As a result, examinees are not allowed to review their responses to previous items when a CAT implementation is adopted. Furthermore, the item exposure rates of some items might be high although the item sets that examinees come across differ from each other. Unfolding plenty of items, no matter how many examinees see them, can influence the accuracy and validity of test scores if the examinees that will take the test in the future have an opportunity to see the test items before testing (Rotou, Patsula, Steffen, & Rizavi, 2007). Another limitation that is caused by the fact that examinees take different test items is that it turns out to be impossible to examine each test form with quality assurance purposes before testing (Luecht & Nungester, 1998).

Computer Adaptive Multi-stage Testing (ca-MST) has not only taken the advantages of computer technology as well as adaptive forms but also found a way to overcome the problems of delivering a different set of items to each examinee. The ca-MST has been widely implemented thanks to these qualities, and it is one of the issues that are frequently studied in the field of measurement and evaluation nowadays. Because of this reason, there are some assessments that have replaced CAT versions with ca-MST versions, such as the National Assessment of Educational Progress (NAEP) and the Graduate Record Examinations (GRE) (Zeng, 2016; Zheng, Nozawa, Gao, & Chung, 2012). The major distinction between ca-MST versions and CAT versions is that ca-MST prescribes examinees to take a set of pre-constructed sub-test which matches their tentative ability estimates all the time (Hendrickson, 2007). However, when a CAT is used, only a single item is selected to match the ability estimates (Zeng, 2016). These pre-assembled sub-tests are called modules and ca-MST make use of these fundamental building blocks (Leucht & Sireci, 2011). In brief, it can be stated that the main difference between ca-MST and CAT is that ca-MST is a module adaptive test, not an item level adaptive test. The literature review shows that CAT and ca-MST are frequently compared against the backdrop of some certain qualities. As a consequence, ca-MST stays between linear test forms (paper and pencil testing and computer-based testing) and conventional item-level CAT (Hendrickson, 2007; Leucht & Nungester, 1998; Sadeghi & Khonbi, 2017; Sarı, Sarı, & Huggins Manley, 2016).

According to Leucht and Sireci (2011), ca-MST has some other advantages besides paying regard to content specifications and item exposure issues. These advantages include, but are not limited to, enabling examinees to review test items included in the same test, simplifying the test format, obtaining test results close to CAT versions especially when long tests and different contents are in question, simplifying the expensive programs that are used for test development and administration, being able to fix “information structure” that is necessary for each panel and reproducing it among panels, making it possible to examine the quality of panels before the test is administered to the test-takers (Hadadi & Leucht, 1998; Leucht & Nungester, 1998; Leucht, 2000; van der Linden, 2005; Patsula, 1999; Schnipke & Reese, 1999; Zenisky & Jodoin, 1999; Zenisky & Hambleton, 2014).

1.1. ca-MST Components

There are four basic test design/administration concepts in ca-MST: (1) modules, (2) panels, (3) stages, and (4) pathways (Leucht, 2000). Modules are units that are homogeneous in terms

of item difficulty. Each module can be structured in line with a specific content and statistical characteristics before examinees take the test (Leucht & Nungester, 1998). The length of the modules is based on the nature of the test, so a module can change between a small size (five to ten items) and large size (50 to 100 items). They can also differ in length according to stages and average difficulty (Leucht, 2000). There is a certain statistical target that should be met by each series of modules in terms of a psychometric perspective; this target can be described as a prescribed level of measurement precision within a specific region of the score scale (i.e., an IRT test information target) (Leucht & Sireci, 2011).

Test assembly requires a process where item modules are grouped so as to form test administration units that are called "panels" in accordance with stage and difficulty level (Leucht & Nungester, 1998; Leucht & Sireci, 2011). A panel is a specific combination of the modules that have to meet the pre-supposed requirements of content and other qualitative test features besides other explicit statistical targets (Leucht & Nungester, 1998). There is a natural hierarchical arrangement which designates panels which own multiple modules and modules that own multiple items (Leucht & Sireci, 2011). The ca-MST panels are divided into two or more stages and each module included in the panel is assigned to a specific stage. Two- or three-stage designs are widely used (Park, 2015). A ca-MST stage may have more than one module, and each module may address a different proficiency level (e.g., one easy, one moderate, and one hard module, each of which is aimed at a particular range of examinee abilities) (Leucht & 1998). A panel configuration refers to a simple sequence of integers that indicate the number of stages and amount of adaptation that are possible in a specific panel design (Leucht & Sireci, 2011). A ca-MST design that consists of more than two stages puts forth that how an examinee performs on the second stage of the test helps routing the examinee to a third stage module later on.

A simple panel structure can be seen in [Figure 1](#) below. This is called a 1-3-3 panel design. In this design, modules with different difficulty levels are created and specified in the figure. The first stage includes a module with a difficulty level of medium (M) or average. On the other hand, the item difficulty levels of modules in the second and third stages range from easy (E) to hard (H). Here, there are seven possible pathways designated for examinees: M1+E2+E3, M1+E2+M3, M1+M2+E3, M1+M2+M3, M1+M2+D3, M1+H2+M3, and M1+H2+H3 (Leucht & Sireci, 2011).

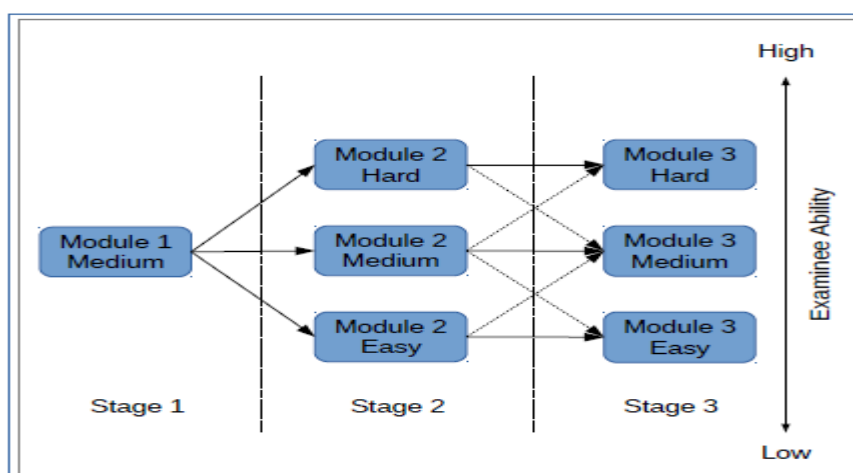


Figure 1. Structure of 1-3-3 panel design of ca-MST

Only one module from each stage is administered for each examinee during an actual test delivery process. Examinees having different abilities are routed to different modules. For

example, an examinee whose ability estimation is at a high level may be assigned a module that includes more difficult items or negatively skewed test (module) information functions (Zenisky & Jodoin, 1999). In [Figure 1](#), Medium-1, Hard-2 and Hard-3 modules are created for a high proficiency examinee.

Principally, some variables should be considered for a ca-MST design. Zenisky (2004) classifies the variables that are necessary for a ca-MST design as follows: i) basic structure variables such as total number of items and total number of stages included in the test, ii) variables of test and module assembly such as difficulty of the first-stage module, the number of the relative difficulty of the modules in the following stage, and content balance and other limitations; and iii) variables related to administration which influence the efficacy and implementation of ca-MST, including strategies of routing and ability estimation methods. In addition to the variables listed above, some other variables that are considered in ca-MST implementations include panel design considerations such as the quality of the item pool, distribution of the difficulty and item discrimination for each content, and the number of modules included in each stage (Han & Guo, 2013; Leucht, Brumfield, & Breithaupt, 2006; Park, 2015; Xing & Hambleton, 2004).

1.2. Aim of the Study

In this study, basic structure as well as test and module assembly variables such as different panel designs, change in b parameter difficulty level of modules, different distributions of item discrimination in the stages and test length have been examined. The aim of this study is to analyze the influence of the specified conditions on measurement precisions. As it is seen in the literature, there are studies which have focused on the effect of some variables such as the test length, dichotomous items and polytomously-scored items on ca-MST test performance (Jodoin, Zenisky, & Hambleton, 2006; Leucht & Nungester, 1998; Patsula, 1999; Xing & Hambleton, 2004). Patsula (1999) carried out a study that examined panel designs including different module numbers and underlined the fact that it is important to examine the number of modules included in each stage instead of the length of modules. Another variable that influences measurement precision is test information function (TIF) which is a degree of measurement precision demanded in the various regions of the ability scale included in the test to be administered. TIF also plays an important role in ensuring the consistency of results obtained from ca-MST against time and across panels (Leucht, 2000; Leucht & Nungester, 1998).

While TIF values are sometimes expected to be maximum at specified decision points in accordance with the aim of the test (i.e., when the test aims at classification), they are expected to be flat in between specific intervals (i.e., a test needs to assess ability across a wide range of theta scale) (Verschoor & Eggen, 2014; Park, 2015). As each module is constructed before the test is implemented in ca-MST design, it is not possible to control TIF values later. This case is directly related to the reliability of the test results. If the number of test items included in the routing module is low, there might be some items that can be answered correctly by guessing. In this case, it turns out to be more important to make ability estimations regarding examinees' performances with fewer mistakes in the following stages. In such cases, if the modules included in the following stages which examinees are routed focus on a narrow region of abilities, this may cause misrouting (Kim & Moses, 2014). This is especially very important in two-stage designs as there is a single adaptation point. In tests that consist of different numbers of stage (three stages or more), it is important to examine the effect of the intervals of TIF values specified for each stage on measurement precision as well as test assembly. In the literature, there are studies in which modules are constructed considering different TIF distributions or difficulty parameter values and the results are compared accordingly (Kim, Chung, Dodd, & Park, 2012; Kim & Moses, 2014; Kim, Moses, & Yoo, 2015).

Another essential point in constructing the modules is paying attention to item discrimination. There is no doubt that there is a relationship between item discrimination and test reliability. In the literature, there are studies which examine the effect of item discrimination on creating item pool and routing module (Boztunç Öztürk, 2019; Xing & Hambleton, 2004).

1.3. Significance of the Study

It is recommended in the literature to examine the change in ranges of difficulty levels related to b parameters for modules in ca-MST implementations (Leucht, Brumfield, & Breithaupt, 2016) and to deal with this change together with the impact of different levels of item discrimination (Kim & Moses, 2014; Kim et al., 2015). During the literature review, the researcher has not come across any study that examines the change in item discrimination of modules included in different stages. Therefore, it is thought that it will be worthwhile to examine the interaction of modules that are constructed depending on small and large b parameter (difficulty) level differences with different values of average discrimination within the framework of this study. Another significance of this study is that it will take the advantage of three-stage panel design (1-3-3 and 1-2-2) unlike other studies in the literature (Kim & Moses, 2014; Kim et al., 2015). The reason why these panel designs are chosen is that 1-2-2 ca-MST structure is popular for classification testing, while 1-3-3 design is the most commonly preferred research for ability estimation testing (Jodoin et al., 2006; Park, 2015; Zenisky, 2004). Furthermore, the researcher aims at contributing to the literature by examining the module length together with module difficulty and module discrimination values. Also, some suggestions will be provided for test operators to use in practice in light of the findings that will be obtained at the end of this study

2. METHOD

2.1. ca-MST Panel Assembly

This study examines the panel design of 1-3-3, which is the most frequently preferred one in the literature (Chen, 2010; Hambleton & Xing, 2006; Jodoin et al., 2006; Leucht & Nungester, 1998; Leucht et al., 2006; Park, 2015; Patsula, 1999; Zenisky, 2004). Patsula (1999) has stated that the change in the number of modules, not stages, produces a difference in terms of measurement precision. This study also addresses 1-2-2 panel design, which is also three-stage but has a different number of modules (Chen, 2010; Patsula, 1999; Zenisky, 2004). The second variable that is considered within the framework of this study is module length assignment. It is preferred to have a condition where each stage includes equal numbers of items. Chen (2010) has underlined that ca-MST studies generally make use of items ranging from 33 to 60 in number. Within the framework of this study, the module length is chosen to be 10, 15 and 20 items, whereas test length is decided to be 30, 45 and 60 items in total for the purpose of observing change in tests that have an average length on one side and long tests on the other side.

The third variable that is varied in this study is item discrimination. The study aims at designating at which stages the average discrimination values of items that are included in the modules can be high or low in a three-stage ca-MST implementation. It is seen that item discrimination has an average value of 0.75-0.85 and SD value of 0.27-0.30 in studies that have been carried out with parameters obtained from a real pool (Kim & Moses, 2014; Kim et al., 2015, Patsula, 1999; Zheng & Chang, 2015). Hambleton and Xing (2004) carried out a study in which they identified item quality as poor ($\bar{x}=0.60$), average ($\bar{x}=1.00$) and best corresponding to average ($\bar{x}=1.40$) according to a parameter value. In this study, a parameter was addressed as average ($\bar{x}=0.80$; $SD=0.25$) and high ($\bar{x}=1.40$; $SD=0.25$) for each stage. Within the framework of this study, a parameter average of items included in the modules in each stage for a three-stage model are addressed with five different conditions which can be

listed respectively as average-average-average, high-average-average, average-high-average, average-average-high and high-high-high.

Small b parameter difficulty level and large b parameter difficulty level were selected while constructing the modules included in stages for both panel designs. Literature review shows that there are studies which are conducted with two-stage tests and examine difficulty difference conditions. Kim, Moses and Yoo (2015) carried out a simulation study in which they designated the theta values as ($b = -0.5$, $b = 0.0$ and $b = 0.5$) in small difficulty difference condition and as ($b = -0.5$, $b = 0.0$ and $b = 0.5$) in large difficulty difference conditions for the second stage. They specified the difficulty difference between easy modules as 0.5 in small and large conditions. The same design was used for the difficulty difference between difficult modules. In addition, medium module was set to be .00 in both difficulty difference conditions by the authors. On the other hand, in the study that was carried out by Kim and Moses (2014), the difficulty difference of two conditions was set to be 0.70 for both easy and difficult modules.

In this study, in which a three-stage test is constructed, considering the fact that the difficulty values between the modules can increase in line with the number of stages (Schnipke & Reise, 1997), the difference between difficulty was set to be 0.5 in both 2. and 3. stages for small and large difficulty difference conditions. Moreover, the difficulty difference between easy modules (or difficult modules) was set to be 0.5 in the 2. and 3. Stages for two conditions. For 1-3-3 design; under the small-difference difficulty condition, the average of item difficulty parameters was set to be .00 for routing; for the second stage, -0.5 for easy, .00 for medium and +0.5 for difficult; for the third stage, -1.00 for easy, .00 for medium and +1.00 for difficult. Under large b parameter difference, the average of item difficulty parameters was set to be .00 for routing; for the second stage, -1.00 for easy, .00 for medium and +1.00 for difficult; for the third stage, -1.5 for easy, .00 for medium and +1.5 for difficult. As a consequence, when small b parameter difference condition is in question, the difference of range of parameters in the second stage is set to be 1.00, while it is set to be 2.00 in the third stage. In the case of large b parameter difference, on the other hand, the difference of range of b parameters is set to be 2.00 in the second stage, whereas it is set to be 3.00 in the third stage. For 1-2-2 panel design, the same item pool has been used while only the module with an average difficulty level at the second and third stage has been removed. For example, routing module with a module length of 10 items has been used for both small and large difference in 1-3-3 and 1-2-2 panel design. The easy and hard modules included in second stage of 1-3-3 panel design are common with the easy and hard modules included in 1-2-2 panel design under all conditions. The variables that are included in the study are summarized in [Table 1](#).

Table 1. *The variables that are included in the study*

Variable	Levels
Panel Design	“1-3-3”; “1-2-2”
Module Length	10-15-20
a parameter (item discrimination) sequence in stages	C1(“0.80”-“0.80”-“0.80”) C2(“1.40”-“0.80”-“0.80”) C3 (“0.80”-“1.40”-“0.80”) C4(“0.80”-“0.80”-“1.40”) C5 (“1.40”-“1.40”-“1.40”)
b parameter (difficulty) difference condition in stages	Small differences (1.00 theta differences in stage two; 2.00 theta differences in stage three) Large differences (2.00 theta differences in stage two; 3.00 theta differences in stage three)

2.2. Data Simulation

MSTGen (Han, 2013) was used for ca-MST application within the context of variables that are given in Table 1. More than one simulation were realized within the scope of the conditions specified in the program while constructing each module, and then, test information function (TIF) graphics were examined for that module before including the most suitable module according to the specified values in the scope of the study. For example, for 1-3-3- panel design small b difference condition; routing module is constructed in a way to reflect one TIF center (theta point of 0.00), the second stage is constructed in a way to reflect three TIF centers (theta points of -.05, .00, +0.5) and the third stage is constructed in a way to reflect three TIF centers (theta points of -1.00, .00, +1.00). 5000 simulees derived from normal distribution (N (0,1)) are simulated in this study. Maximum Fisher Information (MFI) module selection method was used to choose the modules. The method of Expected a Posteriori (EAP) was preferred for ability estimation of examinees. Moreover, 100 replications were carried out.

2.3. Data Analysis

In the study, only one panel implementation was realized while 60 conditions (2 panel designs \times 3 module lengths \times 5 item discrimination sequences \times 2 b parameter differences) were examined. For all conditions, Pearson product-moment correlation coefficient, RMSE (Root Mean Square Error) and AAD (Average Absolute Difference) were calculated. Also, the equations of Pearson product-moment correlation coefficient, RMSE and AAD are presented below.

$$r_{\hat{\theta}_i \theta_i} = \frac{n \sum_{i=1}^n \hat{\theta}_i \theta_i - \sum_{i=1}^n \hat{\theta}_i \sum_{i=1}^n \theta_i}{\sqrt{\left[n \sum_{i=1}^n \hat{\theta}_i^2 - \left(\sum_{i=1}^n \hat{\theta}_i \right)^2 \right] \left[n \sum_{i=1}^n \theta_i^2 - \left(\sum_{i=1}^n \theta_i \right)^2 \right]}}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2}{n}}, \quad AAD = \frac{\sum_{i=1}^n |\hat{\theta}_i - \theta_i|}{n}$$

What each symbol represents in the formulas is given below.

n = the size of the sample

$\hat{\theta}_i$ = estimated level of ability for person i

θ_i = the known level of ability person i

When calculating Conditional RMSE, the groups were initially formed in each theta range of theta ability level and six groups were obtained from -3 theta to 3 theta values. Then, RMSE values of six theta θ change points were calculated.

3. RESULT / FINDINGS

The findings are given under two headings. Under the heading of overall outcomes, there are some explanations regarding goodness of fit values given in Table 2. Under the heading of conditional outcomes, graphs and explanations regarding the change of RMSE according to theta change points are presented.

3.1. Overall Outcomes

Correlation, RMSE and AAD values that have been obtained in relation to the 60 conditions that have been addressed within the framework of this study are given in Table 2.

Table 2. Corralation, RMSE and AAD results of ability estimation

Panel Design	b-parameter (difficulty) difference	a parameter(item discrimination) sequence*	Correlation			RMSE			AAD		
			Module Length			Module Length			Module Length		
			10	15	20	10	15	20	10	15	20
1-3-3	Small	C1	0.95	0.97	0.97	0.32	0.26	0.24	0.25	0.21	0.19
		C2	0.95	0.97	0.98	0.30	0.24	0.21	0.24	0.19	0.16
		C3	0.96	0.97	0.98	0.28	0.23	0.20	0.22	0.18	0.16
		C4	0.97	0.98	0.98	0.26	0.23	0.19	0.20	0.18	0.15
		C5	0.98	0.98	0.99	0.22	0.19	0.16	0.17	0.15	0.13
	Large	C1	0.95	0.97	0.97	0.32	0.26	0.24	0.25	0.21	0.19
		C2	0.96	0.97	0.98	0.30	0.23	0.21	0.23	0.19	0.16
		C3	0.97	0.98	0.98	0.26	0.21	0.19	0.20	0.17	0.15
		C4	0.97	0.98	0.98	0.27	0.22	0.21	0.21	0.18	0.17
		C5	0.98	0.98	0.99	0.22	0.18	0.16	0.17	0.14	0.13
1-2-2	Small	C1	0.95	0.96	0.97	0.32	0.27	0.24	0.25	0.21	0.19
		C2	0.96	0.97	0.98	0.30	0.24	0.21	0.24	0.19	0.17
		C3	0.96	0.97	0.98	0.29	0.23	0.21	0.23	0.18	0.16
		C4	0.97	0.97	0.98	0.27	0.23	0.20	0.21	0.18	0.16
		C5	0.97	0.98	0.99	0.23	0.19	0.16	0.18	0.15	0.13
	Large	C1	0.95	0.96	0.97	0.32	0.27	0.25	0.25	0.22	0.20
		C2	0.96	0.97	0.98	0.30	0.24	0.21	0.24	0.19	0.17
		C3	0.96	0.97	0.98	0.27	0.23	0.20	0.21	0.18	0.16
		C4	0.96	0.97	0.98	0.29	0.24	0.21	0.23	0.19	0.17
		C5	0.97	0.98	0.99	0.23	0.19	0.17	0.19	0.15	0.13

*C1(“0.80”-“0.80”-“0.80”) C2(“1.40”-“0.80”-“0.80”); C3(“0.80”-“1.40”-“0.80”); C4(“0.80”-“0.80”-“1.40”); C5(“1.40”-“1.40”-“1.40”)

The findings related to the values of goodness of fit that are given in [Table 2](#) are presented below.

3.1.1. When the design was 1-3-3 and there was a small difference in b parameters;

i) In ca-MST test structure whose module length was composed of 10 items, the lowest correlation value was found to be 0.95 in the case of C1 where item discrimination was selected to be the lowest across all stages. Similar to this result, the highest RMSE value (0.32) as well as the highest AAD value (0.25) were obtained. On the other hand, the highest correlation value of 0.97 was observed in the case of C5 condition where the items in all modules had the highest mean of item discrimination. The lowest RMSE value (0.22) besides the lowest AAD value (0.17) was obtained in this condition. On the other hand, among the conditions where the values of item discrimination were altered in stages, the highest correlation value (0.97), the lowest RMSE value (0.26) and the lowest AAD value (0.20) were obtained in the case of C4 condition where the items included in the last stage had the highest level of mean discrimination value.

ii) The lowest correlation value (0.97) as well as the highest RMSE value (0.26) and AAD value (0.21) in C1 condition were obtained under the condition where the module length covered 15 items. The highest correlation value (0.98) as well as the lowest RMSE value (0.19) and AAD value (0.15) were obtained under the condition of C5, where the items in all stages had the highest average value of item discrimination. On the other hand, when the change in item discrimination was considered, the highest correlation value (0.98) as well as the lowest RMSE value (0.23) and AAD value (0.18) were obtained under the condition of C4 where the items in the last stage had the highest value of item discrimination.

iii) In the test structure where the module length consisted of 20 items and when it was decided to have a small difference between b parameters as was the case when the module length was composed of 10 and 15 items, the lowest correlation value (0.97) as well as the highest RMSE (0.24) and AAD value (0.19) were obtained under C1 condition. The highest value of correlation (0.99), the lowest value of RMSE (0.16) and AAD (0.13) were obtained under C5 condition. When the change in item discrimination was considered, there was the highest value of correlation (0.98) as well as the lowest value of RMSE (0.19) and AAD (0.15) under the condition of C4 where the items included in the last stage had the highest value of item discrimination.

3.1.2. When the design was 1-3-3 and there was a large difference in b parameters

i) When the module length was decided to cover 10 items, C1 condition gave the lowest correlation value (0.95) while C5 condition gave the highest correlation value (0.98). In relation to the obtained correlation values, the highest RMSE value (0.32) and the lowest AAD value (0.25) were observed under C1 condition. On the other hand, the lowest RMSE value (0.22) and the lowest AAD value (0.17) were obtained in the case of C5 condition where item discrimination values were chosen to be equal and the highest in all stages. When the change in item discrimination values across stages was observed, the highest level of measurement precision was obtained under the condition of C3. The highest correlation value (0.97) as well as the lowest RMSE (0.26) and the lowest AAD value (0.20) were obtained under the condition of C3 where average value of item discrimination was chosen to be the highest in the second stage.

ii) Under the condition where the module length was composed of 15 items the lowest correlation value (0.97), the highest RMSE value (0.26) and AAD value (0.21) were obtained when a parameter sequence was C1. On the other hand, the highest correlation value (0.98) as well as the lowest RMSE value (0.18) and AAD value (0.14) were obtained in C5 sequence. Under the conditions related to the change in the values of item discrimination across stages,

measurement precision was found to be the highest under C3 condition. C3 condition produced a correlation value of 0.98, RMSE value of 0.21 and AAD value of 0.17.

iii) On the other hand, under the condition where the module length was composed of 20 items, the lowest value of correlation (0.97), the highest value of RMSE (0.24) and the highest value of AAD (0.19) were obtained in C1 sequence. Contrary to this condition, the highest value of correlation (0.99), the lowest value of RMSE (0.16) and the lowest value of AAD (0.13) were obtained under the condition of C5. When the change in the values of item discrimination across stages was considered, the highest value of measurement precision was obtained under C3 condition with the highest value of correlation (0.98), the lowest value of RMSE (0.19) and the lowest value of AAD (0.15).

Furthermore, when it comes to 1-3-3 panel design, it is observed that the more the module length increases, the more the correlation values increase in all different sequences of a parameters for both small difference and large difference conditions. When item discrimination sequence conditions are examined, the lowest goodness of fit values was obtained under C1 conditions whereas the highest values were obtained under C5 condition regardless of module length and item difficulty difference. When the sequences in which the average a parameter distribution showed variation across stages were examined, the highest level of measurement precision and a small difficulty difference were obtained under the condition of C4, whereas large difficulty difference was obtained under the condition of C3. In the tests with the same module length and with the conditions of both small difficulty difference and large difficulty difference, under the condition of C2, where the value of a parameter was chosen to be the highest in routing module, the measurement precision was found to be the lowest.

3.1.3. When the design was 1-2-2 and there was a small difference in b parameters

i) When the module length was composed of 10 items, the lowest value of correlation (0.95), the highest value of RMSE (0.32) and the highest value of AAD (0.25) were obtained under the condition of C1. On the other hand, the lowest value of RMSE (0.23) and the lowest value of AAD (0.18) were obtained when it comes to C5 condition, where the average a parameter values were equal and the highest, the highest value of correlation (0.97). On the other hand, when it comes to the conditions where the values of item discrimination were altered across stages, the condition of C4 had the highest measurement precision with the highest value of correlation (0.97), the lowest value of RMSE (0.27) and the lowest value of AAD (0.21).

ii) When the module length was composed of 15 items and there was a small difficulty difference, the lowest level of correlation (0.96) as well as the highest value of RMSE (0.27) and AAD (0.21) were obtained under the condition of C1. On the other hand, the highest value of correlation (0.98), the lowest value of RMSE (0.19) and the lowest value of AAD (0.15) were obtained under the condition of C5. When the change in item discrimination is considered, the condition of C4 produced the highest value of correlation (0.97) as well as the lowest value of RMSE (0.23) and the lowest level of AAD (0.18).

iii) When the module length was selected to be composed of 20 items, the lowest value of correlation (0.97) besides the highest value of RMSE (0.24) and the highest value of AAD (0.19) were obtained under C1 condition. In the condition of C5, where average a parameter value was chosen to be the highest in all modules, there came out the highest value of correlation (0.99), the lowest value of RMSE (0.16) and the lowest value of AAD (0.13). These results were similar to the results of those conditions where module lengths were chosen to be 10 and 15 items respectively. Furthermore, similar to the other module lengths, C4 condition, where the last stage included items with high values of item discrimination, produced the highest value of correlation (0.98), the lowest value of RMSE (0.20) and the lowest value of AAD (0.16).

3.1.4. When the design was 1-2-2 and there was a large difference in b parameters

i) When the module length was composed of 10 items and there were large difficulty difference conditions, the results were found to be similar to those that were obtained in the case of small difficulty difference conditions. The lowest value of correlation (0.95) and the highest value of RMSE (0.32) as well as the highest value of AAD (0.25) were obtained under the condition of C1. C5 condition, on the other hand, produced the highest value of correlation (0.97), the lowest value of RMSE (0.23) and the lowest value of AAD (0.19). The highest measurement precision in the change of values related to item discrimination across stages was observed in the condition of C3. C3 condition produced a correlation value of 0.96, RMSE value of 0.27 and AAD value of 0.21.

When the module length was composed of 15 items, the lowest value of correlation (0.96) as well as the highest RMSE value (0.27) and AAD value (0.22) were obtained under C1 condition. On the other hand, C5 condition resulted in the highest correlation value (0.98), the lowest RMSE value (0.19) and the lowest AAD value (0.15). The highest measurement precision was obtained under C3 condition in the change of item discrimination values across the stages. C3 condition produced a correlation value of 0.97, RMSE value of 0.23 and AAD value of 0.18.

When the module length was selected to be composed of 20 items and there was a large difficulty difference, the lowest value of correlation (0.97) besides the highest value of RMSE (0.25) and the highest value of AAD (0.20) were obtained under C1 condition. Contrary to this condition, the highest value of correlation (0.99) as well as the lowest value of RMSE (0.17) and the lowest value of AAD (0.13) were obtained under C5 condition. In C3 condition, where the value of average item discrimination was chosen to be highest in the second stage, there came out the highest value of correlation (0.98), the lowest value of RMSE (0.20) and the lowest value of AAD (0.16).

When 1-2-2 panel design was in question, it was observed that measurement precision increased with the increase in the module length for both small difficulty difference and large difficulty difference cases, which meant that the obtained goodness of fit values got better. When item discrimination sequence conditions are examined in general, the lowest goodness of fit value in C1 condition and the highest goodness of fit value in C5 condition were obtained regardless of module length and item difficulty difference.

In 1-2-2 panel design, when small difficulty difference condition was in hand, the highest value of measurement precision was obtained under C4 condition where the items having higher values of item discrimination were included in the last stage. However, when there was a large difficulty difference, higher values of goodness of fit were obtained under C3 condition where items having high values of item discrimination in all module lengths were included in the medium stage, when compared to the conditions where items with high values of item discrimination were included in the first and last stages. This was similar to the results obtained with 1-3-3 panel design.

Considering the effect of panel design, when all the other conditions are compared to each other, it is clear in some conditions that the correlation values are higher whereas error values are lower in 1-3-3 panel design in each of the conditions included in this study.

3.2. Conditional Outcomes

Conditional RMSE values are examined at six theta change points within the framework of this study. Figure 2 below shows the conditional RMSE values.

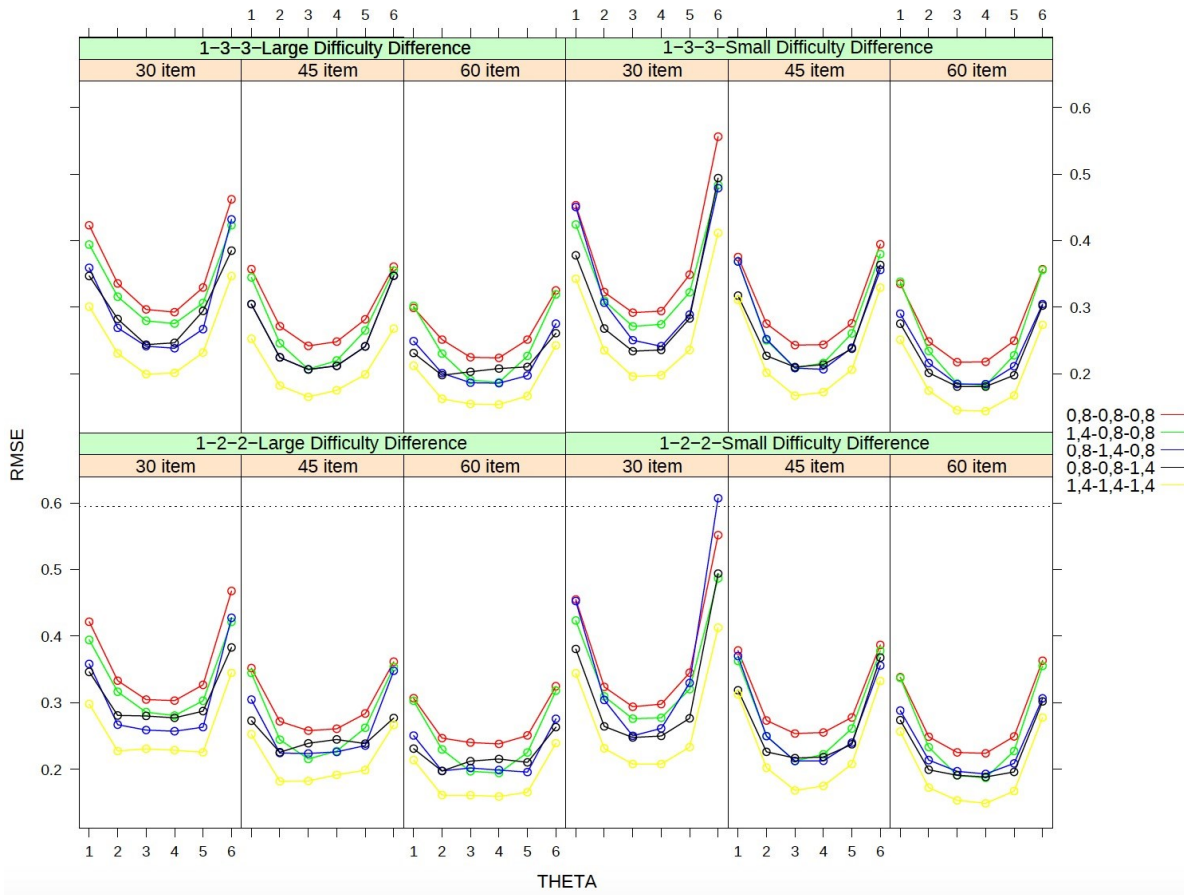


Figure 2. Conditional RMSE's of All Test Designs

When Figure 2 is examined, it is obvious that distribution of errors is lower at the level of extreme ability in all cases when compared to the ones at the level of medium ability. It is also observed that errors at all ability levels decrease with the increase in the module length. Especially in the module lengths with 20 items, it is seen that the difference between the errors at extreme and medium level abilities decrease. When the change of RMSE values in small and large b difference conditions are observed, it can be stated that errors tend to be higher at extremely high or low ability levels in small difficulty difference conditions with the same module length. These results are also valid for both 1-3-3 and 1-2-2 panel designs.

Moreover, at all ability levels, the distribution of errors in general are lower under C5 condition, where a parameter values are chosen to be higher in all stages. However, as it is seen in Figure 2, RMSE values at extreme ability levels are generally observed to be closer to each other under C1 and C2 conditions. Also, when the results obtained from C3 and C4 conditions are examined, close RMSE values are obtained at extreme ability levels.

4. DISCUSSION

The effect of different panel designs, module lengths, different sequence of a parameter value across stages and change in b parameter range on measurement precision in ca-MST implementations have been investigated within the scope of this study. The values of correlation, RMSE and AAD for 60 conditions addressed for that purpose have been calculated.

When the effect of test length is examined, the research result showed that there occurs a decrease in RMSE values at all ability levels as the test length increases. There are studies in the literature that have obtained similar results. Kim and Plake (1993) carried out a study in which they found out that RMSE values decrease as the test length increases in two-stage tests that are composed of 40-45-60 items, respectively. Within the framework of this study, all the modules have been designated to include equal numbers of items. This means that there is an increase in the number of items included in routing module as well as the following modules as the test length increases. The studies that focus on the length of routing module in the literature have given similar results to this study, which puts forth that errors decrease as the test length increases (Kim & Plake, 1993; Kim et al., 2015; Loyd, 1984). Moreover, when conditional RMSE values that are obtained for each ability level are examined, it becomes clear that there are more errors in tests with lower values of test length at extreme ability levels, whereas measurement precision increases as the module length increases.

The study has also focused on investigating the difference between b parameters of modules. When the overall outcomes are examined, differences b parameter did not make any impact on the outcomes. In extremely high or low ability levels, the condition of small difficulty difference has produced higher levels of error irrespective of other conditions. Especially when the difficulty of modules in the second stage is closer to the difficulty level of routing module, poor measurement can be obtained for the individuals with extreme ability levels (Lord, 1971; Patsula, 1999). As a consequence, the condition of large difficulty difference can ensure a higher level of measurement precision when estimating the abilities of individuals with extreme levels of ability. Kim et al. (2015) carried out a study in which they investigated small and large difficulty difference conditions when various ability estimations are in question in two-stages tests. Similar to the results of this study, they have concluded that lower levels of error are obtained (in some ability estimations) under the condition of large difficulty difference in extreme ability level. Test developers can be recommended in the light of the study results to include very easy and/or very difficult items in the ca-MST item pool for the purpose of measurement precision. However, it can be difficult to develop very difficult test items when compared to easy items or the ones with a medium level of difficulty (Kim & Moses, 2014).

When the average values of item discrimination belonging to the items included in the modules are considered, it is obvious that the lowest error is gained under the condition of C5, where a parameter values at all stages is equal and the highest. It is an expected result to have more reliable measurement as item discrimination values increase. Another question which is discussed within the scope of this study is at which stage the items with higher values of item discrimination should be included in order to reduce the errors. In line with this question, an important result of the study is that a high degree of measurement precision is obtained when small difficulty difference condition is in hand under the condition of C4, where the items at the last stage have high values of item discrimination. Under the condition of large difficulty difference, on the other hand, a high degree of measurement precision is obtained under the condition of C3, where the medium stage consists of items with high values of item discrimination. At the same time, the difference the b parameters were chosen as 2.00 theta for both the last stage of the small difficulty difference condition and the second stage of the large difficulty difference condition. At the end of the study, it was also discovered that measurement precision does not increase even if items with high values of item discrimination are used when the difference between b parameters becomes larger.

When using items with high values of item discrimination in the routing module, second stage and last stage in terms of ability levels are considered and, it is observed that the individuals with medium levels of abilities get errors closer to each other under the three conditions (C2, C3 and C4). However, when the test is short, the condition of C2 gives high values of errors at

the medium ability level. When extremely high or low ability levels are in question, including items with high values of item discrimination in the medium and last stages gives similar results, whereas including these items in the last stage produces lower levels of errors. The results of this study are parallel with the results of the study carried out by Chang and Ying (1999) as well as Zheng et al. (2012). It can be recommended to test operators to make use of item pools with high values of item discrimination as it will increase measurement precision (Xing & Hambleton, 2004). However, it can be stated that when this condition cannot be ensured, including items with high values of item discrimination in the last stage can contribute to measurement precision. It can be inferred from the results of this study that including items with high values of item discrimination in the routing module does not have any impact on measurement precision.

When the impact of panel design is considered, measurement precision of 1-3-3 panel design is higher than that of 1-2-2 panel design in some conditions. This can be explained via the fact that the items that are appropriate for medium level of ability are included in 1-3-3 panel design. When there is an increase in the number of modules, measurement precision also increases (Patsula, 1999). However, when RMSE values obtained from both 1-3-3 and 1-2-2 panel designs are examined in terms of ability levels, the errors obtained at the medium ability level are fewer than the errors at the extreme ability levels.

It can be recommended to the researchers in light of the results of this study to carry out similar studies with different panel designs (1-2-4; 1-2-3; 1-2-3-4) including different modules or stages. The effect of content distribution is not addressed in this study, so the effect of conditions with contents of different weights can be investigated. The number of items included in the modules are fixed in this study. The future studies should have a new condition to include different number of items in the modules. Moreover, it can be suggested to analyze the effect of module selection methods.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Melek Gülşah Şahin  <https://orcid.org/0000-0001-5139-9777>

5. REFERENCES

- Boztunç Öztürk, N. (2019). How the length and characteristics of routing module affect ability estimation in ca-MST?. *Universal Journal of Educational Research*, 7(1), 164-170. <https://doi.org/10.13189/ujer.2019.070121>
- Chang, H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 211-222. <https://doi.org/10.1177/01466219922031338>
- Chen, L. Y. (2010). An investigation of the optimal test design for multi-stage test using the generalized partial credit model (unpublished doctoral dissertation). The University of Texas at Austin. Retrieved from <https://repositories.lib.utexas.edu/handle/2152/ETD-UT-2010-12-344>
- Hadadi, A., & Leucht, R. M. (1998). Some methods for detecting and understanding test speededness on timed multiple-choice tests. *Academic Medicine*, 73, 47-50. https://journals.lww.com/academicmedicine/Citation/1998/10000/TESTING_THE_TES_T_Some_Methods_for_Detecting_and.42.aspx

- Hambleton, R.K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education*, 19(3), 221-229. https://doi.org/10.1207/s15324818ame1903_4
- Han, K. T. (2013). MSTGen: simulated data generator for multistage testing. *Applied Psychological Measurement*, 37(8), 666-668. <https://doi.org/10.1177/0146621613499639>
- Han, K. T., & Guo, F. (2013). *An approach to assembling optimal multistage testing modules on the fly* (Report No. RR-13-01). Virginia: GMAC.
- Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, 26, 44-52. <https://doi.org/10.1111/j.1745-3992.2007.00093.x>
- Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test design for credentialing exams with multiple purposes. *Applied Measurement in Education*, 19(3), 203-220. http://doi.org/10.1207/s15324818ame1903_3
- Kim, H., & Plake, B.S. (1993, April). *Monte carlo simulation of two-stage testing and computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta.
- Kim, J., Chung, H., Dodd, B.G., & Park, R. (2012). Panel design variations in the multistage test using the mixed-format tests. *Educational and Psychological Measurement*, 72(4), 574-588. <https://doi.org/10.1177/0013164411428977>
- Kim, S., Moses, T., & Yoo, H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*, 52(1), 70-79. <https://onlinelibrary.wiley.com/doi/abs/10.1111/jedm.12063>
- Kim, S., & Moses, T. (2014). An investigation of the impact of misrouting under two-stage multistage testing: A simulation study (Report No. RR-14-01). Princeton, NJ: English Testing Service.
- Leucht, R. M. (2000, April) *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Leucht, R., Brumfield, T., & Brithaupt K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, 19(3), 189-202. https://doi/abs/10.1207/s15324818ame1903_2
- Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, 35(3), 229-249.
- Leucht, R. M., & Sireci, S. G. (2011). *A review of models for computer-based testing*. (Report No. RR-2011-12). New York: CollegeBoard. <https://doi.org/10.1111/j.1745-3984.1998.tb00537.x>
- Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika*, 36(3), 227-242. <https://doi.org/10.1007/BF02297844>
- Loyd, B. (1984, February). Efficiency and Precision in two-stage adaptive testing. Paper presented at the Annual Meeting of Eastern Educational Research Association, West Palm Beach, FL.
- Park, R. (2015). Investigating the impact of a mixed-format item pool on optimal test designs for multistage testing (Doctoral dissertation). The University of Texas at Austin. Retrieved from <https://repositories.lib.utexas.edu/handle/2152/31011>
- Patsula, L.N. (1999). A comparison of computerized adaptive testing and multistage testing. (Doctoral dissertation). The University of Massachusetts Amherst. Retrieved from

- https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=4283&context=dissertations_s_1
- Rotou, O., Patsula, L., Steffen, M., & Rizavi, S. (2007, March). *Comparison of multistage tests with computerized adaptive and paper and pencil tests*. (Report No: RR-07-04). Princeton, NJ: English Testing Service.
- Sarı, H.İ., Yahşi Sarı, H., & Huggins Manley, A.C. (2016). Computer adaptive multistage testing: Practical issues, challenges and principles. *Journal of Measurement and Evaluation in Education and Psychology*, 7(2), 388-406. <https://doi.org/10.21031/epod.280183>
- Schnipke, D.L., & Reese, L.M. (1999). *A comparison of testlet-based test designs for computerized adaptive testing*. (Report No: 97-01). Princeton, NJ: Law School Admission Council.
- Sadeghi, K., & Khonbi, Z.A. (2017). An overview of differential item functioning in multistage computer adaptive testing using three-parameter logistic item response theory. *Language testing in Asia*, 7(1), 1-16. <https://doi.org/10.1186/s40468-017-0038-z>
- van der Linden, W.J. (2005). *Linear models for optimal test design*. New York: Springer.
- Verschoor, A., & Eggen, T. (2014) Optimizing the test assembly and routing for multistage testing. In D. Yan., A. A. von Davier, & C., Lewis, (Ed.), *Computerized Multistage Testing Theory and Applications* (pp:135-150). Taylor & Francis Group.
- Xing, D., & Hambleton, R., K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, 64(1), 5-21. <https://doi.org/10.1177/0013164403258393>
- Zeng, W. (2016). Making test batteries adaptive by using multistage testing techniques (Doctoral dissertation). The University of Wisconsin-Milwaukee. Retrieved from <https://dc.uwm.edu/cgi/viewcontent.cgi?article=2241&context=etd>
- Zenisky, A.L., & Hambleton, R., K. (2014). Multistage test desing: Moving research results into practice. In D. Yan., A. A. von Davier, & C., Lewis, (Ed.), *Computerized Multistage Testing Theory and Applications* (pp. 21-37). Taylor and Francis Group.
- Zenisky, A., L., & Jodoin, M., G. (1999). Current and future research in multistage testing. (Report No:370). Amherst, MA: University of Massachusetts School of Education.
- Zenisky, A. L. (2004). Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment. (Doctoral dissertation). University of Massachusetts Amherst. Retrieved from <https://scholarworks.umass.edu/dissertations/AA13136800>
- Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. (2012). *Multistage adaptive testing for a large-scale classification test: Design, heuristic assembly, and comparison with other testing modes*. (Report No:2012-6). Iowa City, IA.: ACT.
- Zheng, Y., & Chang, H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, 39(2), 104-118. <https://doi.org/10.1177/0146621614544519>

Investigation of Measurement Invariance of Science Motivation and Self-Efficacy Model: PISA 2015 Turkey Sample

Metehan Gungor ^{1,*}, Kubra Atalay Kabasakal ²

¹ Ministry of National Education, 06450, Ankara, Turkey

² Hacettepe University, Education Faculty, Measurement and Evaluation Department, 06800, Ankara, Turkey

ARTICLE HISTORY

Received: Oct 16, 2019

Revised: Apr 10, 2020

Accepted: May 01, 2020

KEYWORDS

Structural equation modeling,
Measurement invariance,
Instrumental motivation,
Science self-efficacy

Abstract: Measurement invariance analyses are carried out in order to find evidence for the structural validity of the measurement tools used in the field of educational sciences and psychology. The purpose of this research is to examine the measurement invariance of Science Motivation and Self-Efficacy Model constructed by Instrumental Motivation to Learn Science and Science Self-Efficacy subscales found in the PISA 2015 Student Questionnaire across different groups in the Turkish sample survey. The analysis was carried out with the data obtained from 4583 students that met the analysis assumptions. The measurement invariance of the model in terms of gender and statistical regional groups was examined by the structural equation modeling (SEM) technique. Firstly, the data was examined to determine whether the assumptions for the analyses were met. Then, measurement models were verified by performing confirmatory factor analysis (CFA). The measurement invariance across genders and statistical regions was tested by multi-group confirmatory factor analysis (MGCFA). Unweighted Least Squares (ULS) method was used as the estimation method in CFA and MGCFA stages. In order to make final decisions about the stage of measurement invariance models hold, Comparative Fit Index (CFI) was used. The results of the study show that the research model ensures all stages of invariance across gender groups and regions. Science Motivation and Self-Efficacy Model illustrates that it is valid to make comparisons between scores of male and female students or students from different regions of Turkey. According to the findings, the research model could provide complete measurement invariance.

1. INTRODUCTION

Education indicators (budget allocated for education, using education technologies, quality of people working within the field of education, literacy rate, etc.) provide significant information as to the development level of countries. Performances of students that are included in a specific education system can be accepted to be a good indicator of the quality of this education system in a country. Measurement and evaluation instruments are frequently used while making comparison related to the performances of students. Implementing measurement and evaluation at the national and international level plays an important role in developing educational policies of countries. There are a great number of assessment and evaluation implementations around the world. The most popular ones include large scale exams such as Programme for

CONTACT: Metehan Gungör ✉ gungormetehan@gmail.com 📍 Ministry of National Education, 06450, Çankaya/Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS). PISA is a large-scale triennial assessment and it focuses on how 15-year-old students make use of their already-existing knowledge and skills to solve daily life problems. Each implementation is grounded on one of the fields among reading skills, mathematical literacy and science literacy. While the main subject field of PISA was reading skills in 2000, which is the year when PISA was implemented for the first time, the conceptual foundation of the sixth cycle in 2015 was science literacy. PISA is a useful assessment instrument in terms of evaluating the effectiveness of education systems via collecting data on the basis of students, teachers and schools, and using the results obtained at the end of analyzing these data. Taking advantage of PISA to monitor the dispositions in the knowledge and skills of students coming from different countries and different demographic regions of each participant country leads the drive for developing knowledge (Organisation for Economic Co-Operation Development [OECD], 2017). PISA can be regarded as a watershed in the discourse on education in many countries. Countries make use of the results of PISA while they are developing their own educational policies. Some new practices on assessment and curriculum standards have been reformed and PISA-like competencies have been incorporated into their systems. Therefore, it is important to stay careful when interpreting PISA results regarding comparability. It should be tested whether sub-scales that are used in assessment measure the same construct in each sub-group. Otherwise, the interpretation that is based on the results of the assessment will not be valid. Construct equivalence is a basic assumption that should be met if the developers or executors of any assessment aim at comparing the scores of different groups or interpreting these scores in compliance with the intended use (Gierl, 2000). When an assessment instrument is designed to compare participants coming from two or more cultures, the construct to be measured via the test should be equivalent for the comparison to be meaningful (Hambleton, 1994). The necessity to examine if the structures to be measured via tests are equivalent or not makes the issue of measurement invariance a leading topic within the scope of assessment and evaluation implementations.

1.1. Measurement Invariance

Measurement invariance is described by Byrne and Watkins (2003) as “the level of items being perceived and interpreted the same among groups.” On the other hand, Mellenbergh (1989) as well as Meredith and Millsap (1992) starts out from the concept of “biasness” and describes measurement invariance as “the state of the conditional probability of obtaining a specific observed score related to an ability being independent of group membership in mathematical terms.” In other words, measurement invariance is measuring a psychological construct with the same level of correctness in all sub-groups (Sireci, Patsula, & Hambleton, 2005). Measurement invariance is a special property that should be tested in order for the between-group comparisons of the psychological construct that will be measured to be meaningful (Cheung & Rensvold, 2002) and for the deductions and interpretations that will be made at the end of comparisons to be valid (Somer, Korkmaz, Dural, & Can, 2009). Measurement invariance analysis can be carried out to find proof for the structural validity of tests that are developed to draw between-group comparisons. The test of measurement invariance is a kind of covariance structure analysis and it is designed on the basis of measuring a specific structure on different groups (Başusta, 2010). The most common method of testing measurement invariance is Multi-Group Confirmatory Factor Analysis (MGCFA) that falls under the umbrella term of Structural Equation Modeling (SEM) (Jöreskog & Sörbom, 1999; Kline, 2011; Koh & Zumbo, 2008). While SEM includes measurement errors in the model, it also considers the direct and indirect effects of the variances in the created model. Hence, it makes it possible to test, estimate and develop multivariate complex models (Raykov & Marcoulides, 2006). Four hierarchical nested models should be tested while examining measurement invariance with

MGCFA under the umbrella term of SEM. These four hierarchical models can be listed as (1) configural invariance, (2) weak invariance, (3) strong invariance and (4) strict invariance respectively (Meredith, 1993; Wu, Li & Zumbo, 2007).

1.1.1. *Configural invariance*

This is the first step of measurement invariance test. The groups are tested to see if they have the same factor structure or not at this stage. For that purpose, equivalence of factors and pattern of factor loading is analysed at this stage (Taris, Bok & Meijer, 1998). If configural invariance is ensured at the end of the analysis, this means that the same structure is measured in the comparison groups. If the analysis shows that the conditions of configural invariance are not met, this means that different structures are measured among groups (Wu, Li & Zumbo, 2007). Kline (2011) states that if the necessary conditions are not met at this stage, measurement invariance cannot be ensured at more constrained stages.

1.1.2. *Weak invariance*

The equivalence of measurement unit or factor loadings are analysed at this stage. It is tested if the groups have the same measurement unit concerning the latent variable or not at this stage of weak invariance. Therefore, this stage, which can be described as the test of the measurement unit, is called metric invariance. In this model, factor loadings are also restricted in addition to the conditions that are valid at the stage of configural invariance (Vandenberg & Lance, 2000). If the weak invariance cannot be ensured, it can be discussed that factors do not mean the same in different groups (Gregorich, 2006).

1.1.3. *Strong invariance*

It is tested if the constant of regression that is obtained when the factor scores of the groups to be compared is zero is equal or not at this stage of strong invariance. Because of this reason, strong invariance is also called as scalar invariance (Vandenberg & Lance, 2000). The equivalence of the observed variables and factor loadings are also examined at this stage, which requires between-group equivalence of factor variance and covariances. If the necessary conditions are met, it means that the means of the observed variables and factor loadings can be compared (Gregorich, 2006).

1.1.4. *Strict invariance*

While invariance is tested, parameter restrictions as well as error variances are limited at this stage (Vandenberg & Lance, 2000; Wu, Li & Zumbo, 2007). This is the last step of measurement invariance test. Ensuring this stage is proof of measurement invariance. Assessment tools that claim to be measuring the same construct among the groups should meet the conditions of strict invariance. Measurement invariance can be ensured only if this stage is ensured. The stages and the related conditions in question are summarized in [Table 1](#).

Table 1. *Measurement invariance stages.*

Degree of Invariance	Condition of Invariance	Group Comparison
Configural Invariance	Item/Factor groups	-
Weak Invariance	Item/Factor groups and factor loadings	Factor variance and covariances
Strong Invariance	Item/Factor groups, factor loadings and item intercepts	Factor variance and covariances, factor and observed variable averages
Strict Invariance	Item/Factor groups, factor loadings, item intercepts, and item residual variance	Factor variance and covariances, factor and observed variable averages, observed variance and covariances

[Kıbrıslıoğlu Uysal & Akın Arıkan, 2018]

1.2. Instrumental Motivation to Learn Science

Motivation is a psychological construct that affects student success and it provides people with the necessary power to carry out a specific activity (Schunk, Meece & Pintrich, 2014). When motivation is addressed within the scope of learning, it is the power that stimulates, maintains and directs the behaviour towards a specific goal (Dilts, 1998). If the student thinks that the knowledge that s/he has acquired in a lesson will be useful in her/his life and career, s/he can make a great effort in this lesson even if the topics in the lessons are not interesting for her/him (İlhan, 2015). This effort is influential on this student's performance. Such a motivation is called instrumental motivation. Instrumental motivation means that students discern that what they have learnt will be useful in their future studies and career plans, and so they are eager to learn science (Wigfield & Eccles, 2000). Student motivation is an indispensable part of a qualified educational life. Although it is widely accepted that motivation is an indispensable part of education, it is not known well how to use motivation in instructional design and what it means in this context. This results from the fact that motivation is a construct. Instrumental motivation in learning science under the heading of motivation to learn science was measured via a four-point Likert type sub-scale consisting of four items in the cycle of 2015. The items included in the sub-scale try to measure if the students think that science lesson will be useful in their future educational life and career plans (OECD, 2016). The items included in the Instrumental Motivation to Learn Science Scale, which is the subject matter of this study, is given in Table 2.

Table 2. Items that constitute Instrumental Motivation to Learn Science Scale in PISA 2015 student questionnaire.

Code	Item
ST113Q01TA	Making an effort in my <school science> subject(s) is worth it because this will help me in the work I want to do later on.
ST113Q02TA	What I learn in my <school science> subject(s) is important for me because I need this for what I want to do later on.
ST113Q03TA	Studying my <school science> subject(s) is worthwhile for me because what I learn will improve my career prospects.
ST113Q04TA	Many things I learn in my <school science> subject(s) will help me to get a job.

1.3. Science Self-Efficacy

Efficacy belief is one of the concepts that underpin the Social Learning Theory developed by Bandura. Bandura (1997) describes self-efficacy as the judgments of individuals as to what they are able to do and their belief as to the ability to fulfill a specific task successfully or display a behaviour accomplished. Self-efficacy is individuals' own judgments about how well they are able to fulfil an activity that is necessary for the solution to possible problems (Bıkmaz, 2002). Self-efficacy is not related to people's knowing what to do, but to their belief about what they are able to do or learn (Schunk & Pajares, 2009). With reference to these descriptions, it seems possible to describe science self-efficacy as individuals' own judgments about how well they are able to do the specific tasks that are necessary for finding solutions for problems in science. Science self-efficacy is one of the fields on which many studies focus on the literature (Bakan Kalaycıoğlu, 2015; Bircan & Sungur, 2016; Britner & Pajares, 2001; Uzun, Gelbal & Öğretmen, 2010; Zedlin, Britner & Pajares, 2007). It was concluded at the end of a study carried out by Uzun, Gelbal and Öğretmen (2010) in order to examine the relationship between success in science and cognitive qualities that science self-efficacy is the most important variable used to explain the male and female students' success in science. Science self-efficacy was measured in PISA in 2006 and 2015 with the same four-point Likert-type sub-scale that consisted of eight items (OECD, 2016). Whereas PISA is implemented in the same subject field once every nine

years, similar and same items are used for students. The purpose of using some common items in these implementations is to examine the trends in education. The items that constitute Science Self-Efficacy Scale, which is the subject matter of this study, are given in [Table 3](#).

Table 3. *Items that constitute Science Self-Efficacy Scale in PISA 2015 student questionnaire.*

Code	Item
ST129Q01TA	Recognise the science question that underlies a newspaper report on a health issue.
ST129Q02TA	Explain why earthquakes occur more frequently in some areas than in others.
ST129Q03TA	Describe the role of antibiotics in the treatment of disease.
ST129Q04TA	Identify the science question associated with the disposal of garbage.
ST129Q05TA	Predict how changes to an environment will affect the survival of certain species.
ST129Q06TA	Interpret the scientific information provided on the labelling of food items.
ST129Q07TA	Discuss how new evidence can lead you to change your understanding about the possibility of life on Mars.
ST129Q08TA	Identify the better of two explanations for the formation of acid rain.

1.4. Literature Review

The interpretations of the scores that are obtained from the measurement tools may vary among different groups. If the scores obtained from the same test are not comparable among different groups (gender, culture, socio-economic level, etc.), the differences in the mean scores of the groups or the correlation patterns of the test with external variables are potentially artificial and they may be misleading to a great extent (Reise, Widaman & Pugh, 1993). A test that is implemented with different groups may ensure measurement invariance while it does not serve for the measurement invariance among genders. This may result from between-group real differences whereas it may also result from the assessment tool itself (Başusta & Gelbal, 2015). For that reason, measurement invariance studies are carried out with large scale international tests such as PISA, TIMSS and PIRLS, whose results are preferred to make comparisons (Akyıldız, 2009; Alivernini, 2011; Asil & Gelbal, 2012; Ayvalli & Biçak, 2018; Başusta & Gelbal, 2015; Ercikan & Koh, 2005; Ertürk & Erdiñç Akan, 2018; Gülleroğlu, 2017; Karakoç Alatlı, Ayan, Polat Demir & Uzun, 2016; Kıbrıslıoğlu Uysal & Akın Arıkan, 2018; Nagengast & Marsh, 2014; Oliden & Lizaso, 2013; Ölçüoğlu & Çetin, 2016; Scherer, Nilsen & Jansen, 2016; Uyar & Doğan, 2014; Uyar & Kaya Uyanık, 2019; Wu, Li & Zumbo, 2007). For example, Ölçüoğlu and Çetin (2016) modelled some variables that affect the maths success of 8-grade students that participated in TIMSS 2011 in Turkey, and they examined the measurement invariance with MGCFA among the seven regions in Turkey. The sample of this study was composed of 6928 14-year-old students chosen from 239 schools in Turkey. The results showed that only configural invariance and weak invariance were maintained in sub-groups of regions. According to this result, the scale could not meet the conditions of invariance and strict factorial invariance cannot be detected. Therefore, according to the findings of the study, it wouldn't yield valid results to make a comparison between regions with the the scores which have been obtained via the items that constitute the subject matter of the study and that are deemed to have an effect on maths success of students. Gülleroğlu (2017) carried out a study with the data of PISA 2012 in order to examine the measurement invariance of affective qualities towards maths according to the variable of gender. The sample of the study was composed of 1598 students that took Form B in the test and were chosen from 15-year-old 4848 students in 170 schools in Turkey. The measurement invariance of factors which can be listed as interest in mathematics, anxiety for maths, self-perception towards maths and self-efficacy were examined with MGCFA. The researcher reported at the end of the study that all the variables apart from self-efficacy met the conditions of configural invariance. The researcher, who examined the measurement invariance through hierarchical measurement invariance, concluded that strict

invariance was not ensured in all the five factors that were analyzed. Therefore, strict measurement invariance could not be provided in all five factors that were the subject matter of the study. Ertürk and Erdiñç Akan (2018) carried out a study of measurement invariance according to gender with the data of TIMSS 2015. The study aimed at examining the measurement invariance of some variables related to the success of maths according to gender. The sample of the study consisted of 6456 4-grade students who participated in the test in Turkey. The latent variables that were chosen for the study were liking mathematics, interest in mathematics and confidence in mathematics, which were all thought to have an impact on maths. Each variable was examined separately and MGCFA was used to examine the measurement invariance hierarchically. The differences between the values of CFI (comparative fit index) among the invariance were tested. It was found out at the end of the study that all the variables that were tested in the study met the conditions of configural invariance whereas only the variable of liking mathematics met the conditions of strict invariance. Uyar and Dođan (2014) carried out a study with PISA 2009 Turkey sample in which they established a model called learning strategies and they examined the measurement invariance of the model according to the statistical regions in Turkey. The researchers reported that the model met the conditions of strict measurement invariance. For instance, Uyar and Kaya Uyanık (2019) established a different model with the affective scales that are also the subject matter of this study and that were included in PISA 2015, and they examined the gender-based measurement invariance with the sample of Turkey. As a result, the model that was constructed by the researchers could only meet the conditions of the stage of weak invariance. On the contrary to these studies, Ayvalli & Biçak (2018) carried out a study with one of the affective tests of PISA 2012 Turkey sample and reported that strict measurement invariance was ensured after doing analysis between genders. Similarly, Bařusta and Gelbal (2015), who carried out a study with the data of PISA 2006 implementation, reported that strict measurement invariance was ensured between gender groups. It can be concluded from the results of these studies that it is possible to have different findings as to whether similar and same tests provide measurement invariance or not.

Polat and Madra (2018) carried out a study based on gender by using the data obtained from both PISA and TIMSS 2015 and they found out at the end of the study that female students in Turkey were far behind male students about turning advantageous qualities such as self-confidence, sense of belonging to school, motivation, liking to learn into success. There are inconsistent results in the studies based on gender about the success in the field of maths and science (Ađaç & Masal, 2015; Batyra, 2017; Larson, Stephen, Bonitz & Wu, 2014). Starting from study results, the variables that are chosen for this study are Instrumental Motivation to Learn Science and Science Self-Efficacy, which are thought to have an impact on science literacy. One should be sure that the same assessment tool is used in all of the groups in order to interpret the research findings correctly. If the sub-scales included in the PISA Student Questionnaire provide measurement invariance, this means that the same qualities are measured across different groups. Only the data obtained in this way can be comparable across groups. There are a number of measurement invariance studies that use gender and regions as variables (Bařusta & Gelbal, 2015; İmrol, 2017; Kıbrıřlıođlu, 2015; Ölçüođlu & Çetin, 2016; Uyar & Dođan, 2014). Different results have been reported in these studies.

1.5. Aim of the Study

Aim of the study is to examine the measurement invariance of Science Motivation and Self-Efficacy Model constructed by Instrumental Motivation to Learn Science and Science Self-Efficacy subscales in the PISA 2015 Student Questionnaire across different groups in Turkish sample. Measurement invariance of a model that was constructed with the scales used in PISA was examined in this study. The differences between the groups to which the scales are applied

can result from the real differences between the groups, whereas they may also arise from the scales themselves. In this study, it is aimed to provide evidence of the validity of inferences based on differences between groups by investigation of comparability.

Turkey’s variation in performance between schools is particularly large and is about twice the OECD average between-school variance. Therefore, it is thought that the investigation of the comparability of the region groups is especially important. In this context, it is expected that the study results will contribute to the literature of measurement invariance.

2. METHOD

2.1. Research Method

This study is a descriptive study as it aims at identifying whether Science Motivation and Self-Efficacy Model constructed by Instrumental Motivation to Learn Science and Science Self-Efficacy subscales in the Student Questionnaire of PISA 2015 is invariant by gender and statistical regions of Turkey.

2.2. Population and Sample

The sixth cycle of PISA, which is implemented by the Ministry of National Education, Directorate General for Measurement, Assessment and Examination Services, was carried out in 2015 in a computer-based way with the participation of 5895 students in Turkey. Population of PISA 2015 Turkey implementation was determined to be 15-year-old 1.324.089 students while the sample to be able to participate in the implementation was found to be 925.366 students (Ministry of National Education [MNE], 2016).

The sample of this study is composed of 4583 students that met the analysis assumptions. Some statistical regions have been excluded from the scope of this study in accordance with the results of CFA that was conducted for the variables of gender and statistical region separately. The information about the regions that have been excluded from the study is given in FINDINGS section. The figures of the sample with which the study was carried out are given in [Table 4](#).

Table 4. *Sample of the study by gender and statistical regions.*

Gender	n	%
Female	2318	50.6
Male	2265	49.4
Region	n	%
Istanbul (TR1)	837	26.3
Aegean (TR3)	562	17.6
West Anatolia (TR5)	456	14.3
Meditarrenean (TR6)	669	21
Middle East Anatolia (TRB)	179	5.6
Southeast Anatolia (TRC)	485	15.2

Note. TR1, TR3, TR5, TRB and TRC are the codes of the given regions.

2.3. Data Analysis

Because of the reason that all multi-variable statistical techniques are based on assumptions to a certain extent (Çokluk, Şekercioğlu & Büyüköztürk, 2016), the assumptions of (1) missing values, (2) extreme values, (3) normality, (4) multicollinearity, (5) linearity, (6) homogeneity and (7) sample size are tested.

All the analysis carried out within the framework of this study were done by taking the advantage of open-source R software (R Development Core Team, 2017). The scale items which were not answered by the students were accepted to be missing values in this study. As

there was not a specific pattern among the data of students who had missing answers (the item that had the highest missing value was the item coded ST129Q03TA with 374 data, which represented 6.3% of the data set), data gathered from 1063 students were excluded from the data set. It was examined in the study whether there was a univariate extreme value or not. For that purpose, z scores of 12 items in two different scales were calculated. It was found out that all z scores were between the values of -3 and +3. Afterwards, Mahalanobis distances of the variables were examined. The critical values of χ^2 were examined when $p < 0.001$. The value of 32.9095 was obtained for the degree of freedom (df) of 12 and 249 data that were above this value were excluded from the study. The values of skewness and kurtosis of the variables were taken into consideration in order to decide whether the data had a normal distribution or not. It was decided at the end of the statistical analysis that the data had normal distribution. The items that are the subject matter of this study were examined for multicollinearity analysis, and the item coded ST113Q02TA (MOT-2) had the highest VIF value of 4.107 among the items belonging to the Scale of Instrumental Motivation to Learn Science. Similarly, the item coded ST129Q05TA (SCIEFF-5) had the highest VIF value of 2.403 among the items belonging to the Scale of Science Self-Efficacy. Moreover, CI values of items were examined and the highest CI value among the items of the scale was found to be 22.594. With reference to the results of these analyses, it can be stated that there is not a problem of multicollinearity among the items used in this study (Gujarati, 1995; Kline, 2011). Scatter matrix was used for the linearity assumption in the study. It is expected for linearity that diagrams formed by the variable pairs should be ellipsis or ellipsis-like shapes, but linearity assumption could not be ensured in this study. Homoscedasticity can also be examined with Box-M test in multi-variable statistics. When Box-M test is found to be significant ($p < 0.05$), it can be concluded that homoscedasticity assumption cannot be ensured. Box-M test had significant results in this study. When multi-variate normality is ensured, the relation among variables can be said to be homoscedastic (Tabachnick & Fidell, 2013). It can be stated that statistics related to linearity and homoscedasticity assumptions are not enough to ensure normality assumption. Although there are debates about the adequate size of the sample, it is stated that the smallest sample for SEM analysis should be over 150 (Anderson & Gerbing, 1988; Kline, 2005). In this study, the assumptions of missing values, extreme values, normality, multicollinearity as well as size of the sample were examined. The data obtained from 4583 students ensuring the assumptions were found to be enough for SEM analysis. In this study, CFA was used to confirm Science Motivation and Self-Efficacy Model established with the two scales. The model and coefficients obtained according to the results of CFA are given in path diagram in **Figure 1**.

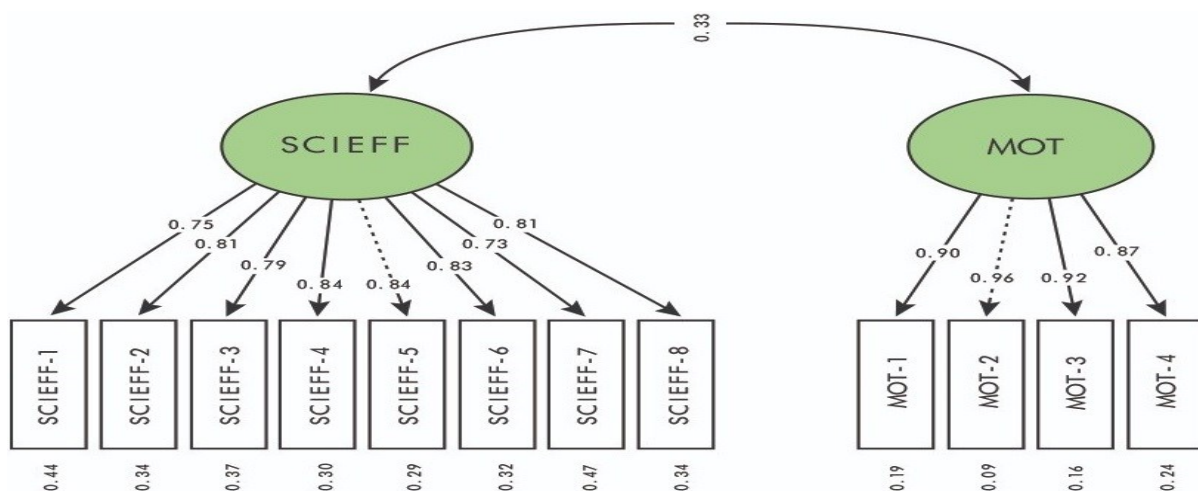


Figure 1. Science Motivation and Self-Efficacy Model path diagram.

In addition to the values of the model above, goodness of fit statistics of the model ($\chi^2 = 401.661$, $\chi^2/df = 7.578$, $RMSEA = .038$, $SRMR = .034$, $TLI = .995$, $CFI = .996$) were between acceptable intervals. Researchers can carry out measurement invariance tests through different methods according to the type of the scale and variables, whether the data set has a normal distribution or not, and the size of the sample. Categorical and ordinal variables are used in this study. Moreover, univariate normality assumption is ensured while multicollinearity and homoscedasticity assumptions are not ensured thoroughly. Because of the aforementioned reasons, the method of ULS estimation, which is reported to give good results for MGCFA under the roof of CFA and SEM, was preferred in this study (Forero, Maydeu-Olivares & Gallardo-Pujol, 2009; Koğar & Yılmaz Koğar, 2015). The analysis was carried out with the open-source statistical software called R ‘lavaan’ (Rosseel, 2012), which gives the opportunity to make estimations with the method of ULS.

Measurement invariance in MGCFA was examined by means of testing four nested hierarchical model or hypothesis. These four hierarchical models are respectively listed as ‘configural invariance’, ‘weak invariance’, ‘strong invariance’ and ‘strict invariance’ (Byrne, Shavelson & Muthen, 1989; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). While model-data fit was being examined at these stages, the values of χ^2 , χ^2/df , $RMSEA$ ($RMSEA < .08$), $SRMR$ ($SRMR < .08$), TLI ($TLI > .95$), CFI ($CFI > .95$), ΔCFI were taken into consideration (Browne & Cudeck, 1993; Hu & Bentler, 1999; Kaplan, 2000; Schermelleh-Engel, Moosbrugger & Müller, 2003; Schumacker & Lomax, 1996). When the fit was thought to be at an adequate level, the next step was started. First of all, the change in Chi-squared difference between two nested models ($\Delta\chi^2$) was used to find out whether comparing the two models would be significant or not, as is suggested by Hirschfeld and von Brachel (2014) in their study. However, $\Delta\chi^2$ is also a function of sample size and its usefulness has been discussed in many studies. Chi-squared difference test rejects the null hypothesis with too much power as the sample size increases. Cheung and Rensvold (2002) warned researchers that $\Delta\chi^2$ has less value in making practical decisions about measurement invariance. One of the alternatives to $\Delta\chi^2$ which was suggested by Cheung and Rensvold (2002) was the change in the CFI value (ΔCFI). Cheung and Rensvold (2002) stated the appropriate cut-offs for change in fit indices to determine. Wishing to extend Cheung and Rensvold’s research, Wu, Li and Zumbo (2007) provided extensive research about the practice of using the change in fit statistics to test for measurement invariance. Based on the studies mentioned above, ΔCFI was conducted to make final decisions about which stage of measurement invariance model holds. ΔCFI was examined among the two models that were more restricted when compared to each other. ΔCFI values smaller than or equal to -0.01; indicate invariance is not satisfied (Cheung & Rensvold, 2002).

3. FINDINGS

Measurement invariance was examined through MGCFA under the roof of SEM at this stage of the study. Before examining measurement invariance by gender and statistical regions, CFA was carried out for these variables separately and the model-data fit was investigated. According to the results of CFA, the goodness of fit statistics were acceptable for gender groups whereas they were outside of the acceptable range in West Marmara ($RMSEA = 0.00$, $TLI = 1.001$, $\chi^2 = 49.153$), East Marmara ($RMSEA = 0.00$, $TLI = 1.003$, $\chi^2 = 38.173$), Middle Anatolia ($RMSEA = 0.00$, $TLI = 1.001$, $\chi^2 = 39.439$), West Black Sea ($RMSEA = 0.00$, $TLI = 1.006$, $\chi^2 = 32.644$), East Black Sea ($RMSEA = 0.00$, $TLI = 1.007$, $\chi^2 = 37.088$), Northeast Anatolia ($RMSEA = 0.00$, $TLI = 1.012$, $\chi^2 = 29.361$), and so these regions were excluded from the scope of the study.

Measurement invariance was tested in accordance with the hierarchical sort order of configural invariance, weak invariance, strong invariance and strict invariance. Results of the two change detection tests and the value of ΔCFI were considered among the two invariance models. The examinations were kept going on until the stage where the study model provided the invariance in the related group.

3.1. Measurement Invariance of the Model by Gender

Within the scope of this model, a model was created with the two scales belonging to the PISA 2015 cycle and the measurement invariance of this model was examined by gender. At the end of the examination, which was done via the method of MGCFA, the Model of Science Motivation and Self-Efficacy met the conditions of all stages of invariance in PISA 2015 Turkey sample. According to this result, it can be stated that the item-factor structure in this study had an equal distribution among males and females. Also, factor loadings, variances, covariances and error variances are found out to be equal by gender. It can be concluded that the study model provided measurement invariance. According to Hirschfeld and von Brachel (2014), the change in the value of χ^2 between the models should be tested. The results of the test regarding the change in the value of χ^2 between the four hierarchical models are given in Table 5.

Table 5. Test of change in χ^2 between the four hierarchical models.

	χ^2	<i>df</i>	$\Delta\chi^2$	Δdf	<i>p</i>
Configural	426.641	106	-	-	
Weak	452.303	116	25.66	10	0.004
Strong	625.290	138	172.99	22	<.01
Strict	1114.154	140	488.86	2	<.01

Moreover, the change in χ^2 value of the models created in four different stages was examined. As is seen in Table 5, the change in Chi-square was found to be 25.66 and this change is statistically significant at the level of $p < .01$. Although the goodness of fit statistics were between acceptable intervals, the change in Chi-square was found to be significant. However, it is known that Chi-squared tests are highly sensitive to sample size. For this reason, the study was continued with the analyses of hierarchical models. The stages that were tested during the examination of measurement invariance by gender and the data belonging to the goodness of fit statistics are given in Table 6.

Table 6. Measurement invariance of the model by gender.

	χ^2	<i>df</i>	χ^2/df	<i>RMSEA</i>	<i>SRMR</i>	<i>TLI</i>	<i>CFI</i>	ΔCFI
Configural	426.641	106	4.02	0.036	0.035	0.995	0.996	-
Weak	452.303	116	3.90	0.036	0.036	0.995	0.996	0.000
Strong	625.294	138	4.53	0.039	0.036	0.994	0.994	-0.002
Strict	1114.154	140	7.98	0.055	0.036	0.988	0.988	-0.006

When the goodness of fit statistics and ΔCFI ($\Delta CFI > -.01$) values given in Table 6 are considered ($RMSEA < .08, SRMR < .08, TLI > .95, CFI > .95$), it can be concluded that measurement invariance is ensured between genders.

3.1. Measurement Invariance of the Model by Statistical Regions

The model created with the two sub-scales was examined among statistical regional groups. At the end of the examination that was done through the method of MGCFA, the Model of Science Motivation and Self-Efficacy met the conditions of all stages of invariance in PISA 2015 Turkey sample. Before having the data of the goodness of fit statistics of the models, the change

in the value of χ^2 between the models were tested. The results of the test regarding the change in the value of χ^2 between the four hierarchical models are given in [Table 7](#).

Table 7. Test of change in χ^2 among the four hierarchical models.

	χ^2	df	$\Delta\chi^2$	Δdf	p
Configural	446.719	318	-	-	-
Weak	501.565	368	54.846	50	0.296
Strong	698.656	478	197.091	110	<.01
Strict	997.433	488	298.77	10	<.01

The change in χ^2 value between the models which were obtained at the end of each stage and which were more restricted when compared to each other was examined. p value that was obtained at the stage of strong invariance is statistically significant at the level of 0.01. When [Table 7](#) is examined, it is clear that the change between the Chi-square values of weak and strong invariance models is found to be 197.091 and this change is statistically significant at the level of $p < .01$. Although the goodness of fit statistics were between acceptable intervals, the change in Chi-square was found to be significant. However, it is known that Chi-squared tests are highly sensitive to sample size. For this reason, the study was continued with the analyses of hierarchical models, as Wu, Li and Zumbo (2007) suggested. The stages that were tested during the examination of measurement invariance by statistical regions and the data belonging to the goodness of fit statistics are given in [Table 8](#).

Table 8. Measurement invariance of the model by statistical regions.

	χ^2	df	χ^2/df	RMSEA	SRMR	TLI	CFI	ΔCFI
Configural	446.719	318	1.40	0.028	0.042	0.997	0.998	-
Weak	501.565	368	1.63	0.026	0.044	0.997	0.998	0.000
Strong	698.656	478	1.46	0.030	0.045	0.997	0.996	-0.002
Strict	997.433	488	2.04	0.044	0.046	0.996	0.991	-0.005

When the goodness of fit statistics and ΔCFI ($\Delta CFI > -.01$) values given in [Table 8](#) are considered ($RMSEA < .08, SRMR < .08, TLI > .95, CFI > .95$), it can be concluded that measurement invariance is ensured among the statistical regions.

4. DISCUSSION and CONCLUSION

A model was created in this study with the two sub-scales included in the PISA 2015 Student Questionnaire and the measurement invariance of this model was examined among different groups. It was found out at the end of the study that the model could provide measurement invariance by gender and statistical regions. The Model of Science Motivation and Self-Efficacy shows that valid comparisons can be made among the scores of male and female students as well as students in different regions of Turkey. The study results comply with the results of the study carried out by Kıbrıslıoğlu Uysal and Akın Arıkan (2018), who examined the measurement invariance of the Science Self-Efficacy Scale used in PISA 2006 and 2015 by gender. The researchers reported at the end of the study that the scale that was used in the two PISA cycles met the conditions of all the stages of measurement invariance by gender. It seems necessary to be more careful while preparing assessment tools to be used in affective domains in science and to carry out measurement invariance examinations for these tools in different groups. Moreover, it would be helpful for researchers to make comparisons by gender and statistical regions with the scores obtained through the assessments in affective domains in science. On contrary to this, Uyar and Kaya Uyanık (2019) used the affective qualities about science and gender as variables in their study that was conducted with PISA 2015 Turkey

sample. The results of their analyses showed that their model provides only configural and weak invariance between genders in Turkey sample. In addition, Uyar and Doğan (2014) used same variables to investigate measurement invariance of a model on learning strategies in ‘Learning by strategies’ part of PISA 2009 Student Questionnaire. In their study, it is reported that while the model only provided configural and weak invariance stages in the groups of gender and school types, it provided all measurement invariance stages among regions of Turkey.

In this study, measurement invariance by gender and statistical regions of the created model was examined. As a result, it is found that strict measurement invariance was provided by either of the variables. Measurement invariance was examined through MGCFA under the roof of SEM in this research. It is possible to have different results when different methods are used in measurement invariance examinations (Yandı, Köse & Uysal, 2017). It is recommended that the research model should be tested with different methods. Moreover, four hierarchical nested models were tested while examining measurement invariance with MGCFA. There are different notions in the literature about the comparison of the nested models and the rejection of the null hypothesis. At those stages, $\Delta\chi^2$ and ΔCFI were used in addition to the goodness of fit statistics. It is possible to have different results, when $\Delta\chi^2$ and ΔCFI values are used together. In this study, ΔCFI values were used to make final decisions about which stage of measurement invariance model holds. ΔCFI values were examined among the two models that were more restricted when compared to each other, as Wu, Li and Zumbo (2007) suggested. More research results are needed about which criteria can be accepted while making comparisons in the investigation of measurement invariance. It is thought that simulation studies could provide this contribution to the field.

Although there are a number of studies about the cognitive domain in literature, the researchers that carry out in the affective domain state that they do research in a more virgin field (Boyd, Dooley & Felton, 2006). There is a need for study results that will contribute to developing scales in this field as well as analyzing the differences among groups in the country. It is believed that this study will contribute to the literature in this field with its results. In this study, measurement invariance was conducted only for gender and statistical regions. Future research on the invariance of the construct across different demographic groups would be concerning.

Acknowledgements

This paper was produced from first author’s master thesis.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Metehan Güngör  <https://orcid.org/0000-0003-4409-2229>

Kübra Atalay Kabasakal  <https://orcid.org/0000-0002-3580-5568>

5. REFERENCES

- Ağaç, G., & Masal, E. (2015). An investigation of the relation between 8th grade students’ beliefs, abstract thought and achievement; The case of mathematics. *International Online Journal of Educational Sciences*, 7(1), 134-144.
- Akyıldız, M. (2009). PIRLS 2001 testinin yapı geçerliliğinin ülkelerarası karşılaştırılması [The comparison of construct validities of the PIRLS 2001 test between countries]. *Van Yuzuncu Yil University Journal of Education*, 6(1), 18-47.

- Alivernini, F. (2011). Measurement invariance of a reading literacy scale in the Italian context: A psychometric analysis. *Procedia Social and Behavioral Sciences*, 15, 436-441.
- Anderson, C. J., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 3, 411-423.
- Asil, M., & Gelbal, S. (2012). PISA öğrenci anketinin kültürler arası eşdeğerliği [Cross-cultural equivalence of the PISA student questionnaire]. *Education and Science*, 37(166), 236-249.
- Ayvallı, M., & Biçak, B. (2018). An investigation into the measurement invariance of PISA 2012 mathematical literacy test. *European Journal of Education Studies*, 4 (11), 39-58.
- Bakan Kalaycıoğlu, D. (2015). The influence of socioeconomic status, self-efficacy, and anxiety on mathematics achievement in England, Greece, Hong Kong, the Netherlands, Turkey, and the USA. *Educational Sciences: Theory & Practice*, 15(5), 1-11.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W. H. Freeman.
- Başusta, B. N. (2010). Ölçme eşdeğerliği [Measurement equivalence]. *Journal of Measurement and Evaluation in Education and Psychology*, 1(2), 58-64.
- Başusta, B. N., & Gelbal, S. (2015). Gruplararası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA anketi örneği [Examination of measurement invariance at groups' comparisons: A study on PISA student questionnaire]. *Hacettepe University Journal of Education*, 30(4), 80-90.
- Batyra, A. (2017). Gender gaps in student achievement in Turkey: Evidence from Trends in International Mathematics and Science Study (TIMSS) 2015. *Education Reform Initiative & Aydın Doğan Foundation*.
- Bıkmaz, H. F. (2002). Fen öğretiminde öz-yeterlik inancı ölçeği [Self-efficacy belief instrument in science teaching]. *Educational Sciences & Practice*, 1(2), 197-210.
- Bircan, H., & Sungur, S. (2016). The role of motivation and cognitive engagement in science achievement. *Science Education International*, 27(4), 509-529.
- Boyd, L. B., Dooley, E. K., & Felton, S. (2006). Measuring learning in the affective domain using reflective writing about a virtual international agricultural experience. *Journal of Agricultural Education*, 47(3), 24-32.
- Britner, S. L. & Pajares, F. (2001). Self-efficacy beliefs, motivation, race, and gender in middle school science. *Journal of Women and Minorities in Science and Engineering*, 7(4), 271-285.
- Browne, M. W., & Cudeck, R. (1993). *Alternative ways of assessing model fit, testing structural equation models*, K. A. Bollen & J. S. Long (Eds.), Newbury Park, CA: Sage.
- Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105(3), 456-466.
- Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-cultural Psychology*, 34(2), 155-175.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233–255.
- Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2016). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları*. Ankara: Pegem.
- Dilts, R. (1998). Motivation. Retrieved August 17, 2019 from <http://nlpu.com/Articles/artic17.htm>
- Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, 5, 23-35.
- Ertürk, Z., & Erdiñç Akan, O. (2018). TIMSS 2015 matematik başarısı ile ilgili bazı değişkenlerin cinsiyete göre ölçme değişmezliğinin incelenmesi [The investigation of

- measurement invariance of the variables related to TIMSS 2015 mathematics achievement in terms of gender]. *Journal of Theoretical Educational Science, UBEK-2018*, 204-226.
- Forero, G. C., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A monte carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, 16, 625-641.
- Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education*, 25(4), 280-296.
- Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups?: testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, 44(11), 78-94.
- Gujarati, D. N. (1995). *Basic econometrics (3rd Ed.)*. New York, NY: Mc-Graw Hill.
- Gülleroğlu, D. H. (2017). PISA 2012 matematik uygulamasına katılan Türk öğrencilerin duyuşsal özelliklerinin cinsiyete göre ölçme değışmezliğinin incelenmesi [An investigation of measurement invariance by gender for the Turkish students' affective characteristics who took the PISA 2012 math test]. *Gazi University Journal of Gazi Educational Faculty*, 37(1), 151-175.
- Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229-244.
- Hirschfeld, G., & Brachel, v. R. (2014). Multiple-group confirmatory factor analysis in R – A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research & Evaluation*, 19(7), 1-12.
- Hu, L., & Bentler, M. P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55.
- İlhan, K. (2015). *Eğitimde pozitif psikoloji uygulamaları*. B. Ergüner Tekinalp & Ş. Işık (Ed.). Ankara: Pegem Akademi Yayıncılık.
- İmrol, F. (2017). Investigation of measurement invariance of motivation and self-belief constructs towards mathematics in PISA 2012 Turkey sample (Master thesis). Retrieved August 14, 2019 from <https://tez.yok.gov.tr/UlusalTezMerkezi/>
- Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8 user's reference guide*. Chicago: Science Software International.
- Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Newbury Park, CA: Sage.
- Karakoç Alatlı, B., Ayan, C., Polat Demir, B., & Uzun, G. (2016). Examination of the TIMSS 2011 fourth grade mathematics test in terms of cross-cultural measurement invariance. *Euroasian Journal of Educational Research*, 66, 389-406.
- Kıbrıslıoğlu, N. (2015). The investigation of measurement invariance PISA 2012 mathematics learning model according to culture and gender: Turkey – China (Shangai) - Indonesia (Master thesis). Retrieved August 14, 2019 from <https://tez.yok.gov.tr/UlusalTezMerkezi/>
- Kıbrıslıoğlu Uysal, N., & Akın Arıkan, Ç. (2018). Measurement invariance of science self-efficacy scale in PISA. *International Journal of Assessment Tools in Education*, 5(2), 325-338.
- Kline, R. B. (2005). *Principles and practices of structural equation modeling (2nd Ed.)*. New York: Guilford Press.
- Kline, R. B. (2011). *Principles and practices of structural equation modeling (3rd Ed.)*. New York: Guilford Press.

- Koçar, H., & Yılmaz Koçar, E. (2015). Comparison of different estimation methods for categorical and ordinal data in confirmatory factor analysis. *Journal of Measurement and Evaluation in Education and Psychology*, 6(2), 351-364.
- Koh, H. K., & Zumbo, D. B. (2008). Multi-group confirmatory factor analysis for testing measurement invariance in mixed item format data. *Journal of Modern Applied Statistical Methods*, 7(2), 471-477.
- Larson, L. M., Stephen, A., Bonitz, V. S., & Wu, T.-F. (2014). Predicting science achievement in India: role of gender, self-efficacy, interests, and effort. *Journal of Career Assessment*, 22(1), 89-101.
- Mellenberg, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research: Applications of Item Response Theory*, 13(2), 123-144.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543.
- Meredith, W., & Millsap, E. R. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, 57(2), 289-311.
- Ministry of National Education [MNE]. (2016). PISA 2015 national report. Ankara.
- Nagengast, B., & Marsh, H. (2014). *Motivation and engagement in science around the globe: Testing measurement invariance with multigroup structural equation models across 57 countries using PISA 2006*. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp 317-345). UK: Taylor & Francis.
- OECD [Organisation for Economic Co-operation and Development]. (2016). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematics and Financial Literacy*. PISA, OECD, Paris: OECD Publishing.
- OECD [Organisation for Economic Co-operation and Development]. (2017). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematics and Financial Literacy and Collaborative Problem Solving, revised edition*. PISA, Paris: OECD Publishing.
- Oliden, E. P., & Lizaso, M. J. (2013). Invariance levels across language versions of the PISA 2009 reading comprehension tests in Spain. *Psicothema*, 25(3), 390-395.
- Ölçüoğlu, R. & Çetin, S. (2016). TIMSS 2011 sekizinci sınıf öğrencilerinin matematik başarısını etkileyen değişkenlerin bölgelere göre incelenmesi [The investigation of the variables that affecting eight grade students' TIMSS 2011 math achievement according to regions]. *Journal of Measurement and Evaluation in Education and Psychology*, 7(1), 202-220.
- Polat, E., & Madra, A. (2018). PISA 2015 ve TIMSS 2015 ışığında Türkiye’de cinsiyete dayalı başarı farkı. *Education Reform Initiative & Aydın Doğan Foundation*, 1-18.
- Raykov, T., & Marcoulides, A. G. (2006). *A first course in structural equation modeling*. Lawrence Erlbaum Associates, Inc.
- Reise, P. S., Widaman, F. K. & Pugh, H. R. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin*, 114(3), 552-566.
- R Development Core Team (2017). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved August 14, 2019 from <http://www.R-project.org/>
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1-36.
- Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: an investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Front. Psychol.*, 7(110), 1-16.

- Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23-74.
- Schumacker, E. R., & Lomax, G. R. (1996). *A beginner's guide to structural equation modeling*. Manwah, NJ: Lawrence Erlbaum Associates.
- Schunk, D. H., Meece, J. L. & Pintrich, P. R. (2014). *Motivation in education: Theory, research and applications*. New Jersey: Pearson Education, Inc.
- Schunk, D. H., & Pajares, F. (2009). *Self-efficacy theory*. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of Motivation at School* (pp. 35-53). New York, NY: Routledge.
- Sireci, S. G., Patsula, L. N., & Hambleton, R. K. (2005). *Statistical Methods for Identifying Flaws in the Test Adaptation Process*. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. New Jersey, London: Lawrence Erlbaum Associates, Publishers.
- Somer, O., Korkmaz, M., Dural, S., & Can, S. (2009). Ölçme eşdeğerliğinin yapısal eşitlik modellemesi ve madde cevap kuramı kapsamında incelenmesi [Detection of measurement equivalence by structural equation modeling and item response theory]. *Turkish Journal of Psychology*, 24(64), 61-75.
- Steenkamp, J.B., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78-90.
- Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics*. Boston: Pearson.
- Taris, W. T., Bok, A. I., & Meijer, Z. Y. (1998). Assessing stability and change of psychometric properties of multi-item concepts across different situations: a general approach. *The Journal of Psychology*, 132(3), 301-316.
- Uyar, Ş., & Doğan, N. (2014). Türkiye örnekleminde öğrenme stratejileri modelinin farklı gruplarda ölçme değişmezliğinin incelenmesi [An investigation of measurement invariance of learning strategies model across different groups in PISA Turkey sample]. *International Journal of Turkish Education Sciences*, 3, 30-43.
- Uyar, Ş., & Kaya Uyanık, G. (2019). Fen bilimlerine yönelik öğrenme modelinin ölçme değişmezliğinin incelenmesi: PISA 2015 örneği [Investigation measurement invariance of learning model towards science: PISA 2015 sample]. *Kastamonu Education Journal*, 27(2), 297-507.
- Uzun, N. B., Gelbal, S., & Öğretmen, T. (2010). TIMMS-R fen başarısı ve duyuşsal özellikler arasındaki ilişkinin modellenmesi ve modelin cinsiyetler bakımından karşılaştırılması [Modeling the relationship between TIMSS-R science achievement and affective characteristics and comparing the model according to gender]. *Kastamonu Education Journal*, 18(2), 531-544.
- Yandı, A., Köse, A. İ., & Uysal, Ö. (2017). Examining measurement invariance with different methods: Example of PISA 2012. *Mersin University Journal of the Faculty of Education*, 13(1), 243-253.
- Vandenberg, J. R., & Lance, E. C. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-70.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25(1), 68-81.
- Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multigroup confirmatory factor analysis: a demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, 12(3), 1-26.
- Zedlin, A. L., Britner, S. L., & Pajares, F. (2007). A comparative study of the self-efficacy beliefs of successful men and women in mathematics, science, and technology careers. *Journal of Research in Science Teaching*, 45(9), 1036-1058.

The Development of a Scale to Evaluate Foreign Language Skills at Preparatory Schools

Recep S. Arslan ^{1,*}

¹Pamukkale University, Faculty of Education, Kınıklı Campus, 20070, Denizli, Turkey

ARTICLE HISTORY

Received: Dec 18, 2019

Revised: Apr 01, 2020

Accepted: May 05, 2020

KEYWORDS

Scale development,
English language teaching,
Teaching language skills

Abstract: The aim of the present study is to develop a valid and reliable scale evaluating the effectiveness of language preparatory programs in the acquisition of language skills. In the development of Foreign Language Skills Scale (FLSS) in this study, research sample consisted of 326 preparatory school students for the exploratory factor analysis (EFA) and 350 preparatory school students for the confirmatory factor analysis (CFA). Based on the data obtained from the first sample, an EFA was carried out on the FLSS. EFA has identified that 27 items of the scale have factor loads between 0.519 and 0.729, while they explain 65.376% of the total variance and are distributed under five factors. These factors are named as *writing skill*, *speaking skill*, *listening skill*, *core skills*, and *reading skill*. A CFA was applied on the data obtained from the second sample that consisted of 350 students. As a result of the CFA, it was confirmed that the FLSS consisted of 27 items and five factors. For all the items in the scale, item-subscale, item-test correlation coefficients and mean differences between the upper and the lower 27% of the participants were calculated, and it is determined that each item is consistent with not only the subscale it is under but also the whole test. In addition, the Cronbach's Alpha reliability coefficients of the total scale's and five sub-scales' internal consistency is quite high. The FLSS is expected to offer a comprehensive evaluation of the acquisition of four language skills in foreign language teaching programs.

1. INTRODUCTION

In the Turkish context there is an emerging need for individuals with a sound knowledge of at least one foreign language, which is usually English. With respect to higher education, the increasing demand for English, in turn, makes it necessary for the universities to offer intensive English programs being either compulsory or voluntary since either the medium of instruction at a number of state universities in Turkey is in English or some courses are offered in English. To this end, preparatory programs offer intensive English courses for tertiary level students before they are admitted to their own field of studies in faculties. Due to their crucial role in enabling tertiary level students to gain a proficient knowledge of English so that they can follow their courses in English effectively, it has, therefore, become essential to evaluate whether preparatory schools serve such ends or not (Coşkun, 2013; Ekşi, 2017).

The Higher Education Council (2016) responsible for the coordination of universities in the Turkish context states the aim of foreign language education as “to teach students basic

CONTACT: Recep S. Arslan ✉ rsarslan@pau.edu.tr 📍 Pamukkale University, Faculty of Education, Department of Foreign Languages Teaching, Kınıklı Campus, 20070, Denizli, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

principles of the foreign language that they are taught, to enhance their foreign language vocabulary, and to ensure that they can understand what they read and listen in a foreign language and they can express themselves orally or in writing” as declared in the Official Gazette dated 23.03.2016, with number 29662. However, curriculum design as well as its implementation and evaluation is left to universities. Regardless of compulsory or elective foreign language instruction offered in preparatory programs at tertiary level in Turkey, the Common European Framework of Reference for Languages (CEFR) that “describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively” (Council of Europe, 2001, p.1) is taken into consideration by almost all state and private foundation universities in designing preparatory programs. CEFR places students in six varying levels, including A1 level as breakthrough/beginner/basic user, A2 level as waystage/elementary/basic user, B1 level as threshold/intermediate/independent user, B2 level as vantage/upper intermediate/independent user, C1 level as effective operational proficiency/advanced user, and C2 level as mastery/proficiency/proficient user under understanding (listening and reading), speaking (spoken interaction/spoken production) and writing with illustrative scales for each skill (Council of Europe, 2001). North (2007) also suggests the existence of six levels plus mid-parts of the scale which came to be known as *plus levels* such as B1+ between levels B1 and B2 and B2+ between levels B2 and C1.

To date, the CEFR, which has been set out to be a framework for the elaboration of language syllabi or examinations, was noted to be the most useful for the planning and development of curricula as well as designing tests and certification (North, 2007). Therefore, evaluation of foreign language teaching programs based on the CEFR guidelines is crucial not only for administrators but also for English language practitioners to get a clearer understanding of and give feedback on the process as it would help administrators and instructors see success, reveal strengths and weaknesses, and make necessary improvements (Black, Harrison, Lee, Marshall, & Wiliam, 2004). It is, therefore, of paramount importance to evaluate language programs systematically and effectively in order to improve the quality expected from such efforts (Coşkun, 2013; Ekşi, 2017; Kiely & Rea-Dickins, 2005; Peacock, 2009).

A number of studies have attempted to evaluate preparatory programs at tertiary level in the Turkish setting from different perspectives. A few researchers evaluated language programs using Context, Input, Process and Product (CIPP) model and mostly reported that the language curriculum components were viewed positively; however, some improvements as to physical conditions, content, materials and assessment in the curriculum needed to be made (Akpur, Alcı, & Karataş, 2016; Coşkun, 2013; Karataş & Fer, 2009; Tunç, 2010). Moreover, Karcı-Aktaş and Gündoğdu (2020) applied ‘Bellon and Handler model’ to evaluate the English preparatory curriculum of a state university and stated the problems as lack of philosophy or goals of the English preparatory curriculum, inefficacy of the skills courses, communication problems between the administration and other participants, need to improve the physical facilities, need for professional English language teaching, and the necessity to involve all stake-holders in decision making processes.

Some other studies also evaluated language programs using survey techniques and also came up with similar results. Language programs were viewed as effective in general; however, content, course materials, and teaching equipments (Güllü, 2007), physical contexts and the necessity to develop communicative skills (Tekin, 2015), and objectives, teaching materials, assessment, evaluation, and general structure (Uysal, 2019) were stated among problematic issues. In addition, curriculum needed revision in line with students’ needs (Sağlam & Akdemir, 2018), curriculum needed to include academic or English for specific purposes courses (Balcı, Üğüten, & Çolak, 2018; Özkanal & Hakan, 2010) or technical English (Özkanal, 2009), a

preference towards teaching academic skills rather than general English was needed (Keser & Köse, 2019), there were some motivational and attendance problems (İşcan, 2017), speaking skill needed to receive more attention and also content, materials and activities were to be modified (Öner & Mede, 2015), speaking and listening skills considered weak also needed to be included more in the program (Yılmaz, 2009), and four language skills were to be tested through contextualised and communicative test items for backwash effect (Paker, 2013).

All these studies have attempted to evaluate foreign language teaching preparatory programs in terms of objectives, content, course materials, teaching equipments, physical contexts, and language components in general (Akpur, Alıcı, & Karataş, 2016; Balcı, Ügüten, & Çolak, 2018; Coşkun, 2013; Karataş & Fer, 2009; Özkanal & Hakan, 2010; Tunç, 2010); however, no study has yet attempted to investigate learners' success level in specific language skills; namely, speaking, reading, writing, or listening skills. Development of a scale to evaluate the effectiveness of preparatory programs in the acquisition of language skills has, therefore, been essential, and it is in this context that the present study aims to develop a scale which can be used to maintain a comprehensive overview of the process of acquisition of language skills within the field of foreign language teaching in an intensive modular preparatory program at a Turkish state university.

2. METHOD

This study used the basic survey model as a scale development study.

2.1. Context of the Study

This study was conducted in the School of Foreign Languages at Pamukkale University in Turkey in the 2018-2019 academic year. The preparatory school founded in 2004 has been offering intensive English language instruction since 2007-2008 academic year for about 1000 students each year. Students are enrolled in various departments such as Business Administration, International Trade and Finance, English Language Teaching, English Language and Literature, Textile Engineering, and Electric and Electronics Engineering, where medium of instruction is in English in either all or in some selected courses. With the idea that the modular system can be effective as students can be placed according to their level of English proficiency, and they can also receive appropriate education designed in line with the CEFR guidelines, the preparatory program has been based on a modular system since 2015-2016 academic year (Erarslan, 2019). Pamukkale University Preparatory Program is also based on the descriptors of the Common European Framework of Reference for Languages (CEFR) including A1, A2, B1, and B1+ levels. Students admitted to the program for at least two modules and at most four modules depending on their level of entry to the program are all supposed to complete the program at B1+ level. Volunteering students have the chance to attend B2 level as well. Each module lasts 8 weeks and the program runs 24 hours weekly with 192 hours of courses in total in a module. The weekly schedule includes such language skills courses as listening (2 hours), speaking (3 hours), writing (5 hours) and reading (5 hours) as well as a core language course for 9 hours. Students in the program go through formative and summative assessment through quizzes, performance assignments, one midterm examination and one final examination for each module.

2.2. Samples

The study was carried out during the Spring Term of 2018-2019 academic year. Convenience sampling method was used to reach the sample since all the participants were already attending the preparatory program and they were easy to reach for research purposes. In this study, different samples were chosen from different levels to conduct a scale development study.

During the scale development phase of the study 326 students studying at Pamukkale University preparatory school participated in exploratory factor analysis (EFA) conducted on the data obtained from the samples. Of the participants 111 (34%) were B1 level, and 204 (62.6) were B1+ level and 11 (3.4 %) were B2 level students. 142 (43.6%) were female and 183 (56.1%) were male students. 1 student (.3 %) did not mention the gender. The validity and reliability work of Foreign Language Skills Scale was obtained at the end of the pilot study conducted on the selected sample. For Comrey and Lee (1992), 300 is good for a sufficient sample size for factor analysis while Kline (1994) finds 200 individuals enough for a sample size with reliable factors.

A confirmatory factor analysis (CFA) was also carried out on the data obtained from the sample group of 350 students. The number of participants per item was more than 10 individuals as the scale consisted of 27 items. Of the participants 105 (29.2 %) were A1 level, 99 (27.5 %) were A2 level, 109 (30.3%) were B1 level, and 47 (13.1%) were B1+ level. 194 (53.9%) were male and 165 (45.8%) were female male students while 1 student (.3 %) did not mention the gender.

2.3. Data Collection

The validity and reliability analyses of the scale were conducted at the end of the pilot study with the selected sample.

2.3.1. Foreign language skills scale for preparatory schools

Foreign Language Skills Scale (FLSS) for Preparatory Schools was developed similar to the scaling approach based on grading totals developed by Likert (1932). During the scale development, first, literature on CEFR and evaluation of language programs was reviewed. Since review of the related literature did not show any measurement tools evaluating language skills in English Language Teaching Preparatory Programs based on CEFR, no specific sample was used while developing the scale items. Based on the review of literature, a number of 67 items were developed for the scale in line with the CEFR descriptors. An item pool of 67 items related to evaluation of language program was then submitted for the opinions of 35 experts in preparatory schools or English Language Teaching departments to consult their views on the development of items in order to validate the item pool of the scale.

During the pilot study stage, the items in the pools were examined by two English language teaching experts and one measurement and evaluation expert as well. According to expert views, researchers removed 34 items of the pilot scale as to the experts such items did not measure what was intended for or such items were found ambiguous. After the pilot study, there were 33 scale items based on 5-point Likert-type; namely, Strongly Agree (5), Agree (4), Neither Agree or Disagree (3), Disagree (2) and Strongly Disagree (1).

2.4. Data Analysis

In order to examine the validity and reliability analyses of the instrument, the data obtained from the first and second samples were uploaded onto the SPSS 22.00 and AMOS 16 software programs and analyzed. Firstly, for the purpose of determining the construct validity of the scale, KMO (Kaiser-Meyer-Olkin) and Bartlett's tests were carried out on the data obtained from the first sample to see the data's suitability for factor analysis. KMO value was obtained to determine if data structure suits factor analysis based on the sampling adequacy. Bartlett's Test of Sphericity was obtained to see the multivariate normal distribution of the data. In determining whether the data are appropriate for factor analysis, Kaiser-Meyer-Olkin value is to be greater than .70. For Bartlett's sphericity test, it was checked whether $p < .05$. .30 for the contribution value to common variance; .40 was used as a criterion for factor load value. While deciding the number of factors, the scree plot graph was used. Based on the obtained values, an EFA was carried out on the data.

Additionally, for each item in the scale, between item-subscale and the item-test correlation coefficient scores were calculated with a purpose to see whether each item was consistent with the subscale and whole scale. In addition, the statistical difference between item scores' means between groups of the upper and lower 27% were examined with 0.05 alpha level. Subsequently, a CFA was applied on the data obtained from the second sample. During the confirmatory factor analysis phase, data set of another 350 students was examined, and extreme and missing values were checked. In order to calculate the reliability coefficient of the scale, the Cronbach's Alpha reliability coefficient method was used.

3. FINDINGS

3.1. Findings on Validity

3.1.1. Exploratory factor analysis

Construct validity was applied to the measurement scale in order to determine the extent to which the FLSS as the measurement instrument can measure the variable it aims to measure without confusing it with other variables (Balçı, 2009; Gorsuch, 1983). To determine the construct validity of the FLSS, firstly, Kaiser-Meyer-Olkin and Bartlett's test analyses were conducted on the data collected from the first sample, and the values were obtained as KMO= 0.940; Bartlett's test value $\chi^2 = 5390.619$; $sd=351$ ($p=0.000$). As KMO values of higher than 0.60 are seen to be sufficient for factor analysis in the social and educational sciences (Büyüköztürk, 2002), it was decided that factor analysis could be conducted on the 33 item in the scale.

In Exploratory Factor Analysis, Principal Component Analysis (PCA) is a technique that is used to reveal whether or not the items in a scale could be divided into a lower number of factors that eliminate each other (Büyüköztürk, 2002). In order to classify the factors that were formed by collecting the items, Varimax orthogonal rotation technique was preferred as a rotation method since it was not expected that there would be a high degree of correlation among the factors that emerged in the principal component analysis (Kline, 1994). Items that have factor load values under 0.30 and those that are distributed under more than one factor with less than a difference of 0.10 between their factors loads need to be removed from the scale (Balçı, 2009; Büyüköztürk, 2002). As a result of the analyses in this study, the eigenvalues of the items had to be at least 1.00, while their factor loads at least 0.50. Items that were distributed under multiple factors were eliminated, 6 items were removed, and the analyses were carried out on the remaining 27 items.

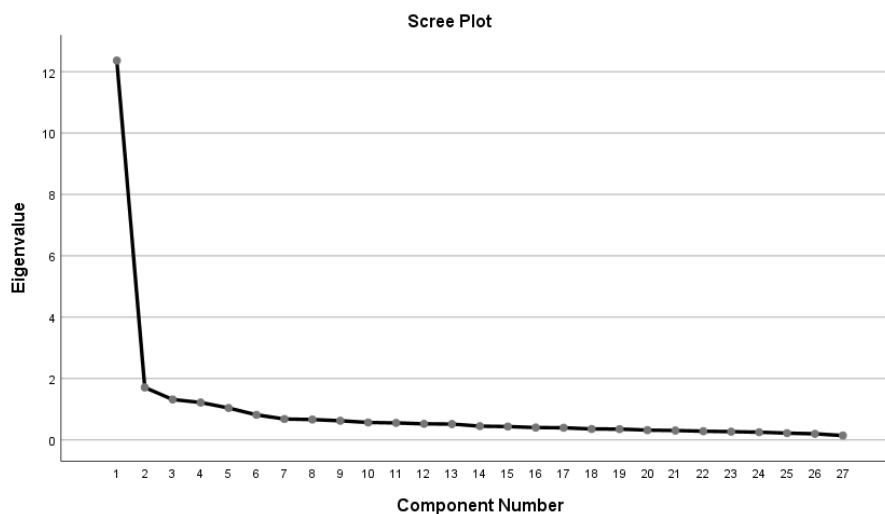


Figure 1. Eigenvalues based on the factors

As can be seen from the scree plot graph in Figure 1, 27 items can be collected under five factors. Without subjecting the remaining 27 items to rotation, it was found that the factor loads varied between 0.614 and 0.770. After subjecting the items to the Varimax orthogonal rotation technique, these factor loads were found to vary between 0.663 and 0.780. Additionally, it was identified that the items and factors in the scale explained 65.37% of the total variance. As it was stated that this ratio needs to be at least 40% (Kline, 1994; Scherer, Wiebe, Luther, & Adams, 1988), the obtained value was found sufficient. This finding obtained by EFA is shown in Figure 1 based on the eigenvalues. When Figure 1 is examined, it is seen that after five factors there is a routinized variation, and therefore, these factors have significant contribution to the variance.

Furthermore, the factors were named by examining the contents of the items gathered under these five factors. There were eight items in the first factor named *writing skill*. There were five items in each of the factors named *speaking skill*, *listening skill* and *reading skill*. In addition, there were 4 items in the factor named *core skills*. Table 1 presents findings on the item loads of the remaining 27 items based on the factors, factor eigenvalues and variance explanation ratios.

Table 1. FLSS common variances, item factor loads, variances explained by sub-scales and item analysis results

Items	Common Variance	Factor 1 Writing Skill	Factor 2 Speaking Skill	Factor 3 Listening Skill	Factor 4 Core Skills	Factor 5 Reading Skill
Q54	.696	.784				
Q51	.709	.726				
Q52	.691	.713				
Q56	.624	.708				
Q53	.594	.671				
Q55	.614	.658				
Q46	.671	.655				
Q47	.577	.609				
Q35	.705		.731			
Q36	.665		.710			
Q40	.683		.703			
Q37	.646		.667			
Q38	.729		.647			
Q31	.678			.731		
Q29	.662			.700		
Q30	.628			.623		
Q33	.673			.616		
Q28	.559			.608		
Q61	.710				.761	
Q63	.697				.741	
Q60	.707				.737	
Q65	.606				.609	
Q15	.711					.755
Q16	.734					.746
Q18	.571					.519
Q21	.519					.479
Q19	.591					.474
Eigenvalue		5.04	3.67	3.12	2.93	2.89
Explained variance		18.68	13.58	11.57	10.85	10.70
Total Variance			65.37			

As seen in Table 1, the factor loads of the items in the factor *writing skill* of the scale varied between 0.609 and 0.784. The eigenvalue of this factor in the general scale was 5.04, and its contribution to the general variance was 18.68%. The factor loads of the items in the factor *speaking skill* varied between 0.647 and 0.731. The eigenvalue of this factor was 3.67, and its contribution to the general variance was 13.58%. The factor loads of the items in the factor *listening skill* varied between 0.608 and 0.731. The eigenvalue of this factor was 3.12, and its contribution to the general variance was 11.57%. The factor loads of the items in the factor *core skills* varied between 0.609 and 0.761. The eigenvalue of this factor was 2.93, and its contribution to the general variance was 10.85%. And finally, the factor loads of the items in the factor *reading skill* varied between 0.474 and 0.755. The eigenvalue of this factor was 2.83, and its contribution to the general variance was 10.70%.

In addition, the relationship between the four factors in the FLSS was determined and for this reason, the correlations among the factors were checked. The findings are shown in Table 2.

Table 2. Correlation analysis results among the factors of the FLSS

Factors	Writing Skill	Speaking Skill	Listening Skill	Core Skills	Reading Skill
Writing Skill	-				
Speaking Skill	0.641**	-			
Listening Skill	0.667**	0.684**	-		
Core Skills	0.625**	0.642**	0.692**	-	
Reading Skill	0.606**	0.602**	0.592**	0.574**	-

** $p < 0.01$

As seen in Table 2, based on the correlation values among the factors of the FLSS, the five factors were found to be significantly related, while there was no problem of autocorrelation.

3.1.2. Item Discrimination

The correlation coefficients between the Item and Subscale correlation and Item and Test correlation were also calculated, and the discrimination rate of each item was determined in order to reveal the degree to which each item served the general purpose of the subscale it was in and the entire scale (Balçı, 2009; Baykul, 2000). Table 3 presents the items, item-factors, item-subscale correlations and item-test correlations.

As seen in Table 3, the item-subscale correlations were in the ranges of 0.665-0.748 for the first factor, 0.642-0.735 for the second factor, 0.593-0.683 for the third factor, 0.623-0.683 for the fourth factor and 0.608-0.692 for the fifth factor. Each item had a significant and positive relationship with the general scale ($p < 0.001$).

When the item-test correlation coefficients for the whole scale were examined, the lowest correlation value was found as 0.570, while the highest one was 0.739. Each item had a significant and positive relationship with the overall scale ($p < 0.001$). These coefficients that were calculated were the validity coefficients of all items, and they indicated the consistency of the items with the entire scale. In other words, these referred to the degree to which the scale served its general objective (Baykul, 2000).

The statistically significantly difference between item scores' means between groups of the upper and lower 27% were examined. It was found that all the items in FLSS were discriminated and the mean difference between the lower and upper groups was at a significant level of 0.05.

Table 3. Item discrimination analysis results

Initial Item No	Updated Item No	Item	Factor	Item-Subscale Correlation	Item-Test Correlation	Upper/Lower 27% <i>t</i>
Q54	26	I can enrich the text I write by using conjunctions	1	.748	.635	11.481**
Q51	23	I can write a paragraph.	1	.758	.692	16.394**
Q52	24	I can express my feelings and thoughts in writing	1	.743	.701	14.743**
Q56	28	I can write the sections of a paragraph such as topic sentence, supporting sentences, and concluding sentence.	1	.712	.649	12.908**
Q53	25	I can write coherent texts.	1	.665	.602	10.508**
Q55	27	I can use examples, quotes, or statistics to support my ideas when I write a paragraph.	1	.739	.696	14.032**
Q46	21	I can write sentences with meaning relations such as cause-effect, contrast, and comparison.	1	.689	.662	12.673**
Q47	22	I can rewrite a given sentence with the same meaning.	1	.692	.662	11.839**
Q35	12	I can answer any question when somebody asks me.	2	.738	.664	11.698**
Q36	13	I can communicate with non- native speakers of English.	2	.718	.641	11.847**
Q40	17	I can express personal information about myself.	2	.642	.619	13.305**
Q37	14	I can communicate with native speakers of English.	2	.705	.637	12.858**
Q38	15	I can participate in a conversation.	2	.755	.739	12.935**
Q31	10	I can deduce the meaning of a word I do not know from the context when I listen to a conversation	3	.658	.594	11.506**
Q29	8	During the listening process, when I am asked, I can catch the details such as who, where, and when,	3	.660	.627	12.841**
Q30	9	I can understand the main idea of any conversation I listen to.	3	.683	.669	12.927**
Q33	11	During the the listening process, I can catch phrases such as 'the door of the room', and 'students in the class'.	3	.623	.634	11.268**
Q28	7	I can take notes when somebody speaks.	3	.593	.570	12.200**
Q61	30	My reading skill has improved.	4	.697	.597	11.783**
Q63	32	My listening skill has improved.	4	.668	.594	12.565**
Q60	29	My speaking skill has improved.	4	.692	.617	13.599**
Q65	31	My writing skill has improved.	4	.608	.616	12.186**
Q15	1	I can guess the meaning of words I do not know in a reading text.	5	.689	.610	12.498**
Q16	2	I can answer questions related to a reading text.	5	.687	.637	12.773**
Q18	3	When answering a question about a reading text, I can easily find the section related to the question.	5	.666	.670	12.326**
Q21	5	I can deduce from a text I read.	5	.627	.646	10.455**
Q19	4	I can understand the main idea of a text I read.	5	.668	.702	13.602**

** $p < 0.01$

3.1.3. Confirmatory factor analysis

The dimensions of the FLSS were determined to consist of five factors as a result of the EFA. To confirm these factors, the scales that consisted of 27 items was applied on the second sample and a CFA was carried out on the data. CFA is based on the relationship among observable and unobservable variables and testing them as hypotheses (Pohlmann, 2004).

According to the results that were obtained, the χ^2/df ratio was calculated as 1.893. A χ^2/df ratio of 5 or lower is considered to be sufficient for model data fit (Schumacker & Lomox, 2004; Wang, Lin & Luarn, 2006). Moreover, a χ^2/df ratio of smaller than 3 shows a high model-data fit (Schumacker & Lomox, 2004). The χ^2/df value obtained as 1.893 in this study was a significant indicator that the measurement instrument had single dimension. Another important index, the RMR value was calculated as 0.021. It is known that the RMR index needs to be between 0 and 1 (Golob, 2003).

Other fit indices were also computed to evaluate the fit of the model. The calculated goodness of fit indices values were as: IFI=0.951; CFI=0.951; GFI=0.888; NFI=0.902; AGFI=0.864, and RFI=0.890. It is generally acceptable that the indices to be in the range of 0.80-0.90 and the values higher than 0.90 refer to a good fit (Yap & Khong, 2006; Wang et al., 2006). The RMSEA analysis result was determined as 0.049. RMSEA values of lower than 0.10 show an acceptable level of model-data fit, while those lower than 0.05 are an indicator of a good fit (Bayram, 2013). Based on the χ^2/df , RMSEA and RMR values obtained from the data in the study, it may be stated that the measurement instrument consisted of five factors. Figure 2 shows the standardized Structural Equation Modelling parameter values on the obtained findings.

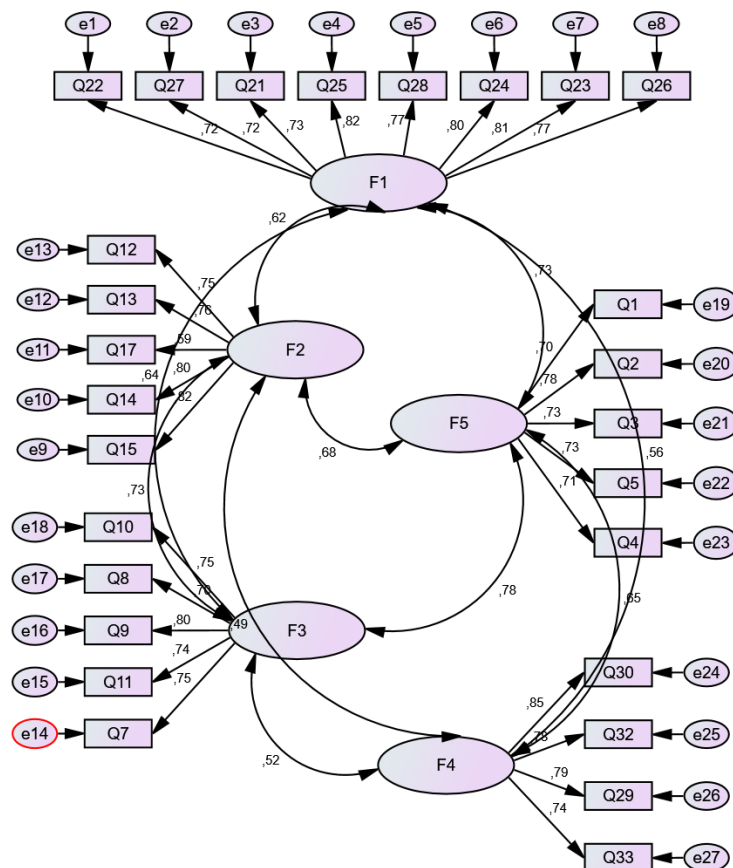


Figure 2. Confirmatory factor analysis results of the scale

As a result of the Confirmatory Factor Analysis, it was confirmed that the FLSS consisted of 27 items and five factors.

3.2. Findings on reliability

Reliability is a concept that is related to whether or not a measurement instrument provides the consistent and sensitive results in times of repeated application (Balçı, 2009; Baykul, 2000). As a result of the EFA, it was determined that the FLSS consisted of a total of 27 items and five factors. In order to identify the reliability indices of these five factors in relation to internal consistency, their Cronbach's Alpha reliability coefficients were calculated. The Cronbach's Alpha reliability coefficients of the factors were as 0.913 for *writing skill*, 0.879 for *speaking skill*, 0.838 for *listening skill*, 0.834 for *core skills* and 0.853 for *reading skill*. The Cronbach's Alpha value for the whole scale was 0.957.

The Cronbach's Alpha coefficient takes values in the range of 0.00 to 1.00. As the coefficient gets up to 1.00, the reliability of the measurement instrument increases, while as it gets closer to 0.00, the reliability decreases. In the educational and social sciences, in general, Cronbach's Alpha coefficients of 0.60 or higher are seen to be acceptable. On the other hand, the reliability indices used for preparing and applying psychometric tests is expected to be 0.70 or higher (Büyüköztürk, 2002). According to the findings obtained, the internal consistency coefficients for the factors and the entire scale were quite high in this study.

4. DISCUSSION and CONCLUSION

With a purpose to develop a scale in order to evaluate language skills in preparatory language teaching programs, 326 students studying at Pamukkale University preparatory school were asked to participate in the explanatory factor analysis phase of the scale development. Prior to the application of the scale, an item pool of 67 items was developed for the scale. An EFA was conducted on the data related to 67 items of the scale and 34 items that were found statistically insignificant were removed from the scale after calculations based on item-factor and item-test correlations. According to the results of EFA, it was decided that factor analysis could be conducted on the 33 items in the scale since Kaiser-Meyer-Olkin (KMO) and Bartlett's test values were obtained as KMO= 0.940; Bartlett's test value $\chi^2 = 5390.619$; $sd=351$ ($p=0.000$). As a result of the analyses, items that were distributed under multiple factors were eliminated, 6 items were removed, and the analyses were carried out on the remaining 27 items. A confirmatory factor analysis was carried out on the data obtained from the sample group of 350 students. The dimensions of the FLSS were determined to consist of five factors as a result of the EFA. To confirm these factors, the scale that consisted of 27 items was applied on the second sample and a CFA was carried out on the data. The χ^2/df value obtained as 1.893 in this study was a significant indicator that the measurement instrument had a single dimension. Another important index, the RMR value was calculated as 0.021. Other fit indices were also computed to evaluate the fit of the model. The calculated goodness of fit indices values were as: IFI=0.951; CFI=0.951; GFI=0.888; NFI=0.902; AGFI=0.864, and RFI=0.890. Based on the χ^2/df , RMSEA and RMR values obtained from the data in the study, the measurement instrument can be considered to consist of five factors.

In order to identify the reliability indices of these five factors in relation to internal consistency, their Cronbach's Alpha reliability coefficients were calculated. The Cronbach's Alpha reliability coefficients of the factors were as 0.913 for *writing skill*, 0.879 for *speaking skill*, 0.838 for *listening skill*, 0.834 for *core skills* and 0.853 for *reading skill*. The Cronbach's Alpha value for the whole scale was 0.957. These findings show the internal consistency coefficients for the factors and the entire scale quite high in this study.

Accordingly, in this particular study the Foreign Language Skills Scale that consisted of five factors and included 27 items was found to be a valid and reliable scale based on the statistical

data. This scale is expected to contribute to the field of foreign language teaching being a unique one that specifically addresses the evaluation of four main language skills in foreign language teaching programs. By using this scale, curriculum designers can evaluate the process of teaching language skills within the field of foreign language teaching and determine whether it is necessary to make changes, modifications or eliminations in the light of program goals and specific objectives. Since the main goal of foreign language teaching is to equip learners with an overall competency in understanding what they read and listen and also in expressing themselves orally or in writing in a foreign language, all items included in the scale would also help all those parties involved in such ventures to see how the actual practice fits the proposed goals of such programs in the acquisition of listening, speaking, reading and writing skills. Scores obtained as result of the application of this scale may either approve the programs as successful ones or may reveal the weaknesses and prompt immediate actions to tackle possible problems. Moreover, the application of the FLLS can also provide language instructors with valid data as to their own performance in teaching four language skills, and may, therefore, suggest whether they should revise their methods, materials, and activities.

However, the FLSS is not free from limitations. Since the scale consists of only 27 items, it assesses a limited number of subskills; thus, other scale attempts can be made to develop more comprehensive scales. Moreover, as the FLSS attempts to evaluate foreign language programs in terms of four language skills only, it excludes evaluation of other essential components of language programs such as the effect of course materials followed, course hours allocated, nature of programs (e.g. general or academic), teaching equipments used, physical contexts, roles of instructors and administrators, and involvement of stakeholders in decision-making processes of curriculum design. Therefore, more comprehensive scales that can investigate foreign language programs from such diverse points are timely.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Recep S. Arslan  <http://orcid.org/0000-0002-2475-5884>

5. REFERENCES

- Akpur, U., Alci, B., & Karataş, H. (2016). Evaluation of the curriculum of English preparatory classes at Yıldız Technical University using CIPP model. *Educational Research and Reviews*, 11(7), 466-473.
- Balcı, A. (2009). *Sosyal bilimlerde araştırma: yöntem, teknik ve ilkeler*. Ankara: PegemA Yayıncılık.
- Balcı, Ö., Durak Ügüten, S., & Çolak, F. (2018). Zorunlu İngilizce hazırlık programının değerlendirilmesi: Necmettin Erbakan Üniversitesi yabancı diller yüksekokulu örneği [The evaluation of compulsory preparatory program: The case of Necmettin Erbakan University school of foreign languages]. *Kuramsal Eğitim Bilim Dergisi*, 11(4), 860-893.
- Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: klasik test teorisi ve uygulaması*. Ankara: ÖSYM Yayınları.
- Bayram, N. (2013). *Yapısal eşitlik modellemesine giriş* (2. baskı). Bursa: Ezgi Kitabevi.
- Black, P, Harrison, C, Lee, C., Marshall, B., & Wiliam, D. (2004) Working inside the Black Box: assessment for learning in the classroom. *Phi Delta Kappan*, 86(1), 8-21.
- Büyüköztürk, Ş. (2002). *Sosyal bilimler için veri analizi el kitabı*. Ankara: PegemA Yayıncılık.

- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Coşkun, A. (2013). An investigation of the effectiveness of the modular general English language teaching preparatory program at a Turkish university. *South African Journal of Education*, 33(3), 1-18.
- Council of Europe (CoE). (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Strasbourg, France: Council of Europe. Cambridge University Press. Retrieved from <https://www.coe.int/en/web/common-european-framework-reference-languages>
- Ekşi, G. Y. (2017). Designing curriculum for second and foreign language studies. In A. Sarıçoban (Ed.), *ELT methodology*. Anı Yayıncılık: Ankara.
- Erarslan, A. (2019). Progressive vs modular system in preparatory school English language teaching program: A case of system change at a state university in Turkey. *Dil ve Dilbilimi Çalışmaları Dergisi*, 15(1), 83-97.
- Golob, T. F. (2003). Structural equation modeling for travel behavior research. *Transportation Research*, 37(1), 1-25.
- Gorsuch, R. L. (1983). *Factor Analysis*. Hillsdale: Lawrence Erlbaum Associates.
- Güllü, A. S. (2007). *An evaluation of English program at Kozan Vocational School of Çukurova University: Students' point of view* (Unpublished master's thesis). Çukurova University, The Graduate School of Social Sciences, English Language Teaching Department, Adana.
- İşcan, S. (2017) The efficacy of modular EFL syllabus in prep classes. *International Journal of Managment and Applied Science*, 3(8), 91-94.
- Karataş, H & Fer, S. (2009). Evaluation of English curriculum at Yıldız Technical University using CIPP model. *Education and Science*, 34, 47-60.
- Karcı-Aktaş, C. & Gündoğdu, K. (2020) An extensive evaluation study of the English preparatory curriculum of a foreign language school. *Pegem Eğitim ve Öğretim Dergisi*, 10(1), 169-214.
- Keser, A. D., & Köse, G. D. (2019). Determining exit criteria for English language proficiency in preparatory programs at Turkish universities. *The Online Journal of Quality in Higher Education*, 6(2), 45-49.
- Kiely, R. & Rea-Dickins, P. (2005). *Program evaluation in language education*. Pelgrave Macmillan.
- Kline, P. (1994). *An easy guide to factor analysis*. New York: Routledge.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 1-55.
- North, B. (2007). The CEFR illustrative descriptor scales. *The Modern Language Journal*, 91(4), 656- 659.
- Öner, G. & Mede, E. (2015). Evaluation of A1 level program at an English preparatory school in a Turkish university: a case study. *International Association of Research in Foreign Language Education and Applied Linguistics. ELT Research Journal*, 4(3), 204-226.
- Özkanal, Ü. (2009). *The Evaluation of English preparatory program of Eskisehir Osmangazi University Foreign Languages Department and a model proposal* (Unpublished doctoral thesis). Anadolu University, Eskişehir.
- Özkanal, Ü., & Hakan, A. G. (2010). Effectiveness of university English preparatory programs: Eskisehir Osmangazi University foreign languages department English preparatory program. *Journal of Language Teaching and Research*, 1(3), 295-305.
- Paker, T. (2013). The backwash effect of the test items in the achievement exams in preparatory classes. *Procedia-Social and Behavioral Sciences*, 70, 1463-1471.

- Peacock, M. (2009). The evaluation of foreign-language-teacher education programmes. *Language Teaching Research*, 13(3), 259-78.
- Pohlmann, J. T. (2004). Use and interpretation of factor analysis in the journal of educational research: 1992-2002. *The Journal of Educational Research*, 98(1), 14-23.
- Sağlam, D., & Akdemir, E. (2018). İngilizce hazırlık öğretim programına ilişkin öğrenci görüşleri [Opinions of students on the curriculum of English preparatory program]. *Yükseköğretim ve Bilim dergisi/Journal of Higher Education and Science*, 8(2), 401-409.
- Scherer, R. F., Wiebe, F. A., Luther, D. C. & Adams, J. S. (1988). Dimensionality of coping: factor stability using the ways of coping questionnaire, *Psychological Reports*, 62(3), 763-770. PubMed PMID: 3406294.
- Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling* (2nd ed.). NJ: Lawrence Erlbaum Associates, Mahwah.
- Tekin, M. (2015). Evaluation of a preparatory school program at a public university in Turkey. *Uluslararası Sosyal Araştırmalar Dergisi. The Journal of International Social Research*, 8(36), 718-733.
- The Higher Education Council (2016). Yükseköğretim Kurumlarında Yabancı Dil Öğretimi ve Yabancı Dille Öğretim Yapılmasında Uyulacak Esaslara İlişkin Yönetmelik [Regulation of the Higher Education Institutions on foreign language teaching and rules to be obeyed]. Retrieved from <https://www.mevzuat.gov.tr/Metin.Aspx?MevzuatKod=7.5.21475&MevzuatIliski=0&sourceXmlSearch=Y%C3%BCksek%C3%B6%C4%9Fretim%20Kurumlar%C4%B1nda%20Yabanc%C4%B1%20Dil%20%C3%96%C4%9Fretimi%20ve%20Yabanc%C4%B1%20Dille%20%C3%96%C4%9Fretim%20Yap%C4%B1lmas%C4%B1nda%20Uyulacak%20Esaslar>
- Tunç, F. (2010). *Evaluation of an English Language Teaching Program at a Public University Using CIPP model* (Unpublished masters' thesis). Ankara: Middle East Technical University. Retrieved from <http://etd.lib.metu.edu.tr/upload/12611570/index.pdf>
- Uysal, D. (2019) Problems and solutions concerning English language preparatory curriculum at higher education in view of ELT instructors. *International Journal of Contemporary Educational Research*, 6(2), 452-467.
- Wang, Y., Lin, H., & Luarn, P. (2006). Predicting consumer intention to use mobile service. *Information Systems Journal*, 16(2), 157-179.
- Yap, B.W., & Khong, K.W. (2006). Examining the effects of customer service management (CSM) on perceived business performance via structural equation modelling. *Applied Stochastic Models in Business and Industry*, 22, 587-605.
- Yılmaz, F. (2009). English language needs analysis of university students at a voluntary program. *The Journal of Social Sciences Research*, 4(1), 148-166.

The Development of Teachers' Knowledge of the Nature of Mathematical Modeling Scale

Reuben S. Asempapa ^{1,*}

¹School of Behavioral Sciences & Education, Penn State Harrisburg, 777 West Harrisburg Pike, W331D Olmsted Building, Middletown, PA 17057-4898.

ARTICLE HISTORY

Received: Dec 12, 2019

Revised: Apr 30, 2020

Accepted: May 15, 2020

KEYWORDS

Nature of modeling,
MMKS,
Factor analysis,
Scale development,
Teachers' knowledge

Abstract: This study addresses a gap in the literature on mathematical modeling education by developing the mathematical modeling knowledge scale (MMKS). The MMKS is a quantitative tool created to assess teachers' knowledge of the nature of mathematical modeling. Quantitative instruments to measure modeling knowledge is scarce in the literature partially due to the lack of appropriate instruments developed to assess such knowledge among teachers. The MMKS was developed and validated with a total sample of 364 K-12 teachers from several public-schools using three phases. Phase 1 addresses content validity of the scale using reviews from experts and interviews with knowledgeable teachers. Initial psychometric properties and piloting results are presented in phase 2 of the study, and phase 3 reports on the findings during the field test, factor structure, and factor analyses. The results of the factor analyses and other psychometric measures supported a 12-item, one-factor scale for assessing teachers' knowledge of the nature of mathematical modeling. The reliability of the MMKS was moderately high and acceptable ($\alpha = .84$). The findings suggest the MMKS is a reliable, valid, and useful tool to measure teachers' knowledge of the nature of mathematical modeling. Potential uses and applications of the MMKS by researchers and educators are discussed, and implications for further research are provided.

1. INTRODUCTION

For the past 30 years, mathematical modeling or modeling with mathematics education has experienced rapid growth at several educational levels across the world and especially in the USA. With the development and enactment of the Common Core new mathematics standards in the USA (National Governors Association Center for Best Practices [NGA Center] & Council of Chief State School Officers [CCSSO], 2010), the assessment guidelines for modeling education report (Consortium for Mathematics and Its Application [COMAP] & Society for Industrial and Applied Mathematics [SIAM], 2016), and modeling standards from other countries across the world including Australia, Germany, Japan, The Netherlands, and Singapore (Ang, 2015; Geiger, 2015; Ikeda, 2015; Kaiser, Blum, Borromeo Ferri, & Stillman, 2011), bring new mathematical practices that accentuate the relevance of mathematical modeling in mathematics education. This new promise of engaging students with mathematical

CONTACT: Reuben S. Asempapa, Ph.D. ✉ rsa26@psu.edu 📧 School of Behavioral Sciences & Education, Penn State Harrisburg, Teacher Education, 777 West Harrisburg Pike, W331D Olmsted Building, Middletown, PA 17057-4898.

ISSN-e: 2148-7456 /© IJATE 2020

modeling fundamentally requires teachers to be effective and well-informed about practices associated with mathematical modeling.

Mathematical modeling enables most of our students to value why we teach and learn mathematics and see the relevance and usefulness of mathematics around us (Asempapa & Foley, 2018; Blum & Borromeo Ferri, 2009). However, sample instruments measuring the knowledge of mathematical modeling among teachers remains scarce, thereby affecting the teaching, learning, and research of mathematical modeling education. The interest in this research study connected to teachers' knowledge of the nature of mathematical modeling stems from the relevance of mathematical modeling to teaching, learning, and doing mathematics not only in the USA, but also elsewhere in the world, where modeling is emphasized heavily in most mathematics curricula. Therefore, creating a tool to examine the know-how of teachers regarding the nature of mathematical modeling remains important considering the growing significance and popularity of mathematical modeling education all over the world.

As already mentioned, evidence of instrument validity and reliability regarding the knowledge of teachers on the nature of mathematical modeling is scant in the literature (Kaiser, Schwarz, & Tiedmann, 2010; Ziebarth, Fonger, & Kratky, 2014). Although a large body of literature exists on mathematical modeling in areas such as (a) the instruction, learning, and studying of modeling (Blum, 2015; Blum & Borromeo Ferri, 2009; Boaler, 2001; Organisation for Economic Co-operation and Development [OECD], 2003; Pollak, 2011); (b) pedagogies of mathematical modeling (Lesh, 2012; Lesh & Doerr, 2003); and (c) assessment of modeling tasks (Asempapa & Foley, 2018; Leong, 2012), the emphasis on theoretical and empirical research about assessment tools on the knowledge of teachers regarding the nature of mathematical modeling practices is limited. Recent emphasis on mathematical modeling has often ignored the important role quantitative measurement instruments play in conducting high quality research.

The need for valid measures and instruments with a clearly defined purpose and supporting validity evidence are fundamental to conducting high quality large-scale quantitative studies (Benjamin et al. 2017). The lack of validated quantitative instruments poses a challenge for most researchers in evaluating if a tool is appropriate for a study and whether it can produce accurate and reliable data (Benjamin et al. 2017; Ziebarth, Fonger, & Krathy, 2014). Thus, the development of the mathematical modeling knowledge scale (MMKS) is necessary and important, and it will provide researchers in the USA and the international community with a validated quantitative tool that is woefully lacking in the mathematics education literature. For these reasons, this current research study was planned to develop the MMKS—a measurement tool—that assesses teachers' knowledge of the nature of mathematical modeling to address a gap in this field. The primary goal in developing the MMKS was to identify questions that would be quicker and more suitable to answer yet would be powerful indicators of teachers' knowledge of the nature of mathematical modeling. Therefore, the purpose of this research was to create, examine the fidelity of, and verify the factor structure related to the development of the MMKS.

2. THEORETICAL FRAMEWORK and RELATED LITERATURE

2.1. The Nature of Mathematical Modeling and Its Process

Mathematical modeling usually means the ability to move back and forth between the real world and the mathematical world (Blum, 2015; Crouch & Haines, 2004; Pollak, 2011). Although mathematical modeling is highlighted and emphasized in most standards and curricula worldwide, missing in the literature is a single agreed-upon approach or definition; rather there are various approaches presented by authors of shared understandings (Lesh & Doerr, 2003; Kaiser & Sriraman, 2006). The various approaches are based on different theoretical

frameworks, and there is no consensus on approaches to mathematical modeling in the literature (Kaiser & Sriraman, 2006). For instance, in the GAIMME report modeling is defined as “a process that uses mathematics to represent, analyze, make predictions or otherwise provide insight into real-world phenomena” (COMMAP & SIAM, 2016, p. 8). According to Borromeo Ferri (2018), mathematical modeling is a process that involves transitioning back and forth between reality and mathematics and using mathematics to understand and solve a specified real-world problem.

Alternatively, the process of mathematical modeling can be described as using several learning situations; from deductively arranged authentic problem modeling activities (English & Sriraman, 2010) to inductively organized inquiry-based problem-solving activities leading the learner to formulate general patterns (Sokolowski & Rackly, 2011). Moreover, Blum and Borromeo Ferri (2009) described mathematical modeling as the “process of translating between the real world and mathematics in both directions (p. 45). Despite the lack of a direct and single agreed approach or definition for mathematical modeling, the convergent view of mathematical modeling can be described as a process that includes the following: (a) identify a problem in real life, (b) make choices and assumptions concerning the problem, (c) utilize a mathematical model, and (d) translate the results into the context of the original problem. A typical mathematical modeling process or procedure adapted for this study is shown in Figure 1.

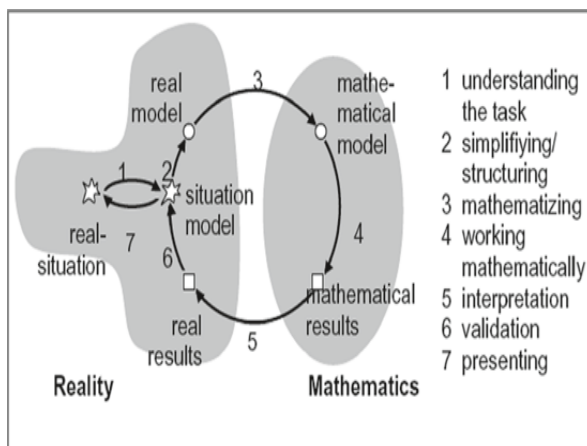


Figure 1. A typical mathematical modeling process (adapted from Blum & Leiss, 2007, p. 225).

Most mathematics educators have attempted to teach or communicate the concept of mathematical modeling through the mathematical modeling process. However, Perrenet and Zwaneveld (2012) argued that this is a challenge for instruction on mathematical modeling because of the lack of agreement about the mathematical modeling process regarding its essence, vision, and inherent complexity. For the purpose of this research study, the researcher’s conceptualization of mathematical modeling is based on the definition provided by Blum and Borromeo Ferri (2009). Despite the lack of unanimity on the approaches and definition of mathematical modeling in the literature, the mathematical modeling process demonstrates that individuals must solve a real-life problem utilizing their mathematical knowledge. A possible strategy for testing the efficacy of teaching and learning with mathematical modeling is through the creation of a scale that constitute the knowledge of teachers pertaining to the nature of mathematical modeling practices. In developing the scale, a series of phases were undertaken based on different samples. The phases contributed to construction of items that adequately reflected the domain of interest, relatively free of social desirability bias, and sufficiently represented the underlying construct. Therefore, these phases helped in the initial development and validation of the MMKS using a construct validity approach to scale development (DeVillis, 2017; Messick, 1995, 1998).

2.2. Teachers' Knowledge of the Nature of Mathematical Modeling

According to Ma (1999) “the quality of teachers subject matter knowledge directly affects student learning” (p. 144). Ponte and Chapman (2008) explained that a robust knowledge is insufficient for being an important or valuable teacher, however instructors or teachers with mediocre know-how makes teaching uneasy on students. This implies that it is essential for us to develop and improve the pedagogies of teaching mathematical modeling. Although there has been several research studies on the content knowledge of teachers in mathematics, the area of mathematical modeling is still scarce. Moreover, research indicates the knowledge of teachers regarding mathematical modeling is deficient, but appropriate and well-timed given the elevated attention on modeling practices in most mathematics standards and reports (COMAP & SIAM, 2016; NGA Center & CCSSO, 2010).

Philosophical and experimental knowledge into the pedagogy, instruction, and learning (Ma, 1999; Shulman, 1986, 1987) have highlighted the significance of the understanding of the content in teaching. Additionally, several documents have shown the variation in knowledge of teachers regarding the teaching of mathematics (Ball, 1990; Ma, 1999). The work of Hill, Schilling and Ball (2004), supports this argument, and this warrants a shift and modification in addressing teachers' knowledge and willingness on mathematical modeling. Because teachers' experiences contribute an important part in instruction and teaching (Lortie, 2002), their actions, dispositions, and attitudes toward mathematics and its relevance in the community, which involves mathematical modeling practices is important. Therefore, it is essential we design and develop research studies centered on teachers that focus on the content knowledge pertaining to the nature of mathematical modeling.

In recent years, the knowledge of teachers regarding mathematical modeling practices has received much discussion in the literature (Borromeo Ferri, 2018; Kaiser, Schwarz, & Tiedmann, 2010; Paolucci & Wessels, 2017). However, within mathematics education, defining the knowledge of mathematical modeling could seem as a complex construct because of the discrepancy in the components associated with the mathematical modeling process usually used as a criterion in teaching mathematical modeling. In conjunction with the above information, it seems important to identify and explain the phrase “knowledge of the nature of mathematical modeling.” Teachers knowledge of the nature of mathematical modeling was conceptualized as their understanding, interpretations, familiarizations, and minimal competencies associated with the Common Core standard of mathematical practice—model with mathematics—and teaching and learning of mathematical modeling (Borromeo Ferri, 2018; Blum, 2015; Lesh, & Doerr, 2003; NGA Center & CCSSO, 2010). Based on recent research and literature, the domain of the construct—knowledge of the nature of mathematical modeling—involved the mathematical modeling process, real-world connections, and mathematical modeling tasks (Blum & Leiss, 2007). Because establishing a questionnaire about mathematical modeling knowledge would be too broad and difficult to achieve with a simple scale, the manner in which teachers' comprehend or understand mathematical modeling was conceptualized as the familiarity with mathematical modeling applications, practices, and procedures. Therefore, Blum's and Leiss's (2007) modeling procedure or method was used as a contextual framework and domain for the development of the MMKS, which provides educators and researchers a heuristic guideline for exploring mathematical modeling.

3. PHASE 1: GENERATION and DEVELOPMENT OF ITEMS

3.1. Item Generation and Format

Phase 1 addressed issues regarding the evidence on face and content validity for the scale items that has the potential to assess the understanding of teachers about practices that engage students in mathematical modeling. In doing so, the researcher employed DeVellis's (2017)

recommendations in scale development. These recommendations include (a) measured construct; (b) generated items; (c) measurement scale format; (d) reviews by experts; and (e) incorporating valid items. Upon examination of relevant literature and standards (Ball, Thames, & Phelps, 2008; Blum & Borromeo Ferri, 2009; English, Fox, & Watters, 2005; Gould, 2013; Lesh, & Doerr, 2003; NGA Center & CCSSO, 2010; Pollak, 2011; Sriraman & English, 2010; Wolfe, 2013), an initial 22 items were generated to constitute the knowledge of teachers regarding practices about modeling with mathematics. The intention of this approach to selecting and generating these items was to promote an all-inclusive content-valid construct (Messick, 1995) as a strong content and applicable of the proposed knowledge of the nature of mathematical modeling. Sample scale items are provided in Appendix A.

To identify appropriate questions that fit the identified domain, experts and teachers from the Midwest in the USA were consulted at the inception of the scale. During the pilot phase, the researcher used 21 items, and the final design of the MMKS was reduced to 12 binary option (true or false questions), with an open-ended item, and other demographic items. The researcher used the true or false item type because this is the first attempt to develop an instrument of this kind to measure a complex construct—nature of mathematical modeling—which has the potential to generate quick but useful information from participants. Because the focus of this article was on scale development and evaluation of the items, no discussion on the open-ended question was presented. The 12 true or false items were graded with possible scores of 0–12.

3.2. Inclusion of Items and Content Validity

A further important aspect of the scale's development and validation was that the items were reviewed by experts. DeVellis (2017) explained that, the initiative to evaluate things for a newly constructed instrument should be extended to 6–10 experts. The experts evaluated each item's importance and suitability for the domain and offered suggestions and opinions on their view of the products and the MMKS. Ten experts from renowned midwestern universities reviewed the MMKS before the field test phase. These experts comprised three doctoral professors with modeling experience, three professors with analysis, assessment, and measurement skills, and four professors with diverse research interests in mathematical modeling at a reputable research-based university.

In order to assist in the iterative process of qualitative content analysis during the creation of the measure, comprehensive input was received from numerous experts regarding participant directions, scope of item sampling and item quality, and construction of the rating scale. All the experts offered suggestions for the revision of the items. Most of the experts and researcher came together to debate on the inclusion of items based on criteria and theoretical significance. After three iterations, we reached agreement on the final set of items. Before the initial version of the MMKS was submitted to a structured pilot study, a somewhat more detailed evaluation was conducted, using interviews with knowledgeable teachers (usually known as cognitive interview). (Fowler, 2014; Tourangeau, Rips, & Rasinski, 2000). During the cognitive interview, four teachers including primary, middle, and high school teachers were used to provide face/content validation for the items. Final design of the MMKS used for the field test demonstrated that the items were logically arranged, reasonable, comprehensible, and truly representative of the construct—knowledge pertaining to the nature of mathematical modeling.

4. PHASE 2: PILOT STUDY and PRELIMINARY PSYCHOMETRICS

4.1. Testing Items with a Development Sample

Trying out items is the exclusive approach of ensuring that the written survey items connect to the participants as expected (DeVellis, 2017). The goals of pre-testing guarantee that single items follow all the fundamental principles for quality questionnaire design. These goals

include the holistic testing of the questionnaire, ensuring smooth cohesion of procedures, maintaining appropriate survey routines, and developing excellent questionnaire codes (DeVellis, 2017). As a result, a try out for the MMKS was conducted via a pilot study with teachers from a big public-school in the midwestern part of the United States. After determining which relevant items to be used, the scale was then tried out or tested on a sample similar to the target population. The target population for this current study was K–12 teachers of mathematics, which included elementary (primary) middle and high school teachers. This population was suitable and appropriate for the current study because mathematical modeling is a standard of mathematical practice for these group of teachers. Table 1 demonstrates the MMKS design stages from the initial phase to the field-test stage.

Table 1. *MMKS from the Initial Phase to the Field-Test Phase*

Domain(s)	Development Stages		
	MMKS–Initial Version	MMKS–Pilot Study	MMKS–Field-Test
No. of items During (After)	22 (22)	21(13)	13
Demographic Items During (After)	18	19 (14)	14
Total items During (After)	40	40 (27)	27
Authenticity and quality	Items reviewed and conducting interviews.	Items revised and psychometric analyses.	Further psychometric analyses.

As per DeVellis (2017), the sample composition should be broad enough to remove the heterogeneity of the sample and aid with the appropriateness of the items. Experts have suggested several sample sizes for scale model pilot studies. Sample size from 25 to 75 was proposed by Converse and Presser (1986); Fowler (2014) suggested a size between 15–35; and when asking for a single point calculation, Johanson and Brooks (2010) suggested a size of 30 for the sample. While there are some risks involved with small sample size, pre-testing is better than not. Therefore, a size of the sample between 15 to 75 was considered appropriate during this phase.

Phase 1 findings resulted in the creation of a proposed collection of 21 items to evaluate the knowledge of teachers on the nature of mathematical modeling. These 21 items were produced by interviewing scholars knowledgeable and with theoretical and experimental experiences in survey production and mathematical modeling. Consequently, the next step was to investigate some of the psychometric measures of these 21 questions or items. Phase 2 therefore investigated whether these 21 items could reliably capture or operationalize the factor—knowledge of modeling—as suggested and conceptualized by the researcher. Phase 2 of this analysis was motivated by the following research questions.

Research Question 1 (RQ1): Depending on the eligible questions or items produced, which ones created maximum level of understanding on teachers’ knowledge of the nature of mathematical modeling, and should be part of the scale?

Research Question 2 (RQ2): Could the current 21 questions or items established via RQ1 and content validity processes reliably and validly operationalize the nature of mathematical modeling knowledge as suggested and conceptualized by the researcher?

4.2. Methods

4.2.1. Site and Participants

Participants enlisted for this investigation were mathematics teachers from a large government-funded school site in the U.S. Midwest. Maximum responses checked were 102, but 71 completed all survey items on the MMKS once data has been filtered and formatted. The response rate in the school district was about 19.6 percent compared to the number of mathematics teachers ($n = 520$). According to Converse and Presser (1986) having a size for the sample between 25 to 75 is adequate for trying out items, and Johanson and Brooks (2010) suggested a size of 30 for a sample, so the 71 respondents in this phase was considered adequate at this phase of the study. The majority of the 71 completed surveys were K–5 elementary teachers ($n = 36$, 50.7%) and were master’s degree holders ($n = 25$, 35.2%). The age range of respondents varied, about 77% were 35 years of age and older, and about 60% were Caucasian or White. As far as gender was concerned, 15% were classified as males and 85% as females. Such demographics represent a general trend in the USA of K–12 teachers of mathematics.

4.2.2. Data Collection and Analysis

Phase 2 utilized purposeful sampling, a non-probabilistic method of sampling. Data were gathered via a self-administered internet-based questionnaire This started the procedure of recognizing defined items, conceptual framework on modeling, applicable literature, and conceptual modeling information description. Surveys were sent by email to the study respondents and their answers were gathered and downloaded via the Qualtrics program. The researcher utilized both qualitative and quantitative methods such as elimination of redundant elements or items, measures of tendency and variability, reliability, and factor analyses to identify and evaluate the selected questions or items. Respondents responses were coded as incorrect response = 0 and correct response = 1. The total scale score was determined and the reliability of the internal consistency was evaluated by computing item-total-correlations.

4.3. Results

4.3.1. Item Analysis

Item review of the formatted data was carried out to determine the quality and authenticity of the items. The analyses involved evaluating the matrix of association or correlation, the overall correlations and the scale accuracy, quality and consistency. Established associations or correlations under .30 were supposed to be excluded (Field, 2009; Osterlind, 2010). Additionally, items which reduced the overall consistency in reliability in general should be excluded if conceptual deletion was appropriate. The outcome of the item analyses resulted in the retention of 12 items. All the items retained had theoretical and statistical significance with .30 and higher associations or correlations and, if removed, could not have increased Cronbach’s alpha as a whole. Phase 2 was intended to offer proof supporting the establishment of the MMKS. The 71 surveys containing the 12 items therefore produced a .80 Cronbach’s coefficient alpha, indicating that the MMKS offered accurate and functional measuring questions or items.

4.3.2. Exploratory Factor Analysis

Authenticity of the construct was achieved by examining homogeneity of the item via item-total correlation and factorial validity (DeVellis, 2017; Meyers, Gamst, & Guarino, 2013). Despite the relative small sample size of 71, the ratio was nearly 1:6 (Kline, 2000; Meyers, Gamst, & Guarino, 2013); consequently, during the pilot study, analysis of exploratory factor (EFA) was used to affirm the validity of the 12 MMKS items. The measure of accuracy for the sample (KMO = .81) and Bartlett’s test of sphericity ($p < .01$) demonstrated the applicability of exploratory factor analysis (Meyers, Gamst, & Guarino, 2013; Warner, 2013). The factorial

validity used principal axis factoring (PAF) with a rotation by varimax approach. PAF examines the interrelationship between objects, offers a basis for eliminating items, helps to classify structures and associated domains. (DeVellis, 2017; Meyers, Gamst, & Guarino, 2013).

Analysis of exploratory factor (EFA) was used to determine structures of one and two factors. However, after analyzing the items described in the factor loadings and variances of the component, the one-factor structure produced the best simple fit. Due to the theoretical significance, total variance accounted, the criteria of eigenvalue suggested by Kaiser (> 1.00) and the plot of the eigenvalues of factors “leveling off” of its own values, the one-factor approach was favored. Together the one-factor structures explained about 29.0% of the variance and was labeled *knowledge of modeling*. Using parallel analysis (O’Connor, 2000) as a standard methodology to evaluate the threshold for derived factors provided, a one-factor solution was also achieved explaining approximately 28.5 percent of the total variability. For every question or item from the MMKS, the factor loadings for the one-factor model was moderate to relatively high from .29 to .81.

5. PHASE 3: FIELD-TEST and FURTHER PSYCHOMETRICS

The pilot study and initial findings outlined in Phase 2 resulted in a reasonable collection of items to evaluate the knowledge of the nature of mathematical modeling among teachers. These items were generated by consensus between leading experts with expertise in mathematical modeling methods, modeling pedagogy, and measurement assessment. In this research effort, the next extra logical step was examining the psychometric measures of the 12 questions or items. Consequently, Phase 3 investigated whether these 12 items could effectively and validly operationalize the information collected on the MMKS. The research question in this study’s Phase 3 included:

Research Question 3 (RQ3): Could the current 12 items established via RQ2 and construct validity procedures reliably and validly operationalize knowledge on the nature of mathematical modeling as proposed and conceptualized by the researcher?

5.2. Methods

5.2.1. Site and Participants

The field test setting comprised of teachers in midwestern U.S. public school districts. Teachers teaching mathematics from Kindergarten to high school in the U.S. were the target group in this phase of the study. The field test consisted of nine districts that were among the largest in the USA of public schools and the study respondents teach mathematics to students. Additionally, the respondents lived within the identified school districts classified as rural, small-town, suburban, and urban.

A purposeful sampling technique was used during this phase to identify the sample frame and fit the geographic strata. Fourhundred seventy three teacher responses were obtained by the Qualtrics system, but after data cleaning and coding, 364 completed data points were utilized in analyzing the data. This sample size classified 21% as males and 79% as females. The mean age for the respondents was about 40.42 years ($SD = 10.84$). The oldest respondent was aged 67, and the youngest was aged 22. Roughly 66.5% ($n = 242$) of respondents were elementary teachers, 17.3% ($n = 63$) were teachers from middle grades, and the remaining 16.2% ($n = 59$) were teachers from the high school. The data was split into dual data points for both EFA and confirmatory factor analysis (CFA) because the completed data was large enough, which is a standard procedure for developing scales (Brown, 2015; Costello & Osborne, 2005). The EFA was allotted randomly to one hundred and eighty-two data set, and the remaining data ($n = 182$) was used for the CFA.

5.2.2. Data Collection and Analysis

As defined by Fowler (2014), the field test used a cross-sectional survey design. Data were obtained through a self-managed web-based survey that did not require respondents to exchange responses with an interviewer. This approach is likely to validate the compilation of confidential data (Fowler, 2014). The MMKS used 12 binary (true or false) items, one short answer question, and some demographic information to collect survey data (see Appendix A). The researcher gathered data through Qualtrics system and analyzed it using the statistical packages SPSS and SAS, widely utilized in social science research. The data analysis focused on the evaluation of the MMKS' structure (key factors) and psychometric measures (accuracy, reliability, authenticity, and validity) issues. The analyzes carried out included descriptive analysis, measures of normality, reliability analysis, item-total-correlation, EFA, and CFA.

5.3. Results

5.3.1. Item Analysis

Although the distribution of scores from the respondents was somehow skewed, it was assumed that there would not be much ceiling effect because of the large sample size. Overall, the average score of the respondents was ($M = 9.17$, $SD = 2.81$) and the mean female teacher score ($M = 9.31$) was substantially higher than the mean male teacher score ($M = 8.06$). An item discrimination index was not performed; however, the observation of the distributions of data between groups on the construct indicated the items correctly differentiated between the respondents. To evaluate the reliability of the questions or items, an item analysis was conducted. Correlations or associations between items estimated and below .30 were supposed to be excluded (Field, 2009; Osterlind, 2010). Additionally, items that usually reduced Cronbach's alpha should be excluded if conceptual deletion was acceptable.

The deletion benchmark for items was a correlation value below .30 (Osterlind, 2010), beginning with least correlations or associations. The correlation values analyzed indicated item Q3 had relatively low values in comparison to other items (see [Tables 2](#) and [Table 3](#)). Upon eliminating item Q3, however, the alpha value of Cronbach would only have improved by a value of .001. All 12 questions or items on the scale had item-to-total correlation values that exceeded .30 ($r = .30$). Therefore, because of their theoretical significance, all items were kept, with item-correlations higher than .30. The 364 surveys comprising the 12 items culminated in a Cronbach's alpha of .84, indicating that the MMKS produced accurate and functional measuring items. [Table 2](#) offers information on the MMKS items regarding Cronbach's alpha and item-total-correlations.

Table 2. Descriptive statistics on the MMKS scores—Field-Test

Item	<i>M</i>	<i>SD</i>	<i>SE</i>	ITC	α if Item is Deleted
Q1	.78	0.41	0.02	.51	.83
Q2	.87	0.34	0.02	.62	.82
Q3	.72	0.45	0.03	.39	.84
Q4	.72	0.45	0.02	.45	.83
Q5	.73	0.45	0.02	.41	.83
Q6	.82	0.39	0.02	.50	.83
Q7	.91	0.29	0.01	.77	.81
Q8	.78	0.41	0.02	.48	.83
Q9	.87	0.33	0.01	.67	.82
Q10	.75	0.44	0.02	.46	.83
Q11	.76	0.43	0.02	.47	.83
Q12	.84	0.37	0.02	.46	.83

Note: $n = 364$; ITC = item-total correlation

5.3.2. Exploratory Factor Analysis

An EFA was carried out to ascertain the number of common factors that are acceptable and acceptable MMKS indicators by the amount and scope of the factor loadings (Brown, 2015). The EFA used principal axis factoring (PAF) with a rotation by varimax approach. The KMO = .92 tested showed that the sample was appropriate for EFA (Field, 2009). A KMO near 1 with small partial correlation values demonstrate a common factor for the variables. The sphericity test by Bartlett was statistically significant ($p < .001$), which showed that the items were appropriate and suitable for performing EFA using a PAF approach.

An assessment of the extracted factor based on the Kaiser eigenvalue criteria (> 1.00) and the scree plot analysis showed no significant difference in the number of factors. Consequently, for further validity proof, a parallel analysis (O'Connor, 2000) was performed. Parallel analysis is a statistical method for facilitating the choice of factors in the EFA. This is achieved by comparing parallel randomly generated data points representing the number of original data items and factors. Afterwards, one derives eigenvalues from the generated random data points and contrasts it with the original. O'Connor (2000) explained that components or factors are kept provided the original i th eigenvalue is higher than the random data. The performed parallel analysis provided a one-factor solution accounting for 47.3% of the explained total variance. Examination of the factors revealed that all item factor loadings surpassed .30. Therefore, the one-factor solution with all 12 items were kept on the scale.

5.3.3. Factor Structure

Following Preacher's and MacCallum's (2003) recommendations, several measures were utilized in deciding on the factors to keep. The researcher employed three strategies: scree plot, Kaiser's eigenvalue test (> 1.00), and parallel analysis tests. (Horn, 1965). Visual examination of the factor item content was used for all evaluated solutions to verify that the extracted factor was relevant. The EFA scree plot of the 12 items showed a sharp decline until after the first factor. It supports the parallel analysis for the one-factor solution discussed in the previous paragraph. The factor extracted from the EFA had items with factor loadings exceeding .30 (Tabachnick & Fidell, 2007).

5.3.4. Confirmatory Factor Analysis

The factor structure was evaluated using the SAS PROC CALIS analytical technique for CFA. This was done to determine whether the measurement hypothesis was compatible with actual data during the field test using the MMKS scores. The data set had an item-to-respondent ratio of 1:15, ideal for CFA. CFA was performed on the data because CFA could determine the underlying factor structure of the scale and test the validity of the MMKS. According to Brown (2015), CFA's hypothesis-driven existence is a fundamental feature. By previous empirical analysis utilizing EFA during the try out phase, and based on theoretical grounds, a one-factor solution and underlying structure of the MMKS was tentatively defined. All expectations and assumptions for performing a CFA on the MMKS data was met. The assumptions included, adequate sample size, the right definition of a priori model, multivariate normality, multicollinearity, and the items-to-factor ratio.

Because the MMKS was one-dimensional, a CFA was performed for the entire scale of the overall measurement model. Due to the huge lack of agreement in the literature on preferred fit indices, the model fit was evaluated using these goodness-of-fit indices. (Hu & Bentler, 1999; Kline, 2000). The fit indicators also included the chi-square, Tucker-Lewis index (TLI), goodness of fit index (GFI), the root mean square error of approximation (RMSEA), the normed fit index (NFI), the comparative fit index (CFI), and the standardized root mean square residual (SRMR). A one-factor model was established on the basis of previous evidence and theory as well as the results of the EFA. The one-factor CFA model was subsequently carried out on the

12 items during the field test, with 182 valid results. The one-factor model fit measurement produced the following results: chi-square χ^2 (53) = 91.99, $p < .001$; $TLI = .96$; $GFI = .95$; $RMSEA = .05$ and $90\% CI = [.03, .06]$; $NFI = .92$; $CFI = .97$; and $SRMR = .04$.

Kenny (2015) stated that for CFA or structural equation models (SEM), CFI, TLI, RMSEA and SRMR are at the moment the most famous fit of measurements or statistics commonly reported. Additionally, the following are the recommended cut-offs that indicate a good model fit: $CFI \geq .90$; $TLI \geq .95$; $RMSEA < 0.08$; and $SRMR < 0.08$ (Kenny 2015; Kline, 2016). Thus, in comparison with the fit statistics commonly reported and as recommended by Kenny (2015), the construct's one-factor model fits the data from the above CFA results. This provided validity proof for the MMKS and validated the scale. The moderate to relatively high standardized factor loadings in Table 3 provided additional proof of validity for the MMKS items. This yielded extra inherent or intrinsic proof of construct authenticity for the instrument. The 12 items accounted for about 47% of the total MMKS variation, and all factor loadings were $> .30$.

Table 3. The standardized factor loading values on the MMKS—Field-Test

Items	SE	FL	p
Q1	0.05	.53	.00
Q2	0.04	.67	.01
Q3	0.05	.43	.00
Q4	0.05	.46	.01
Q5	0.04	.47	.00
Q6	0.05	.57	.01
Q7	0.03	.86	.01
Q8	0.05	.54	.00
Q9	0.03	.76	.01
Q10	0.04	.48	.00
Q11	0.05	.51	.01
Q12	0.05	.52	.01

Note: $n = 364$; FL = factor loadings; each FL value in the table was more than .30

6. DISCUSSION

Mathematical modeling is now a highly crucial component of mathematics education at different levels around the world and especially in the USA. Implementing modeling tasks and lessons during mathematics class have important influence on students doing mathematics. Recent literature indicates that an increasing number of teachers and researchers are involved in using and involving students in classroom mathematical modeling activities (COMAP & SAIM, 2016; Doerr, Ärlebäck, & Costello, 2014). Nonetheless, involving students with classroom activities and events that incorporate mathematical modeling practices is challenging for most teachers of mathematics. In this context, and to help comprehend the understanding teachers have about the nature of mathematical modeling, it became necessary to develop this instrument. Since there are no current instruments assessing the knowledge of teachers on mathematical modeling and in the spirit of creating a useful, reliable and credible scale, Messick's (1995, 1998) unified assessment of the legitimacy of validating a construct was implemented. Proof of validity in the Messick model implies gathering data for accurate analysis of scores or results that are intended for a particular purpose and at a specified time point (Downing, 2003).

The validity model of Messick illustrates construct validity because almost all social science evaluations deal with constructs — “intangible collections of abstract concepts and principles”

(Downing, 2003, p. 831)—such as the knowledge of the nature of mathematical modeling. Establishing the legitimacy of the construct requires a continuous procedure of collecting evidence. This indicates that the scores of the measurement procedure represent the anticipated structure. Cronbach (1998) defined the process as a justification for validation, which provides evidence for score interpretation. In this study, the validity of the construct was demonstrated utilizing content validity, consequential, factor structure, and factor analyses evidence. This was accomplished through the three phases to justify the worthiness and validity of the MMKS for future applications.

Although the development of the MMKS was evidently supported by theoretical significance, reliability, and factorial validity, and all 12 items were well correlated, only item (Q3) did not perform optimally under psychometric measures. The goal of item Q3 was to determine whether teachers could identify the difference between the modeling and problem-solving processes. Teachers' responses to this item was poor and this could have resulted in the weak correlations between item Q3 and the other items. However, the final MMKS's model retained 12 items because of their theoretical relevance. The Cronbach's alpha ($\alpha = .84$) of the MMKS was fairly decent for the unidimensional prototype during the field test. This means that the model determined 84% of the variation in the MMKS scores to reflect the construct being examined and an error rate of approximately 16% in the scores associated or identified with the MMKS. Therefore, based on these values, the proportion of variance on the scores in the MMKS that is due to extraneous or measurement error was relatively small, and it is within acceptable range (Field, 2009; Meyers, Gamst, & Guarino, 2013).

Additionally, this study investigated what the MMKS revealed about how teachers conceptualize the nature of mathematical modeling practices. Based on their MMKS scores, most of the teachers demonstrated reasonable levels of professional knowledge of the nature of mathematical modeling in this data set. In terms of gender, the researcher found female teachers to be relatively more knowledgeable about the nature of mathematical modeling practices than their male colleagues. Overall, the final one-dimensional model results of the MMKS showed a great model that suits the underlying proposed prototype by the one-factor and 12-item structure. The findings obtained from the content and construct validity works showed that the MMKS was reliable and useful. This research is the only first step in developing a quantitative measure to evaluate the knowledge of teachers regarding the nature of mathematical modeling. As far as the psychometric characteristics of MMKS are concerned, the supporting evidence confirms the proposed dimension, quality, and credibility of the construct. Although the study does not provide adequate specifics on convergent and discriminant validity, the MMKS was initially developed to achieve greater applicability with acceptable sample size.

7. CONCLUSION and IMPLICATIONS

The goal of this study was to generate reliable items and evaluate the factor structure of the MMKS in measuring teachers' knowledge of the nature of mathematical modeling. The approaches used in this work could be used in conjunction with other techniques such as dimensionality analysis, convergent and discriminant analyses. This can provide further confirmation evidence to boost awareness and implementation of the findings of this research to educational research. Future work should concentrate on how to build certain subscales that can capture or classify a specific contribution of different factors to explaining the knowledge of teachers in mathematical modeling practices. Additional collection of data must continue, particularly for convergent and discriminant validity. Other and future studies must analyze settings with a larger population of both public and private schools. Such data would help philosophically endorse the theoretical concepts of mathematical modeling and be more inclusive in the variety of measures and respondents.

Although the content, internal structure, and construct validity were determined during this study, establishing and defining certain aspects of the validity evidence for future research (generalizability and external validity) would be helpful and important. Because the MMKS has been developed with binary options, an item response theory (IRT) technique can be a wonderful complement to help establish the validity evidence of MMKS items in future research. The IRT methodology is based on the use of specific scale items to evaluate the construct being examined. The IRT approach claims that the characteristics of both the respondent and the item affect a person's reaction to an item. (Furr & Bacharach, 2014). Finally, future research can improve the MMKS using a Likert scale with multiple options for enough knowledge retention and interpretation.

Taking into account the information gathered from this research and provided in this article, the MMKS appears to be valuable in addressing interesting research concerns and information creation to expand the reach of mathematical modeling education. It is important that we build teacher's mathematical modeling knowledge to fulfill the school mathematics vision set out by the Common Core, national council of teachers of mathematics (NCTM), COMAP, SIAM, and other international standards. The finalized MMKS presented in this study represents a reliable and adaptable survey with which educators and researchers can monitor and assess both practicing and preservice teachers' development of their knowledge on the nature of mathematical modeling practices. Furthermore, for the successful integration and application of mathematical modeling into teaching school mathematics, the MMKS has the potential to support practicing teachers feel comfortable in their teaching.

This scale will allow researchers and mathematics educators to undertake mathematical modeling research using different methods for teacher programs and preservice courses. Although some work needs to be done with the MMKS in capturing teachers' comprehensive knowledge on mathematical modeling practices, the MMKS in its current form represents a useful and reliable tool for mathematics educators and researchers. The scale provides users with valuable information regarding the pedagogical content knowledge of mathematical modeling and its practices. This article offers a first step in the development of a quantitative tool that evaluates teachers' knowledge of the nature of mathematical modeling. It is a promising tool to guide researchers and educators as well as to inform teachers which areas they need to improve in their mathematical modeling practices. It is hoped that this scale will provide researchers and mathematics educators with the opportunity to accurately assess the knowledge of teachers about the nature of mathematical modeling practices.

Acknowledgements

A Dissertation Research Grant from the Patton College of Education at Ohio University funded or supported this research. In addition, I would like to thank Prof. Gregory D. Foley for his comments and suggestions in writing this article, and most of all for his continued mentorship as a mathematics educator.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Reuben Selase Asempapa  <http://orcid.org/0000-0003-4168-9409>

8. REFERENCES

- Ang, K. C. (2015). Mathematical modelling in Singapore schools: A framework for instruction. In N. H. Lee & K. E. D. Ng (Eds.), *Mathematical modelling: From theory to practice* (pp. 57–72). Singapore: World Scientific.
- Asempapa, R. S., & Foley, G. D. (2018). Classroom assessment of mathematical modeling tasks. In M. Shelley & S. A. Kiray (Eds.), *Education Research Highlights in Mathematics, Science, and Technology 2018* (pp. 6–20). Ames; IA. International Society for Research in Education and Science (ISRES).
- Ball, D. L. (1990). The mathematical understandings that prospective teachers bring to teacher education. *Elementary School Journal*, 90, 449–466.
- Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, 59, 389–407.
- Benjamin, T. E., Marks, B., Demetrikopoulos, M. K., Rose, J., Pollard, E., Thomas, A., & Muldrow, L. L. (2017). Development and validation of scientific literacy scale for college preparedness in STEM with freshmen from diverse institutions. *International Journal of Science and Mathematics Education*, 15(4), 607–623.
- Blum, W. (2015). Quality teaching of mathematical modelling: What do we know, what can we do? In S. J. Cho (Ed.), *Proceedings of the 12th International Congress on Mathematical Education: Intellectual and attitudinal challenges* (pp. 73–96). New York, NY: Springer.
- Blum, W., & Borromeo Ferri, R. (2009). Mathematical modelling: Can it be taught and learnt? *Journal of Mathematical Modelling and Application*, 1, 45–58.
- Blum, W., & Leiss, D. (2007). How do students and teachers deal with modelling problems? In C. Haines, P. Galbraith, W. Blum, & S. Khan (Eds.), *Mathematical modelling: Education, engineering and economics–ICTMA 12* (pp. 222–231). Chichester, United Kingdom: Horwood.
- Boaler, J. (2001). Mathematical modelling and new theories of learning. *Teaching Mathematics and Its Applications*, 20, 121–127.
- Borromeo, F. R. (2018). Learning how to teach mathematical modeling in school and teacher education. Picassoplatz, Switzerland: Springer. <https://doi.org/10.1007/978-3-319-68072-9>
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.
- Consortium for Mathematics and Its Applications [COMAP] and Society for Industrial and Applied Mathematics [SIAM] (2016). *Guidelines for assessment and instruction in mathematical modeling education*. Retrieved from <http://www.comap.com/Free/GAIMME/>
- Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Beverly Hills, CA: Sage.
- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10, 1–9. <https://doi.org/10.1.1.110.9154>
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.
- Crouch, R. & Haines C. (2004). Mathematical modeling: Transitions between the real world and the mathematical model. *International Journal of Mathematics Education Science Technology*, 35(2), 197–206.
- DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Thousand Oaks, CA: Sage.

- Doerr, H. M., Ärlebäck, J. B., & Costello Staniec, A. (2014). Design and effectiveness of modeling-based mathematics in a summer bridge program. *Journal of Engineering Education*, 103(1), 92–114.
- Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education*, 37, 830–837.
- English, L. D., Fox, J. L., & Watters, J. J. (2005). Problem posing and solving with mathematical modeling. *Teaching Children Mathematics*, 12, 156–163.
- English, L., & Sriraman, B. (2010). Problem solving for the 21st century. In B. Sriraman & L. D. English (Eds.), *Theories of mathematics education: Seeking new frontiers—advances in mathematics education* (pp. 263–290). New York, NY: Springer.
- Field, A. (2009). *Discovering statistics using SPSS*. Los Angeles, CA: Sage.
- Fowler, F. J. (2014). *Survey research methods* (5th ed.). Thousand Oaks, CA: Sage.
- Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction*. (2nd ed.). Thousand Oaks, CA: Sage.
- Geiger, V. (2015). Mathematical modelling in Australia. In N. H. Lee & K. E. D. Ng (Eds.), *Mathematical modelling: From theory to practice* (pp. 73–82). Singapore: World Scientific.
- Gould, H. T. (2013). *Teachers' conceptions of mathematical modeling*. Retrieved from <http://academiccommons.columbia.edu/item/ac:16149>
- Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal*, 105, 11–30.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria to fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1–55.
- Ikeda, T. (2015). Mathematical modelling in Japan. In N. H. Lee & K. E. D. Ng (Eds.), *Mathematical modelling: From theory to practice* (pp. 83–96). Singapore: World Scientific.
- Johanson, G. A., & Brooks, G. P. (2010). Initial scale development: Sample size for pilot studies. *Educational and Psychological Measurement*, 70(3), 394–400.
- Kaiser, G., Blum, W., Borromeo Ferri, R., & Stillman, G. (2011). Trends in teaching and learning of mathematical modelling—Preface. In G. Kaiser, W. Blum, R. Borromeo Ferri, & G. Stillman (Eds.), *Trends in teaching and learning of mathematical modelling: ICTMA14* (pp. 1–8). New York, NY: Springer.
- Kaiser, G., Schwarz, B., & Tiedemann, S. (2010). Future teachers' professional knowledge on modeling. In R. Lesh, P. L. Galbraith, C. R. Haines, & A. Hurford, (Eds.), *Modeling students' mathematical modeling competencies: ICTMA 13* (pp. 433–444). New York, NY: Springer.
- Kaiser, G., & Sriraman, B. (2006). A global survey of international perspectives on modelling in mathematics education. *ZDM—The International Journal on Mathematics Education*, 38(3), 302–310.
- Kenny, D. A. (2015). *Measuring model fit*. Retrieved from davidakenny.net/cm/fit.htm
- Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). New York, NY: Routledge.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.
- Leong, R. K. E. (2012). Assessment of mathematical modeling. *Journal of Mathematics Education at Teachers College*, 3, 61–65.
- Lesh, R. (2012). Research on models & modeling and implications for common core state curriculum standards. In R. L. Mayes & L. L. Hatfield (Eds.), *WISDOM^e monograph: Quantitative reasoning and mathematical modeling: A driver for STEM integrated*

- education and teaching in context* (Vol. 2, pp. 197–203). Laramie, WY: University of Wyoming.
- Lesh, R., & Doerr, H. M. (Eds.). (2003). *Beyond constructivism: Models and modelling perspective on mathematics problem solving, learning, and teaching*. Mahwah, NJ: Erlbaum.
- Lortie, D. C. (2002). *School teacher* (2nd ed.). Chicago, IL: University of Chicago Press.
- Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States*. Mahwah, NJ: Erlbaum.
- Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749.
- Messick, S. (1998). Test validity: A matter of consequences. *Social Indicators Research*, 45, 35–44.
- Meyers, L. S., Gamst, G., & Guarino, A. J. (2013). *Applied multivariate research: Design and interpretation* (2nd ed.). Los Angeles, CA: Sage.
- National Council of Teachers of Mathematics. (2014). *Principles to actions: Ensuring mathematical success for all*. Reston, VA: Author.
- National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. Washington, DC: Author. Retrieved from http://corestandards.org/assets/CCSSI_Math%20Standards.pdf
- O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, 32(3), 396–402.
- Organisation for Economic Co-operation and Development (2003). *The PISA 2003 assessment framework—mathematics, reading, science and problem solving, knowledge, and skills*. Paris, France: OECD Press.
- Osterlind, S. J. (2010). *Modern measurement: Theory, principles, and applications of mental appraisal*. Upper Saddle River, NJ: Pearson.
- Paolucci, C., & Wessels, H. (2017). An examination of preservice teachers' capacity to create mathematical modeling problems for children. *Journal of Teacher Education*, 68(3), 330–344.
- Perrenet, J., & Zwaneveld, B. (2012). The many faces of the mathematical modeling cycle. *Journal of Mathematical Modelling and Application*, 1(6), 3–21.
- Pollak, H. O. (2011). What is mathematical modeling? *Journal of Mathematics Education at Teachers College*, 2, 64.
- Ponte, J. P., & Chapman, O. (2008). Preservice mathematics teachers' knowledge and development. In L. D. English (Ed.), *Handbook of international research in mathematics education* (2nd ed., pp. 223–261). New York, NY: Routledge.
- Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, 2(1), 13–43.
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15, 4–14.
- Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57, 1–22.
- Sokolowski, A., & Rackley, R. (2011). Teaching harmonic motion in trigonometry: Inductive inquiry supported by physics simulations. *Australian Senior Mathematics Journal*, 24(2), 45–54.
- Sriraman, B., & English, L. D. (2010). Surveying theories and philosophies of mathematics education. In B. Sriraman & L. D. English (Eds.), *Theories of mathematics education*:

- Seeking new frontiers—advances in mathematics education* (pp. 7–34). New York, NY: Springer.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York, NY: Cambridge University Press.
- Warner, R. M. (2013). *Applied statistics: From bivariate through multivariate techniques* (2nd ed.). Thousand Oaks, CA: Sage.
- Wolfe, N. B. (2013). Teachers' understanding of and concerns about mathematical modeling in the common core standards (Doctoral dissertation). Retrieved from <https://search.proquest.com/docview/1432193702>
- Ziebarth, S., Fonger, N., & Kratky, J. (2014). Instruments for studying the enacted mathematics curriculum. In D. Thompson, & Z. Usiskin (Eds.), *Enacted mathematics curriculum: A conceptual framework and needs* (pp. 97–120). Charlotte, NC: Information Age.

APPENDIX A

SECTION 1: This section focuses on assessing teachers' knowledge of the nature of mathematical modeling. Consider how they can be used in the classroom. The items below describe the nature of mathematical modeling. Please respond to these items to the best of your ability.

Q1. The practice of mathematical modeling involves a single-step process.

- True
- False

Q2. Mathematical modeling is a process of translation between the real world and mathematics.

- True
- False

Q3. The mathematical modeling process is the same as mathematical problem solving

- True
- False

Q4. Mathematical modeling discourages students' interest in mathematics

- True
- False

Q5. Mathematical modeling involves problem posing before problem solving

- True
- False

Q6. Mathematical modeling connects mathematical representations.

- True
- False

Q7. Solving mathematical modeling tasks always require the use of technology

- True
- False

Q8. Mathematical modeling assists students in their social interactions

- True
- False

Q9. Mathematical modeling supports productive struggle in learning mathematics

- True
- False

Q10. Mathematical modeling tasks are of low cognitive demand.

- True
- False

Q11. Mathematical modeling facilitates meaningful mathematical discourse, which elicits evidence of student thinking.

- True
- False

Q12. Mathematical modeling is accomplished by simply covering the content standards in the Common Core State Standards for Mathematics (2010) marked with a ★

- True
- False

Q13. Write a brief definition of mathematical modeling.

SECTION 2: Demographic Information and Experience with Mathematical Modeling.

Q14. What is your gender?

- Male
- Female
- Other

Q15. What is your age in years? _____

Q16. What is your race or ethnicity? _____

Q19. In which grade level(s) do you teach? _____

Q20. What is your highest degree earned? _____

Q23. Do you teach mathematical modeling activities? _____

Q27. Please comment on your experiences with mathematical modeling.

Thank you for taking time out of your busy schedule to complete this questionnaire!

Parametric or Non-parametric: Skewness to Test Normality for Mean Comparison

Fatih Orcan ^{1,*}

¹Department of Educational Measurement and Evaluation, Trabzon University, Turkey

ARTICLE HISTORY

Received: Dec 06, 2019

Revised: Apr 22, 2020

Accepted: May 24, 2020

KEYWORDS

Normality test,
Skewness,
Mean comparison,
Non-parametric,

Abstract: Checking the normality assumption is necessary to decide whether a parametric or non-parametric test needs to be used. Different ways are suggested in literature to use for checking normality. Skewness and kurtosis values are one of them. However, there is no consensus which values indicated a normal distribution. Therefore, the effects of different criteria in terms of skewness values were simulated in this study. Specifically, the results of t-test and U-test are compared under different skewness values. The results showed that t-test and U-test give different results when the data showed skewness. Based on the results, using skewness values alone to decide about normality of a dataset may not be enough. Therefore, the use of non-parametric tests might be inevitable.

1. INTRODUCTION

Mean comparison tests, such as t-test, Analysis of Variance (ANOVA) or Mann-Whitney U test, are frequently used statistical techniques in educational sciences. The techniques used differ according to the properties of the data sets such as normality or equal variance. For example, if the data is not normally distributed Mann-Whitney U test is used instead of independent sample t-test. In a broader sense, they are categorized as parametric and non-parametric statistics respectively. Parametric statistics are based on a particular distribution such as a normal distribution. However, non-parametric tests do not assume such distributions. Therefore, they are also known as distribution free techniques (Boslaung & Watters, 2008; Rachon, Gondan, & Kieser, 2012).

Parametric mean comparison tests such as t-test and ANOVA have assumptions such as equal variance and normality. Equal variance assumption indicates that the variances of the groups which are subject to test are the same. The null hypothesis for this assumption indicated that all the groups' variances are equal to each other. In other words, not rejecting the null hypothesis shows equality of the variances. The normality assumption, on the other hand, indicates that the data were drawn from a normally distributed population. A normal distribution has some properties. For example, it is symmetric with respect to the mean of the distribution where the mean, median and mode are equal. Also, normal distribution has a horizontal asymptote (Boslaung & Watters, 2008). That is, the curve approaches but never touches the x-axis. With

CONTACT: Fatih Orcan ✉ fatihorcan@trabzon.edu.tr 📧 Trabzon University, Fatih Collage of Education, Room: C-110, Trabzon, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

normality assumption, it is expected that the distribution of the sample is also normal (Boslaung & Watters, 2008; Demir, Saatçioğlu & İmrol, 2016; Orçan, 2020). In case for comparison of two samples, for example, normality assumption indicates that each independent sample should be distributed normally. Departure from the normality for any of the independent sample indicates that the parametric tests should not be used (Rietveld & van Hout, 2015) since the type I error rate is affected (Blanca, Alarcon, Arnua, et al., 2017; Cain, Zhang, & Yuan, 2017). That is, parametric tests are robust in terms of type I error rate (Demir et al., 2016) and as the distribution of the groups apart from each other type I error rate raises (Blanca et al., 2017)

For independent samples, test of normality should be run separately for each sample. Checking the normality of the dependent variable for entire sample, without considering the grouping variable (the independent variable), is not the correct way. For example, if a researcher wants to compare exam scores between male and female students, the normality of exam scores for male and female students should be tested separately. If one of the groups is normally and the other is non-normally distributed the normality assumption is violated. Only if both groups' tests indicate normal distribution then parametric tests (i.e., independent sample t-test) should be considered. On the other hand, for one sample t-test or paired samples t-test (testing difference between pairs), normalities of the dependent variables are tested for entire sample at once.

Normality could be tested with variety of ways, some of which are Kolmogorov-Smirnov (KS) test and Shapiro-Wilk (SW) test. These are two of the most common ways to check normality (Park, 2008; Razali & Wah, 2011). Both tests assume that the data is normal, H_0 . Therefore, it was expected to not to reject the null (Miot, 2016). KS test is recommended to use when the sample size is large while SW is used with small sample sizes (Büyüköztürk et al., 2014; Demir et al., 2016; Razali & Wah, 2011). Park (2008) pointed that SW test is not reliable when sample size is larger than 2000 while KS is useful when the sample size is larger than 2000. However, it was also pointed that SW test can be powerful with large sample sizes (Rachon et al., 2012). Besides, it was stated that KS test is not useful and less accurate in practice (Field, 2009; Ghasemi & Zahediasl, 2012; Schucany & Tong NG, 2006).

In addition, KS and SW tests, other ways are also available for checking the normality of a given data set. Among them, few graphical methods are also available: Histogram, boxplot or probability-probability (P-P) plots (Demir 2016; Miot, 2016; Park, 2008; Rietveld & van Hout, 2015). For example, shape of the histogram for a given data set is checked to see if it looks normal or not. Even though it is frequently used, the decisions made based only on it would be subjective. Nevertheless, using histogram with other methods to check the shape of the distribution can be informative. Therefore, it will be useful to use graphical methods with other methods.

Another way to check the normality of data is based on checking skewness and kurtosis values. Although the use of skewness and kurtosis values are common in practice, there is no consensus about the values which indicate normality. Some suggest skewness and kurtosis up to absolute value of 1 may indicate normality (Büyüköztürk, Çokluk, & Köklü, 2014; Demir et al., 2016; Huck, 2012; Ramos et al., 2018), while some others suggest much larger values of skewness and kurtosis for normality (Iyer, Sharp, & Brush, 2017; Kim, 2013; Perry, Dempster & McKay, 2017; Şirin, Aydın, & Bilir, 2018; West et al., 1996). Lei and Lomax (2005) categorized non-normality into 3 groups: "The absolute values of skewness and kurtosis less than 1.0 as slight nonnormality, the values between 1.0 and about 2.3 as moderate nonnormality, and the values beyond 2.3 as severe nonnormality" (p. 2). Similarly, Bulmer (1979) pointed skewness, in absolute values, between 0 and .5 shows fairly symmetrical, between .5 and 1 shows moderately skewed and larger than 1 shows highly skewed distribution.

Standard error of skewness and kurtosis were also used for checking normality. That is, z-scores for skewness and kurtosis were used as a rule. If z-scores of skewness and kurtosis are smaller than 1.96 (for %5 of type I error rate) the data was considered as normal (Field, 2009; Kim, 2013). Besides, for larger sample sizes it was suggested to increase the z-score from 1.96 up to 3.29 (Kim, 2013).

Sample size is also an important issue regarding normality. With small sample size normality of a data cannot be quarantined. In an example, it was shown that sample of 50 taken from normal distribution looked nonnormal (Altman, 1991, as cited in Rachon et al., 2012). Blanca et al. (2013) examined 693 data sets with sample sizes, ranging between 10 and 30, in terms of skewness and kurtosis. They found that only 5.5% of the distributions were close to normal distribution (skewness and kurtosis between negative and positive .25). It was suggested that even with small sample size the normality should be controlled prior to analysis.

Since parametric tests are more powerful (Demir et al. 2016) researchers may try to find a way to show that their data is normal. Sometimes only SW or KS test are used while sometimes values such as skewness and kurtosis are used. In fact, based on Demir et al. (2016) study, 24.8% of the studies which test normality used skewness and kurtosis values while 24.1% of them used KS or SW tests. Even though the difference between the percentages is small, more researchers used skewness and kurtosis to check normality. There might be different reasons why researchers use skewness and kurtosis values to check normality. One of which might be related to get broader flexibility on the reference values of skewness and kurtosis. As indicated, different reference points on skewness and kurtosis were available in the literature. Therefore, it seems that it is easier for the researchers to show normality by using skewness and kurtosis values.

Based on the criteria chosen to check normality it is decided to use parametric or nonparametric tests. If the criterion is changed, the test to be chosen might also change. For example, if one use “skewness smaller than 1” instead of “z-score of skewness” criteria t-test instead of U-test might need to be used. In fact, normality test results might change with respect to the test which is used to utilized (Razali & Wah, 2011). Therefore, the aim of this study is to see how much difference might occur on decisions made on the used of t-test and U-tests under different skewness criteria. It was not aimed to point whether parametric or non-parametric tests are more or less useful then the other one. For this purpose, a simulation study was conducted with different design factors.

2. METHOD

2.1. Study Design Factors

Three different design factors were used to simulate independent sample testing proses. The first design factor was sample size. In order to simulate data from small to large sample four different values were considered (60, 100, 300 and 1000). It was indicated that sample size of 30 is small, while around 400 is large (Abbott, 2011, as cited in Demir et al., 2016). Later, percentages of the independent groups (25%, 50% or 75%) within the sample were changed and only one of the independent groups' normality was altered as the second design factor. For the third design factor, non-normality was added to the selected group. For non-normality, five conditions were utilized. The conditions were choosen to represent normal to non-normal distributions. The non-normality values were summarized at [Table 1](#). For example, under $Sk=0$, the skewness values were constrained to be between .00 and .10 while kurtosis values were between .00 and .20. For $SK=2*SE$ group, maximum values of skewness and kurtosis were constrained to be smaller than 1.96 time of their standard errors. These values were considered to represent normal ($Sk=0$), non-normal ($Sk=1$) and severe non-normal ($Sk=1.75$) distributions.

Data generation procedure was different for one sample and independent sample tests. First, the procedure for independent sample test was described. Namely, data were generated to simulated one factor structure which was estimated by five items. The values of the factor loadings were adapted from Demirdağ and Kalafat (2015) and set to .70, .78, .87, .77 and .53. The loadings represent small (.53) to large (.87) values.

2.2. Data Generation Procedure

To simulate independent sample testing, first, normally distributed factor scores with mean of 0 and standard deviation of 1 was generated in R. Then, Fleishman's power transformation method (Fleishman, 1978) was used to get non-normal factor scores. This is one of the recognized method to simulate non-normality (Bendayan, Arnau, Blanca & Bono, 2014). Only one of the two independent groups was non-normal.

Table 1. Skewness and Kurtosis Values Used for Data Generation

Condition	Skewness		Kurtosis	
	Min	Max	Min	Max
Sk=0	.00	.10	.00	.20
Sk=2*SE	1.70*SES	1.96*SES	1.50*SEK	1.96*SEK
Sk=1	.90	1.00	.80	1.00
Sk=1.5	1.40	1.50	1.50	2.50
Sk=1.75	1.60	1.75	5.00	-

Sk: Skewness; SES: Standard Error of Skewness, SEK: Standard Error of Kurtosis

For example, for 25% of the sample (group 1) was non-normal and 75% of the data (group 2) was normal. That is, for the specified percent of total sample was non-normally distributed and the rest of sample was normal. To ensure this structure, first a normal distributed data set was generated for a given sample size. After getting a normally distributed data set another data set with non-normal distribution was generated. Later these two data sets were merged to get one data set in which the grouping variable was also available. Before saving the merged data set equal variance assumption was tested in R. If the assumption was satisfied the merged data sets were saved for independent sample tests. In total of 500 data sets were generated for each condition. Therefore, totally 30,000 ($500 \times 4 \times 3 \times 5$) data sets were generated for independent sample tests.

For the dependent sample (one sample) test, the same factor structure was used. Fleishman's power transformation method was used to get non-normal factor scores. The simulated scores were considered as if they were score differences between pre-test and post-test results. For the dependent sample tests, only sample size and level of non-normality was used as design factors. The replication number was 500. Namely, 500 data sets were simulated for each of the given conditions. In total, 10,000 ($500 \times 4 \times 5$) data sets were generated for the dependent sample tests.

2.3. Data Analysis

The simulated data sets were also tested in R. To run the t-test and Mann-Whitney U test (U-test) *t.test* and *wilcox.test* functions were used. Type I error rates for both test was set to .05. In other words, significancies of the U-test and t-test were tested at the .05 alpha level. For independent sample t-test equal variance was assumed since it was controlled within data generation process. Simulated data sets were analyzed under both t-test and U-test. For empirical studies only the *p*-values of the tests were used to decide about the null hypothesis. Therefore, only the *p*-values for the t-test and U-test were checked under this study too. Consequently, the numbers of t-test and U-test which showed the same result based on the *p*-values (significant or not significant) were counted. In other words, *p*-values larger than .05 and

smaller than .05 for both t-test and U-test were counted. These results showed how much of conclusion made on the null hypothesis were the same between t-test and U-test.

3. RESULT / FINDINGS

3.1. One Sample Test Results

Based on the simulation conditions given above, one sample test results were given below. Based on the results, skewness (i.e., non-normality) of the data has effect on t-test and U-test. Figure 1 shows the discrepancy between one sample t-test and Mann-Whitney U test. As the skewness of the data was increased the dissimilarity between the tests was increased. For example, when skewness was 1, under sample size of 100, t-test and U-test were given different results for 10% of the time. However, under the same condition when the skewness was increased to 1.5 the difference was increased to 30%.

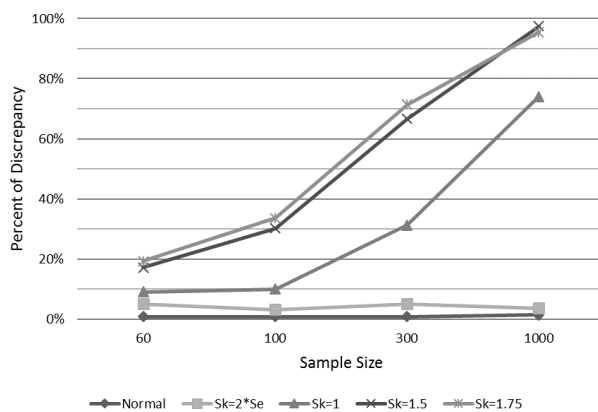


Figure 1. Discrepancy between t-test and U-test for One-Sample tests

The discrepancies were also dependent to sample sizes. As the sample size was increased the differences between t-test and U-test also increased for skewed data sets. For example, under the skewness of 1, when the sample size was increased from 100 to 300, the difference between the tests was increased from 10% to 31%.

When the data sets were normal the discrepancies between the tests were just about 1%. That is, when the data were normal, regardless of sample size, t-test and U-test gave the same results for 99% of the times. Figure 1 also shows the results for skewness equal to two times of its standard error (2*SES). Under this condition, the t-test and U-test were given the same results for 95% of the time on average. Table 2 gives the results of one sample tests in detail. For example, when sample size was 60 and skewness was 1.75 the discrepancy between t-test and U-test was 19%. As it is seen from the Table 2, for skewed data 2*SES rule gave the least discrepancies where the values were between 3 and 5 percents.

Table 2. Discrepancy Values (%) between t-test and U-test for One Sample Tests

Sample Size	Normal	Skewed Data			
		2*SES	1	1.5	1.75
60	1	5	9	17	19
100	1	3	10	30	34
300	1	5	31	67	71
1000	1	4	74	97	95

3.2. Independent Sample Test Results

Two independent groups were compared under this simulation study. Based on the results, sample size had an effect on the p -values for skewed data only as it was the case for one-sample test results. As the sample size was increased discrepancy between the p -values of tests also increased for skewed data. For example, under 25% of non-normal and skewness was 1, as the sample size was increased from 100 to 1000 the dissimilarity on the p -values increased from 4% to 20%. Left panel of Figure 2 shows the result for 25% of non-normal data while right panel shows the result for 50% (balanced) of non-normal data. Based on the results, under normally distributed data the p -values did not change much and the discrepancy was 2% at maximum. Thus, when the data were normal, regardless of sample size, t-test and U-test gave the same results for more than 98% of the times. Figure 2 also shows the results for skewness equal to two times of standard error of skewness ($2*SES$). Under this condition, the t-test and U-test gave the same results for more than 97% of the times in terms of p -values. Sample size did not affect the results under this condition. For example, as shown at left side of Figure 2, discrepancies for the p -values of the tests were about 3% for both sample sizes of 100 and 1000.

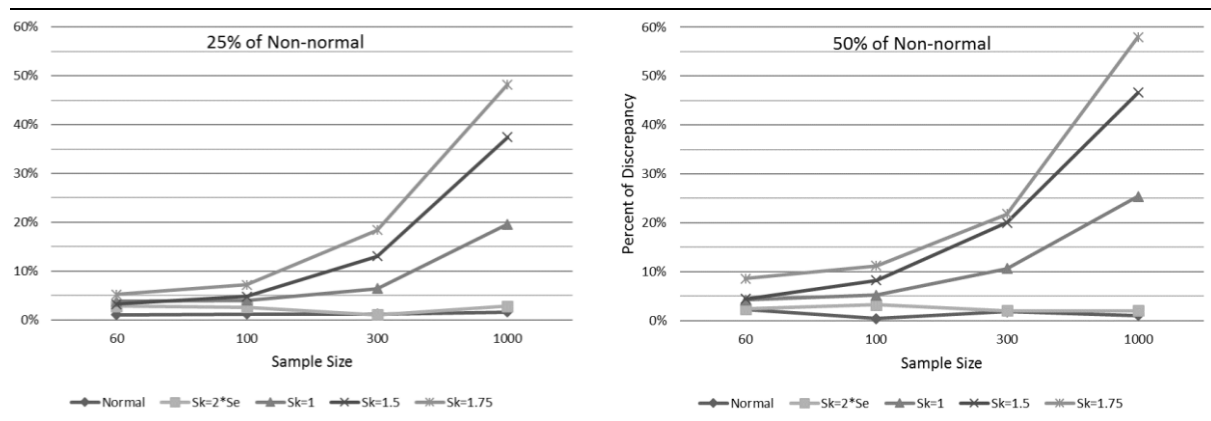


Figure 2. Discrepancy between t -test and U -test for 25% and 50% (balanced) of non-normal data

On the other hand, skewness also had effect on the p -values. As skewness was increased the difference between the p -values also increased. For example, on the left panel of Figure 2, as skewness was increased from 1 to 1.75 the difference between the p -values increased from 6% to 18%, under sample size of 300. Also, as sample size was increased the range of p -values also increased for skewed data. For example, the range was about 3% for sample size of 100 but 12% for 300 and 28% for 1000.

Percent of skewed data has also affected the results of t and U tests. Figure 3 shows the percent effects for sample sizes of 60 and 1000. When the sample size was small (60) the results of 25%, 50% and 75% non-normal data did not change much. Under these conditions, the discrepancies between the p -values were between 3% and 9%. However, as the sample size was increased, the effect of the percentages became more prominent. Interestingly, discrepancies between 25% and 75% of non-normality were similar. However, 50% of non-normality showed different and larger discrepancy as sample size was increased. On the other hand, when skewness was equal to two times its standard error ($2*SES$), percent of skewed data did not affect the results and discrepancies were between %1 and 3%. The results for independent tests were given at Table 3 in detail.

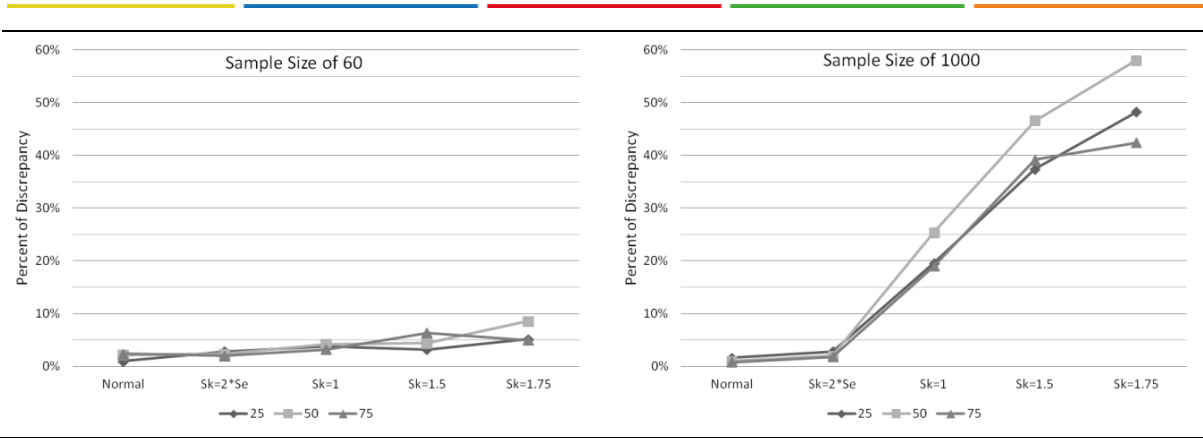


Figure 3. Discrepancy between *t*-test and *U*-test for sample sizes of 60 and 1000

Based on the results it was obvious that under skewed data sets *t*-test and *U*-test gave different results in terms of the *p*-values. The differences get clear as sample size and skewness of data were increased. However, under the 1.96 standard error rule, neither the sample size nor the percent of skewness were effective. Therefore, the results of this condition were investigated in detail.

Table 3. Discrepancy Values (%) between *t*-test and *U*-test for Independent Sample Tests

Sample Size	% of Skewness	Skewed Data				
		Normal	2*SES	1	1.5	1.75
60	25	1	3	4	3	5
	50	2	2	4	4	9
	75	2	2	3	6	5
100	25	1	3	4	5	7
	50	0	3	5	8	11
	75	1	2	2	7	13
300	25	1	1	6	13	18
	50	2	2	11	20	22
	75	1	3	8	15	15
1000	25	2	3	20	37	48
	50	1	2	25	47	58
	75	1	2	19	39	42

Table 4 shows average values of discrepancies between *t*-test and *U*-test with respect to SW tests. When the sample size was 60 about 92.8% of data was normal based on SW tests. Under this condition, when *t*-test was supposed to be used, 97.5% (90.5/92.8) of the *U*-test and when *U*-test was supposed to be used, 98.6% (7.1/7.2) of the *t*-test gave the same results. Even though SW tests results were different percent of similarities were alike across the sample sizes. For example, under sample size of 1000, when *t*-test was supposed to be used 97.5% (83.5/85.6) of the *U*-tests gave the same result in terms of *p*-values.

4. DISCUSSION and CONCLUSION

Checking the normality assumption is one of the critical steps for mean competition studies. Based on the results either parametric or non-parametric tests were considered to test mean differences. Literature suggests different approaches to check the assumption. Some of which are Kolmogorov-Smirnov test, Shapiro-Wilk test, checking skewness and kurtosis values or

basically looking the histogram of the dependent variables. Based on the test chosen the results of normality test might be different (Razali & Wah, 2011).

Table 4. *Discrepancy Average Discrepancy Values (%) for 2*SE Rule*

SW test Results		Sample Size			
		60	100	300	1000
Normal	Same	90.5	92.3	91.7	83.5
	Different	2.3	2.5	2.0	2.1
Non-normal	Same	7.1	5.0	6.3	14.3
	Different	.1	.1	.0	.1
Total of the Same		97.6	97.3	98.0	97.8

The use of skewness and kurtosis values to check normality is common in practice. Some suggest that the values can be up to as large as 2 in absolute values. On the other hand, standard errors of skewness and kurtosis were also used for normality tests. It was suggested that skewness and kurtosis values smaller than 1.96 times of their standard errors indicates normality (Kim, 2013; Field, 2009). However, there is no agreement on the values which indicate normality of a dataset. Therefore, this current study simulated different conditions to check the effect of skewness and kurtosis values on the decision made for mean comparison tests (a.k.a., t-test and U-test).

Based on the one-sample test results (see Table 2) when the data were normal or $Sk < 1.96*SES$, t-tests and U-tests were showed similar results with respect to *p*-values. Therefore, under these conditions, t-test can be used without any concerns. The results for normally distributed data were as expected. Nevertheless, under $Sk < 1.96*SES$ condition, *p*-values of t-tests and U-tests were worth to point again. When skewness is smaller than its 1.96 standard error, t-tests and U-tests indicated the same results. Therefore, if $Sk < 1.96*SES$, t-tests can be used to test mean differences. However, when skewness is around 1 or larger, the t-tests and U-tests pointed different conclusions. Therefore, test of normality has to be considered carefully. There needs to be other evidences to show normality of data. If no evidence is found for normality and skewness is around or larger than 1, given the limitation of this study, U-tests should be used to test mean differences.

Similar results were obtained for two-sample tests as well. That is, when the data were normal or $Sk < 1.96*SES$, t-tests and U-tests were showed similar results with respect to *p*-values. Therefore, if $Sk < 1.96*SES$, t-tests can be used to test mean differences. However, if no other evidence found and skewness is around or larger than 1, U-tests should be used to test mean differences. This suggestion especially important for larger sample sizes. As the sample size was increased the effect of skewness become clear and the discrepancies between t-test and U-test increased.

On the other hand, a more detailed results for the $1.96*SE$ rule were given at Table 4. Based on the table, when SW test indicated that the data was normal, on average 97.6% of the t-test and U-test were the same in terms of *p*-values. Similarly, when SW test indicated that the data was not normal, on average 99.0% of the tests were the same in terms of *p*-values. Therefore, in order to use t-test for mean comparison the $1.96*SE$ rule can be used. Regardless of SW test results, if skewness and kurtosis of a given dataset are smaller than their 1.96 standard errors (about 2 standard errors), t-test can be preferred over U-test. However, based on the results of the simulation, when skewness and kurtosis of a given dataset are larger than 1 another proof to show normality (e.g., Shapiro-Wilk) is needed. Therefore, if no other proof is granted non-parametric U-test should be used for mean comparison. In other words, “skewness around and larger than 1” rules should not be used to decide between t-test and U-test.

For example, let's say that, a researcher wanted to test if there is difference on math achievement scores between male and female students. For this purpose, about 300 student's scores were collected in a data set. The researcher tested normality of the scores for each gender groups by Shapiro Wilk test. Let's say that, the test indicated that the data were non-normal. After the test, the researcher checked the skewness and kurtosis values. The values were about 1.5. Since the values were smaller than 2, the researcher decided to use the parametric test (e.g., t-test). In this case, there is 16% of chance (average of 13%, 20%, 15%) that the results of the t-test were different than U-test. Therefore, using only the skewness and kurtosis values to decide about the normality of a data set is too risky. That means that if only skewness and kurtosis values are used for normality it is possible that researchers may decide to use a wrong method to test their hypotheses. For example, they may decide to use t-test when U-test is supposed to be used. Regarding that, as far as this study showed, as skewness and sample size increased t-test and U-tests gave different conclusions in term of rejecting H_0 . Therefore, it can be concluded that skewness and kurtosis values alone should not be used.

The literature also says that violation of normally assumption may not have serious effects on the results (Glass, Peckham, & Sanders, 1972, Blanca, Alarcon, Arnua, et al., 2017). However, uses of non-parametric tests are still very common in practice. Therefore, test of normally is still checked before mean comparison tests. The current study showed that the results changes based on the test chosen. The results of this study are limited with comparison of two means and predefined simulation conditions. Therefore, the results are limited to the conditions used within the study. For example, Ghasemi and Zahediasl (2012) and Kim (2013) and suggested the use of $2.58*SE$ or $3.29*SE$ rules under large sample size. Another study which simulated these conditions may also be useful. Under this study only the normality assumption was examined. Besides this, a simulation study where data are normal but equal variance assumption is violated can be informative as well.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Fatih Orcan  <http://orcid.org/0000-0003-1727-0456>

5. REFERENCES

- Abbott, M.L. (2011). *Understanding educational statistics using Microsoft Excel and SPSS*. United States: Wiley & Sons, Inc.
- Altman, D.G. (1991). *Practical statistics for medical research*. London: Chapman and Hall
- Bendayan, R., Arnau, J., Blanca, M.J. & Bono, R. (2014). Comparison of the procedures of Fleishman and Ramberg et al. for generating non-normal data in simulation studies. *Anales de Psicología*, 30(1), 364-371. <https://dx.doi.org/10.6018/analesps.30.1.135911>
- Bulmer, M. G. (1979). *Principles of statistics*. Mineola, New York: Dover Publications Inc.
- Büyüköztürk, Ş., Çokluk, Ö. & Köklü, N. (2014). *Sosyal bilimler için istatistik* (15th Edition). Ankara: Pegem Akademik.
- Blanca, M.J., Arnau, J., Lopez-Montiel, D., Bono, R. & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, 9(2), 78–84. <https://dx.doi.org/10.1027/1614-2241/a000057>
- Blanca, M.J., Alarcon, R., Arnua, J., Bono, R. & Bendayan, R. (2017) Non-normal data: Is ANOVA still a valid option? *Psicothema*, 29(4), 552-557. <https://dx.doi.org/10.7334/psicothema2016.383>

- Boslaugh, S. & Watters, P.A. (2008). *Statistics in a nutshell*. Sebastopol, CA: O'REILLY.
- Cain, M.K., Zhang, Z. & Yuan, K. (2017) Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behav Res*, 49, 1716–1735. <https://dx.doi.org/0.3758/s13428-016-0814-1>
- Demir, E., Saatcioğlu, Ö. & İmrol, F. (2016). Uluslararası dergilerde yayımlanan eğitim araştırmalarının normallik varsayımları açısından incelenmesi, *Current Research in Education*, 2(3), 130-148. Retrieved from <https://dergipark.org.tr/tr/pub/crd/issue/28292/300531>
- Demirdağ, S., & Kalafat, S. (2015). Meaning in life questionnaire (MLQ): The study of adaptation to Turkish, validity, and reliability. *İnönü Üniversitesi Eğitim Fakültesi Dergisi*, 16(2), 83-95. <https://dx.doi.org/10.17679/iuefd.16250801>
- Field, A. (2009). *Discovering Statistics Using SPSS* (3rd Edition). London: SAGE Publications Ltd
- Fleishman, A.I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43, 521-532. <https://dx.doi.org/10.1007/BF02293811>
- Ghasemi, A. & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *Int J Endocrinology & Metabolism*, 10(2), 486-489. <https://dx.doi.org/10.5812/ijem.3505>
- Glass, G., Peckham, P. & Sanders, J. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Measurement*, 42, 237-288.
- Huck, S.W. (2012). *Reading statistics and research* (6th Edition). Boston, MA: Pearson
- Iyer, D.N., Sharp, B.M. & Brush, T.H. (2017). Knowledge creation and innovation performance: An exploration of competing perspectives on organizational systems. *Universal Journal of Management*, 5(6), 261-270. <https://dx.doi.org/10.13189/ujm.2017.050601>
- Kim, H. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Open lecture on statistics (NA)*, 52-54. <https://dx.doi.org/10.5395/rde.2013.38.1.52>
- Lei, M. & Lomax, R.G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling*, 12(1), 1-27. https://dx.doi.org/10.1207/s15328007sem1201_1
- Miot, H.A. (2016). Assessing normality of data in clinical and experimental trials. *Jornal Vascular Brasileiro* 16(2) 88-91. <https://dx.doi.org/10.1590/1677-5449.041117>
- Orçan, F. (2020). Sosyal bilimlerde istatistik SPSS ve Excel uygulamaları (1st Edition). Ankara: Anı Yayıncılık.
- Park, H.M. (2008). *Univariate analysis and normality test using sas, stata, and spss*. Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University
- Perry, J.L., Dempster, M. & McKay, M.T. (2017) Academic self-efficacy partially mediates the relationship between scottish index of multiple deprivation and composite attainment score. *Frontiers in Psychology*, (8), NA. <https://dx.doi.org/10.3389/fpsyg.2017.01899>
- Razali N.M. & Wah, Y.B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests, *Journal of Statistical Modeling and Analytics*, 2(1), 21-33. Retrieved from: <https://www.researchgate.net/publication/267205556>
- Rachon, J., Gordan, M. & Kieser, M. (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples, *BMC Medical Research Methodology*, (12), 81. <https://dx.doi.org/10.1186/1471-2288-12-81>

- Ramos, C., Costa, P.A., Rudnicki, T., et al. (2018). The effectiveness of a group intervention to facilitate posttraumatic growth among women with breast cancer. *Psycho-Oncology*, (27), 258–264. <https://dx.doi.org/10.1002/pon.4501>
- Rietveld, T. & van Hout, R. (2015). The t test and beyond: Recommendations for testing the central tendencies of two independent samples in research on speech, language and hearing pathology. *Journal of Communication Disorders*, (58), 158-168. <https://dx.doi.org/10.1016/j.jcomdis.2015.08.002>
- Schucany, W.R. & Tony N.G., H.K. (2006). Preliminary goodness-of-fit tests for normality do not validate the one-sample student t. *Communications in Statistics – Theory and Methods*, 35, 2275-2286. <https://dx.doi.org/10.1080/03610920600853308>
- Şirin, Y.E., Aydın, Ö. & Bilir, F.P. (2018). Transformational-transactional leadership and organizational cynicism perception: physical education and sport teachers sample. *Universal Journal of Educational Research*, 6(9), 2008-2018. <https://dx.doi.org/10.13189/ujer.2018.060920>
- West, S.G., Finch, J.F. & Curran, P.J. (1995). Structural equation models with nonnormal variables: problems and remedies. In RH Hoyle (Ed.). *Structural equation modeling: Concepts, issues and applications*. Newbery Park, CA: SAGE.

Factors Affecting Academic Self-efficacy of Syrian Refugee Students: A Path Analysis Model

Hasibe Yahsi Sari ^{1,*}, Selahattin Gelbal ¹, Halil Ibrahim Sari ²

¹Department of Educational Sciences, Faculty of Education, Hacettepe University, Ankara, Turkey

²Kilis 7 Aralik University, Muallim Rifat Faculty of Education, Department of Educational Sciences, Kilis, Turkey

ARTICLE HISTORY

Received: Marc 11, 2020

Accepted: May 24, 2020

KEYWORDS

Resilience,
Self-efficacy,
Perceived social support,
Path analysis,
Life satisfaction
Syrian refugee

Abstract: In this study, the effect of resilience, perceived social support, life satisfaction and self-regulation variables on the academic self-efficacy of Syrian refugee undergraduate students were examined with a path analysis model. The sample consisted of Syrian undergraduate students living in Turkey. The sample of the research was randomly selected and participation was voluntarily. Data collection tools used were demographic information form, Arabic versions of academic self-efficacy, resilience, perceived social support, life satisfaction and self-regulation scales. In the data analysis, self-regulation and perceived social support selected as the exogenous variables, academic self-efficacy was selected as the endogenous variable, and resilience and life satisfaction were selected as the mediator variables. In the study, the direct and indirect effects from exogenous variables to academic self-efficacy were examined. The findings of the research revealed that self-regulation and perceived social support directly affected academic self-efficacy, life satisfaction had a mediating effect on perceived social support, and resilience had self-regulation. It is concluded that in order to increase the academic self-efficacy of refugee students, self-regulation and social support from the society should be increased, as well as life satisfaction and resilience against difficulties.

1. INTRODUCTION

The political internal disturbances, called the Arab Spring, which started at the end of 2010, spread to many Middle Eastern countries, and finally showed its effects in Syria. Political events affected Syria deeply, and the country completely went to civil war. During the years, the events in the country have become an international problem rather than being an internal issue of Syria. The Syrians refugees were forced to flee in neighborhood countries, especially in Turkey, because of the negative living conditions of the ongoing civil war and the influence of the terrorist organizations that emerge every day. According to the January 2020 reports of the United Nations Refugee Agency (UNHCR); about 5.5 million people left Syria and 6.6 million were moved within Syria. According to UNHCR data, 64.4% of the 5.5 million Syrians, forced

CONTACT: Hasibe Yahsi Sari ✉ hsbyahsi@gmail.com 📍 Department of Educational Sciences, Faculty of Education, Hacettepe University, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

to leave their country, and sheltered to Turkey. The data published on January 30, 2020, by the Directorate General of Migration Management in Turkey (DGMM) showed that the total number of registered Syrian refugees in Turkey was 3,576,344 and the 63 491 of those people stay in Temporary Accommodation Centers.

One of the most important rights of Syrian refugees in Turkey is to reach free education. Many of the young Syrians have begun or continued to university by this granted right. However, they have difficulties in their education life in Turkey due to the traumas they experience in the war environment, language problems, adaptation problems to the new culture, and negative perspective of the Turkish society. Due to having such external problems, Syrian students do not improve their academic self-efficacy that is a necessity to be successful in the school life (cite). This situation has a negative effect on their academic achievement and their belief in their academic skills (Bong & Clark, 1999; Demirdag, 2015). Therefore, Syrian students' academic self-efficacy and the potential factors effecting self-efficacy should be explored.

Bandura (1997) defined self-efficacy as the beliefs of individuals about their ability to plan and execute the necessary actions in the process of achieving their goals. According to Ekici (2012), academic self-efficacy is the perception of the individual that he/she can perform a given academic task at a determined success level. Based on these definitions, it is possible to define academic self-efficacy as individuals' beliefs about their skills of planning and executing the actions needed in the process of achieving an academic goal (Zimmerman, 1995). According to Bandura (1997), the factors that affect the perception about self-efficacy in individuals are direct experiences related to success, indirect experiences based on observation, verbal persuasion, and psychological-physiological situations. There is a high positive correlation between students' academic achievements and academic self-efficacy (Bandura, 1997; Ekici, 2012; Phan, 2012). In order to increase the academic success of refugee students, factors affecting academic self-efficacy should be emphasized.

Chung, AlQarni, Al Muhairi, and Mitchell (2017) studied the relationship between self-efficacy, trauma, posttraumatic stress and psychiatric diseases of 790 Syrian refugees living in Turkey. They found that traumatic events like war, armed conflict, etc., which affect adults very much, affect students much as well. Akkaya, Çilingir and Levent (2018) studied the Syrians in higher education levels and investigated the problems they experienced in Turkey. In the study, the problems faced by foreign students at higher education level were expressed as language problems, academic self-efficacy. Bayramdurdyeva (2019) examined the factors that affect the success of 48 international students (20 girls and 28 boys) from Asia, Europe, Africa, the Middle East, and North American. In the study, it was concluded that factors affecting the success of international students were family support, good-disciplined, friend/social environments, self-confidence and making use of the time well. On behalf of the good future of Turkey and the refugee students, the high academic achievement of students will facilitate the solution of the problems. Thus, it is necessary to support refugees come out of the war in terms of educational, social, economic and psychological problems.

Bandura (1997) states that self-efficacy is a necessity for self-regulation ability or vice versa. The studies (Aldan Karademir, Deveci, & Çayli, 2018; Garcia & Pintrich, 1996; Kayacan & Selvi, 2017) showed that there is a positive relationship between self-efficacy and self-regulation ability. Self-regulation is the management of emotions, thoughts, and movements in accordance with the goals wanted to achieve by the students (Kayacan & Selvi, 2017; Zimmerman & Schunk, 1989).

Masten (2001) argues that resilience consists of ordinary resources and processes, not rare features, and also it is the result of the well-performed basic compliance system. Resisting all risks and disadvantages like war, trauma, disability, etc., the ability to overcome them, and achieving positive outcomes regardless of the difficulties of life are defined as resilience

(Rutter, 2006). In other words, resilience is the individuals' adaptation to daily life skills correctly despite all stressful events. Considering the Syrian refugees living conditions in a war environment, resilience is very important for them to maintain their social life skills in a healthy way. The interaction of risk factors and protective factors is involved in the development of resilience skills (Masten, 2001). Many studies in the literature showed that wars are the risk factors affecting resilience levels (Hubbard, Realmuto, Northwood, & Masten, 1995; Masten & Coastworth, 1998; Peltonen, Qouta, Diab, & Punamaki, 2014; Pieloch, McCullough, & Marks, 2016; Demir & Aliyev, 2019). Peltonen et al. (2014) examined the resilience level and protective factors of 482 Palestinian students attending school during the war. As a result of the research, it was found that children with high resilience levels had better friendships compared to the traumatized group with low resilience levels. It was observed that social support and peer relations become protective factors in difficult conditions such as war.

On the other hand, Demir and Aliyev (2019) examined the sources of resilience in Syrian migrants who were victims of war in terms of risk and protective factors. According to the results of the research, while risk factors have more social sources, individual factors have individual sources. The risk factors mentioned in the research were mistrust to others, anger management, being pessimistic, financial difficulties, the influence of media, witnessing to death, disruption of education, social prejudice and exclusion, problems with the new settlement, language problem, change of living space, death of the family members, and living separated from family members. In case protective factors were; social support, career intentionality, patience, self-confidence, willingness to learn, perseverance, spirituality, financial support, host community support, immigrant support, and support from family members.

There are several studies related to resilience and self-efficacy in the literature (Arslan & Balkıs, 2016; Can & Cantez, 2018). Can and Cantez (2018) found the moderate significant relationship between university students' happiness, psychological resilience and self-efficacy levels. Arslan and Balkıs (2016) investigated the mediating role of self-efficacy and psychological resilience in the relationship between emotional abuse perceived from parents and problem behaviors in adolescence. As a result of the study, it was found that self-efficacy has a partial mediating role in the relationship between emotional abuse perceived by parents and psychological resilience. In addition, Turgut (2018) examined the relationship between the psychological resilience, academic achievement and academic self-efficacy levels of the students studying at the nursing faculty. In the results of the research, it has been determined that there is a positive but weak relationship between the psychological resilience and academic self-efficacy of the students; while there is a positive but weak relationship between academic self-efficacy and general academic average.

Getting high social support after the war reduces the effect of trauma (Karaman, Karadas, & Vela, 2019; Kuterovac-Jagodic, 2003). The literature showed the relation between the perceived social support and resilience (Güney, 2016; Süleymanov, Sonmez, Demirbas Unver, & Akbaba, 2017; Gez, 2018). There are many definitions of social support in the literature. Çakır and Palabıyıköğlü (1997) defined social support as getting the help of the nearby people. Social support can be physical or cognitive. Perceived social support is related to dimensions like the need for support after the problems experienced by the individual, how much these problems can be solved with the support received, the individual's expectations, etc., and this perception varies from person to person. Deryahanoğlü, Demirdöken, Canaydin and Yamaner (2019) investigated the levels of academic self-efficacy and perceived social support of university students. As a result of the research, there was no significant difference in the academic self-efficacy scores according to the nationality of the participants, but a significant difference in the sub-dimension of the family and friends of social support was detected. The study of

Danielsen, Samdal, Hetland, and Wold (2009) investigated the effects of perceived social support on school satisfaction, the mediation between scholastic competence and self-efficacy, students' life satisfaction. As a result of the research, it was found that the social support received from teachers was highly related to life satisfaction getting from the school. In addition, it was also found that there was a positive relationship between the levels of students' self-efficacy and life satisfaction. Gez (2018) conducted a study on Syrian children and adolescents and investigated the levels of psychological resilience and types of perceived social support in terms of demographic variables. As a result of the research, it was stated that there was a relationship between psychological resilience and perceived social support. In addition to the results, when the emotions and thoughts of Syrian children and adolescents are examined, many findings related to the risk factors for psychological resilience is observed. However, it is seen that the majority of the participants do not feel alone, they trust themselves, believe that they will be successful and want to return to their country. It is thought that this result is related to perceived social support.

The related literature showed the relation between perceived social support and life satisfaction (Danielsen et al., 2009; Diener & Fujita, 1995). Life satisfaction is the life expectancy of individuals to reach the expected level. Increasing perceived social support levels of individuals provides an increase in the positive effect of life satisfaction on individuals. In their study, Diener and Fujita (1995) emphasize that the resources (family, friendships, environment, etc.) in establishing social relations used by the individual, are important in subjective well-being. They also state that individuals' life satisfaction levels will be high if their aims and goals are compatible with their individual and social resources. In their study, Hırtlak et al. (2017) aimed to determine the relationship between the quality of faculty life, academic self-efficacy and life satisfaction. The results of the study illustrated that there was a positive significant relationship between students' academic self-efficacy and life satisfaction levels.

While there are many studies investigated the academic self-efficacy of undergraduate students in the literature (Alemdağ, Erman, & Yilmaz, 2014; Azar, 2010; Çuhadar, Gündüz, & Tanyeri, 2013; Şeker, 2017), studies on the variables affecting the academic self-efficacy of Syrian refugee students have not been found. This research is very important in terms of the gap in the literature about refugees while encountering studies on academic self-efficacy and self-regulation skills (Garcia & Pintrich, 1996; Kayacan & Selvi, 2017; Aldan Karademir et al., 2018). Cortes and Buchanan (2007) stated that the common characteristics of six Colombian children who are not affected by the war and have a high resilience score are feelings of being individual, self-regulation, social bond, hope, and spiritual bond. In the light of these researches; it was concluded that there is a relationship between life satisfaction, perceived social support, resilience, self-regulation and academic self-efficacy.

Considering the similar studies in the literature, in this study, the effects of self-regulation and perceived social support on the academic self-efficacy, and mediating effects of resilience and life satisfaction in these relations in Syrian refugee students were examined. Based on the research findings in the literature, the purpose of this research is to examine the stated relationships with the path analysis model. For this purpose, the following research questions were answered in the analysis section.

1. What are the direct and indirect effects of perceived social support on academic self-efficacy for refugee students?
2. What are the direct and indirect effects of self-regulation on academic self-efficacy for refugee students?
3. What are the mediating effects of life satisfaction and resilience in the relations between academic self-efficacy and perceived social support, and self-regulation?

2. METHOD

2.1. Sample and Participants

The data were collected from Syrian undergraduate students attending a four-year program in universities in Turkey. The participants consisted of 365 students, and the students voluntarily participated in the study. The data were collected during the 2020 spring semester, and this process was completed in three weeks. The form consisted of 44 survey items in total and some demographic questions. There were 210 (55.1%) female and 171 (44.9%) male students. 321 (88%) students were single and 44 (12%) students were married. The ages of them ranged from 18 to 39, and the average age of participants was 21.9 with a standard deviation of 3.11. There were 173 freshmen (47.4%), 102 sophomore (28%), 34 junior (9.3%), and 56 senior (15.3%) students in the sample.

2.2. Measures

2.2.1. Academic Self Efficacy (ASE)

The English version of the ASE survey was developed by Chemers, Hu, and Garcia. (2001), and adapted to the Arabic language by Almohazie (2018). The survey aims to measure students' academic self-efficacy and their beliefs on academic success. The original survey consisted of eight items but the translated version included nine items. Since we used the Arabic version, we administered nine items to all respondents. All items had seven response options from 1 = *Very untrue* to 7 = *Very true*. We grouped all survey items (e.g., 8 items), and calculated summated scores for each of the respondent. The Cronbach alpha value in the translated study was .92, and it was .91 in our study.

2.2.2. Self-regulation

The English version of this survey was created by Velayutham, Aldridge, and Fraser (2011), and adapted to the Arabic language by Alzubaidi, Aldridge, and Khine (2016). The scale aims to measure students' four types of domains: *learning goal orientation*, *task values*, *self-efficacy*, and *self-regulation*. However, in this study, we only used the self-regulation subscale. In this subscale, there were eight items, and all items had five response options from 1 = *Never* to 5 = *Very much*. We grouped all self-regulation items (e.g., 8 items) then, calculated summated scores for each of the respondents. The Cronbach alpha value in the translated study was .85, and it was .85 in our study as well.

2.2.3. Perceived Social Support Scale (PSSS)

The English version the PSSS was originally developed by Zimet, Dahlem, Zimet ve Farley (1988), and adapted to the Arabic language by Merhi and Kazarian (2012). The aim of the PSSS is to measure levels of social support that students receive from people around them. The survey is comprised of three subscales as *support from family*, *friends*, and *significant others* with four items in each subscale, and for a total of 12. All survey items had seven response options from 1 = *Definitely disagree* to 7 = *Definitely agree*. We calculated summated scores across the 12 items and obtained single PSSS scores (e.g., observed scores) for each of the students. In the adaptation study, the Cronbach alpha values were .82, .86 and .85, for the support from family, friends and significant others, respectively. In our study, they were .83, .88 and .89 for the three subscales, respectively.

2.2.4. Satisfaction with Life Scale

The satisfaction with life survey (SWLS) was developed by Diener, Emmons, Larsen and Griffin (1985) to measure the pleasure of life students received from their life. The SWLS translated to Arabic by Abdallah (1998). The survey is comprised of a single dimension with five items in total. All survey items in the survey had seven response options from 1 = *Strongly*

disagree to 7 = *Strongly agree*. We calculated summated scores across the five items and obtained single *SWLS* scores (e.g., observed scores) for each of the respondents. In the adaptation study, the Cronbach alpha value was .79, and it was .83 in our study.

2.2.5. Conner-Davidson Resilience Scale (CD-RISC)

The English version of the CDRS-10 was developed by Conner and Davidson (2003) and adapted into Arabic by Elias (2016). The survey aims to measure the levels of coping with the students face after tragedy, or trauma. There are two versions of the same survey as CDRS-10 and CDRS-25. In this study, we used the one with 10 items. All survey items had five response options from 0 = Not true at all to 4 = True nearly all the time. The Cronbach alpha value as an internal consistency was .83 in both the adaptation study and our study. The scores of all items measuring were summed to calculate observed scores across all students.

2.3. Data Analysis

Based on the literature, firstly, we developed the theoretical path model given in [Figure 1](#). In this model, perceived social support and self-regulation are exogenous variables (e.g., no arrows pointing to them), academic self-efficacy is endogenous variable and life satisfaction and resilience are mediating variables between endogenous and exogenous variables. We hypothesized that there should be direct and indirect effects from the two exogenous variables to the endogenous variable, and the resilience and life satisfaction are mediating these effects.

However, due to encountering model fit problems in this model, we had to modify the hypothesized model by removing insignificant paths. Besides, looking at the modification indices, we added a path from life satisfaction to resilience. This new model was called as final path model (see [Figure 2](#)).

Table 1. Bivariate correlations, means and standard deviations amongst the observed variables

Variable	1	2	3	4	5
1. Academic Self-efficacy	--				
2. Self-regulation	.62*	--			
3. Perceived Social Support	.32*	.24*	--		
4. Life Satisfaction	.26*	.14*	.45*	--	
5. Resilience	.54*	.55*	.21*	.18*	--
<i>Mean</i>	42.19	28.40	59.02	19.66	
<i>Standard Deviation</i>	11.51	5.85	15.74	6.56	

* $p < .05$

The bivariate correlations amongst all variables are given in [Table 1](#). We run both hypothesized and final models in Mplus software version 7 (Muthen & Muthen, 1998-2012), and used the bootstrap with 5000 iterations to obtain 90% confidence intervals for the effects. The sizes and 90% confidence intervals of total, direct and indirect effects of exogenous variables on endogenous variables are given in [Table 2](#).

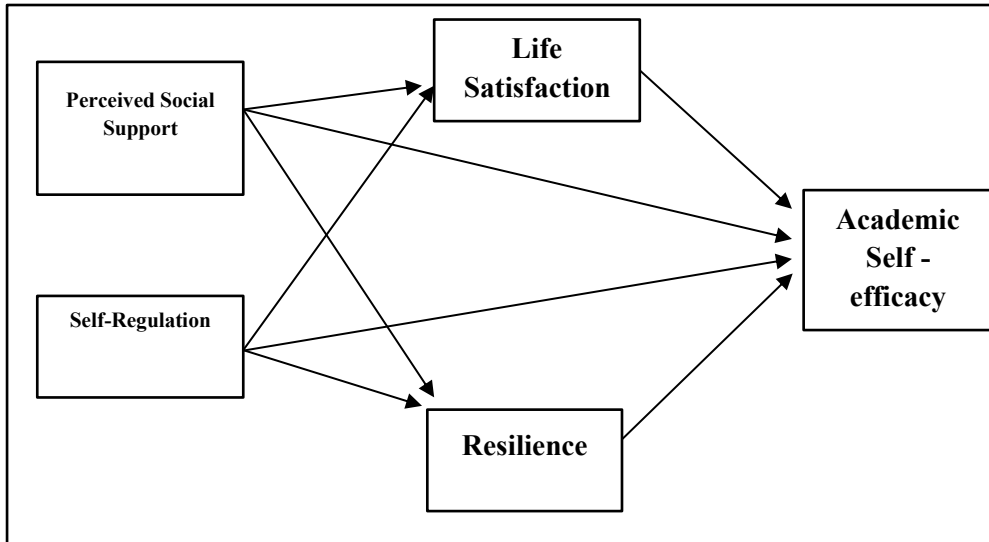


Figure 1. Hypothesized path model.

3. RESULT / FINDINGS

3.1. Model Fit Results of Hypothesized Path Model

The model fit statistics for the hypothesized model given in Figure 1 were $\chi^2(1)=3.44, p<.05$, RMSEA=.09 with 90%CI[.00, .18], CFI=.99, TLI=.95, SRMR=.02. The chi-square and RMSEA statistics were somewhat unacceptable. Besides, the path from self-regulation to life satisfaction was insignificant. Therefore, we removed this path from the model and added a path from life satisfaction to resilience.

3.2. Model Fit Results of Final Path Model

The model fit statistics for the final path model given in Figure 2 were $\chi^2(2)=1.55, p>.05$, $\chi^2/df=.77$, RMSEA=.01 with 90% CI[.00, .09], CFI=1.00, TLI=1.00, SRMR=.01. These values indicated very good model fit.

3.3. Results of Direct, Indirect and Total Effects

Self-regulation had a significant effect on academic self-efficacy with a total effect of 1.12. As given in Table 2, the .85 effect was direct and the .27 effect was indirect. The indirect effect was mediated through resilience. The total and direct effects were large and indirect effect medium in size.

Perceived social support had a significant effect on the academic self-efficacy with a total effect of .12. As given in Table 2, the .08 effect was direct and the .04 effect was indirect. There were two specific indirect effects as presented in Figure 2. These were a) from perceived social support to life satisfaction, from life satisfaction to academic self-efficacy ($B=.03, p<.05$), and b) from perceived social support to life satisfaction, from life satisfaction to resilience, and from resilience to academic self-efficacy ($B=.01, p<.05$). All specific indirect effects, direct effect and total effect were small in size but significant.

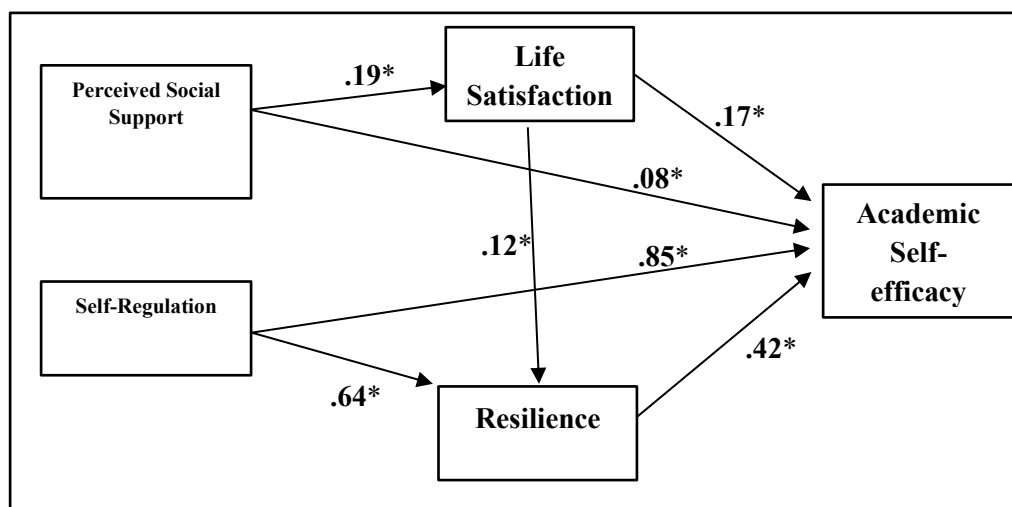
Life satisfaction had a significant effect on academic self-efficacy with a total of .22. As specified in Figure 2, the .17 effect is direct and the .05 effect is indirect. Therefore, both direct and indirect effects were small in size. The indirect effect was mediated through resilience. All effects were significant. Resilience had a significant effect on academic self-efficacy with a total of .42. Thus, the effect of resilience on academic self-efficacy was small to medium. As specified in Figure 2, this was an entirely direct effect, and there was no indirect effect of resilience on the academic self-efficacy.

Table 2. The Sizes and 90% Bootstrapping coefficients, Confidence Intervals for Total, Direct and Indirect Effects of Variables in The Selected Path Model

Endogenous Variables	Exogenous Variables			
	Self-Regulation	Perceived Social Support	Life Satisfaction	Resilience
Academic Self-efficacy	.85* [.70, 1.01]	.08* [.03,.14]	.17* [.04,.29]	.42* [.29,.55]
	.27* [.18, .35]	.04* [.01,.07]	.05* [.02,.08]	--
	1.12* [.98, 1.27]	.12* [.07,.18]	.22* [.08,.35]	.42* [.29,.55]
Self-regulation	--	--	--	--
	--	--	--	--
	--	--	--	--
Perceived Social Support	--	--	--	--
	--	--	--	--
	--	--	--	--
Life Satisfaction	--	.19* [.15,.22]	--	--
	--	--	--	--
	--	.19* [.15,.22]	--	--
Resilience	.64* [.55,.73]	--	.12* [.05,.18]	--
	--	.02* [.00,.03]	--	--
	.64* [.55,.73]	.02* [.00,.03]	.12* [.05,.18]	--

Note. Direct effects in regular text, total *indirect* effects in italics, total effects in bold. The symbol -- means the effect is not in the model; * $p < .05$; all effects are unstandardized effects.

Self-regulation had a significant effect on resilience with a total effect of .64. This effect was the entirely indirect effect, and large in size. There was no direct effect from self-regulation to resilience. Perceived social support had a significant effect on resilience with a total effect of .02. This was entirely indirect and very small in size. There was no direct effect of perceived social support on resilience. Life satisfaction played a mediator role on this effect. Perceived social support had a significant effect on life satisfaction with a total effect of .19. This was entirely direct and small in size.

**Figure 2.** Final path model.

* < .05

4. DISCUSSION and CONCLUSION

In this study, the relationship between the variables of academic self-efficacy, resilience, perceived social support, life satisfaction and self-regulation and their direct and indirect effects on academic self-efficacy were investigated for Syrian refugee undergraduate students with the path analysis model. It was supported that the developed model is compatible with the data. The findings of the research showed that self-regulation, perceived social support and life satisfaction positively affected students' academic self-efficacy directly and indirectly. In other words, the level of academic self-efficacy was high for individuals who had higher levels of self-regulation skills, perceived social support, and life satisfaction.

According to the first finding of the research, self-regulation skill has direct and indirect effects on academic self-efficacy. In addition, the direct effect that predicted academic self-efficacy mostly is the effect of self-regulation ($B = .85$). These findings support that there is a positive relationship from self-efficacy and self-regulation skills consistent with the literature (Garcia & Pintrich, 1996; Kayacan & Selvi, 2017; Aldan Karademir et al., 2018). Bandura (1997) states that self-regulation skills are needed for academic self-efficacy. Self-regulation is that students manage their own emotions, thoughts, and movements in accordance with the goals they want to achieve (Kayacan & Selvi, 2017; Zimmerman & Schunk, 1989). A student with high self-regulation skills can control his/her cognitive, affective, and psychomotor skills for the purposes he/she wants to achieve. This controllability results from the self-regulation skill, driven by an internal impulse. This skill provides an increase in academic self-efficacy from an academic point of view. As Zimmerman (1995) mentioned in the definition of academic self-efficacy, individuals should have the skills to plan and carry out the actions they need in the process of achieving an academic goal. Findings regarding the positive relationship between self-regulation skill and academic self-efficacy supports the definition of Zimmerman (1995).

There is also an indirect effect from self-regulation to academic self-efficacy. The indirect effect of self-regulation skill on academic self-efficacy sourced from the mediating effect of the resilience variable. Masten (2001) argues that resilience arises as a result of basic compliance systems working well and one of these basic systems is the self-regulation system. Cortes and Buchanan (2007) conducted a study in a war environment and as a result of their research, they stated that the common characteristics of the children who have higher levels of resilience are self-regulation, feeling of being individual, social bond, hope, and spiritual bond. The findings in the literature (Masten, 2001; Cortes & Buchanan, 2007; Keskin & Akça, 2019) also support the finding that self-regulation predicts resilience.

According to the second finding of the study, there is a positive effect directly and indirectly from perceived social support to academic self-efficacy. Perceived social support, the need of support after the problems experienced by the individual, how much he/she could overcome his/her problems with the support received and his/her expectations etc. relates to the perception of social support. The study of Deryahanoğlu et al. (2019), conducted to university students for analyzing academic self-efficacy and perceived social support levels, showed that there was no significant difference in academic self-efficacy scores according to the nationality status of the participants, but there was a significant difference in academic self-efficacy scores in the family and friends sub-dimension of social support. The findings of Deryahanoğlu et al. (2019) supports the current findings. Danielsen et al. (2009) investigated the effects of perceived social support on students' life satisfaction with the mediation of school satisfaction, scholastic competence, and self-efficacy. As a result of the research, it was found that the social support received from teachers was highly related to life satisfaction received from the school.

Life satisfaction and resilience are mediated in indirect effects from perceived social support variables to academic self-efficacy. The positive relationship between perceived social support and life satisfaction aligned with (Diener & Fujita, 1995; Danielsen et al., 2009). The increase

in perceived social support leads to an increase in the positive effect of life satisfaction on individuals. Diener and Fujita (1995) emphasize that the resources (family, close friendships, environment, etc.) used by the individual in establishing social relations are important in subjective well-being. They also state that individuals' satisfaction will be high if their goals are compatible with their individual and social resources.

Another mediator variable in indirect effects from perceived social support variable to academic self-efficacy is resilience. The finding of the positive relationship between perceived social support and resilience is supported by the related literature (Peltonen et al., 2014; Güney, 2016; Süleymanov et al., 2017; Gez, 2018).

According to the third finding of the study, there is a positive direct and indirect effects from the life satisfaction to academic self-efficacy. The life satisfaction also played mediating role between perceived social support and academic self-efficacy. This means that as the perceived social support increased, life satisfaction increased, and this led to increase in academic self-efficacy. The increase in life satisfaction also increased the level of resilience, and this increase led to increase in academic self-efficacy as well. This finding aligns with the results of the research conducted by Hırlak, Taşlıyan, Fidan and Güler (2017) and showed a positive relationship between students' academic self-efficacy and life satisfaction levels. Similarly, in the study of Danielsen et al. (2009), it was found that student' academic self-efficacy levels were related to their life satisfaction.

The findings of the research show that resilience is a positively related variable that directly affects academic self-efficacy. The positive relationship between resilience and academic self-efficacy is consistent with the findings of Turgut (2018). It can be seen in many studies in the literature that wars are the risk factors affecting the resilience (Hubbard et al., 1995; Masten & Coastworth, 1998; Peltonen et al., 2014; Pieloch et al., 2016; Demir & Aliyev, 2019). Chung et al. (2017) founded that more than half of the individuals have post-traumatic stress in their study on the refugees living in Turkey. All of these findings indicated the importance of resilience for the Syrian refugees studying in Turkey after the war.

According to Bandura (1997), factors affecting self-efficacy perception in individuals are direct experiences related to success, indirect experiences based on observation, verbal persuasion and psychological-physiological situations. These factors explain all the variables of life satisfaction, perceived social support, resilience and self-regulation; and the direct and indirect effects on academic self-efficacy. To sum up, the data fit with the chosen model perfectly. There is a positive relationship between academic self-efficacy and perceived social support, self-regulation, life satisfaction and resilience. Self-regulation, perceived social support and life satisfaction affect academic self-efficacy both directly and indirectly. Life satisfaction directly affects academic self-efficacy and also affects it with the mediation of resilience. There were also mediating effects of the resilience between self-regulation and academic self-efficacy; and between the perceived social support and academic self-efficacy. Moreover, there were mediating effects of the life satisfaction variable between perceived social support and academic self-efficacy. The highest direct effect is resulted from self-regulation, while the lowest direct effect is resulted from the perceived social support. The chosen path model shows that the variable that predicts academic self-efficacy mostly is the self-regulation variable.

The results of this study showed important recommendations or implications for university faculty and administrators. First, the study showed that Syrian students need to improve levels of self-regulations. Since this directly affected their academic self-efficacy, university consulting services should get involved in this step, and organize group or individual meetings with refugee students. The study also showed that perceived social support of students was important for raising academically successful students. Thus, the problems the students face in daily life such as language or communication problems should be minimized. In this context,

each refugee student can be matched with a Turkish peer, and help each other in school assignments and in daily life. This would also increase their life satisfaction and resilience and lead to increase academic self-efficacy.


The initial purpose of this study was to include student GPA as one of the endogenous variables. However, most of the students participated in the study were the first-year students. Since their GPA was not available at the time the data were collected, we could not include the GPA as the endogenous variable. A further study should examine the effects of studied variables on the GPA. Also, the study did not investigate the effects of demographic variables (e.g., gender, marital status, grade etc.) to the academic self-efficacy. However, these variables might have potential effects on it. Thus, a future study should be conducted with those demographic variables.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Hasibe Yahşi Sarı  <https://orcid.org/0000-0002-0451-6034>

Selahattin Gelbal  <https://orcid.org/0000-0001-5181-7262>

Halil İbrahim Sarı  <https://orcid.org/0000-0001-7506-9000>

5. REFERENCES

- Abdallah, T. (1998). The Satisfaction with Life Scale (SWLS): Psychometric properties in an Arabic-speaking Sample, *International Journal of Adolescence and Youth*, 7(2), 113-119.
- Akkaya, A.Y., Çilingir, G.A., & Levent, G.T. (2018) A study on academic challenges of Syrian students at the University of Van Yuzuncu Yıl. *Journal of Social Policy Studies*, 18(40/2), 413-448.
- Aldan Karademir, Ç., Deveci, Ö., & Çaylı, B. (2018). Investigation of secondary school students' self-regulation and academic self- efficacy. *e-Kafkas Journal of Educational Research*, 5(3), 14-29.
- Alemdağ, C., Erman, Ö., & Yılmaz, A.K. (2014). Preservice physical education teachers' academic motivation and academic self-efficacy. *Hacettepe Journal of Sport Sciences*, 25(1), 23-35.
- Almohazie, M.F. (2018). *Reliability and validity of an Arabic translation of academic self-efficacy scale (ase) on students at King Faisal University* (Unpublished dissertation, Wayne State University, Detroit, The United States). Retrieved from https://digitalcommons.wayne.edu/oa_dissertations/1910
- Alzubaidi, E., Aldridge, J.M. & Khine, M.S. (2016). Learning English as a second language at the university level in Jordan: Motivation, self-regulation and learning environment perceptions. *Learning Environments Research*, 19(1), 133-152.
- Arslan, G., & Balkis, M. (2016). The relationship between emotional maltreatment, problem behaviors, psychological resilience, and self-efficacy in adolescents. *Sakarya University Journal of Education*, 6(1), 8-22.
- Azar, A. (2010). In-service and pre-service secondary science teachers' self-efficacy beliefs about science teaching. *ZKU Journal of Social Sciences*, 6(12), 235-252.
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavior change. *Psychological Review*, 84, 191-215.
- Bayramdurdyeva, G. (2019). Factors affecting the success of international students. *International Journal of Humanities and Education*, 5(11), 509-524.

- Bong, M., & Clark, R.E. (1999). Comparison between self-concept and self-efficacy in academic motivation research. *Educational Psychologist*, 34(3), 139-153.
- Çakır, Y., & Palabıyıkoglu, R. (1997). Reliability and validity study of multidimensional scale of perceived social support. *Journal of Kriz*, 5(1), 15-24.
- Can, M., & Cantez, E. (2018). Investigation of happiness, resilience and self-efficacy levels in university students. *Aydın İnsan ve Toplum Dergisi*, 4(2), 61-76. Retrieved from <http://static.dergipark.org.tr/article-download/8efd/5d8d/42ff/5c7f65914ebcd.pdf?>
- Chemers, M.M., Hu, L.T., & Garcia, B. F. (2001). Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational Psychology*, 93(1), 55-64.
- Chung, M.C., AlQarni, N., Al Muhairi, S., & Mitchell, B. (2017). The relationship between trauma centrality, self-efficacy, posttraumatic stress and psychiatric co-morbidity among Syrian refugees: is gender a moderator? *Journal of Psychiatric Research*, 94, 107-115.
- Connor, K.M. & Davidson, J.R.T. (2003). Development of a new resilience scale: The Connor-Davidson resilience scale (CD-RISC). *Depression and Anxiety*, 18, 71-82.
- Cortes, L., & Buchanan, M.J. (2007). The experience of Columbian child soldiers from a resilience perspective. *International Journal for the Advancement of Counselling*, 29(1), 43-55.
- Çuhadar, C., Gündüz, Ş., & Tanyeri, T. (2013). Investigation of relationship between studying approach and academic self-efficacy of computer education and instructional technologies department students. *Mersin University Journal of The Faculty of Education*, 9(1), 251-259. Retrieved from <http://static.dergipark.org.tr/article-download/imported/1002000349/1002000258.pdf?>
- Danielsen, A.G., Samdal, O., Hetland, J., & Wold, B. (2009), "School-related social support and students' perceived life satisfaction", *The Journal of Educational Research*, 102(4), 303-318.
- Demir, Ö., & Aliyev, R. (2019). Resilience among Syrian university students in Turkey. *Turkish Journal of Education*, 8(1), 33-51.
- Demirdağ, S. (2015). Comparing academic self-efficacy of students based on skills, educational setting and quality of education. *Journal of Research in Education and Teaching*, 4(1), 315-323.
- Deryahanoğlu, G., Demirdöken, Ç., Canaydin, A., & Yamaner, F. (2019). Analysis of the level of academic self-efficacy and social support of the sports sciences faculty students. *The Journal of International Social Research*, 12(66), 1407-1413.
- Diener, E., & Fujita, F. (1995). Resources, personal striving and subjective well-being, *Journal of Personality and Social Psychology*, 69(1), 120-132.
- Diener, E., Emmons, R.A., Larsen, R.J. & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49, 71-75.
- Diener-Biswas, R., Diener, E., Tamir, M. (2004). The psychology of subjective wellbeing. *Daedalus*, 133(2), 18-25.
- Ekici, G. (2012). Academic self-efficacy scale: Academic self-efficacy scale: the study of adaptation to Turkish, validity and reliability. *Hacettepe University Journal of Education*, 43(43), 174-185.
- Elias, R.R. (2016). *Hope, parenting styles, and resilience in Lebanese university youth* (Unpublished master's dissertation). Beirut, Lebanon. Retrieved from <https://scholarworks.aub.edu.lb/bitstream/handle/10938/11034/t-6412.pdf?sequence=1>
- Garcia, T., & Pintrich, P. R. (1996). The effects of autonomy on motivation and performance in the college classroom. *Contemporary Educational Psychology*, 21(4), 477-486.
- Gez, A. (2018). *Investigation of the relationship between psychological resilience and perceived social support of Syrian children and adolescents*. (Unpublished master's dissertation). Mersin, Turkey.

- Göç İdaresi Genel Müdürlüğü (2020). Statistics, temporary protection. Retrieved from <https://www.goc.gov.tr/gecici-koruma5638>
- Güney, E. (2016). *Research of resiliency traits and perceived social support levels of high school students whose parents are divorced and nondivorced*. (Unpublished master's dissertation). Konya, Turkey.
- Hırlak, B., Taşlıyan M., Fidan, E., & Güler, B. (2017). The relationship between the quality of faculty life, academic self-efficacy and satisfaction with life: A field research on students of faculty of economics and administrative sciences. *Journal of Social & Humanities Sciences Research (JSHSR)*, 4(9), 86-104.
- Hubbard, J., Realmuto, G.M., Northwood, A.K., & Masten, A.S. (1995). Comorbidity of psychiatric diagnoses with post traumatic stress disorder in survivors of childhood trauma. *Journal of the American Academy of Child and Adolescent Psychiatry*, 34, 1167-1173.
- Karaman, M.A., Karadas, C., & Vela, J.C. (2019). Development of perceived school counselor support scale: Based on the ASCA mindsets and behaviors. *International Journal of Assessment Tools in Education*, 6(2), 202-217.
- Kayacan, K., & Selvi, M. (2017). The effect of inquiry based learning enriched with self regulated activities on conceptual understanding and academic self-efficacy. *Kastamonu Education Journal*, 25(5), 1771-1786.
- Keskin, B.B., & Akça, F. (2019). The importance of resilience: Family, school, community size predictions. *International Journal of Scientific and Technological Research*, 5, 170-182.
- Kuterovac-Jagodić, G. (2003). Posttraumatic stress symptoms in Croatian children exposed to war: A prospective study. *Journal of Clinical Psychology*, 59(1), 9-25.
- Masten, A.S. (2001) Ordinary magic: Resilience processes in development. *American Psychologist*, 56, 227–238.
- Masten, A.S. & Coastworth, J.D. (1998). The development of competence in favorable and unfavorable environments: Lessons from research on successful children. *American Psychologist*, 53, 205-220.
- Merhi, R. & Kazarian, S.S. (2012). Validation of the Arabic translation of the multidimensional scale of perceived social support (Arabic-MSPSS) in a Lebanese community sample. *Arab Journal of Psychiatry*, 23(2), 159-168.
- Muthén, L.K., & Muthén, B.O. (1998). Statistical analysis with latent variables. *Mplus User's guide*, 2012.
- Peltonen, K., Qouta, S., Diab, M., & Punamaki, R.-L. (2014). Resilience among children in war: The role of multilevel social factors. *Traumatology*, 20(4), 232- 240.
- Phan, H.P. (2012). Relations between informational sources, self-efficacy and academic achievement: A developmental approach. *Educational Psychology*, 32, 81-105.
- Pieloch, K.A., McCullough, M.B., & Marks, A.K. (2016). Resilience of children with refugee statuses: A research review. *Canadian Psychology/Psychologie canadienne*, 57(4), 330–339. <https://doi.org/10.1037/cap0000073>
- Rutter, M. (2006). *Implications of resilience concepts for scientific understanding*. Annals of the New York Academy of Sciences, 1094, 1-12.
- Şeker, S.S. (2017). The examination of prospective music teachers' academic self-efficacy and academic motivation levels. *Abant İzzet Baysal University Journal of Faculty of Education*, 17(3), 1465-1484.
- Süleymanov, A., Sonmez, P., Demirbas Unver, F., & Akbaba, S.M. (2017). *International migration and children*. Transnational Press: London, England, 2017, ISBN: 978-1-910781-56-2

- Turgut, N. (2018). *Psychological Resilience Academic Achievement, And SelfEfficacy Levels in Nursing Students*, (Unpublished Master's thesis, Near East University, Lefkoşa, North Cyprus). Retrieved from <http://docs.neu.edu.tr/library/6689569321.pdf>
- United Nations High Commissioner for Refugees (2020). *Regional Strategic Overview*. Retrieved from <https://data2.unhcr.org/en/documents/download/73116>
- Velayutham, S., Aldridge, J.M., & Fraser, B.J. (2011). Development and validation of an instrument to measure students' motivation and self-regulation in science learning. *International Journal of Science Education*, 15, 2159–2179.
- Zimet, G.D., Dahlem, N.W., Zimet, S.G., & Farley, G.K. (1988). The multidimensional scale of perceived social support. *Journal of personality assessment*, 52(1), 30-41.
- Zimmerman B.J. (1995). *Self-efficacy and educational development. Self-efficacy in Changing Societies*, Cambridge University Press: New York, NY.
- Zimmerman, B.J., & Schunk, D.H. (Eds.). (1989). *Springer series in cognitive development. Self-regulated learning and academic achievement: Theory, research, and practice*. Springer-Verlag: New York. <https://doi.org/10.1007/978-1-4612-3618-4>

Psychometric Characteristics of Written Response Instruments Used in Postgraduate Theses Completed in Special Education

Gamze Sarikas¹, Safiye Bilican Demir^{2,*}

¹Serkan Argin Secondary School, Bursa, Turkey

²Faculty of Education, Kocaeli University, Kocaeli, Turkey

ARTICLE HISTORY

Received: Jan 14, 2020

Revised: May 17, 2020

Accepted: May 26, 2020

KEYWORDS

Written response instrument,
Psychometric characteristics,
Postgraduate thesis,
Special education

Abstract: The purpose of this study is to examine the written response instruments used in postgraduate theses completed in the field of special education in Turkey between 2015 and 2018 and explore the psychometric profiles of these instruments. In the study, a total of 137 master's theses and 37 dissertations were reviewed using the Data Collection Instrument Review Form. Categorical and frequency analyses were used in the analysis of the data. Also, the relevant categories were discussed by citing remarkable examples of errors or deficiencies. According to the research findings, a total of 387 written response instruments were used in the theses reviewed. Most of the written response instruments were developed by the researchers themselves and of these instruments, the most frequently used were the personal information forms and scales. In the theses, most of the written response instruments were not introduced or only partly introduced and the validity and reliability of these instruments were either not reported or only partly reported. The results of the research showed that there were crucial deficiencies and errors in reporting the basic methodological information about the written response instruments used in the theses and this situation was repeated in the theses. In parallel with the results of the research, the related problems and their causes were discussed, and suggestions offered.

1. INTRODUCTION

The aim of scientific research is to produce knowledge. In this process, answers to the researched problem are sought in accordance with standard scientific principles. The production of scientific knowledge takes place only when the stages of scientific research methods are carried out completely. Universities are one of the institutions that produce and share scientific information. Universities support social development by producing information and technology through postgraduate studies and by providing qualified human resources. Theses produced as a result of postgraduate studies are of great importance in the development of a field. The theses put forward specific solution suggestions for a research problem in accordance with the scientific process steps. As such, theses show that the prospective researcher possesses the knowledge and the proficiency to carry out independent research and can produce scientific information that will contribute to the field of study (Tavşancıl et al., 2010). In addition to providing scientific standards, these research reports are valuable in that they are reviewed by

CONTACT: Safiye Bilican Demir ✉ safiye.demir@kocaeli.edu.tr 📧 Kocaeli University, Faculty of Education, Department of Educational Sciences, Umuttepe Campus, 41001, Kocaeli, Turkey.

ISSN-e: 2148-7456 /© IJATE 2020

a scientific jury. Therefore, postgraduate theses play a key role in the growth and development of a discipline (Evrekli et al., 2011). In this regard, it is clear that development and change in any field is closely related to scientific research in that field (Seçer et al., 2014).

The increase in the number of universities in Turkey in recent years has particularly resulted in an increase in postgraduate education, which in turn has led to a quantitative increase in scientific research in education (Özenç & Özenç, 2013). The same is true of special education (Diken et al., 2009). Both the increase in the number of students receiving special education and the legal regulations relating to the field are seen as factors in the increase in scientific studies in the field of special education. According to the World Health Organization (WHO), approximately 12% of individuals in the 6-18 age group have special needs. Some sources put this figure as high as 14% (Metin, 2012). The number of students receiving special education services has increased by 585% from 2002 to 2013 (Melekoğlu, 2014). The increase in the number of students with special education needs has led to an increase in graduate programs training teachers to work in this field resulting in an increase in the number of preservice teachers graduating from these programs and more teaching staff working in universities' special education departments (Ağca, 2014). The importance of postgraduate programs not only in terms of producing scientific information but also in terms of training teachers who will run graduate and postgraduate programs in special education is starting to emerge. As a result, the developments in various aspects of special education are having a positive effect on the increase in scientific studies made in this field.

The quantitative increase in postgraduate studies both in the field of special education and in other fields of education can be considered a positive step for the development of the related field. However, the extent to which knowledge obtained from studies contributes to science and how scientific this knowledge is will always be a topic of debate. Knowledge can be produced through scientific studies in any field, but not all information may be truly scientific (Benligiray, 2009). In this respect, the quantitative increase in scientific studies is not always the guarantee of qualitative development. At this point, the method used to obtain the results obtained from scientific research becomes at least as significant as the results. Therefore, to produce scientific knowledge, the steps of the scientific process must be carried out completely. In this way, scientifically sound information not only reveals the facts, but it also allows scientific debate to continue by being a point of reference for new studies. However, it is known that researchers working in social sciences in Turkey experience important problems particularly when it comes to methodology (Köklü & Büyüköztürk, 1999).

Some of the methodology-related problems can be addressed in the context of data collection instruments. Researchers use various data collection techniques to obtain information about the subject of interest. Which data collection technique the researcher will use varies depending on the research problem, the nature of the data, or the source of the data. Although there are different classifications in the literature; in general, data collection approaches can be classified as written response instruments, interview, observation, available data, and document analysis (Büyüköztürk et al., 2016; Karasar, 2016; Tavşancıl et al., 2010). This study discusses the written response instruments that are used frequently in research. In the written data collection approach, communication between the researcher and the participant is made in writing and the researcher may use various data collection instruments such as questionnaire, scale, test, or inventory to collect data. No matter what type of data collection instrument the researcher uses, it is expected that this instrument's characteristics such as purpose, item number, and scoring format, etc. be reported in a scientifically appropriate manner. At the same time, the psychometric properties of the results/scores obtained from this data collection instrument should be at the desired level. Otherwise, the scientific validity of the data collection instrument and the results obtained using it will be regarded with suspicion. For the accuracy of the data

obtained using the collection instruments to be satisfactory, two fundamental characteristics are required and these are known as "reliability," which is an evidence of the stability of the results of the scores obtained from the data collection instruments, and "validity," which is an evidence of the degree to which the instrument is able to measure the characteristics it is supposed to measure (Horst, 1966). However, related studies show that although articles or theses in the field of social sciences have been published, they are not error-free and that they even contain deficiencies and errors, especially in terms of research methods (Başol & Akin, 2006; Tavşancıl et al., 2010).

Today, many different scientific studies are carried out in the field of education. When considering the theoretical or practical effects of these studies, it is necessary to classify them, determine the emerging trends, and evaluate their results (Kutluca & Demirkol, 2016; Varışoğlu et al., 2013). Staton and Wulff (1984) state that the most suitable way to do this is to review the studies in any given field periodically. These types of review can act as road maps for researchers still unfamiliar with the terminology associated with the scientific method in terms of methods used and topic selection (Cohen et al., 2007). In addition, the results of this kind of research may be considered valuable in terms of guiding research in the related field, in saving researchers' time, and facilitating access to research information. This is because an excess of work done in a field can sometimes create problems. For example, researchers who want to do research in a field often find it difficult to access all the studies done in that field or they spend more time accessing them (Göktaş et al., 2012). In this context, revealing the content and meta-analysis results of studies made in a particular field or topic by reviewing them at regular intervals makes it easier for researchers to assess the latest situations regarding their fields (Karadağ, 2009; Lee et al., 2009).

It is noteworthy that many studies have been conducted in recent years with the aim of identifying developments in different disciplines so as to determine research trends. For example; scientific studies in such fields as educational sciences (Arık & Türkmen, 2008; Doğan & Tok, 2018; Erdem, 2011; Şenyurt & Özer-Özkan, 2017; Tavşancıl et al., 2010; Yalçın, 2016), educational technologies (Alper & Gülbahar, 2009; Tosuntaş et al., 2019), curriculum (Ozan & Köse, 2014), education management (Aydın & Uysal, 2011; Turan et al., 2014), preschool education (Yılmaz & Altinkurt, 2012), science and mathematics education (Çiltaş et al., 2012; Kutluca et al., 2018; Yaşar & Papatğa, 2015) have been examined and general trends have been revealed across a very broad spectrum including research subject, research model, target group, data collection instruments, data analysis techniques, publication year, and number of authors.

As in the fields mentioned above, it is seen that trend studies have also been made in the field of special education (Aslan & Özkubat, 2019; Çoşkun et al., 2014; Demirok et al., 2015; Diken et al., 2008; Diken et al., 2016; Doğru et al., 2015; Güner-Yıldız et al., 2016; Küçüközyiğit et al., 2016; Özkubat et al., 2014; Ünlü et al., 2020; Tiryaki, 2017; Tiryakioğlu, 2014). For example, with the aim of identifying trends in special education, Ünlü et al. (2020) investigated doctoral dissertations in special education in terms of various variables. The findings obtained at the end of the research have exposed that the subject of intellectual disability was studied most in the thesis, "single subject design" was preferred mostly and a large part of the thesis was completed in Anadolu University. Also, according to the findings it has been revealed that the number of thesis in special education has been increasing in late years and more than half of all thesis were completed after 2011. Aslan and Özkubat (2019) reviewed papers published in the booklets of national congresses on special education held in Turkey between 2007 and 2017. Within the scope of the study, 1,742 papers were given a content analysis review looking at year, number of authors, sample group, research model, data collection instruments, data analysis method, and research topic categories. The results of the review showed that the

majority of the papers have one or two authors, the research topics were concentrated in inclusive education, a large number of studies used the descriptive screening model as their method, data were mostly collected by interview, and descriptive analysis was used for data analysis.

In another analysis, Tiryaki (2017) reviewed theses completed in the field of special education between the years 2000 and 2006 in terms of year of completion, institution, research subject, research model, target group, and data collection instruments. The prominent results of that review showed that in the years in question a large proportion of the theses focused on special education and language skills, there was a high proportion of qualitative research, and that data were collected using observation and interview techniques. Küçüközyiğit et al. (2016) conducted a content analysis review of 155 theses completed in the field of education for the visually impaired. That study's conclusions emphasized that there was limited study in the related field at doctoral degree and that most of the theses were descriptive and used the single-subject research model. Yıldız et al. (2016) reviewed 113 articles about special education published in journals. The results of that study showed that most of the reviewed articles had one author and that the single-subject research method was frequently preferred. Another study (Demirok et al., 2015) reviewed 400 articles published in international journals in the context of various variables such as subject, research type and sample group. The findings of that study showed that most papers were published in 2012, most of the publications had two authors, most of the articles were made using students with physical disabilities, and that experimental methods were used in more than half the articles. Studies by Çoşkun et al. (2014) and Özkubat et al. (2014) aimed to identify the research trends relating to completed theses in the field of special education. To this end, theses were assessed in terms of research field, topic, method, and pattern used, sample selection and size, data collection, and analysis. When the study's findings are analyzed, it can be seen that approximately 80% of the theses are master's theses, the topics were mostly related to skill teaching, single-subject experimental patterns are the most commonly used research pattern, most studies were quantitative, observation was the most commonly used data collection instruments, and that descriptive analysis techniques were frequently preferred.

In the studies summarized above, the scientific studies in the field of special education were subjected to a content analysis review of topic, quality and quantity, and the methods and techniques used, and an attempt was made to determine the direction of trend. For the field of special education, which is closely related to the other disciplines in the education, these kinds of studies are valuable for the development of this field. However, there is also a need for research that looks at the problem of method in studies in the field of special education and that makes a detailed psychometric review of the problem. Although there are studies that reveal the general trends in data collection instruments used in the field of special education, no studies have been found that make a detailed review to determine if these instruments meet basic psychometric standards. The role of postgraduate theses produced in any field in the development of that field is clear. At this point, the results obtained from this study are valuable in that they provide information on the psychometric properties of the data collection instruments used in postgraduate studies published in the field of special education.

This study looks at the methodology problem seen in the field of special education in terms of data collection instruments and it examines the written response instruments used in postgraduate theses completed in the field of special education as well as the psychometric characteristics of these instruments.

2. METHOD

2.1. Research Design

In this study, a qualitative research based on descriptive content analysis was used to describe the current situation of related theses.

2.2. Population

The population of the study consists of a total of 189 postgraduate theses comprising 152 master's theses and 37 dissertations completed in the field of special education between 2015 and 2018 and scanned at the Higher Education Council's National Thesis Center in Turkey. Since there are not many postgraduate theses in the population considering the researcher's conditions, sampling was not done and, instead, all the theses in the population were included in the scope of the study. The distribution of the theses in the population by distribution by years is given in [Table 1](#).

Table 1. *Distribution by year of theses*

Year	Master's Thesis		Dissertation
	f	%	f
2015	41	26.97	16
2016	48	31.58	8
2017	35	23.03	8
2018	28	18.42	5
Total	152	100.00	37

According to [Table 1](#), more postgraduate theses were completed in 2015 and 2016 than in the other years. A total of 41 (26.97%) of master's theses and 16 (43.24%) of dissertations were completed in 2015; 48 (31.58%) of the master's theses and eight of the dissertations were completed in 2016.

2.3. Data Collection Instruments

The study's data were obtained using the "Data Collection Instrument Review Form" for the written response instruments of the thesis review form developed by Tavşancıl et al. (2010). In this form, written response instruments were reviewed in detail under the categories of development-adaptation status, category (questionnaire, scale, and achievement/ability test), data collection instrument presentation status, development-adaptation steps, and evidence for the instruments' validity and reliability. In the thesis review form, the development steps for the researcher-developed data collection instruments were checked to see if they had been followed completely and the following steps for the questionnaire, the achievement/ability test, and the scale were considered (Tavşancıl et al., 2010):

For the questionnaire developed by the researcher:

- 1) Literature review
- 2) Review of the same or similar data collection instruments
- 3) Item generation
- 4) Seeking expert opinion and explaining the experts' characteristics
- 5) Pilot study
- 6) Performing item analyses and/or qualitative analyses (clarity of items, etc.)
- 7) Deciding the final version of the questionnaire

For the achievement/ability test developed by the researcher:

- 1) Literature review

- 2) Review of the same or similar data collection instruments
- 3) Preparing a test blueprint
- 4) Item generation
- 5) Seeking expert opinion and explaining the experts' characteristics
- 6) Pilot study
- 7) Performing item analyses and/or qualitative analyses (clarity of items, etc.)
- 8) Deciding the final version of the achievement/ability test

For the scale developed by the researcher:

- 1) Literature review
- 2) Review of the same or similar data collection instruments
- 3) Composition work (Creating items by examining a student's attitude, feelings, thoughts, etc. by getting the student to write a composition.)
- 4) Content analysis
- 5) Item generation
- 6) Seeking expert opinion and explaining the experts' characteristics
- 7) Pilot study
- 8) Performing item analyses and/or qualitative analyses (clarity of items, etc.)
- 9) Deciding the final version of the scale

The presentation status of the questionnaire, achievement/ability test, and scale developed by another researcher was reviewed using the following steps (Tavşancıl et al., 2010):

- 1) Who developed the data collection instrument?
- 2) When was the data collection instrument developed?
- 3) For which target group was the data collection instrument developed
- 4) The purpose for which the data collection instrument was developed
- 5) Number of questions in the data collection instrument
- 6) The structure of the data collection instrument (graded, categorically scored, etc.)
- 7) How the data collection instrument is rated

The following steps were considered in the presentation of data collection instruments adapted by other researchers (Tavşancıl et al., 2010).

- 1) Adapted by whom
- 2) Adapted when
- 3) Adapted for which target group
- 4) Adapted for what purpose
- 5) Number of questions in the data collection instrument
- 6) Structure of the data collection instrument
- 7) Scoring method

In addition, in the thesis review form, the validity and reliability coefficients obtained for data collection instruments were determined and it was checked to see whether or not these coefficients were appropriate to the structure of the data collection instrument and whether the level was high or low. When coding for this, coefficients of 0.70 and above are considered sufficient (Tavşancıl et al., 2010).

To determine the reliability of the review form, the agreement between the coding made by different coders was 85%; the agreement between codes made by the same encoder at different times was calculated as 95% (Tavşancıl et al., 2010). Within the scope of this research, the agreement between the coding made by two different coders for the review form was calculated as 97% and the agreement between the coding made by the researcher at two different times was 86%.

2.4. Data Analysis

Categorical and frequency analyses were used in the analysis of the data. In categorical analysis, the message is first divided into units, and then these units are grouped into categories according to specific criteria. In frequency analysis, the frequency of occurrence of units and elements is determined numerically (Bilgin, 2006). Also, theses in the relevant categories are discussed by citing remarkable examples of errors or deficiencies.

3. RESULTS

The findings are given according to the categories in the review form. Firstly, the use of the written data collection technique in postgraduate theses was examined and the findings are given in [Table 2](#).

Table 2. *Distribution for use of written data collection technique*

Data Collection Techniques	Master's Thesis	Master's Thesis		Dissertation
		f	%	
Written	Used	133	87.50	31
Response	Unused	19	12.50	6
Instrument	Total	152	100.00	37

According to [Table 2](#), the written data collection technique was used in 133 (87.50%) of the reviewed master's theses and in 31 of the dissertations. The written response instruments were used in a total of 272 master's theses and 115 dissertations.

The written response instruments were reviewed on the following bases: developed by the researcher, developed by another researcher, adapted by the researcher, and adapted by another researcher; the distribution is given in [Table 3](#).

Table 3. *Distribution of written data collection instruments by development and adaptation*

Measuring Instrument Development/ Adaptation Status	Master's Thesis		Dissertation	
	f	%	f	%
Researcher Developed	161	59.19	68	59.13
Researcher Adapted	-	-	-	-
Developed Instrument Used	52	19.12	17	14.78
Adapted Instrument Used	59	21.69	30	26.09
Total	272	100.00	115	100.00

When [Table 3](#) is examined, it can be seen that 161 (59.19%) of the written data collection instrument used at the master's degree were developed by the researcher, 52 (19.12%) were developed by other researchers, and 59 (21.69%) were adapted by other researchers. There are no adapted instruments used by the researcher. At doctoral degrees, 68 (59.13%) of the instruments were developed by the researcher, 17 (14.78%) were developed by other researchers and 30 (26.09%) were adapted by other researchers. Among the instruments used at doctoral degrees, there are no adapted instruments used by the researcher. At both degrees, the written response instruments used by the researcher were frequently those developed by the researcher.

3.1. Findings for Written Response Instruments Developed by Researcher

3.1.1. Category

Various data collection instruments have different properties in terms of measurement technique and the steps that need to be followed when developing them differ. Taking this into consideration, the researcher-developed instruments were grouped in three categories, namely, "questionnaire," "achievement/ability test," and "scale", and they were examined in accordance with the instrument development steps required by their respective categories. The data collection instruments that are not included in these categories are reviewed under the "other" category. These include checklists, forms for the social validity of the research, reinforcement determination lists, evaluation forms, personal information forms, and similar data collection instruments. The distribution by category of the written response instruments developed by the researcher is presented in [Table 4](#).

Table 4. Distribution by category of written response instruments

Category	Master's Thesis		Dissertation	
	f	%	f	%
Questionnaire	36	22.36	7	10.30
Achievement-Ability Test	7	4.35	8	11.76
Scale	12	7.45	4	5.88
Other	106	65.84	49	72.06
Total	161	100.00	68	100.00

According to [Table 4](#), of the 161 data collection instruments developed by the researcher at master's degrees, 36 (22.36%) were questionnaires, seven (4.35%) were achievement/ability tests, 12 (7.45%) were scales, and 106 (65.84%) were other. Of the 68 data collection instruments developed by the researcher at doctoral degrees, seven (10.30%) were questionnaires, eight (11.76%) were achievement/ability tests, four (5.58%) were scales, and 49 (72.06%) were other. The distribution of the development steps as reported or not for the data collection instruments developed by the researcher is presented in [Table 5](#).

Table 5. Distribution for reporting of development steps

Category	Development Steps	Master's Thesis	Dissertation
		f	f
Questionnaire	Reported	6	2
	Not reported	30	5
	Total	36	7
Achievement-Ability Test	Reported	1	5
	Not reported	6	3
	Total	7	8
Scale	Reported	12	4
	Not reported	-	-
	Total	12	4

When [Table 5](#) is examined, it can be seen that of all the instruments developed by the researcher at master's level, development steps were reported for six of the 36 questionnaires, one of the seven achievement/ability tests, and all 12 scales; at doctoral level, development steps were reported for one of the seven questionnaires, five of the eight achievement/ability tests, and all four scales. Accordingly, it is remarkable that the development steps for most of the data collection instruments developed by the researcher at both levels were not reported. The

distribution of complete/incomplete development steps for those data collection instruments where the development steps have been stated is given in [Table 6](#).

Table 6. *Distribution of complete/incomplete development steps*

Category	Development Steps	Master's Thesis f	Dissertation f
Questionnaire	Complete	1	-
	Incomplete	5	2
	Total	6	2
Achievement-Ability Test	Complete	-	-
	Incomplete	1	5
	Total	1	5
Scale	Complete	3	1
	Incomplete	9	3
	Total	12	4

According to [Table 6](#), of the 12 scales with reported development steps at master's level, three are complete; the development steps for only one scale at doctoral level are complete. However, six of the master's level questionnaires, nine of the scales, and one achievement/ability test were found to have incomplete development steps. In dissertations, two of the questionnaires, five of the achievement/ability tests, and three of the scales were found to have incomplete development steps. Incomplete development steps are a situation that is frequently encountered in all categories of instruments at both degrees.

When the data collection instruments determined to be missing in the development steps are reviewed, for two surveys at master's level and five at doctoral level, the most common missing steps are "item generation", "stating the rate of feedback", "conducting item analyses and/or qualitative analyses on the data obtained from the application, determining the psychometric properties", and "stating whether or not monitoring was carried out"; while the most common missing steps in one achievement and ability test level at master's level and five achievement and ability tests at doctoral level were determined as "preparing a test blueprint", "establishing a pool of items", "seeking expert opinion and explaining the characteristics of experts", and "pilot study". It was determined that the most common missing steps in seven scales at master's level and three at doctoral level were the "composition work", "content analysis", and "establishing a pool of items" steps.

3.1.2. *Proof of validity*

The first examination of the psychometric properties of the written response instruments developed by the researcher was carried out using validity prediction methods. Whether or not proof of validity for the results obtained by researcher-developed data collection instruments in the relevant category was stated and what kind of proof of validity was presented were examined and the findings are given in [Table 7](#). When [Table 7](#) is examined, it can be seen that while proof of validity is not provided for most of the questionnaires or any of the achievement/ability tests developed by the researcher at master's level, proof of validity is provided for most of the scales; and at doctoral level it can be seen that proof of validity is stated for two questionnaires, four achievement/ability tests, and two scales. At both degrees, proof of validity is not reported for questionnaires. At doctoral level, proof of validity was reported for four achievement/ability tests; the validity prediction method used for these was construct validity based on exploratory factor analysis (EFA) only; in addition, in only one achievement/ability test, were EFA and confirmatory factor analysis used together. The

distribution of validity prediction methods used in researcher-developed scales is given in [Table 8](#).

Table 7. *Distribution for reported proof of validity by instrument category*

Category	Proof of Validity	Master's Thesis	Dissertation
		f	f
Questionnaire	Reported	1	2
	Not reported	35	0
	Total	36	2
Achievement-Ability Test	Reported	-	4
	Not reported	7	4
	Total	7	8
Scale	Reported	11	2
	Not reported	1	2
	Total	12	4

Table 8. *Distribution of validity prediction methods used in scales developed by researcher*

Validity prediction Methods		Master's Thesis	Dissertation
		f	f
Content Validity	Reported	9	1
	Not reported	3	-
	Total	12	1
Construct validity	Reported	7	1
	Not reported	4	-
	Total	11	1
Criteria Based Validity	Reported	1	-
	Not reported	10	1
	Total	11	1

When [Table 8](#) is examined, it can be seen that at master's level, structure and content validity are mainly used for scales; while at doctoral level, content and construct validity were used for one scale. At both levels, expert opinion was used for content validity and EFA was frequently used for construct validity.

3.1.3. Proof of reliability

The second examination of the psychometric properties of the written response instruments developed by the researcher was carried out using reliability prediction methods. The distribution of reported proof of reliability for these instruments developed by the researcher is given in [Table 9](#).

Table 9. *Distribution for reported proof of reliability*

Category	Proof of Reliability	Master's Thesis	Dissertation
		f	f
Achievement-Ability Test	Reported	-	4
	Not reported	7	4
	Total	7	8
Scale	Reported	10	1
	Not reported	2	3
	Total	12	4

According to [Table 9](#), proof of reliability is reported for 10 of 12 scales at master's level; while at doctoral level, proof of reliability is reported for four of eight achievement/ability tests and one of four scales. It was observed that no proof of reliability was reported questionnaires at both levels and for seven achievement/ability tests at master's level.

In cases where researcher-developed written response instruments were used, instances were observed where proof for the instrument was reported using only one validity prediction method, as were instances where proof was reported using multiple validity prediction methods. Taking this into consideration, every single validity prediction method reported for the achievement/skill tests and scales was examined separately. The distribution of predicted reliability type for the achievement/ability test at doctoral level is given in [Table 10](#).

Table 10. *Distribution of achievement/ability test reliability coefficients by type and level*

Reliability Prediction Method		Dissertaion f
Alpha Reliability	Not predicted	3
	Appropriate and high level	1
	Total	4
Test-Retest Reliability	Not predicted	3
	Appropriate and high level	1
	Total	4
KR-20 Reliability	Appropriate and high level	4
	Total	4

According to [Table 10](#), Cronbach's Alpha reliability, test-retest, and KR-20 reliability were calculated for the achievement/ability tests developed by the researcher at the doctoral level. These calculated reliability coefficients were found to be appropriate to the structure of the measuring instrument used and the predicted value was also high (0.70 and above). The KR-20 coefficient was used most frequently as proof of reliability for achievement/ability tests.

The distribution of reliability coefficients in the scales developed by the researcher by predicted status and predicted reliability type and level is given in [Table 11](#).

Table 11. *Distribution of reliability coefficients in the scales by type and level*

Reliability Prediction Method		Master's Thesis f	Dissertation f
Alpha Reliability	Not predicted	1	3
	Appropriate and high level	10	1
	Appropriate and low level	1	-
	Total	12	4
Test-Retest Reliability	Not predicted	7	3
	Appropriate and high level	5	1
	Total	12	4
Split-half Reliability	Not predicted	10	4
	Appropriate and high level	2	-
	Total	12	4

According to [Table 11](#), Cronbach's Alpha, test-retest, and split-half reliability were calculated for scales at master's level. While Cronbach's Alpha is the most common value calculated for scales, it was observed that the alpha value calculated for one scale was low. At doctoral level,

Cronbach's alpha and test-retest reliability were calculated for the scales. These coefficients were found to be high and appropriate to the scale structure.

3.2. Findings on Written Response Instruments Developed by Other Researchers

3.2.1. Category

It was determined that 52 data collection instruments were developed by other researchers at master's level and 16 at doctoral level. The first criterion taken into consideration when evaluating the presentation of these instrument in postgraduate theses is the instrument's category. The distribution of data collection instruments developed by other researchers by categories is given in [Table 12](#).

Table 12. Distribution of data collection instruments developed by other researchers by category

Category	Master's Thesis		Dissertation
	f	%	f
Achievement-Ability Test	15	28.85	4
Scale	29	55.77	3
Other	8	15.38	10
Total	52	100.00	17

When [Table 12](#) is examined, it can be seen that 15 (28.85%) of the 52 data collection instruments at master's level are achievement/ability tests, 29 (55.77%) are scales and eight (17.30%) are in the other data collection instruments category; while at doctoral level, four of the instruments are achievement/ability tests, three of them are scales and 10 of them are in the other category. It was also determined that no questionnaire developed by another researcher was at the master's or doctorate degrees.

Another examination of the data collection instruments developed by other researchers looked at the how the instrument was introduced. The instrument categories were examined to see if the information that needs to be presented when introducing the instrument was reported. The distribution for the introduction of data collection instruments developed by other researchers in the "achievement/ability test" and the "scale" categories is given in [Table 13](#).

Table 13. Distribution of introduction status by instrument category

Category	Introduction Status	Master's Thesis	Dissertation
		f	f
Achievement-Ability Test	Introduced	3	1
	Not Fully Introduced	10	2
	Not introduced	2	1
	Total	15	4
Scale	Introduced	15	1
	Not Fully Introduced	12	1
	Not introduced	2	1
	Total	29	3

According to [Table 13](#), at master's level, 10 achievement/ability tests were fully introduced while three were not fully introduced or not introduced at all. Of the scales, 15 were introduced, 12 were not fully introduced and two were not introduced at all. It can be seen that at doctoral level, one of the four achievement/ability tests was introduced, two were not fully introduced, and one not introduced at all; while of the three scales, one was introduced, one was not fully introduced, and one not introduced at all.

It was determined that at master's level, the most frequently observed deficiency in the introduction of achievement/ability tests developed by other researchers was due to information about the structure of the instrument (graded scale, etc.) not being given. At doctoral level, it was found that information about the number of questions in the instrument, the structure of the instrument, and the instrument's scoring method was not given.

3.2.2. Proof of validity

The written response instruments developed by other researchers were also examined in terms of the methods used for predicting the validity. Information reported in the theses with respect to validity prediction methods and proof of validity was examined under two categories, namely, "Original proof of validity" and "Proof of validity as reported in the study". Whether or not original proof of validity was given for data collection instruments developed by other researchers was examined and the findings are given in [Table 14](#).

Table 14. *Distribution of reported original proof of validity*

Category	Original Proof of Validity	Master's Thesis	Dissertation
		f	f
Achievement-Ability Test	Reported	10	1
	Not reported	5	3
	Total	15	4
Scale	Reported	14	2
	Not reported	15	1
	Total	29	3

According to [Table 14](#), at master's level, original proof of validity was given for 10 of achievement/ability test. It was determined that the most frequently cited proof was construct validity, content validity, and criterion-based validity. At this level, original proof of validity was given for 14 of the scales and this evidence was frequently based on construct validity, criterion-based validity, and content validity. At doctoral level, original proof of validity was given for one achievement/ability test and two scales. This evidence was often based on construct validity.

Presentation of proof of validity for data collection instruments developed by other researchers within the scope of the study was examined and it was determined that absolutely no proof of validity was reported for the achievement/ability tests within the scope of the study; and that only at master's level was construct validity proof of validity reported within the scope of the study for two scales.

3.2.3. Proof of reliability

Information regarding proof of reliability reported in postgraduate theses was examined as "Original proof of reliability" and "Proof of reliability reported in the study". Presentation of original proof of reliability in theses where an instrument developed by someone else was used was examined and the distribution is given in [Table 15](#).

Table 15. *Distribution of reported original proof of reliability*

Category	Original Proof of Reliability	Master's Thesis	Dissertation
		f	f
Achievement-Ability Test	Reported	10	2
	Not reported	5	2
	Total	15	4
Scale	Reported	22	2
	Not reported	7	1
	Total	29	3

According to [Table 15](#), at master's level original proof of reliability was reported for 10 achievement/ability tests and 22 scales, while at doctoral level it was reported for just two achievement/ability tests and two scales. It was found that the Cronbach's Alpha coefficient, test-retest, split-half reliability, and inter-rater reliability coefficient were the ones most frequently calculated for achievement/ability tests developed by other researchers at master's level and that their level was high. At doctoral level, it was observed that original proof of reliability based on test-retest reliability was obtained and that these values were high. The Cronbach's alpha coefficient, test-retest and two-half reliability evidence were given as proof of original reliability in all 22 scales used at the master's level, and the values obtained from them were found to be high. However, the Cronbach's Alpha reliability coefficient calculated for the two instruments was found to be low. At doctoral level, the original proof of reliability was calculated based on Cronbach's alpha, test-retest, split-half, and KR - 20 reliability.

In addition, proof of reliability for written response instruments developed by other researchers and obtained within the scope of postgraduate theses was examined. While the reliability coefficient calculated for five scales developed by other researchers at master's level was high, it was determined that the reliability coefficient calculated for two scales was low. At doctoral level, it was seen that no proof of reliability was reported in the relevant theses within the scope of the study.

3.3. Findings on Written Response Instrument Adapted by Other Researchers

3.3.1. Category

It was determined that 59 written response instruments at master's level and 30 instruments at doctoral level were adapted to Turkish culture by other researchers. The distribution of instruments adapted by other researchers by the categories of "questionnaire," "achievement/ability test," and "scale" is given in [Table 16](#).

Table 16. *Distribution of adapted instruments by category*

Category	Master's Thesis f	Dissertation f
Achievement-Ability Test	8	2
Scale	36	24
Other	13	4
Total	59	30

In [Table 16](#), it can be seen that eight of the written response instruments at master's level were achievement/ability tests, 36 were scales, and 13 fell into the "other" category; while at doctoral level, two were achievement/ability tests, 24 were scales, and four were in the "other" category.

Another examination, this time of written response instruments adapted by other researchers, sought whether or not these instruments had been introduced and if so whether or not the reported information was complete. The introduction status of the data collection instruments adapted by other researchers was examined and the findings are given in [Table 17](#). When [Table 17](#) is examined, it can be seen that three achievement/ability tests and 27 scales adapted by other researchers were introduced but that four achievement/ability tests and nine scales were not fully introduced at master's level. It was found that one achievement/ability test and 20 scales adapted by other researchers were introduced but that one achievement/ability test and three scales had incomplete introduction information at doctoral level.

Table 17. *Distribution of introduction status of adapted instruments*

Category	Introduction Status	Master's Degree	Dissertation
		f	f
Achievement/Ability Test	Introduced	3	1
	Not Fully Introduced	4	1
	Not introduced	1	-
	Total	8	2
Scale	Introduced	27	20
	Not Fully Introduced	9	3
	Not introduced	-	1
	Total	36	24

At master's level, it was determined that incomplete information was given for the achievement/ability test, most frequently in the "number of questions in the instrument", "structure (graded scale etc.)", and the "scoring method" areas. At doctoral level, incomplete information was given only for "instrument structure (graded scale etc.)". For scales, it was found that at master's level, incomplete information was given for "adapted by whom", "adapted when", and "scoring method"; and that at doctoral level incomplete information was given for "instrument structure (graded scale etc.)", "adapted by whom", "adapted when", "number of questions in the instrument", and "scoring method".

3.3.2. Proof of validity

The first examination of the psychometric properties of written response instruments adapted by other researchers was made using validity prediction methods. For these instruments, the original proof of validity reported by the researcher who developed the instrument, the proof of validity obtained in his/her own studies by the researcher who adapted the instrument to Turkish culture, and the proof of validity reported in their studies by researchers who used the adapted instrument all needed to be reported. Accordingly, the information obtained was presented as "Original proof of validity," "Proof of validity reported by researchers who adapted the instrument to Turkish culture," and "Proof of validity reported in the study". Whether or not original proof of validity for written response instruments adapted by other researchers was reported was examined and the resulting distribution is given in [Table 18](#).

Table 18. *Distribution of original validity prediction methods in adapted scales*

Original Validity Prediction Method		Master's Thesis	Dissertation
		f	f
Construct validity	Reported	7	2
	Not reported	7	-
	Total	14	2
Predictive Validity	Reported	3	-
	Not reported	11	2
	Total	14	2
Criteria Based Validity	Reported	3	-
	Not reported	11	2
	Total	14	2
Content Validity	Reported	13	-
	Not reported	1	2
	Total	14	2

According to [Table 18](#), at master's level, construct validity was given as original proof of validity for seven scales, criterion-based and predictive validity for three scales, and content

validity for one scale. At doctoral level, construct validity was reported for the original validity of the two scales.

Whether or not proof of validity determined by researchers who took written response instruments adapted by other researchers and adapted them to Turkish culture was stated was examined and the resulting distribution is given in [Table 19](#).

Table 19. *Distribution for reported proof of validity for adapted instruments*

Original Validity Prediction Method	Proof of Validity	Master's Thesis f	Dissertation f
Achievement/Ability Test	Reported	1	-
	Not reported	7	2
	Total	8	2
Scale	Reported	17	10
	Not reported	19	14
	Total	36	24

In [Table 19](#), it can be seen that proof of validity determined by researchers who adapted the data collection instrument to Turkish culture was stated for one achievement/ability test and 17 scales at master's level and for 10 scales at doctoral level. It was determined that content and construct validity were used as the validity determination method for the achievement/ability test at master's level. It was found that construct validity and proof based on criterion-based validity were used for scales. In addition to this, it was seen that at doctoral level, construct validity based on factor analysis was stated for almost all the scales, while proof was reported based on criterion-based validity by using similar scales in another instrument.

It was determined that proof of validity obtained within the scope of the study was not reported for written response instruments adapted by other researchers.

3.3.3. Proof of reliability

The second examination of the psychometric properties of written response instruments developed by adapted by other researchers was made using reliability prediction methods. The findings are presented as "Original proof of reliability," "Proof of reliability reported by researchers who adapted the instrument to Turkish culture," and "Proof of reliability reported in the study". For data collection instruments adapted by other researchers, whether or not original proof of validity determined by researchers who developed the instrument for the original culture was stated was examined and the findings are given in [Table 20](#).

Table 20. *Distribution of reported original proof of reliability for adapted instruments*

Category	Original Proof of Reliability	Master's Thesis f	Dissertation f
Achievement-Ability Test	Reported	1	-
	Not reported	7	2
	Total	8	2
Scale	Reported	17	6
	Not reported	19	18
	Total	36	24

According to [Table 20](#), at master's level, original proof of reliability was reported for one achievement/ability test and 17 scales adapted by other researchers. It was determined that at doctoral level, original proof of reliability was not reported for achievement/ability tests but

was reported for just six scales. At master's level, the original reliability prediction method for one achievement/ability test was calculated using the KR - 21 coefficient.

Table 21 shows the distribution by type and level of predicted reliability for stated original reliability coefficients stated as have been predicted by the researchers who developed the instrument for scales adapted by other researchers.

Table 21. *Distribution of prediction of original reliability coefficients in adapted scaled by type and level of predicted reliability*

Original Reliability Prediction Method		Master's Thesis	Dissertation
		f	f
Alpha Reliability	Not predicted	3	1
	Appropriate and high level	11	3
	Appropriate and low level	3	1
	No information about level	-	1
	Total	17	6
Test-Retest Reliability	Not predicted	11	6
	Appropriate and high level	6	-
	Total	17	6
KR-20 Reliability	Not predicted	17	5
	Appropriate and high level	-	1
	Total	17	6
KR - 21 Reliability	Not predicted	16	6
	Appropriate and high level	1	-
	Total	17	6

According to Table 21, Cronbach's Alpha reliability was estimated for 14 of the 17 scales as the original proof of reliability in theses at master's level and the estimated value was high for 11 scales but low for three scales. For the three scales, the relevant coefficients were not reported. Remarkably, high test-retest reliability coefficient was calculated for six of the 17 scales, while KR - 21 coefficient was calculated for one scale. It can be seen that at doctoral level, Cronbach's Alpha reliability was predicted for five of the scales of which three were found to have appropriate scale structure and high values; while one was found to have an appropriate scale structure but a low value. For one scale the relevant coefficient was stated as having been predicted but no information for this value is given and that the KR - 20 coefficient was calculated for one other scale.

In the examination of the psychometric properties of written response instruments adapted by other researchers, the reporting in postgraduate theses of the proof of reliability obtained by researchers who adapted the instrument to Turkish culture was examined. The distribution of reported proof of reliability determined by the researchers adapting the instrument to the Turkish culture is given in Table 22. When Table 22 is examined, stated proof of reliability determined by researchers who adapted the instruments to Turkish culture can be seen for four achievement/ability tests and 24 scales at master's level and for two achievement/ability tests and 13 scales at doctoral level.

Table 22. *Distribution of reported proof of validity determined by researchers who took adapted instruments and adapted them to turkish culture*

Category	Proof of Reliability	Master's Thesis	Dissertation
		f	f
Achievement-Ability Test	Reported	4	2
	Not reported	4	-
	Total	8	2
Scale	Reported	24	13
	Not reported	12	11
	Total	36	24

Table 23 and Table 24 show the distribution of the type and level of the coefficients for achievement/ability tests and scales, respectively. According to Table 23, the reliability coefficients predicted by the researchers who adapted the instrument to Turkish culture were Cronbach's Alpha, the split-half test, and the KR - 21 at master's level and Cronbach's Alpha, test-retest, and parallel test form reliability at doctoral level. It was determined that these calculated coefficients were appropriate to the structure of the instrument and their level was high. It was determined that the reliability of the split-half test for the achievement/ability test was low. In Table 24, it is seen that Cronbach's Alpha test re-test, split-half, and KR - 20, and parallel test reliability were used respectively for scales at master's level. For scales at doctoral level, Cronbach's alpha, test-retest, split-half, and parallel test reliability were used. Of these scales, it was found that the Cronbach's Alpha reliability coefficient was low for two scales at master's level and one at doctoral level and that at doctoral level, the test-retest and split-half reliability coefficients for one scale were low. It was noted that the KR - 20 coefficient, which is not appropriate to the structure of the scale, was calculated as proof of reliability at master's level. The majority of the scales used at both levels were found appropriate to the structure of the instrument and to have high proof of reliability.

Table 23. *Distribution of achievement/ability test adapted to turkish culture by reliability coefficient prediction and type and level of predicted reliability*

Original Reliability Prediction Method		Master's Thesis	Dissertation
		f	f
Cronbach's Alpha Reliability	Not predicted	7	1
	Appropriate and high level	1	1
	Total	8	2
Test-Retest Reliability	Not predicted	8	1
	Appropriate and low level	-	1
	Total	8	2
Split-half Reliability	Not predicted	7	1
	Appropriate and low level	1	1
	Total	8	2
KR - 21 Reliability	Not predicted	7	2
	Appropriate and high level	1	-
	Total	8	2
Parallel Test Reliability	Not predicted	8	1
	Appropriate and high level	-	1
	Total	8	2

Table 24. *Distribution by reliability coefficients type and level of predicted reliability for adapted scales*

Reliability Prediction Method		Master's Thesis f	Dissertation f
Cronbach's Alpha Reliability	Not predicted	12	12
	Appropriate and high level	22	11
	Appropriate and low level	2	1
	Total	36	24
Test-Retest Reliability	Not predicted	25	21
	Appropriate and high level	11	2
	Appropriate and low level	-	1
	Total	36	24
KR-20 Reliability	Not predicted	34	24
	Appropriate and high level	2	-
	Total	36	24
Split-half Reliability	Not predicted	26	22
	Appropriate and high level	10	1
	Appropriate and low level	-	1
	Total	36	24
Parallel Test Reliability	Not predicted	34	23
	Appropriate and high level	2	1
	Total	36	24

When examined within the scope of the study, it was found that for instruments adapted by other researchers, at master's level, proof of reliability was reported for eight out of 36 scales and that for all of them a high Cronbach's Alpha reliability coefficient was obtained. At doctoral level, it was found that there was only one scale with reported proof of reliability and that a high Cronbach's Alpha value was obtained for this scale.

4. DISCUSSION and CONCLUSION

In this study, written response instruments used in postgraduate theses completed in the field of special education between 2015 and 2018 and their psychometric properties were examined. It was seen that most of the written data collection technique at both degrees were researcher-developed and that no researcher-adapted written response instruments were used.

The most frequently used researcher-developed written response instruments were the questionnaire at master's level and the achievement/ability test at doctoral level. Consistent with the findings of this study, it was determined that surveys and achievement tests were frequently used in the studies published in the field of educational sciences both in Turkey and abroad (Doğru et al., 2012; Erdem, 2011; Tavşancıl et al., 2010; Yalçın, 2016; Yalçın et al., 2015).

It was noted that development steps were not reported in more than half of the researcher-developed questionnaire, achievement/ability test, and scale. In a similar study, Tavşancıl et al. (2010) stated that there were significant deficiencies in reporting the development steps of the data collection instruments developed by the researchers in the relevant theses at master and doctoral level; given that this information given in a limited number of theses, it was concluded that the measurement procedures were not done with sufficient quality. In studies by Başol and Akın (2006) and Arık and Türkmen (2009), it was emphasized that in the articles they reviewed, there was not enough information about the data collection instruments used and that this situation could negatively impact the intelligibility of the studies. This indicates that the deficiencies in the introduction of data collection instruments used in studies are still seen today.

When the results of reported proof of validity in researcher-developed instruments are examined, it can be seen that proof of validity was not reported for any of the

achievement/ability tests used at master's level, that it was presented in a large majority of the scales, and that the most frequently used proof of validity for scales were construct and content validity. At doctoral level, it was observed that proof of validity for researcher-developed achievement/ability tests and scales was rarely reported. In findings relating to proof of reliability for researcher-developed instruments, it was seen that at master's level, proof of reliability was not reported for any questionnaire or achievement/ability test but that proof of reliability was reported for a large majority of the scales. Cronbach's Alpha coefficient was frequently used for predicting reliability at the master's level; the calculated coefficients were found to be high and consistent with the scale's scoring structure. Similarly, in the study conducted by Mor-Dirlik and Kula-Kartal (2016), it was stressed that Cronbach's Alpha coefficient was the most commonly used evidence of reliability in both education and psychology. It was found that proof of reliability was given for half the achievement/ability tests and only one scale used at doctoral level. For achievement/ability tests, it was determined that the KR - 20 coefficient was calculated most frequently, and these coefficients were found to be high and consistent with the instrument's scoring method. Tavşancıl et al. (2010) showed that the KR-20 reliability coefficient was frequently reported for achievement tests in postgraduate theses in the field of educational sciences.

The results of the study show that there were significant deficiencies in the validity and reliability of the written response instruments developed by the researcher. This will bring controversy about the accuracy of the results obtained from research in which data collection instruments of dubious validity and reliability were used. As it is known, validity and reliability are the basic psychometric properties required of a data collection instruments. The meaning of the scores obtained from a data collection instrument and the lack of evidence that the instrument makes accurate measurements without confusing research variables with other variables could cause the relevant research results to become questionable for both readers and other researchers in terms of the scientific method. It was noted that although there are many methods for predicting validity for those instruments that had proof of validity presented, only proof of construct-related evidence of validity was reported. However, if validity studies are considered to be the process of collecting evidence for the accuracy of the scores obtained from the data collection instruments, it stands to reason that evidence obtained from different validation methods will contribute to the accuracy of the research results. The same is true for reliability. In related studies (Başol & Akın, 2006; Büyüköztürk & Kutlu, 2006; Tavşancıl et al., 2010), it was emphasized that the failure to present validity and reliability information for data collection instruments was the most serious methodological problem.

Another conclusion of the study was that of the data collection instruments developed by other researchers, the scale was the one used most frequently at master's level and the achievement/ability test at doctoral level. When the presentation status of the developed instruments (structure, scoring method, etc.) was examined, it was seen that at master's level in most of the achievement/ability tests and the scales, the most common missing information were *structure* and *scoring* method. At doctoral level, this includes *the number of questions*, *the structure of the instrument*, and *the scoring* method.

For the data collection instruments developed by other researchers, it was observed that original proof of validity for the achievement/ability test was mostly reported at master's level and that this proof of validity was construct validity based on factor analysis. At doctoral level, it was found that the achievement/ability test and the scale were those instruments for which original proof of validity was not reported. However, it was found that construct validity was reported for most of the written response instruments for which original proof of validity was reported. At master's level, it was reported that Cronbach's Alpha reliability coefficient, which has a high reliability coefficient and is appropriate to the scale structure, was the coefficient calculated the

most for achievement/ability tests and scales developed by other researchers. However, it was seen that the original reliability coefficients were presented based on test-retest, split-half, and inter-rater reliability. At doctoral level, it was noted that the frequency of reporting original proof of reliability for achievement/ability tests and scales was low. In addition, when calculating the reliability coefficient for some scales, it was noted that the KR-20 coefficient, which is not appropriate for the graded structure of these instruments, was calculated. It was determined that proof of reliability obtained within the scope of the study for achievement/ability tests and scales developed by other researchers was not reported.

In the reviewed theses, it was determined that the data collection instruments adapted by other researchers were mostly in the scale and achievement/ability test category and that most of these instruments were introduced. The original proof of validity for scales adapted by other researchers was most often found to be based on construct validity. At neither level original proof of validity was reported for adapted achievement/ability tests. Similarly, within the scope of the study, it was noted that no validation work was carried out for the adapted written response instruments. Original proof of reliability was given for scales adapted by other researchers. Cronbach's Alpha reliability was frequently used for scales at both levels. However, it was found that the KR - 21 coefficient, which is not appropriate for graded scales, was calculated at master's level and that the KR - 20 coefficient was calculated at doctoral level. When proof of reliability status obtained by researchers who adapted instruments to Turkish culture was examined, it was seen that proof of reliability was reported for most scales at both levels. It was determined that Cronbach's Alpha reliability coefficients, which are appropriate to the instrument's structure and have a high level, were used for the data collection instruments at both levels. However, it was observed that at master's level, a low Cronbach's Alpha coefficient was obtained for some scales. In addition, again at master's level, it was found that the KR - 20 reliability coefficient, which is not appropriate for determining the reliability of scales and is applied only for dichotomously scored items, was used. Similarly, in a study by Tavşancıl et al. (2010), it was seen that reliability prediction methods such as KR-20 and KR-21 coefficients, which are not appropriate for data collection instruments consisting of graded items and which can only be used when the item structure is dichotomous, were reported when predicting reliability for scales. For instruments adapted by other researchers, within the scope of the study, it was found that Cronbach's Alpha proof of reliability was presented for scales only.

In the reviewed theses, the existence of several serious repeated mistakes was noted. The most common of these repeated errors is the discrepancy between the written response instrument and its name. Some of the written response instrument used in master's theses were called *questionnaire* but it was seen that these instruments were actually scales that give total scores. For example, in the instruments called “*Frequency of Use of Phonological Awareness in Teaching Activities Questionnaire*” and “*Strengths and Difficulties Questionnaire*” it was seen that the items were scored using a five-point Likert scale able to obtain total scores. Similarly, in the study conducted by Tavşancıl et al. (2010), problems were seen that stemmed from the concepts of questionnaire and scale being used interchangeably. Another remarkable situation relating to written data collection instruments is the fact that the names in some theses are quite general and not understandable. For example, as seen in the “*Collecting Effectiveness Data*”, “*Discretionary Reinforcement Processing Criteria-Dependent Measurement Tool*” and “*Start Level Data Form*” and, likewise, the “*Productivity Data Collection Form*” and “*Start Level Sessions Form*”, it was seen that some of the written response instrument names are very general with no information given as to what structure it measures.

Another common error in the reviewed theses is related to obtaining and interpreting proof of validity. For example; for one scale, “*the findings obtained from a study of the Turkish version's*

psychometric characteristics concluded that the scale was valid and reliable”, and in another study, “work on the original version of the scale and the Turkish form presented proof of validity and reliability” saying that the scale was valid and reliable. This shows that the researcher has incomplete or inaccurate information about how to reflect the basic psychometric properties of the data collection instruments. Some of the researchers, on the other hand, concluded that the written response instrument is valid based on the assumptions of factor analysis. For example, as proof of validity for one scale that was used, “the KMO Barlett coefficient was applied for construct validity and was found to be 0.79”, and it was seen that the researcher accepted validity assumptions as proof of validity.

It has been determined that there are serious deficiencies in introducing the data collection instruments used in the postgraduate theses and reporting their psychometric properties. At this point, it might be a good suggestion for researchers to work on developing their research methodology and academic reporting skills and for official units to be formed where they could receive advice. In addition, a "Thesis Writing Guide" based on standards to be formulated jointly by all universities could be prepared. It is noteworthy that similar errors are repeated in the reviewed theses. In this respect, graded scoring keys or Thesis Review Forms could be developed for research reports that can be used by both the researcher and interested parties.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

Acknowledgement

This paper was prepared as part of a master's thesis completed by the first author under supervision of the second author.

ORCID

Gamze Sarıkaş,  <https://orcid.org/0000-0001-6591-467X>

Safiye Bilican Demir  <https://orcid.org/0000-0001-9564-9029>

5. REFERENCES

- Ağca, Ö. (2014, 25-27 Eylül). *Türkiye’deki üniversitelerin lisans ve lisansüstü programlarının özel eğitim açısından incelenmesi* [Conference presentation abstract]. 24. Ulusal Özel Eğitim Kongresi, Edirne, Türkiye.
- Alper, A., & Gülbahar, Y. (2009). Trends and issues in educational technologies: A review of recent research in TOJET. *The Turkish Online Journal of Educational Technology - TOJET*, 8(2), 124-135.
- Arık, R. S., & Türkmen, M. (2009). Eğitim bilimleri alanında yayınlanan bilimsel dergilerde yer alan makalelerin incelenmesi. Retrieved from <http://www.eab.org.tr/eab/2009/pdf/488.pdf>
- Aydın, A., & Uysal, Ş. (2011). Türkiye’de ve yurt dışında eğitim yönetimi alanında yapılan doktora tezlerinin konu, yöntem ve sonuçlar açısından değerlendirilmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 14(1), 177-201.
- Başol, G., & Akın, U. (2006, 13-15 Eylül). *2001-2006 yılları arasında Türkiye’de eğitim alanında belli başlı indeksli dergilerde yayımlanan araştırma makalelerinin metodolojik bakımdan değerlendirilmesi* [Conference presentation abstract]. XV. Eğitim Bilimleri Kongresi, Muğla, Türkiye.
- Benligiray, S. (2009). Türkiye’de insan kaynakları yönetimi alanında yapılan lisansüstü tezler ve bu tezlerde incelenen temaların analizi: 1983-2008 dönemi [The theme analysis of the

- postgraduate theses written on human resource management in Turkey:1983-2008 period]. *Eskişehir Osmangazi Üniversitesi İİBF Dergisi*, 4(1), 167-197.
- Bilgin, N. (2006). *Sosyal bilimlerde içerik analizi -Teknikler ve örnek çalışmalar*. Ankara: Siyasal Kitapevi.
- Büyüköztürk, Ş., & Kutlu, Ö. (2006). Sosyal bilim araştırmalarında yöntem sorunu. *Sosyal Bilimlerde Süreli Yayıncılık- 2006-I. Ulusal Kurultay Bildirileri* (pp. 113–122). Ankara: TÜBİTAK Yayını.
- Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö.E., Karadeniz, Ş., & Demirel, F. (2016). *Bilimsel araştırma yöntemleri*. Ankara: Pegem Akademi Yayınları.
- Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). New York: Routledge.
- Coşkun, İ., DüNDAR, Ş., & Parlak, C. (2014). Türkiye’de özel eğitim alanında yapılmış lisansüstü tezlerin çeşitli değişkenler açısından incelenmesi (2008-2013) [The analysis of the postgraduate thesis written on special education in terms of various criteria in Turkey (2008-2013)]. *Ege Eğitim Dergisi*, 15(2), 375-396.
- Çiltaş, A., Güler, G., & Sözbilir, M. (2012). Türkiye’de matematik eğitimi araştırmaları: Bir içerik analizi çalışması. [Mathematics education research in Turkey: A content analysis study] *Kuram ve Uygulamada Eğitim Bilimleri*, 12(1), 565-580.
- Demirok, M.S., Bağlama, B., & Beşgül, M. (2015). A content analysis of the studies in special education area. *Procedia - Social and Behavioral Sciences*, 197, 2459-2467.
- Diken, İ.H., Ünlü, E., & Karaaslan, Ö. (2008). *Zihinsel yetersizlik ve yaygın gelişimsel bozukluk alanlarında lisansüstü tez bibliyografyası*. Ankara: Maya.
- Diken, İ. H., Görgün, B., Öğülmüş, K., Kurnaz, E., & Baki, K. (2016). *Zihin yetersizliği ve otizm spektrum bozukluğu alanlarında lisansüstü tez bibliyografisi 2008-2015*. Ankara: Eğiten.
- Doğan, H., & Tok, T. N. (2018). Türkiye’de eğitim bilimleri alanında yayımlanan makalelerin incelenmesi: Eğitim ve Bilim Dergisi örneği. *Curr Res Educ*, 4(2), 94-109.
- Doğru, M., Gençosman, T., Ataalkın, N.A. & Şeker, F. (2012). Fen bilimleri eğitiminde çalışılan yüksek lisans ve doktora tezlerinin analizi [Analysis of the postgraduate and doctoral theses conducted on sciences education]. *Türk Fen Eğitimi Dergisi*, 9(1), 49-62.
- Doğru, S.Y., Özlü, Ö., Kañeşme, C., & Doğru, S. (2015). Özel eğitim üzerine yapılan proje çalışmalarının değerlendirilmesi [Quantification of projects done on special education]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 15(Özel Sayı), 286-296.
- Erdem, D. (2011). Türkiye’de 2005–2006 yılları arasında yayımlanan eğitim bilimleri dergilerindeki makalelerin bazı özellikler açısından incelenmesi: Betimsel bir analiz. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 2(1), 140-147.
- Evrekli, E., İnel, D., Deniz, H., & Balım, A.G. (2011). Fen eğitimi alanındaki lisansüstü tezlerde yöntemsel ve istatistiksel sorunlar [Methodological and statistical problems in graduate theses in the field of science education]. *Ilkogretim Online - Elementary Education Online*, 10(1), 206-218.
- Göktaş, Y., Hasançebi, F., Varışoğlu, B., Akçay, A., Bayrak, N., Baran, M., & Sözbilir, M. (2012). Türkiye’deki eğitim araştırmalarında eğilimler: Bir içerik analizi. *Kuram ve Uygulamada Eğitim Bilimleri*, 12(1), 443-460.
- Güner-Yıldız, N., Melekoğlu, M.A., & Paftalı, A.T. (2016). Türkiye’de özel eğitim araştırmalarının incelenmesi [Special education research in Turkey]. *Ilkogretim Online - Elementary Education Online*, 15(4), 1076-1089. <https://doi.org/10.17051/ieo.2016.06677>
- Horst, P. (1966). *Psychological measurement and prediction*. Belmont: Wadsworth Pub. Co.
- Karadağ, E. (2009). A thematic analysis on doctoral dissertations made in the area of education sciences. *Journal of Kırşehir Education Faculty*, 10(3), 75-87.
- Karasar, N. (2016). *Bilimsel araştırma ve yöntemi*. Ankara: Nobel Yayın Dağıtım.

- Köklü, N., & Büyüköztürk, Ş. (1999). Eğitim bilimleri alanında öğrenim gören lisans-üstü öğrencilerin araştırma yeterlikleri konusunda öğretim üyelerinin görüşleri [Advisors' opinions about the research competencies of masters and doctorate students in educational sciences]. *Eğitim ve Bilim*, 23(112), 18-28.
- Kula-Kartal, S., & Mor-Dirlik, E. (2016). Geçerlik kavramının tarihsel gelişimi ve güvenilirlikte en çok tercih edilen yöntem: Cronbach Alfa katsayısı [Historical development of the concept of validity and the most preferred technique of reliability: Cronbach alpha coefficient]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 16(4), 1865-1879.
- Kutluca, T., & Demirkol, M. (2016). Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi dergisinin bibliyometrik analizi [Bibliometric analysis of journal of Dicle University Ziya Gökalp Faculty of Education]. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, 28, 108-118. <https://doi.org/10.14582/DUZGEF.674>
- Kutluca, T., Birgin, O., & Gündüz, S. (2018). Türk Bilgisayar ve Matematik Eğitimi Dergisi'nde yayımlanmış makalelerin içerik analizi bağlamında değerlendirilmesi [Evaluation of the published articles in turkish journal of computer and mathematics education according to content analysis]. *Türk Bilgisayar ve Matematik Eğitimi Dergisi*, 9(2), 390-412. <https://doi.org/10.16949/turkbilmate.332518>
- Lee, M.H., Wu, Y.T., & Tsai, C.C. (2009). Research trends in science education from 2003 to 2007: A content analysis of publications in selected journals. *International Journal of Science Education*, 31(15), 1999-2020.
- Melekoğlu, M.A. (2014). Special education today in Turkey. A. F. Rotatori, J. P. Bakken, S. Burkhardt, F.E. Obiakor, & U. Sharma (Eds.). *Special education international perspectives: Practices across the globe* (pp. 529-557). Bingley, UK: Emerald Group Publishing.
- Metin, N.E. (2012). Özel Gereksinimli çocuklar. In E. Nilgün Metin (Ed.), *Özel gereksinimli çocuklar I*. Ankara: Maya Akademi.
- Ozan, C., & Köse, E. (2014). Eğitim programları ve öğretim alanındaki araştırma eğilimleri [Research trends in curriculum and instruction]. *Sakarya University Journal of Education*, 4(1), 116-136.
- Özenç, E.G., & Özenç, M. (2013). Türkiye'de üstün yetenekli öğrencilerle ilgili yapılan lisansüstü eğitim tezlerinin çok boyutlu olarak incelenmesi [The multidimensional examination of master-doctorial dissertations made in Turkey about gifted and talented students]. *Türkiye Sosyal Araştırmalar Dergisi*, 171, 13-28.
- Seçer, D., Ay, D., Ozan, C., & Yılmaz, B.Y. (2014). Rehberlik ve psikolojik danışma alanındaki araştırma eğilimleri: Bir içerik analizi [Research trends in the field of guidance and psychological counseling: A content analysis]. *Turkish Psychological Counseling and Guidance Journal*, 5(41), 49-60.
- Staton, A.Q.S., & Wulff, D.H. (1984). Research in communication and instruction: Categorization and synthesis. *Communication Education*, 33(4), 377-391.
- Şenyurt, S., Özer-Özkan, Y. (2017). Eğitimde ölçme ve değerlendirme alanında yapılan yüksek lisans tezlerinin tematik ve metodolojik açıdan incelenmesi [A thematic and methodological analysis of master's dissertations in the field of measurement and evaluation in education]. *Ilkogretim Online - Elementary Education Online*, 16(2), 628-653, 2017. <https://doi.org/10.17051/ilkonline.2017.304724>
- Ünlü, Ö., Güner-Yıldız, N., & Aktar, E. (2020). Investigation of doctoral dissertations in special education in Turkey. *Ilkogretim Online - Elementary Education Online*, 19(2), 923-931. <https://doi.org/10.17051/ilkonline.2020.695825>
- Tavşancıl, E., Çokluk, Ö., Gözen Çıtak, G., Kezer, F., Yalçın Yıldırım, Ö., Bilican, S., Büyükturan, B.E., Şekercioğlu, G., Yalçın, N., Erdem, D., & Özmen, T.D., (2010). *The*

- investigation of theses completed at the institutes of educational sciences (2000-2008)*. Ankara University Scientific Research Project Final Report, Ankara.
- Tiryaki, E.N. (2017). Evaluation of the postgraduate theses written in the field of special education in terms of language education and teaching. *International Online Journal of Educational Sciences*, 9(2), 454-463.
- Tiryakioğlu, Ö. (2014). Content analysis of the articles published in the Ankara University Special Education Journal within the years 2004-2013. *Procedia - Social and Behavioral Sciences*, 143, 1164-1170.
- Tosuntaş Ş.B., Emirtekin E., & Süral İ., (2019). Eğitim ve öğretim teknolojileri konusunda yapılan tezlerin incelenmesi (2013-2018) [Examination of theses on educational and instructional technologies (2013-2018)]. *Journal of Higher Education and Science*, 9(2), 277-286. <https://doi.org/10.5961/jhes.2019.330>
- Turan, S., Karadağ, E., Bektaş, F., & Yalçın, M. (2014). Türkiye’de eğitim yönetiminde bilgi üretimi: Kuram ve uygulamada eğitim yönetimi dergisi 2003-2013 yayınlarının incelenmesi [Knowledge production in educational administration in Turkey: An overview of researches in journal of educational administration: Theory and practice - 2003 to 2013-]. *Kuram ve Uygulamada Eğitim Yönetimi*, 20(1), 93-119. <https://doi.org/10.14527/kuey.2014.005>
- Varışlıoğlu, B., Şahin, A., & Göktaş, Y. (2013). Türkçe eğitimi araştırmalarında eğilimler [Trends in turkish education studies]. *Kuram ve Uygulamada Eğitim Bilimleri*, 13(3), 1767-1781. <https://doi.org/10.12738/estp.2013.3.1609>
- Yalçın, S. (2016). Ölçme ve değerlendirme alanındaki dergilerde yayımlanan makalelerin içerik analizi [Content analysis of research articles in measurement and evaluation journals]. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 49(1), 65-84. https://doi.org/10.1501/Egifak_0000001375
- Yalçın, S., Yavuz, H.Ç., & İlgün Dibek, M. (2016). A review of articles published in educational journals having highest impact factors: Content analysis. *Eğitim ve Bilim*, 40(182), 1-28. <https://doi.org/10.15390/EB.2015.4868>
- Yaşar, Ş., & Papatğa, E. (2015). İlkokul matematik derslerine yönelik yapılan lisansüstü tezlerin incelenmesi [The analysis of the graduate theses related to mathematics courses]. *Trakya Üniversitesi Eğitim Fakültesi Dergisi*, 5(2), 113-124.
- Yılmaz, K., & Altınkurt, Y. (2012). An examination of articles published on preschool education in Turkey. *Kuram ve Uygulamada Eğitim Bilimleri*, 12(4), 3227-3241.