# Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

## Journal of Measurement and Evaluation in Education and Psychology

*ISSN:1309-6575*

*Kış* 2019
*Winter* 2019

*Cilt: 10- Sayı: 4*
*Volume: 10- Issue: 4*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (EPOD) yılda dört kez yayınlanan hakemli ulusal bir dergidir. Yayımlanan yazıların tüm sorumluluğu ilgili yazarlara aittir.

**Iletişim**
e-posta: epod@epod-online.org
Web: http://epod-online.org

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

**Contact**
e-mail: epod@epod-online.org
Web:http://epod-online.org

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

**Hakem Kurulu / Referee Board**

**Hakem Kurulu / Referee Board**

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

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Aralık 2019, Sayı: 10-4)

Journal of Measurement and Evaluation in Education and Psychology (December 2019, Issue: 10-4)

# İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(4); 348-364

# Comparison of Person-Fit Statistics for Polytomous Items in Different Test Conditions *

Asiye ŞENGÜL AVŞAR **

**Abstract**
The validity of individual test scores is an important issue that needs to be studied in psychological and educational assessment. An important factor affecting the validity of individual test scores is aberrant item response behavior. Aberrant item scores may increase/decrease the individuals' scores and as a result individuals' ability can be estimated above/below their true ability. Person-fit statistics (PFS) are useful tools to detect aberrant behavior. There are a great number of parametric and nonparametric PFS in the literature. The general purpose of the study is to examine the effectiveness of the parametric and nonparametric PFS in data sets which consist of polytomous items. This study is fundamental research aimed at determining the effectiveness of PFS using simulated data sets. According to the results, as expected, as the Type I error rates (significance alpha level) increased, detection rates (power) increased. In general, it is seen that as the number of misfitting item score vector and number of items increased, detection rates increased. Generally, nonparametric PFS (N-PFS) (especially $G^p$) detected more aberrant individuals than parametric PFS (P-PFS) $l_z^p$. However, in some tests' conditions $l_z^p$ detected more aberrant individuals than N-PFS for longer tests. The results indicate that N-PFS outperformed P-PFS in most of the test conditions.

*Key Words:* Polytomous items, aberrant item response, person-fit statistics.

## INTRODUCTION

It is known that psychological and educational tests are important in making decisions about individuals and identifying their learning problems, developmental problems, and psychological disturbances. It is clear that test users will focus on individual scores, especially in psychological diagnoses and treatments (Emons, 2003, 2009). Therefore, the validity of individual test scores is an important issue that needs to be studied in psychological and educational assessment.

An important factor that affects the validity of individual scores is aberrant item response behavior. For example, an individual may give incorrect answers to easy items in an exam because of being anxious during a test. This situation can lead to the person's ability estimated below her/his true ability. Another example is a situation that low-skilled individuals copy correct answers from highly skilled individuals sitting around them. This situation can lead the person's ability estimated above her/his true ability. Not taking the test seriously, lacking motivation, concentration problems in cognitive tests, giving fake responses in personality tests also form the basis for aberrant item responses. Thus, the validity of individuals' ability estimates can be negatively affected (Emons, 2003, 2008; Sijtsma & Molenaar, 2002).

Aberrant item scores may increase/decrease the individuals' scores and as a result individuals' estimated ability will be above/below their true ability. According to this, the ability of cheaters and lucky guessers are estimated spuriously high, while the abilities of examinees who are confused at the beginning of test, who never reach to items towards the end, who have language deficiencies are estimated lower than their actual ability levels (Meijer, 1996). Moreover, sometimes random guessers or examinees who respond without an idea about the item content, creatives (examinees who interpret items in a creative way) and examinees (misalign their answer sheets) also have aberrant item scores

[*] This study is a part of 2219 Tubitak Project which was directed by supervisor Dr. W. H. M. Emons.

[**] Assist. Prof., Recep Tayyip Erdoğan University, Faculty of Education, Rize-Turkey, asiye.sengul@erdogan.edu.tr, ORCID ID: 0000-0001-5522-2514

To cite this article:
Şengül-Avşar, A. (2019). Comparison of person-fit statistics for polytomous items in different test conditions. *Journal of Measurement and Evaluation in Education and Psychology, 10*(4), 348-364. doi: 10.21031/epod.525647

*Received: 11.02.2019*
*Accepted: 24.08.2019*

and the abilities of the individuals may be estimated lower or higher than their real ability levels (Meijer, 1996). In all these cases, it is clear that individuals are not evaluated correctly. Therefore, in order to be able to make right decisions according to the test results, it is important to evaluate the validity of individual item-score patterns, which raise concerns about validity.

The purpose of person-fit analysis is to determine the fit of individual response patterns with the postulated model and to identify aberrant-misfitting individual item-score vectors (Meijer & Sijtsma, 2001). To accomplish this goal, person-fit statistics (PFS) are used. PFS reveal atypical test performance with the response patterns that the individuals gave to the test items (Emons, 2008; Meijer & Sijtsma, 2001). PFS play an important role in reaching more valid results since it prevents important decisions about the individual from possibly invalid test results (Emons, 2008). Also, person-fit analysis is a valuable method for validity, which is one of the important psychometric properties of measurement tools.

Many PFS have been developed in the literature. Examples of these statistics include caution indices, norm-conformity indices, and appropriateness measurement (Drasgow, Levine & McLaughlin, 1987; Embretson & Reise, 2000; Levine & Drasgow, 1983; Tatsuoka, 1984; Tatsuoka & Tatsuoka, 1982; as cited in Emons, 2003). PFS are generally divided into parametric and nonparametric statistics (Karabatsos, 2003; Mousavi, Tendeiro, & Younesi, 2016). Parametric PFS (P-PFS) are based on parametric item response theory (PIRT), while nonparametric PFS (N-PFS) are based on group statistics (i.e., item means) or nonparametric item response theory (NIRT) (Karabatsos, 2003). Table 1 shows examples of PFS according to the item type (Tendeiro, 2016).

Table 1. Parametric and Nonparametric PFS According to Item Type

| **P-PFS** | **Explanation** | **Item Type** |
|---|---|---|
| $l_z$ | The standardized log-likelihood of the response vector | Dichotomous |
| $l^*_z$ | Developed $l_z$ (to overcome $l_z$ limitation) | Dichotomous |
| $l_z^p$ | Natural extension of $l_z$ to polytomously scores | Polytomous |
| **N-PFS** | **Explanation** | **Item Type** |
| $r_{pbis}$ | Personal biserial statistic | Dichotomous |
| $C$ | The caution statistic | Dichotomous |
| $G$ | Number of Guttman errors | Dichotomous |
| $G_N$ | Normalized version of $G$ | Dichotomous |
| $A, D, E$ | Agreement, disagreement, and dependability statistics | Dichotomous |
| $U3$, $ZU3$ | van der Flier's *U3* and *ZU3* | Dichotomous |
| $C$ | Caution statistic | Dichotomous |
| $C^*$ | Modified caution statistic | Dichotomous |
| $NCI$ | $NCI = 1 - 2G_{N(normed)}$ | Dichotomous |
| $H^T$ | Sijtsma's $H^T$ person-fit statistic | Dichotomous |
| $G^p$ | Number of Guttman errors for polytomous items (*Gpoly*) | Polytomous |
| $G_N^p$ | Normalized version of *Gpoly* | Polytomous |
| $U3^p$ | Generalization of *U3* person-fit statistic for polytomous items (*U3 poly*) | Polytomous |

In the literature, log likelihood based $l_z$ statistic is the most frequently studied for binary items (Rupp, 2013). It is expressed that the most frequently used P-PFS for polytomous items is $l_z^p$; whereas popular N-PFS include $G^p$, $G_N^p$, and $U3^p$ (Emons, 2008; Rupp, 2013; Syu, 2013).

Statistic $l_z^p$ is the extended version of $l_z$ for polytomous items developed by Drasgow, Levine, and Williams (1985). Statistic $l_z^p$ is assumed to be standard normally distributed under the null model of no aberrance, where large negative values (say less than -1.645) of $l_z^p$ suggest aberrant response behavior (Meijer, 2003). One of the N-PFS is Guttman errors ($G$). Statistic $G$ is the number of item pairs for which the respondent passed/answered the difficult item but failed the easy items for dichotomous items. As for polytomous items, $G$ is also based on item pairs. In particular, a Guttman error occurs when a respondent passed difficult steps on one item and fails easy steps on another item (Meijer, 1996, 2003). Emons (2008) proposed a normed version which takes into account the maximum of the $G^p$ based on the sum score of the test. Both $G^p$'s and $G_N^p$'s minimum value is zero, which means no Guttman error, in other words, no misfit was observed. The maximum value of $G^p$

depends on the total score, while the maximum value of $G_N^p$ is one and means extreme misfit (Emons, 2008). Another N-PFS is $U3^p$ (Emons, 2008), which is the extended version of *U3*. Minimum value of $U3^p$ is zero indicating no misfit, a maximum value of $U3^p$ is one indicating extreme misfit (Emons, 2008).

N-PFS have few advantages over P-PFS. N-PFS methods only require the fit of a nonparametric model and do not require fit of more restrictive parametric models (Emons, 2003). In particular, for N-PFS it is sufficient that the data set fits the Mokken Homogeneity Model (MHM). This model assumes unidimensionality, local independence, and monotonicity (i.e., nondecreasing item characteristic curves). Therefore, these assumptions should be examined before using N-PFS (Emons, 2008).

Person-fit analysis which is emphasized as an important issue in education and psychology has been successfully applied especially in achievement tests and cognitive tests (Meijer & Sijtsma, 2001). Educational studies (examining inconsistencies in curriculum, Harnisch & Linn, 1981), cognitive psychology studies (determining of learning strategies, Tatsuoka & Tatsuoka, 1982), intercultural comparison (comparing and evaluating test scores of groups from different languages, van der Flier, 1982), personality measurement studies (identification of fake answers in the measurement tools developed for the purpose of measuring personality, Dodeen & Darabi, 2009; Ferrando, 2004, 2009, 2012; Reise & Waller, 1993; Woods, Oltmanns, & Turkheimer, 2008; Zickar & Drasgow, 1996), studies on work and organization psychology (identification of individuals with unexpected item vector score in a chosen test, Meijer, 1998), evaluating attitudes (Curtis, 2004), and research on health outputs (Custers, Hoijtink, van der Net & Hel, 2000; Tang et al., 2010) can be presented as examples (as cited in Emons, 2003; Rupp, 2013). Psychological evaluations (Conijn, Emons, De Jong & Sijtsma, 2015; Meijer, Egberink, Emons & Sijtsma, 2008) also can be presented as for PFS studies.

In addition to these studies, a literature review shows that researchers developed new PFS and tested PFS in different test conditions (Emons, 2008; Glass & Dagohoy, 2007; Karabatsos, 2003; Twiste 2011; van der Flier, 1982), determined aberrant behavior via real data test applications (Egberink, 2010; Emmen, 2011; Meijer, 2003; Spoden, 2014), tested which PFS perform best detecting aberrancy (Emons, 2008; Karabatsos, 2003; Syu, 2013; Voncken, 2014). As indicated in the literature review conducted by Rupp (2013), person-fit analyses are researched via both simulated and real data sets. However, the review also shows that the person-fit analyses are studied often for binary items, and only little for polytomous items. Hence, the literature review shows paucity in research on polytomous PFS and need for more studies on the effectiveness of polytomous PFS in various simulated test conditions, especially under small samples and skew distributions of test.

### *Purpose of the Study*

The general purpose of the study is to examine the effectiveness of parametric and nonparametric PFS in data sets which consist of polytomous items. The following questions are addressed, which are in line with the overall objective that is determined:

1. How does the proportion of detected individuals with aberrant item scores vary across test conditions such as sample size, distribution of ability, test length, and proportion of aberrancy which depends on manipulation of items and persons?
2. Which PFS performs best in different test conditions?

## METHOD

This study includes a fundamental research aimed at determining the effectiveness of PFS using simulated data sets.

### *Data Simulation*

In this study, data were simulated under Samejima’s Graded Response Model (GRM), which is a suitable model for items with ordered answer categories. This model is defined by three basic assumptions, including unidimensionality, local independence, and monotonicity between latent trait and item responses (Hambleton, van der Linden & Wells, 2011; Meijer & Tendeiro, 2018).

To formally define the model, the following notation will be used. Let *J* be the number of items indexed by *j*. Each item is assumed to have (M+1) ordered answer categories. Let $X_j$ be the random variable with realizations *xj* (0, …, M). The core of GRM is the item-step response functions (ISRF), which are defined as:

$$P_{jx_j}(\theta)=P\left(X_j \geq x_j \middle| \theta\right)=\frac{e^{\alpha_j(\theta-\delta_{jx_j})}}{1+e^{\alpha_j(\theta-\delta_{jx_j})}};\ x_j=(1,\ 2,\ \ldots,\ M) \qquad (1)$$

In equation 1, $\theta$ is person ability, $\alpha_j$ is the item-slope parameter, and $\delta_{jxj}$ (1, …, M) is the location parameter. This means that each item is modeled by one common discrimination parameter and M location parameters. The location parameters $\delta_{jxj}$ shows where on the ability scale the probability of score $x_j$ (1, …, M) or higher is equal to .50. Because item-step response functions are defined by two parameters, the model is a generalized two parametric logistic model (Embretson & Reise, 2000; Hambleton et al., 2011).

R software was employed to generate simulated data. By using the “catIRT” package (Nydick, 2015) in the R software, data sets that fit for the GRM are produced. Regardless of NIRT analysis (especially for N-PFS), the main reason data are generated based on GRM is that GRM is a special form of the MHM, and data that fit to GRM also fit to the MHM (Emons, 2008; Sijtsma, Emons, Bouwmeester, Nyklícek & Roorda, 2008). In addition, the “fungible” package (Waller & Jones, 2016) was used to generate skewed ability distributions. To compute $l_z^p$, one needs estimates of $\theta$, which can be obtained using weighted maximum likelihood estimation method (WML) (Wang, 2001; Warm, 1989). Dedicated algorithms in R programming language were used for WML estimation. Accompanying R code was obtained from Emons and are available upon request.

#### *Design factors*

In this study, simulations were done as follows:

1. Data were generated under the null model according to GRM using the test conditions envisaged.
2. According to the aim of the research, data were manipulated to mimic aberrant response behavior.
3. Extreme scores when respondents choose the same extreme response options were excluded from the analyses (e.g., strongly agree or strongly disagree) for all items. That is because Emons (2008) emphasized, extreme scores do not provide adequate information for person-fit analyses.
4. Abilities were estimated using WML estimation. While estimating the abilities, true item parameters for generating the data were used.
5. PFS were computed to detect aberrancy in different conditions with “perfit package” developed by Tendeiro (2016) in R.

Test conditions are the independent variables of the study. Test conditions included different levels of sample size (100, 250, 500, and 1,000), different shapes for the distribution of person ability (normal, positively skewed, and negatively skewed), different levels of test length (*J* = 10 and *J* = 30 items), and two levels of aberrancy (low and high). For low level of aberrancy, 20% of respondents showed aberrant response behavior on half of the items; and for high level of aberrancy, 30% of respondents showed aberrant response behavior on all items.

Table 2 shows the descriptive statistics of the simulated ability distribution. For all ability distributions, mean approximately equals zero and standard deviation equals one. Inspection of skewness coefficients shows that under the normal distribution, these coefficients were very close to zero, between of 0.54 to 0.61 for positively skewed distribution, and between of -0.58 to -0.55 for negatively skewed distribution.

Table 2. Descriptive Statistics for Ability Distributions

| | **Mean** | **Sd** | **Median** | **Mad** | **Min.** | **Max.** | **Range** | **Skewness** | **Kurtosis** | **Se** |
|---|---|---|---|---|---|---|---|---|---|---|
| Normal | | | | | | | | | | |
| 100 | -0.03 | 0.87 | -0.11 | 0.84 | -2.15 | 2.07 | 4.22 | 0.17 | -0.10 | 0.09 |
| 250 | -0.01 | 0.94 | -0.07 | 0.94 | -2.99 | 2.13 | 5.12 | 0.01 | -0.32 | 0.06 |
| 500 | -0.02 | 0.95 | -0.03 | 0.90 | -2.99 | 2.67 | 5.65 | -0.03 | 0.02 | 0.04 |
| 1,000 | -0.03 | 0.96 | -0.04 | 0.89 | -3.05 | 3.11 | 6.15 | 0.02 | 0.10 | 0.03 |
| Positively Skewed | | | | | | | | | | |
| 100 | 0.00 | 1.00 | -0.10 | 0.99 | -1.81 | 2.91 | 4.72 | 0.54 | 0.06 | 0.10 |
| 250 | 0.00 | 1.00 | -0.11 | 1.00 | -1.90 | 3.41 | 5.31 | 0.58 | 0.19 | 0.06 |
| 500 | 0.00 | 1.00 | -0.10 | 1.00 | -1.94 | 3.7 | 5.64 | 0.59 | 0.24 | 0.04 |
| 1,000 | 0.00 | 1.00 | -0.11 | 1.00 | -1.97 | 4.04 | 6.01 | 0.61 | 0.31 | 0.03 |
| Negatively Skewed | | | | | | | | | | |
| 100 | 0.00 | 1.00 | 0.10 | 0.99 | -2.89 | 1.81 | 4.70 | -0.55 | 0.01 | 0.10 |
| 250 | 0.00 | 1.00 | 0.10 | 1.00 | -3.34 | 1.91 | 5.25 | -0.55 | 0.12 | 0.06 |
| 500 | 0.00 | 1.00 | 0.11 | 1.00 | -3.64 | 1.95 | 5.59 | -0.57 | 0.18 | 0.04 |
| 1,000 | 0.00 | 1.00 | 0.11 | 1.00 | -3.96 | 1.98 | 5.94 | -0.58 | 0.24 | 0.03 |

Sd: Standard deviation, Mad: Median absolute deviation, Min: Minimum, Max: Maximum, Se: Standard error of mean

To generate item responses under the GRM, the *a* parameters were chosen between 1.50 and 2.00 and *b* parameters were, consistent with the literature, drawn from the uniform distribution in between -2.00 and 1.50 (Bahry, 2012; Cohen, Kim, & Baker, 1993; DeMars, 2002; Jiang, Wang & Weiss, 2016; Syu, 2013). Table 3 shows the item parameters for the 10 items and 30 items test.

Table 3. Item Parameters

| | **Item** | ***a*** | ***b1*** | ***b2*** | ***b3*** | ***b4*** | **Item** | ***a*** | ***b1*** | ***b2*** | ***b3*** | ***b4*** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *J*=10 | 1 | 1.96 | -1.40 | -0.79 | 0.51 | 1.51 | 6 | 1.71 | -1.01 | 0.33 | 1.49 | 2.65 |
| | 2 | 1.73 | -1.80 | -0.66 | 0.63 | 1.39 | 7 | 1.67 | -1.18 | -0.24 | 0.37 | 0.99 |
| | 3 | 1.96 | -1.03 | -0.02 | 0.83 | 1.82 | 8 | 1.88 | -1.75 | -0.28 | 0.37 | 1.38 |
| | 4 | 1.63 | -1.35 | -0.14 | 0.42 | 1.03 | 9 | 1.92 | -1.31 | -0.67 | 0.76 | 1.56 |
| | 5 | 1.67 | -1.63 | -0.27 | 0.80 | 1.81 | 10 | 1.51 | -1.17 | 0.11 | 1.08 | 2.34 |
| | **Item** | ***a*** | ***b1*** | ***b2*** | ***b3*** | ***b4*** | **Item** | ***a*** | ***b1*** | ***b2*** | ***b3*** | ***b4*** |
| *J*=30 | 1 | 1.81 | -1.40 | -0.40 | 0.42 | 1.82 | 16 | 1.53 | -1.16 | -0.23 | 0.93 | 1.95 |
| | 2 | 1.65 | -1.80 | -1.05 | 0.45 | 0.96 | 17 | 1.61 | -1.55 | -0.72 | 0.04 | 1.49 |
| | 3 | 1.67 | -1.03 | -0.04 | 0.96 | 1.59 | 18 | 1.78 | -1.04 | 0.22 | 0.95 | 2.36 |
| | 4 | 1.56 | -1.35 | -0.73 | 0.49 | 1.08 | 19 | 1.95 | -1.86 | -0.51 | 0.08 | 1.24 |
| | 5 | 1.64 | -1.63 | -0.62 | 0.81 | 2.25 | 20 | 1.82 | -1.22 | -0.71 | 0.53 | 1.35 |
| | 6 | 1.55 | -1.01 | 0.15 | 1.59 | 2.23 | 21 | 1.53 | -1.20 | -0.03 | 1.11 | 1.80 |
| | 7 | 1.55 | -1.18 | -0.56 | 0.71 | 1.97 | 22 | 1.67 | -1.21 | 0.01 | 1.40 | 2.78 |
| | 8 | 1.63 | -1.75 | -0.73 | 0.10 | 0.88 | 23 | 1.52 | -1.64 | -0.37 | 0.89 | 1.63 |
| | 9 | 1.53 | -1.31 | -0.51 | 0.82 | 2.15 | 24 | 1.75 | -1.94 | -0.50 | 0.83 | 1.47 |
| | 10 | 1.80 | -1.17 | 0.09 | 1.50 | 2.16 | 25 | 1.55 | -1.43 | -0.69 | 0.81 | 2.01 |
| | 11 | 1.56 | -1.90 | -0.48 | 0.70 | 1.95 | 26 | 1.71 | -1.34 | 0.07 | 1.48 | 2.68 |
| | 12 | 1.75 | -1.35 | -0.40 | 0.78 | 2.14 | 27 | 1.65 | -1.89 | -0.77 | -0.10 | 1.27 |
| | 13 | 1.68 | -1.49 | -0.07 | 0.83 | 2.18 | 28 | 1.93 | -1.85 | -0.58 | 0.78 | 1.84 |
| | 14 | 1.89 | -1.29 | -0.53 | 0.65 | 1.25 | 29 | 1.76 | -1.07 | 0.25 | 1.11 | 2.07 |
| | 15 | 1.85 | -1.14 | -0.29 | 1.06 | 1.96 | 30 | 1.83 | -1.52 | -0.75 | 0.55 | 1.57 |

Baker (2001) suggested the following guidelines for interpreting *a* coefficients: 0 *none*, 0.01-0.34 *very low*, 0.35-0.64 *low*, 0.65-1.34 *moderate*, 1.35-1.69 *high*, > 1.70 *very high*, and ∞ (+ infinity) *perfect*. Hence, the tests in this study consisted of relatively high discriminating items, but these values are unrealistic in practice. Previous studies convincingly showed that the power of PFS relates to the items' discrimination power (Emons, 2008; Meijer, Molenaar, & Sijtsma, 1994; Meijer & Sijtsma, 2001). Higher discrimination power may produce a higher detection rate (Emons, 2008).

There are many kinds of aberrant behavior that may affect test results. One of them is *careless and inattention*. In some test applications, individuals answer items randomly because they are careless, or a random pattern emerges due to misreading or not reading the questions, or due to alignments errors (Emons, 2008). Randomness-like response behaviors from important types of aberrant behavior (Conijn et al. 2015) and will be the subject of this study. To accomplish this goal, aberrant item response vectors were created by simulating random scores from the uniform distribution similar to Emons's (2008) study.

The selected test conditions are based on the literature (Lee, 2007; Lee, Wollack & Douglas, 2009; Liang, Wells & Hambleton, 2014; Ramsay, 1991; Syu, 2013). In particular, variation in the shape of ability distribution, small sample sizes and short tests are often seen in classroom measurement applications. One condition nevertheless consisted of a large sample size (1,000). This condition was chosen to see how PFS function in large samples and can be seen as a benchmark for the other results.

Data were generated using a fully factorial design including 4 (sample size) × 3 (ability distribution) × 2 (test length) × 2 (aberrancy levels) = 48 conditions. In total 100 replications were obtained for each test condition, thus in total 4800 data sets were simulated.

### *Data Analysis*

Empirical Type I error rates and detection rates (power) are the dependent variables of the study. For each PFS ($l_z^p$, $U3^p$, $G_N^p$ and $G^p$), the empirical Type I error rates and detection rates were evaluated at four the theoretical Type I error rates (nominal significance levels) ($\alpha = .01$, $\alpha = .05$, $\alpha = .10$ and $\alpha = .20$). Empirical Type I error rate is the observed proportion of non-aberrant persons identified as aberrant. Also, the detection rate is the proportion of aberrant persons correctly identified as aberrant (Voncken, 2014).

The theoretical Type I error rates which were chose in the study determined from the literature view results. It is stated in the literature that large alpha levels (e.g., .05, .10 and .20) are preferable because PFS have relatively low power detect aberrancy for small test lengths and low alpha levels (Emons, 2008; Emons, Glas, Meijer & Sijtsma, 2003; Meijer, 2003; Spoden, 2014; Voncken, 2014).

To decide whether a pattern shows significant misfit, one needs to have critical values. Certain rules are followed in the calculation of critical values for the PFS. In particular, the critical values for parametric $l_z^p$ is determined, as in Voncken's (2014) study, to be -2.32, -1.645, -1.28, and -0.84. These are critical values from the standard normal distribution for alphas of .01, .05, .10 and .20 (one-tailed tests). Because N-PFS lack theoretical distributions, the critical values have to be determined differently. This study uses critical values of N-PFS that were determined automatically by perfit package in a pilot study. These cut-off values were fixed for every simulation and replication. Researchers are strongly recommended to fix the cut-off score with the command *set.seed ()* before identifying individuals with aberrant item patterns according to the cut-off score in the relevant package (Meijer, Niessen & Tendeiro, 2016; Tendeiro, 2016). Otherwise, different critical values with small differences are reached in each calculation.

## RESULTS

There are two levels of aberrancy in this study. PFS analysis results are given in Table 4 to Table 9. Table 4 gives the findings for normally distributed ability for 10 items.

Table 4. Detection Rates for Normal Distributed Sample for 10 Items with Low and High Aberrancy Level

| **PFS** | **Low Aberrancy** | | | | | | | | **High Aberrancy** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Nominal Significance Levels and Detection Rates** | | | | | | | | **Nominal Significance Levels and Detection Rates** | | | | | | | |
| | **.01** | **D.R.** | **.05** | **D.R.** | **.10** | **D.R.** | **.20** | **D.R.** | **.01** | **D.R.** | **.05** | **D.R.** | **.10** | **D.R.** | **.20** | **D.R.** |
| | N = 100 | | | | | | | | | | | | | | | |
| $l_z^p$ | .03 | **.05** | .03 | .10 | .04 | .10 | .08 | .35 | .00 | .10 | .00 | .30 | .00 | .43 | .03 | .60 |
| $U3^p$ | .01 | **.05** | .04 | .10 | .04 | .30 | .21 | .70 | .00 | .10 | .01 | **.40** | .01 | **.57** | .07 | .67 |
| $G_N^p$ | .01 | **.05** | .03 | .10 | .05 | .30 | .18 | .65 | .00 | .13 | .00 | **.40** | .01 | .53 | .07 | .67 |
| $G^p$ | .01 | **.05** | .03 | **.15** | .08 | **.35** | .16 | **.75** | .00 | **.17** | .00 | .37 | .01 | .50 | .07 | **.77** |
| | N = 250 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | **.18** | .02 | .32 | .02 | .40 | .07 | .48 | .00 | **.17** | .01 | .33 | .01 | .44 | .01 | .67 |
| $U3^p$ | .01 | .04 | .03 | .42 | .06 | .52 | .16 | .64 | .01 | .11 | .01 | .33 | .03 | .49 | .05 | .71 |
| $G_N^p$ | .01 | .08 | .03 | .42 | .08 | **.56** | .16 | .66 | .01 | .13 | .01 | .35 | .02 | .52 | .05 | .72 |
| $G^p$ | .00 | **.18** | .03 | **.48** | .05 | .52 | .12 | **.70** | .00 | .13 | .00 | **.37** | .02 | **.55** | .04 | **.77** |
| | N =5 00 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .11 | .03 | .20 | .04 | .30 | .11 | .42 | .00 | .15 | .00 | .34 | .01 | .47 | .02 | .63 |
| $U3^p$ | .02 | .04 | .06 | .27 | .08 | .40 | .17 | .60 | .01 | .12 | .03 | .38 | .04 | .54 | .09 | **.75** |
| $G_N^p$ | .02 | .11 | .06 | .28 | .08 | .43 | .14 | .58 | .01 | .12 | .03 | .35 | .03 | .52 | .07 | .72 |
| $G^p$ | .01 | **.14** | .04 | **.34** | .06 | **.49** | .14 | **.69** | .00 | **.17** | .01 | **.41** | .02 | **.59** | .07 | **.75** |
| | N = 1 000 | | | | | | | | | | | | | | | |
| $l_z^p$ | .01 | .09 | .02 | .18 | .04 | .30 | .09 | .40 | .00 | .12 | .00 | .33 | .01 | .44 | .02 | .62 |
| $U3^p$ | .01 | .08 | .05 | .23 | .09 | .34 | .14 | .52 | .01 | .12 | .02 | .35 | .04 | .49 | .08 | .65 |
| $G_N^p$ | .02 | .11 | .05 | .25 | .09 | .35 | .15 | .56 | .01 | .11 | .03 | .35 | .04 | .49 | .07 | .63 |
| $G^p$ | .01 | **.15** | .03 | **.28** | .07 | **.45** | .13 | **.61** | .00 | **.14** | .00 | **.37** | .02 | **.52** | .06 | **.71** |

Note. The bolded detection rates denote the conditions in which PFS perform best. D.R.: Detection rates. N: Sample size

Inspection of Table 4 shows that as sample size increased, the detection rate increased in many test conditions. Almost all conditions, detection rates increased with increasing aberrancy levels. In general, $G^p$ showed best performance to detect aberrancy. In addition to these findings, it is found that nonparametric $U3^p$ and $G_N^p$ statistics are very close to each other. When empirical Type I error rates are examined, it is seen that these values exceed their nominal levels especially for low aberrancy level at α = .01 and α = .05. Also, empirical Type I error rates are smaller than their nominal levels in all conditions for high aberrancy level except for α = .01. It can be seen that as increased of aberrancy, empirical Type I error rates decreased.

Table 5 gives the findings for positively skewed ability distribution for 10 items. Table 5 shows empirical Type I error rates and detection rates for PFS for positive distributed ability, for different sample sizes and low and high aberrancy levels. As expected, it is seen that as the Type I error rates increased, the detection rate increased. It is seen that as sample size increased, the detection rate increased in many test conditions for high aberrancy level. Almost all conditions detection rates increased according to the aberrancy level. In general, $G^p$ showed best performance to detect aberrancy. In addition to these findings, it is found that nonparametric $U3^p$ and $G_N^p$ statistics are very close to each other. When empirical Type I error rates are examined, it is seen that these values are smaller than their nominal levels both low and high aberrancy except for α = .01. Empirical Type I error rates are equal to or smaller than their nominal level for α = .01. It can be seen that as increased of aberrancy, empirical Type I error rates decreased.

Table 6 gives the findings for negatively skewed distribution for 10 items. Table 6 shows the detection rates for negatively distributed ability, for different sample sizes and low and high aberrancy. It is seen that as the nominal significance level increased, the detection rates increased almost all test conditions. In general, as sample size increased, the detection rates increased. However, detection rates of $l_z^p$ decreased dramatically for large sample in low aberrancy level when α = .05. Detection rates increased according to the aberrancy level in all test conditions. In general, $G^p$ showed best performance to detect aberrancy. In addition to these findings, it is found that nonparametric $U3^p$ and $G_N^p$ statistics are very close to each other. When empirical Type I error rates are examined, in general, these values are smaller than their nominal levels both low and high aberrancy except for α = .01. Also, empirical Type

I error rates are equal to or smaller than their nominal α = .01. It can be seen that as increased of aberrancy, empirical Type I error rates decreased.

Table 5. Detection Rates for Positively Skewed Distributed Sample for 10 Items with Low and High Aberrancy Level

| PFS | Low Aberrancy | | | | | | | | High Aberrancy | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Nominal Significance Levels and Detection Rates | | | | | | | | Nominal Significance Levels and Detection Rates | | | | | | | |
| | **.01** | **D.R.** | **.05** | **D.R.** | **.10** | **D.R.** | **.20** | **D.R.** | **.01** | **D.R.** | **.05** | **D.R.** | **.10** | **D.R.** | **.20** | **D.R.** |
| | N = 100 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .07 | .01 | .19 | .03 | .29 | .07 | .42 | .00 | .11 | .00 | .28 | .01 | .41 | .03 | .57 |
| $U3^p$ | .01 | .07 | .04 | .24 | .08 | .38 | .16 | .59 | .00 | .09 | .02 | .30 | .04 | .46 | .09 | .66 |
| $G_N^p$ | .01 | .08 | .03 | .26 | .07 | .41 | .15 | .60 | .00 | .10 | .02 | .30 | .03 | .47 | .08 | .67 |
| $G^p$ | .00 | **.12** | .02 | **.31** | .06 | **.46** | .14 | **.64** | .00 | **.12** | .01 | **.34** | .02 | **.53** | .06 | **.71** |
| | N = 250 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .07 | .01 | .20 | .03 | .30 | .07 | .45 | .00 | **.14** | .00 | .31 | .01 | .43 | .02 | .60 |
| $U3^p$ | .01 | .07 | .04 | .28 | .08 | .43 | .16 | .61 | .00 | .11 | .02 | .33 | .04 | .50 | .08 | .69 |
| $G_N^p$ | .01 | .09 | .04 | .30 | .07 | .45 | .16 | .62 | .00 | .11 | .02 | .33 | .03 | .50 | .08 | .70 |
| $G^p$ | .00 | **.14** | .02 | **.35** | .06 | **.49** | .14 | **.66** | .00 | **.14** | .00 | **.39** | .01 | **.54** | .05 | **.73** |
| | N = 500 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .07 | .01 | .20 | .03 | .30 | .07 | .44 | .00 | .14 | .00 | .32 | .01 | .45 | .02 | .61 |
| $U3^p$ | .01 | .08 | .04 | .28 | .08 | .42 | .16 | .61 | .01 | .12 | .02 | .35 | .03 | .51 | .08 | .70 |
| $G_N^p$ | .01 | .10 | .04 | .30 | .08 | .45 | .16 | .62 | .00 | .12 | .02 | .35 | .03 | .51 | .08 | .69 |
| $G^p$ | .00 | **.14** | .03 | **.34** | .06 | **.49** | .14 | **.66** | .00 | **.15** | .00 | **.39** | .01 | **.54** | .05 | **.73** |
| | N = 1 000 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .08 | .01 | .20 | .03 | .30 | .07 | .45 | .00 | .14 | .00 | .33 | .01 | .45 | .02 | .61 |
| $U3^p$ | .01 | .08 | .04 | .29 | .08 | .44 | .17 | .61 | .01 | .13 | .02 | .36 | .04 | .52 | .09 | .71 |
| $G_N^p$ | .01 | .11 | .04 | .31 | .08 | .46 | .16 | .63 | .01 | .13 | .02 | .36 | .03 | .52 | .08 | .71 |
| $G^p$ | .00 | **.15** | .03 | **.36** | .06 | **.49** | .14 | **.66** | .00 | **.17** | .01 | **.40** | .02 | **.56** | .05 | **.74** |

Note. The bolded detection rates denote the conditions in which PFS perform best. D.R.: Detection rates. N: Sample size

Table 6. Detection Rates for Negatively Skewed Distributed Sample for 10 Items with Low and High Aberrancy Level

| PFS | Low Aberrancy | | | | | | | | High Aberrancy | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Nominal Significance Levels and Detection Rates | | | | | | | | Nominal Significance Levels and Detection Rates | | | | | | | |
| | **.01** | **D.R.** | **.05** | **D.R.** | **.10** | **D.R.** | **.20** | **D.R.** | **.01** | **D.R.** | **.05** | **D.R.** | **.10** | **D.R.** | **.20** | **D.R.** |
| | N = 100 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .07 | .01 | .20 | .03 | .29 | .07 | .45 | .00 | .12 | .00 | .28 | .01 | .41 | .02 | .58 |
| $U3^p$ | .01 | .07 | .04 | .24 | .08 | .40 | .16 | .56 | .01 | .09 | .02 | .30 | .04 | .48 | .09 | .67 |
| $G_N^p$ | .01 | .08 | .04 | .26 | .07 | .42 | .15 | .58 | .00 | .09 | .02 | .31 | .04 | .47 | .08 | .67 |
| $G^p$ | .00 | **.13** | .02 | **.33** | .05 | **.46** | .13 | **.64** | .00 | **.13** | .01 | **.36** | .02 | **.52** | .06 | **.72** |
| | N = 250 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .07 | .01 | .20 | .03 | .30 | .07 | .45 | .00 | .14 | .00 | .31 | .01 | .44 | .02 | .60 |
| $U3^p$ | .01 | .07 | .04 | .28 | .08 | .43 | .16 | .61 | .01 | .10 | .02 | .33 | .04 | .50 | .08 | .70 |
| $G_N^p$ | .01 | .10 | .04 | .30 | .07 | .44 | .16 | .62 | .01 | .11 | .02 | .33 | .03 | .50 | .08 | .70 |
| $G^p$ | .00 | **.15** | .03 | **.34** | .06 | **.48** | .14 | **.66** | .00 | **.15** | .01 | **.38** | .02 | **.55** | .05 | **.73** |
| | N = 500 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .08 | .01 | .20 | .03 | .30 | .07 | .44 | .00 | .14 | .00 | .32 | .01 | .45 | .02 | .61 |
| $U3^p$ | .01 | .08 | .05 | .27 | .08 | .42 | .17 | .60 | .01 | .12 | .02 | .36 | .04 | .52 | .08 | .70 |
| $G_N^p$ | .01 | .10 | .04 | .30 | .08 | .44 | .17 | .62 | .01 | .12 | .02 | .36 | .04 | .52 | .08 | .70 |
| $G^p$ | .01 | **.14** | .03 | **.34** | .06 | **.48** | .14 | **.65** | .00 | **.16** | .01 | **.40** | .02 | **.55** | .06 | **.73** |
| | N = 1 000 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .08 | .00 | .08 | .03 | .30 | .07 | .44 | .00 | .14 | .00 | .33 | .01 | .45 | .02 | .61 |
| $U3^p$ | .01 | .07 | .05 | .29 | .09 | .43 | .17 | .61 | .01 | .12 | .02 | .37 | .04 | .53 | .09 | .71 |
| $G_N^p$ | .01 | .10 | .04 | .31 | .08 | .45 | .17 | .62 | .01 | .13 | .02 | .36 | .04 | .52 | .08 | .71 |
| $G^p$ | .00 | **.15** | .03 | **.35** | .06 | **.49** | .14 | **.65** | .00 | **.17** | .01 | **.40** | .02 | **.56** | .06 | **.74** |

Note. The bolded detection rates denote the conditions in which PFS perform best. D.R.: Detection rates. N: Sample size

Table 7 gives the findings for normally distributed ability for 30 items. Table 7 shows the detection rates for normally distributed ability, for different sample sizes and aberrancy levels. As expected, it is seen that as the nominal significance levels increased, the detection rates increased as well. There is no specific trend regarding the effect of sample size on the detection rates. However, when all test conditions are examined, the highest detection rates were observed in the largest sample. For $l_z^p$, detection rates increased with increasing aberrancy levels at all nominal significance levels. In general, $G^p$ showed best performance to detect aberrancy in low aberrancy level, while $l_z^p$ showed best performance to detect aberrancy in high aberrancy level. In addition to these findings, it is found that nonparametric $U3^p$ and $G_N^p$ statistics were very close to each other. When empirical Type I error rates are examined, it is seen that these values never exceed their nominal levels in all test conditions. Empirical Type I error rates are smaller than or equal to their nominal α = .01 for low aberrancy. Also, all empirical Type I error rates are smaller than their nominal levels for high aberrancy. It can be seen that as increased of aberrancy, empirical Type I error rates decreased.

Table 7. Detection Rates for Normal Distributed Sample for 30 Items with Low and High Aberrancy Level

| PFS | Low Aberrancy | | | | | | | | High Aberrancy | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Nominal Significance Levels and Detection Rates | | | | | | | | Nominal Significance Levels and Detection Rates | | | | | | | |
| | **.01** | **D.R.** | **.05** | **D.R.** | **.10** | **D.R.** | **.20** | **D.R.** | **.01** | **D.R.** | **.05** | **D.R.** | **.10** | **D.R.** | **.20** | **D.R.** |
| | N = 100 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | **.25** | .03 | **.45** | .05 | .55 | .11 | .75 | .00 | **.53** | .00 | **.77** | .03 | **.83** | .04 | **.93** |
| $U3^p$ | .00 | .15 | .04 | .40 | .05 | **.70** | .10 | **.80** | .00 | .07 | .00 | .40 | .00 | .70 | .04 | .87 |
| $G_N^p$ | .00 | .15 | .04 | .35 | .05 | **.70** | .11 | .75 | .00 | .07 | .00 | .33 | .00 | .70 | .04 | .87 |
| $G^p$ | .00 | **.25** | .00 | .40 | .05 | .65 | .06 | **.80** | .00 | .07 | .00 | .27 | .00 | .67 | .00 | .90 |
| | N = 250 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | **.26** | .02 | **.46** | .05 | .58 | .08 | .68 | .00 | **.56** | .00 | **.75** | .00 | **.85** | .00 | .92 |
| $U3^p$ | .00 | .18 | .02 | .36 | .05 | .48 | .10 | .76 | .00 | .16 | .00 | .56 | .00 | .76 | .03 | **.95** |
| $G_N^p$ | .00 | .18 | .01 | .36 | .04 | .48 | .11 | .74 | .00 | .12 | .00 | .51 | .00 | .77 | .03 | .92 |
| $G^p$ | .00 | .20 | .01 | .44 | .01 | **.62** | .07 | **.84** | .00 | .15 | .00 | .52 | .00 | .75 | .01 | .93 |
| | N = 500 | | | | | | | | | | | | | | | |
| $l_z^p$ | .01 | .19 | .02 | .44 | .03 | .55 | .07 | .70 | .00 | **.55** | .00 | **.77** | .00 | **.85** | .01 | **.94** |
| $U3^p$ | .01 | .16 | .02 | .47 | .06 | .57 | .10 | .77 | .00 | .07 | .00 | .50 | .01 | .69 | .02 | .90 |
| $G_N^p$ | .01 | .16 | .02 | .48 | .06 | .60 | .12 | .75 | .00 | .07 | .01 | .46 | .01 | .69 | .02 | .87 |
| $G^p$ | .00 | **.26** | .01 | **.49** | .03 | **.65** | .09 | **.85** | .00 | .13 | .00 | .51 | .00 | .76 | .01 | .91 |
| | N = 1 000 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .28 | .01 | .50 | .02 | .64 | .05 | .76 | .00 | **.61** | .00 | **.78** | .00 | **.87** | .00 | **.95** |
| $U3^p$ | .01 | .23 | .02 | .49 | .04 | .64 | .09 | .82 | .00 | .42 | .00 | .63 | .01 | .75 | .01 | .91 |
| $G_N^p$ | .01 | .30 | .02 | .50 | .04 | .65 | .10 | .83 | .00 | .42 | .00 | .62 | .01 | .75 | .01 | .92 |
| $G^p$ | .00 | **.31** | .01 | **.59** | .02 | **.74** | .06 | **.88** | .00 | .41 | .00 | .63 | .00 | .77 | .00 | .92 |

Note. The bolded detection rates denote the conditions in which PFS perform best. D.R.: Detection rates. N: Sample size

Table 8 gives the findings for positively skewed ability distribution for 30 items. Table 8 shows the detection rates for PFS for positively skewed distributed ability for different sample sizes, low and high aberrancy. In general, detection rates increased with increasing aberrancy levels. However, for N-PFS results show higher detection rates for low aberrancy level than for high aberrancy level. This result is seen in test conditions which are consist for sample size 100 and at α = .01 and α = .05 nominal levels, for sample size 250 at α = .01 nominal level. Statistic $G^p$ showed best performance to detect aberrancy at low aberrancy levels except for sample size 100 at α = .01 and α = .05 nominal levels, and for sample size 250 at α = .01 nominal level. It is seen that $l_z^p$ showed best performance to detect aberrancy for all sample sizes and all Type I error rates in high aberrancy level. In addition to these findings, it is found that detection rates for nonparametric $U3^p$ and $G_N^p$ statistics were very close to each other. When empirical Type I error rates are examined, it is seen that these values were not exceed their nominal levels in most of test conditions. Only for $U3^p$, empirical Type I error rate was equal to its α = .01 nominal level for large sample and low aberrancy. Also, it is found that all empirical Type I error rates are smaller than their nominal levels for high aberrancy.

Table 8. Detection Rates for Positively Skewed Distributed Data for 30 Items with Low and High Aberrancy Level

| PFS | Low Aberrancy | | | | | | | | High Aberrancy | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Nominal Significance Levels and Detection Rates | | | | | | | | Nominal Significance Levels and Detection Rates | | | | | | | |
| | **.01** | **D.R.** | **.05** | **D.R.** | **.10** | **D.R.** | **.20** | **D.R.** | **.01** | **D.R.** | **.05** | **D.R.** | **.10** | **D.R.** | **.20** | **D.R.** |
| | N = 100 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | **.27** | .01 | **.49** | .02 | .62 | .06 | .74 | .00 | **.51** | .00 | **.74** | .00 | **.84** | .01 | **.91** |
| $U3^p$ | .00 | .12 | .01 | .38 | .03 | .59 | .08 | .78 | .00 | .11 | .00 | .38 | .00 | .60 | .01 | .86 |
| $G_N^p$ | .00 | .12 | .01 | .39 | .03 | .58 | .08 | .78 | .00 | .10 | .00 | .36 | .00 | .60 | .01 | .86 |
| $G^p$ | .00 | .15 | .00 | .44 | .01 | **.64** | .06 | **.84** | .00 | .11 | .00 | .37 | .00 | .61 | .00 | .87 |
| | N = 250 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | **.29** | .01 | .49 | .02 | .62 | .05 | .76 | .00 | **.57** | .00 | **.79** | .00 | **.87** | .00 | **.94** |
| $U3^p$ | .00 | .19 | .02 | .47 | .04 | .65 | .09 | .82 | .00 | .19 | .00 | .51 | .00 | .72 | .01 | .89 |
| $G_N^p$ | .00 | .20 | .01 | .47 | .03 | .64 | .09 | .82 | .00 | .18 | .00 | .50 | .00 | .71 | .01 | .89 |
| $G^p$ | .00 | .23 | .00 | **.53** | .02 | **.70** | .06 | **.87** | .00 | .20 | .00 | .52 | .00 | .72 | .00 | .91 |
| | N = 500 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .28 | .01 | .50 | .02 | .62 | .06 | .75 | .00 | **.59** | .00 | **.80** | .00 | **.88** | .00 | **.94** |
| $U3^p$ | .00 | .23 | .02 | .52 | .04 | .67 | .10 | .82 | .00 | .28 | .00 | .60 | .00 | .78 | .02 | .91 |
| $G_N^p$ | .00 | .25 | .02 | .52 | .04 | .66 | .09 | .81 | .00 | .27 | .00 | .59 | .00 | .77 | .02 | .91 |
| $G^p$ | .00 | **.30** | .01 | **.58** | .02 | **.73** | .07 | **.87** | .00 | .28 | .00 | .60 | .00 | .78 | .00 | .92 |
| | N = 1,000 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .29 | .01 | .50 | .02 | .61 | .05 | .76 | .00 | **.60** | .00 | **.81** | .00 | **.89** | .00 | **.95** |
| $U3^p$ | .01 | .27 | .02 | .55 | .04 | .68 | .10 | .82 | .00 | .31 | .00 | .64 | .01 | .80 | .02 | .92 |
| $G_N^p$ | .00 | .29 | .02 | .55 | .04 | .68 | .10 | .82 | .00 | .30 | .00 | .62 | .01 | .78 | .02 | .92 |
| $G^p$ | .00 | **.34** | .01 | **.60** | .02 | **.74** | .07 | **.87** | .00 | .32 | .00 | .63 | .00 | .80 | .00 | .93 |

Note. The bolded detection rates denote the conditions in which PFS perform best. D.R.: Detection rates. N: Sample size

Table 9 gives the findings for negatively skewed distribution for 30 items. Table 9 shows the detection rates for PFS for negatively skewed distributed ability, for different sample sizes and for low and high aberrancy levels.

Table 9. Detection Rates for Negatively Skewed Distributed Data for 30 Items with Low and High Aberrancy Level

| PFS | Low Aberrancy | | | | | | | | High Aberrancy | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Nominal Significance Levels and Detection Rates | | | | | | | | Nominal Significance Levels and Detection Rates | | | | | | | |
| | **.01** | **D.R.** | **.05** | **D.R.** | **.10** | **D.R.** | **.20** | **D.R.** | **.01** | **D.R.** | **.05** | **D.R.** | **.10** | **D.R.** | **.20** | **D.R.** |
| | N = 100 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | **.27** | .01 | **.48** | .02 | .60 | .06 | .72 | .00 | **.54** | .00 | **.77** | .00 | **.85** | .01 | **.93** |
| $U3^p$ | .00 | .12 | .01 | .38 | .03 | .58 | .09 | .77 | .00 | .11 | .00 | .38 | .01 | .62 | .01 | .87 |
| $G_N^p$ | .00 | .12 | .01 | .38 | .03 | .58 | .08 | .78 | .00 | .11 | .00 | .38 | .00 | .62 | .01 | .87 |
| $G^p$ | .00 | .13 | .00 | .43 | .01 | **.64** | .06 | **.83** | .00 | .12 | .00 | .40 | .00 | .64 | .00 | .88 |
| | N = 250 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | **.29** | .01 | .51 | .02 | .63 | .06 | .76 | .00 | **.58** | .00 | **.80** | .00 | **.88** | .00 | **.94** |
| $U3^p$ | .01 | .16 | .02 | .46 | .04 | .64 | .09 | .81 | .00 | .20 | .00 | .54 | .01 | .73 | .02 | .90 |
| $G_N^p$ | .00 | .17 | .02 | .46 | .04 | .63 | .09 | .80 | .00 | .19 | .00 | .52 | .01 | .72 | .02 | .90 |
| $G^p$ | .00 | .25 | .01 | **.54** | .02 | **.70** | .06 | **.86** | .00 | .22 | .00 | .55 | .00 | .75 | .00 | .91 |
| | N = 500 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .29 | .01 | .50 | .02 | .62 | .06 | .75 | .00 | **.60** | .00 | **.81** | .00 | **.89** | .00 | **.95** |
| $U3^p$ | .01 | .23 | .02 | .51 | .04 | .66 | .09 | .82 | .00 | .27 | .00 | .61 | .01 | .79 | .02 | .92 |
| $G_N^p$ | .01 | .23 | .02 | .50 | .04 | .65 | .10 | .81 | .00 | .26 | .01 | .60 | .01 | .78 | .02 | .91 |
| $G^p$ | .00 | **.30** | .01 | **.58** | .02 | **.73** | .07 | **.86** | .00 | .30 | .00 | .62 | .00 | .79 | .00 | .92 |
| | N = 1 000 | | | | | | | | | | | | | | | |
| $l_z^p$ | .00 | .29 | .01 | .50 | .02 | .62 | .06 | .76 | .00 | **.61** | .00 | **.82** | .00 | **.90** | .00 | **.95** |
| $U3^p$ | .01 | .25 | .02 | .54 | .05 | .68 | .10 | .82 | .00 | .32 | .00 | .65 | .01 | .81 | .02 | .93 |
| $G_N^p$ | .01 | .26 | .02 | .53 | .05 | .67 | .10 | .81 | .00 | .30 | .01 | .64 | .01 | .80 | .02 | .92 |
| $G^p$ | .00 | **.34** | .01 | **.61** | .02 | **.74** | .07 | **.87** | .00 | .34 | .00 | .66 | .00 | .81 | .00 | .93 |

Note. The bolded detection rates denote the conditions in which PFS perform best. D.R.: Detection rates. N: Sample size

Inspection of Table 9 shows that as expected, as the nominal significance levels increased, the detection rates increased as well. It is also seen in almost all conditions of low aberrancy that as sample size increased, the detection rate increased. Although, it is seen that as sample size increased, the detection rate increased in high aberrancy level for all samples. In general, detection rates increased according to the aberrancy level except for α = .01 and α = .05 for N-PFS. Broadly speaking, across all conditions, $G^p$ showed best performance to detect aberrancy at low aberrancy level while $l_z^p$ showed best performance to detect aberrancy at high aberrancy level. In addition to these findings, it is found that the detection rates of nonparametric $U3^p$ and $G_N^p$ statistics were very close to each other. When empirical Type I error rates are examined, it is seen that these values did not exceed their nominal levels in high aberrancy. However, empirical Type I error rates are smaller than or equal to their nominal α = .01 for low aberrancy. It can be seen that as increased of aberrancy, empirical Type I error rates decreased.

## DISCUSSION and CONCLUSION

The general purpose of the study is to examine the effectiveness of parametric and nonparametric PFS in data sets which consist of polytomous items. According to this aim, data simulated in different test conditions and these data sets were analyzed.

The results confirmed several important effects of significance level, sample size, ability distribution, and aberrance level. As expected, the detection rates increased with increasing nominal significance levels (the theoretical Type I error rates) in all test conditions. Moreover, it is seen that detection rates increased as the number of misfitting item score vector and number of misfitting items increased. Simulation results suggest that the shape of sample distributions has little effect on the detection of aberrancy. So, it can be said that shape of ability distribution (determined in this study's test conditions) is an unimportant factor for the effectiveness of PFS.

In general, sample size affected detection rates. In most of test conditions, it is seen that as sample size increased, detection rates increased. However, this result conflicts with Syu (2013), who studied with parametric $l_z^p$ and nonparametric $G^p$ and $U3^p$. Syu (2013) only found small differences in the detection rates across sample sizes for specific PFS. In addition to this finding, Syu (2013) stated that findings are tentative because sample size is too small for providing sufficient calculations for PFS.

It is seen that in general, empirical Type I error rates smaller than their nominal levels (the theoretical Type I error rates). However, in all shapes of ability distributions for 10 and 30 items, empirical Type I error rates are equal to or smaller than their nominal level at α = .01. Except of this conclusion, it is seen that for normally distributed sample for 10 items, empirical Type I error rates exceed its nominal level at α = .01. In Voncken's (2014) study, detection rates were determined for binary items. In that study it is found that $l_z$*'s empirical Type I rate exceeds its nominal level at α = .01. Also, it is seen that as increased of aberrancy, empirical Type I error rates decreased. These findings are consistent with Voncken (2014).

To summarize, as expected, as the nominal significance level was set higher, tests were longer, and amount of the aberrant proportions increased, the detection rates increased as well. These findings are consistent with other person-fit studies (Emons, 2008; Karabatsos, 2003; Meijer & Sijtsma, 2001; Voncken, 2014).

A comparison of the effectiveness of the different PFS showed the following important trends. It is seen that detection rates were very close to each other for P-PFS and N-PFS (especially $U3^p$ and $G_N^p$). However, in general, $G^p$ was the most effective in detecting aberrant individuals and even performed better than $l_z^p$. These results are consistent with Emons (2008) and Syu (2013). They compared same PFS as used in this study in different test conditions. Like in this study, in their studies $G^p$ showed best performance to detect aberrancy. In Syu's (2013) study it's also stated that for small sample sizes N-PFS perform better than P-PFS.

It is found that for all test conditions detection rates were sufficiently high except at α = .01. Detection rates got their maximum value at α = .20. PFS may have very low detection rates at small significance levels of α = .01, which questions their effectiveness at these significance levels. These findings are consistent with literature. Therefore, it is suggested that researchers should choose liberal significance levels (i.e., α = .20) to reach some power in detecting aberrancy (Emons, 2008; Meijer, 2003; Voncken, 2014).

Based on the result, the following general conclusions about the suitability of different statistics can be drawn. Results also showed that for detecting careless and inattention aberrant behavior long tests are more useful than small tests. However, long tests are not always feasible in practice. This renders PIRT models less useful in many applications because they require large sample sizes and sufficiently longer tests to obtain accurate estimates of the item parameters. NIRT models, and accompanying N-PFS do not suffer from these problems as they use observed group statistics and therefore are particularly useful in small samples and short tests (Junker & Sijtsma, 2001; Meijer, 2004; Molenaar, 2001). When PIRT and NIRT models are compared, NIRT models are less restrictive. The main difference between these models is about item characteristic curves. In PIRT model, these curves which are logistic or normal ogive are determined postulated parametric model (Lee et al., 2009; Sodano & Tracey, 2011). However, in NIRT models these curves do not require any parametric forms, especially MHM assumes only that monotony nondecreasing $\theta$ (Lee et al., 2009; Sijtsma & Molenaar, 2002). And so, it can be said that NIRT models are more flexible than PIRT models.

It must be emphasized that in practice if researchers want to study aberrant response behavior with N-PFS, researcher should investigate MHM assumptions. MHM can fit with skewed data (Şengül Avşar & Tavşancıl, 2017). MHM is an appropriate model for small samples (Junker & Sijtsma, 2001; Molenaar, 2001). These are MHM's important advantages to their parametric counterparts. Of course, if researchers want to study response aberrancy with P-PFS, they should demonstrate fit of the data with the parametric model assumptions. In general, if data do not fit PIRT models, researchers often can use NIRT models and N-PFS for detecting aberrant individuals.

An assumption was that all individuals answered all items in this study. In other words, there were no missing data in data sets. Missing data effects on PFS and missing data handling methods for best recovery PFS can be investigated. Apart from the test conditions determined in the study, the effectiveness of PFS can be determined by simulating different test conditions. Also, PFS which were used in this study can compared with real data applications.

## REFERENCES

Bahry, L. M. (2012). *Polytomous item response theory parameter recovery: an investigation of nonnormal distributions and small sample size* (Master's thesis). Retrieved from ProQuest Dissertations and Theses database. (UMI No. MR90146)

Baker, F. B. (2001). *The basis of item response theory*. United State of America: Eric Clearinghouse on Assessment and Evaluation.

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*(4), 335-350. doi: 10.1177/014662169301700402

Conijn, J. M., Emons, W. H., De Jong, K., & Sijtsma, K. (2015). Detecting and explaining aberrant responding to the outcome questionnaire-45. *Assessment*, *22*(4), 513-524. doi: 10.1177/1073191114560882

DeMars, C. E. (2002, April). *Recovery of graded response and partial credit parameters in multilog and parscale*. Paper presented at the annual meeting of American Educational Research Association, Chicago.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67-86. doi: 10.1111/j.2044-8317.1985.tb00817.x

Egberink, I. J. A. L. (2010). *Applications of item response theory to non-cognitive data.* Groningen: University Library Groningen.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey, NJ: Lawrence Erlbaum Associates.

Emmen, P. (2011). *A person-fit analysis of personality data* (Master thesis). Vrije Universiteit, Amsterdam. Retrieved from https://www.innovatiefinwerk.nl/sites/innovatiefinwerk.nl/files/field/bijlage/patrick_emmen.pdf

Emons, W. H. M. (2009). Detection and diagnosis of person misfit from patterns of summed polytomous item scores. *Applied Psychological Measurement*, *33*(8), 599-619. doi: 10.1177/0146621609334378

Emons, W. H. M. (2003). *Detection and diagnosis of misfitting item-score vectors*. Amsterdam: Dutch University Press.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement, 32*(3), 224-247. doi: 10.1177/0146621607302479

Emons, W. H. M., Glas, C. A. W., Meijer, R. R., & Sijtsma, K. (2003). Person fit in order-restricted latent class models. *Applied Psychological Measurement, 27*(6), 459-478. doi: 10.1177/0146621603259270

Glass, C. A. W., & Dagohoy, A. V. T. (2007). A person-fit test for irt models for polytomous items. *Psychometrika, 72*(2), 159-180. doi: 10.1007/s11336-003-1081-5

Hambleton, R. K., van der Linden W. J., & Wells, C. S. (2011). IRT models for the analysis of polytomous scored data: Brief and selected history of model building advances. In Nering M. L., & Ostini R. (Eds.), *Handbook of polytomous item response theory models* (pp. 21-42). New York, NY: Routledge.

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers In Psychology*, *7*. doi: 10.3389/fpsyg.2016.00109

Junker, B., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*(3), 211-220. doi: 10.1177/01466210122032028

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277-298. doi: 10.1207/S15324818AME1604_2

Lee, Y. S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement*, *31*(2), 121-134. doi: 10.1177/0146621606290248

Lee, Y. S., Wollack, J. A., & Douglas, J. (2009). On the use of nonparametric item characteristic curve estimation techniques for checking parametric model fit. *Educational and Psychological Measurement*, *69*(2), 181-197. doi: 10.1177/0013164408322026

Liang, T., Wells, C. S., & Hambleton, R. K. (2014). An assessment of nonparametric approach for evaluating the fit of item response models. *Journal of Educational Measurement*, *51*(1), 1-17. doi: 10.1111/jedm.12031

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*(1), 3-8. doi: 10.1207/s15324818ame0901_2

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8*(1), 72-87. doi: 10.1037/1082-989X.8.1.72

Meijer, R. R. (2004). *Investigating the quality of items in CAT using nonparametric IRT*. (LSAC Research Report Series No. 04-05). Newton, PA: Law School Admission Council.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person-fit. *Applied Psychological Measurement, 25*(2), 107-135. doi: 10.1177/01466210122031957

Meijer, R. R., & Tendeiro, J. N. (2018). Unidimensional item response theory. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 413-443). UK: John Wiley & Sons

Meijer, R. R., Egberink, I. J., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with harter's self-perception profile for children. *Journal of Personality Assessment, 90*(3), 227-238. doi: 10.1080/00223890701884921

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, *18*(2), 111-120. doi: 10.1177/014662169401800202

Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2016). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment*, *23*(1), 52-62. doi: 10.1177/1073191115577800

Molenaar, I. W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 295-299. doi: 10.1177/01466210122032091

Mousavi, A., Tendeiro, J. N., & Younesi, J. (2016). Person fit assessment using the Perfit package in R. *The Quantitative Methods for Psychology*, *12*(3), 232-242. doi: 10.20982/tqmp.12.3.p232

Nydick, S. W. (2015) *catIrt: An R package for simulating IRT-based computerized adaptive tests. R package version 0.4-2*. Retrieved from http://CRAN.R-project.org/package=catIrt

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611-630. Retrieved from https://link.springer.com/article/10.1007/BF02294494

Rupp, A. A. (2013). A systematic review of the methodology for person-fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*(1), 3-38. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2013_20130326/01_Rupp.pdf

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. USA: Sage Publications.

Sijtsma, K., Emons, W. H., Bouwmeester, S., Nyklícek, I., & Roorda, L. D. (2008). Nonparametric irt analysis of quality of life scales and its application to the world health organization quality of life scale (whoqol-bref). *Quality Of Life Research: An International Journal Of Quality Of Life Aspects Of Treatment,Care And Rehabilitation*, *17*(2), 275-290. doi: 10.1007/s11136-007-9281-6

Sodano, S. M., & Tracey, T. J. (2011). A brief inventory of interpersonal problems–circumplex using nonparametric item response theory: Introducing the iip–c–irt. *Journal of Personality Assessment, 93*(1), 62-75. doi: 10.1080/00223891.2010.528482

Spoden, C. (2014). *Person fit analysis with simulation-based methods* (Doctoral dissertation). Universitäts bibliothek Duisburg-Essen. Retrieved from https://duepublico2.uni-due.de/servlets/MCRFileNodeServlet/duepublico_derivate_00038262/DISSERTATION_Spoden.pdf

Syu, J. J. (2013). *Applying person-fit in faking detection-the simulation and practice of non parametric item response theory* (Doctoral dissertation). National Chengchi University. Retrieved from http://nccur.lib.nccu.edu.tw/bitstream/140.119/58646/1/251501.pdf

Şengül Avşar, A., & Tavşancıl, E. (2017). Examination of polytomous items' psychometric properties according to nonparametric item response theory models in different test conditions. *Educational Sciences: Theory & Practice, 17*(2). doi: 10.12738/estp.2017.2.0246

Tendeiro, J. N. (2016). *Package "PerFit"*. Retrieved from https://cran.r-project.org/web/packages/PerFit/PerFit.pdf

Twiste, L. T. (2011). *Detection of unmotivated test takers through an analysis of response patterns: beyond person-fit statistics* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3478798)

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*(3), 267-298. doi: 10.1177/0022002182013003001

Voncken, L. (2014). *Comparison of the $l_z$* Person-Fit Index and ω copying-index in copying detection* (First year paper). Universiteit van Tilburg. Retrieved from http://arno.uvt.nl/show.cgi?fid=135361

Waller, G. N., & Jones, J. (2016). Package "fungible". Retrieved from https://www.rdocumentation.org/packages/fungible

Wang, S. X. (2001). *Maximum weighted likelihood estimation* (Doctoral dissertation).University of British Columbia. Retrieved from https://open.library.ubc.ca/cIRcle/collections/ubctheses/831/items/1.0090880

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450. doi: 10.1007/BF02294627

# Birey Uyum İstatistiklerinin Farklı Test Koşullarında Çok Kategorili Puanlanan Maddeler İçin Karşılaştırılması

## *Giriş*

Psikolojik ölçme araçları, bireyler hakkında karar vermede ve bireylerin öğrenme problemleri, gelişimsel problemleri ve psikolojik bozukluklarının tanımlanması gibi amaçlarla kullanırlar. Özellikle psikolojik tanı ve tedavilerde bireysel test puanlarına odaklanılacağı açıktır (Emons, 2003, 2009). Bu nedenle bireysel test puanlarının geçerliği eğitimde ve psikolojik değerlendirmelerde araştırılması gereken önemli bir konudur.

Örneğin bir birey sınavda kaygılı olmasından dolayı sınavdaki kolay maddelere yanlış cevap verebilir. Bu durum kişinin yeteneğinin, gerçek yeteneğinin altında kestirilmesine neden olabilmektedir. Bir başka örnek ise düşük yetenekli bireylerin etraflarında bulunan yüksek yetenekli bireylerden kopya çekme durumlarıdır. Bu durumda bireyin yeteneği, gerçek yeteneğinin üstünde kestirilir. Motivasyon eksikliğine dayalı olarak testin ciddiye alınmaması, bilişsel testlerde konsantrasyon problemleri, kişilik testlerinde sahte yanıt verme durumları normal olmayan madde puanlarına kaynaklık etmektedir. Tüm bunların sonucunda bireylerin yeteneğiyle ilgili yapılan kestirimlerin hatalı olacağı açıktır (Emons, 2003, 2008; Sijtsma & Molenaar, 2002).

Uyumsuz madde puanları bireylerin puanlarını arttırarak bireyin yeteneğinin gerçek yeteneği üzerinde kestirilmesine neden olabileceği gibi uyumsuz madde puanları bireylerin puanlarını azaltarak bireyin yeteneğinin gerçek yeteneği altında kestirilmesine neden olabilir. Buna göre kopya çekenler ya da şans başarısı yüksek olan şanslı yanıtlayıcıların puanları yapay olarak yüksek kestirilirken, test uygulamasının başında kaygılı, testi sonuna kadar yanıtlamayan, ya da dil problemi olan bireylerin puanları gerçekte olduğundan yapay olarak düşük kestirilir (Meijer, 1996). Ayrıca bazen madde içeriği ile ilgili bilgisi olmayan, maddeleri kendilerine göre yorumlayan, yanıtlarını yanlış kodlayan (kodlama sırasında kaydırma yapan) bireyler de uyumsuz madde puan örüntülerine sahip olacaklardır. Bu bireyler için kestirilen puanlar, gerçekte olduğundan daha yüksek veya düşük olabilir (Meijer, 1996). Bütün bu durumlarda bireylerin doğru değerlendirilemeyecekleri açıktır. Bu nedenle test sonuçlarına göre bireyler hakkında doğru kararlar verebilmek için bireysel madde puan örüntülerinin geçerliğini değerlendirmek önem taşımaktadır.

Birey uyum analizlerinin amacı seçilen/önerilen ölçme modeline göre bireysel test puanlarının uyum gösterip göstermediğini belirlemek ve bireysel test puan vektörlerini tanımlamaktadır (Meijer & Sijtsma, 2001). Bu amaç için birey uyum istatistikleri (BUİ) kullanılır. BUİ'ler bireylerin test maddelerine verdikleri yanıtlardan beklenmedik test performansını ortaya çıkarır (Meijer & Sijtsma, 2001). BUİ'ler bireyler hakkında önemli kararlar vermede geçersiz puanları ortaya çıkararak daha geçerli sonuçlara ulaşılmasında önemli rol oynarlar (Emons, 2008).

BUİ'ler genellikle parametrik ve parametrik olmayan istatistikler olacak şekilde iki kategoride incelenmektedir (Karabatsos, 2003; Mousavi, Tendeiro, & Younesi, 2016). Parametrik BUİ'ler (P-BUİ) parametrik madde tepki kuramına (PMTK), parametrik olmayan BUİ'ler (PO-BUİ) parametrik olmayan madde tepki kuramına (POMTK) dayalıdır (Karabatsos, 2003). P-BUİ ve PO-BUİ arasındaki temel fark, dayandıkları madde tepki kuramıdır. POMTK modellerinin getirdiği birtakım avantajlar, PO-BUİ'lere de yansımaktadır. PO-BUİ'ler için verinin POMTK modeline uyum göstermesi gerekmektedir (Emons, 2003). Özellikle verinin POMTK modellerinden Mokken Homojenlik Modeline (MHM) uyum göstermesi, diğer bir deyişle tek boyutluluk, yerel bağımsızlık ve madde karakteristik eğrilerinin monotonluğu varsayımlarının sağlanması gerekmektedir (Emons, 2008). Literatürde çok kategorili puanlanan maddeler için en fazla kullanılan P-BUİ'nin $l_z^p$ istatistiği, PO-BUİ'lerin $G^p$, $G_N^p$ ve $U3^p$ istatistikleri olduğu ifade edilmektedir (Emons, 2008; Rupp, 2013).

Birey uyum analizleri eğitimde ve psikolojide önemli bir konu olarak ele alınmaktadır. Özellikle başarı testleri ve bilişsel testlerde başarıyla uygulanmaktadır (Meijer & Sijtsma, 2001). Eğitim çalışmalarında (örneğin müfredattaki tutarsızlıkların belirlenmesinde, Harnisch & Linn, 1981), bilişsel psikoloji çalışmalarında (öğrenme stratejilerinin belirlenmesi, Tatsuoka & Tatsuoka, 1982), kültürler arası karşılaştırmalar (farklı dil gruplarından gelen bireylerin test puanlarının değerlendirilmesi ve karşılaştırılması, van der Flier, 1982), kişilik ölçme çalışmalarında (kişilik ölçme amacıyla geliştirilen ölçme araçlarında sahte yanıtların belirlenmesi, Dodeen & Darabi, 2009; Ferrando, 2004, 2009, 2012; Reise & Waller, 1993; Woods, Oltmanns, & Turkheimer, 2008; Zickar & Drasgow, 1996), örgüt psikolojisi çalışmalarında (bireylerin seçilen test için beklenmedik madde puan vektörlerini açıklama, Meijer, 1998), tutumların değerlendirilmesi (Curtis, 2004), sağlık araştırmaları (Custers, Hoijtink, van der Net & Hel, 2000; Tang ve diğerleri, 2010) örnek olarak verilebilir (akt., Emons, 2003; Rupp, 2013). BUİ'ler psikolojik değerlendirmelerde de (Conijn, Emons, De Jong & Sijtsma, 2015; Meijer, Egberink, Emons & Sijtsma, 2008) başarıyla uygulanmaktadır.

Yapılan literatür taramasında araştırmacıların; yeni BUİ'ler geliştirdikleri ve yeni geliştirilen bu BUİ'leri çeşitli test koşullarında inceledikleri (Emons, 2008; Glass & Dagohoy 2007; Karabatsos, 2003; Twiste 2011; van der Flier, 1982), uyumsuz madde puanlarının gerçek veri setlerinde belirledikleri (Egberink, 2010; Emmen, 2011; Meijer, 2003; Spoden, 2014) ve en iyi performans gösteren BUİ'leri belirledikleri (Emons, 2008; Karabatsos, 2003; Syu, 2013; Voncken, 2014) görülmüştür. Rupp'un (2013) çalışmasında da BUİ ile ilgili literatür taranmıştır. Yapılan bu çalışmada BUİ'lerin özellikle ikili puanlanan maddelerde daha fazla çalışıldığı, çok kategorili puanlanan maddelerde yapılan çalışmaların çok sınırlı olduğu ifade edilmiştir. Bununla birlikte yapılan literatür taramasında simülatif olarak üretilen veriler üzerinde BUİ'lerin özellikle küçük örneklemler ve çarpık dağılımlar gibi çeşitli test koşullarında daha fazla araştırılması gerektiği görülmüştür.

*Çalışmanın amacı*

Bu çalışmanın genel amacı P-BUİ ve PO-BUİ'lerin çok kategorili puanlanan maddelerden oluşan testlerde etkililiklerinin belirlenmesidir. Belirlenen amaç doğrultusunda aşağıdaki araştırma sorularına cevap aranmıştır:

1. BUİ'lere göre belirlenen uyumsuz madde puanlarına sahip kişilerin oranı; örneklem büyüklüğü, yetenek dağılımı, test uzunluğu ve madde ve kişilerin manipülasyonuna bağlı olarak oluşturulan anormallik durumlarına göre nasıl değişmektedir?
2. Farklı test koşullarında en iyi performansı gösteren BUİ hangisidir?

***Yöntem***

Bu araştırma BUİ'lerin, simülatif olarak oluşturulan test koşullarında, etkililiklerinin belirlenmesinin amaçlandığı temel araştırmadır.

*Veri simülasyonu*

Bu araştırmada çok kategorili puanlanan maddeler Samejima'nın Dereceli Tepki Modeline (DTM) göre üretilmiştir. Bu araştırmada POMTK'ya dayalı PO-BUİ'ler araştırmasına rağmen, parametrik DTM'ye göre veri üretilmesinin nedeni DTM'ye uyumlu olan veri setinin aynı zamanda MHM'ye uyumlu olmasıdır (Emons, 2008; Sijtsma, Emons, Bouwmeester, Nyklícek & Roorda, 2008). Verilerin üretilmesinde R programı kullanılmıştır. DTM'ye uygun verilerin üretilmesinde "catIRT" paketi (Nydick, 2015), çarpık dağılımlı veri setlerinin üretilmesinde "fungible" paketi (Waller & Jones, 2016) kullanılmıştır. Bu araştırmada simülatif verilerin üretilmesinde aşağıdaki adımlar izlenmiştir:

1. Belirlenen test koşullarında DTM'ye uyumlu veri setleri üretilmiştir.
2. Araştırmanın amacı doğrultusunda, veri setleri uyumsuz madde puanı içerecek şekilde (düşük ve yüksek oranlarda) manipüle edilmiştir.
3. Manipüle edilen veri setlerinde uç değerler belirlenmiş (tüm maddelerde kesinlikle katılıyorum veya hiç katılmıyorum kategorilerini seçenler) ve analiz dışı tutulmuştur. BUİ'lerin uç değerlerde yeteri kadar bilgi vermemesi (Emons, 2008), uç değerlerin atılmasının temel nedenidir.
4. Yetenekler ağırlıklandırılmış maksimum olasılığa (weighted maximum likelihood estimation) göre kestirilmiştir. Yetenekler kestirilirken veri üretimindeki gerçek madde parametreleri kullanılmıştır.
5. Farklı test koşullarında uyumsuz madde puanlarının belirlenmesi için BUİ'ler, Tendeiro (2016) tarafından geliştirilen "perfit" paketi kullanılarak kestirilmiştir.

Bu araştırmanın bağımsız değişkenleri; dört farklı örneklem büyüklüğü (100, 250, 500 ve 1000), üç farklı örneklem dağılımı (normal dağılan, sağa çarpık dağılan ve sola çarpık dağılan), iki farklı test uzunluğu (10 maddelik ve 30 maddelik test) ve iki farklı uyumsuzluk (düşük ve yüksek düzeylerde) oranıdır. Bağımlı değişkenleri ise deneysel I. Tip Hata oranları ve bu değerler için hesaplanan güç değerleridir. Bu araştırmada dört farklı BUİ ($l_z^p$, $U3^p$, $G_N^p$ ve $G^p$) için I. Tip Hata oranları ve güç değerleri hesaplanmıştır.

Literatürde uyumsuz madde puanlarına neden olabilecek çeşitli davranışlardan bahsedilmiştir. Bu araştırmada *dikkatsiz ve özensiz davranışlar* dikkate alınmıştır. Bazı test uygulamalarında bireyler maddeleri rastgele cevaplarlar, maddeleri yanlış okurlar, maddeleri okumazlar ya da kodlama hatası yaparlar. Bu durumlar dikkatsiz ve özensiz davranışlara örnek olarak verilebilir (Emons, 2008). Bu araştırmada, bu davranışa yönelik uyumsuz madde puan vektörleri Emons'un (2008) çalışmasında olduğu gibi tek biçimli dağılımdan yararlanarak oluşturulmuştur.

## *Sonuç ve Tartışma*

Bu araştırmanın genel amacı, P-BUİ ve PO-BUİ'lerin etkililiklerinin çok kategorili puanlanan maddelerden oluşan test koşullarında etkililiklerinin belirlenmesidir. Araştırma sonucunda beklendiği gibi, hesaplanan BUİ'ler için, I. Tip Hata oranı arttıkça uyumsuz madde puanına sahip bireylerin belirlenme oranı artmıştır. Araştırmada oluşturulan test koşullarında madde sayısı ve uyumsuz madde puan vektörleri arttıkça uyumsuz madde puanı belirleme oranı/güç artmıştır. Simülasyon sonuçları örneklemin dağılım şeklinin uyumsuz madde puanlarını belirlemede küçük bir etkisinin olduğunu göstermiştir. Diğer bir deyişle yetenek dağılımının şekli, uyumsuz madde puanı belirlemede bu araştırmadaki test koşullarına göre önemli bir faktör değildir. Genel olarak örneklem büyüklüğü, uyumsuz madde puanı oranlarını etkilemiştir. Örneklem büyüklüğü artıkça uyumsuz madde puanlarının belirleme oranları artmıştır. Araştırmanın bu bulgusu Syu'nun (2013) bulgularıyla farklılaşmıştır. Syu (2013) çalışmasında $l_z^p$, $G^p$ ve $U3^p$ istatistiklerini araştırmıştır. Syu (2013) oluşturduğu test koşullarında örneklem büyüklüğünün çok küçük farklılıklar oluşturduğunu ancak seçilen koşulların BUİ'lerle ilgili yeterli bilgi veremeyeceğini de belirtmiştir.

Özetlenecek olursa nominal I. Tip Hata oranları artıkça, uzun testler kullanıldıkça ve manipüle edilen uyumsuz madde puanlarının oranı artıkça, uyumsuz madde puanlarının belirlenmesinin oranı da artmaktadır. Bu bulgu literatürdeki diğer araştırma bulgularına paraleldir (Emons, 2008; Karabatsos, 2003; Meijer & Sijtsma, 2001; Voncken, 2014).

Araştırmada genel olarak $G^p$ istatistiğinin en iyi performansa sahip BUİ olduğu görülmüştür. Ancak özellikle uzun testlerde parametrik $l_z^p$ istatistiğinin daha iyi performans gösterdiği de belirtilmelidir. Kısa testlerde ve küçük örneklemlerde $G^p$ istatistiğinin daha iyi performans göstermesi, Emons (2008) ve Syu'nun (2013) araştırma bulgularına paraleldir. Syu (2013) çalışmasında küçük örneklemlerde PO-BUİ'lerin daha iyi performans gösterdiğini belirtmiştir. Ek olarak bu araştırmada BUİ'lerin uyumsuz madde puanlarını belirleme oranları, birbirlerine yakın değerler vermiştir. PO-BUİ'lerde özellikle $U3^p$ ve $G_N^p$ birbirine oldukça yakındır. Uyumsuz madde puanlarını belirleme oranı en fazla $\alpha = .20$ düzeyinde olmuştur. Bu durum literatüre paraleldir (Emons, 2008; Meijer, 2003; Voncken, 2014).

Araştırma sonuçlarına göre dikkatsiz ve özensiz davranışların kaynaklık ettiği uyumsuz madde puanlarının belirlenmesinde uzun testlerin tercih edilmesi önerilebilir. Ancak uzun testler pratikte her zaman çok kullanışlı değillerdir. PMTK modelleri de parametrelerin doğru kestirilmesi için büyük örnekleme duyulan ihtiyaçtan dolayı çok kullanışlı değildir. Bu durumda PMTK modellerine göre daha az sınırlayıcı olan POMTK modellerinden MHM (Junker & Sijtsma, 2001; Meijer, 2004; Molenaar, 2001) kullanılarak uyumsuz madde puan örüntüleri PO-BUİ'lerle belirlenebilir.

Bu araştırma oluşturulan test koşulları dikkate alındığında özellikle küçük örneklem büyüklüklerinde ve kısa testlerde PO-BUİ'lerin kullanılması önerilebilir. Bu araştırmada kayıp veri içeren veri setleri üretilmemiştir. Belirlenen test koşullarında kayıp verilerin BUİ'lerin performanslarını nasıl etkiledikleri araştırılabilir. Araştırmada belirlenen test koşullarının dışında, farklı test koşulları oluşturularak BUİ'lerin etkililikleri belirlenebilir. Ayrıca bu araştırmada kullanılan istatistikler, gerçek veri setlerine kullanılarak araştırmanın bulgularıyla karşılaştırılabilir.

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(4); 365-377

# Investigation of the Variables Affecting the Students' Science Achievement with Multilevel Regression Model *

Ezgi ULUTAN ** Derya ÇOBANOĞLU AKTAN ***

**Abstract**

This study investigated the variables affecting the science achievement of eighth-grade students by multi-level regression analysis. The variables included in this research were students' attitudes, confidence level, value, engagement in science, socioeconomic status, school type, school region, and teacher experience. The study group consisted of 1049 students and 41 teachers. In the first research question, differences in students' science achievement scores among their schools were investigated. According to the results, the students' achievements differed among their schools. Approximately 16.3% of the differences observed in science achievement were stem from the differences among schools, and 83.6% stem from the differences among students. In the second research question, student characteristics that explain the differences among the science achievements of the schools have been examined. Students' socioeconomic level, attitude, and confidence level were only variables that have statistically significant relationship with achievement. Socioeconomic and confidence level variables have a positive effect on achievement, but attitude variable has a negative effect on achievement. In the third research question, student and school characteristics that affect science achievement have been examined simultaneously. The school characteristics that have been included in the regression model were teacher experience, region, and school type. It was determined that none of the regression coefficients for the school characteristics variables were statistically significant in the regression model.

*Key Words:* Multi-level regression analysis, TEOG science exam, affective characteristics of students, school characteristics.

## INTRODUCTION

The rapidly developing technology, the growth of the economy, and the changes in priorities of social life lead to the differentiation of the needs of our lives. Particularly, the rapid progression of technology makes science fields more prominent. Therefore, in recent years, countries started to emphasize science education and encourage students to enter science-related jobs more than the other fields. According to the report of the Scientific and Technological Research Council of Turkey (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu-TUBİTAK), science and technology will be the foundation of the professions which will be needed in the future (TUBİTAK, 2016). To be able to enter the occupational fields related to science, it is very important for individuals to have an interest in science and concrete science education. However, it is noteworthy that nowadays individuals are not inclined toward science-related occupational fields. The lack of employees in these areas is expected to affect the productivity and technological development of countries significantly. For this purpose, the importance of science education and the factors affecting the success of students should be examined, and interest in these fields should be increased. In this context, many studies on the science achievement of students at both national and international levels were done, and the factors affecting the students' success of science were examined in Turkey.

When the studies concerning the national examinations administered in Turkey on science were examined, various variables affecting the science achievement of students have been determined.

* The paper is produced from the first author's master degree thesis.

** Ankara-Turkey, ezgigunal5@gmail.com, ORCID ID: 0000-0002-4554-3362
*** Assist. Prof., Hacettepe University, Faculty of Education, Ankara-Turkey, coderya@gmail.com, ORCID ID: 0000-0002-8292-3815

To cite this article:
Ulutan, E. & Çobanoğlu Aktan, D. (2019). Investigation of the variables affecting the students' science achievement with multilevel regression model. *Journal of Measurement and Evaluation in Education and Psychology, 10*(4), 365-377. doi: 10.21031/epod.533713

*Received: 28.02.2019*
*Accepted: 4.11.2019*

According to the literature, these variables are socioeconomic level, value, self-efficacy, attitude, perception, education level of the family, gender, time allocated to study, teacher characteristics, and school characteristics (Acar, 2009; Anıl, 2011; Atalmış, Avgın, Demir & Yıldırım, 2016; Ötken, 2012; Şahin, 2011; Uzun, Gelbal & Öğretmen, 2010).

In addition to national exams, variables affecting the science achievement of the Turkish students at international exams such as PISA (Program for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study) were also investigated in the literature. These variables are attitude, self-efficacy, value, socioeconomic level, education level of a family, gender, home resources, material resources, computer environment teacher characteristics, and school location (Abazoğlu & Taşar, 2016; Acar & Öğretmen, 2012; Akıllı, 2015; Akyüz, 2006; Anıl, 2009; Atar & Atar, 2012; Berberoğlu, Çelebi, Özdemir, Uysal & Yayan, 2003; Büyüköztürk, Çakan, Tan & Atar, 2014; Pektaş, 2010; Uçar & Öztürk, 2010).

These variables were investigated in various combinations in the related research. For example, Anıl (2011) investigated the factors that predict PISA science achievement of the Turkish students with the parents' level of education, attitude, computer, and family's wealth of culture variables. Pektaş (2010), on the other hand, evaluated the students' TIMSS science scores with the variables of attitude, self-efficacy, value, and education level of the family. In another study, 8th-grade students' science achievement in TIMSS were examined via attitudes, values towards science, and self-efficacy variables (Akıllı, 2015).

These types of studies have only addressed student characteristics. In addition to student characteristics, there are also studies dealing with the characteristics of teachers and schools. For instance, in the TIMSS-2011 study conducted by Abazoğlu and Taşar (2016), teacher characteristics that affect students' science achievement were determined as job satisfaction, computer use in class, and participation in professional development activities. In terms of teacher characteristics, Atar (2014) found that some teacher characteristics measured by TIMSS 2011 were determiners of the students' science and technology achievement. Those teacher characteristics were participation in in-service training programs related to information technologies, importance given by teachers to academic achievement, gender of teachers, and cooperation among colleagues.

The variables such as attitude and self-efficacy discussed in these studies are the individual characteristics of the students, whereas the variables such as teacher experience and school type are characteristics of students' groups. In other words, there are variables related to the students and student groups. That is, the data obtained from the students and their schools show a hierarchical structure such as students, classes and schools. If this hierarchical structure is ignored when examining the predictors of science achievement, the principle of independence required for regression analysis is violated, and the result of the analysis may be biased. In hierarchical data, more complex error structure should be added to the model to take account of the dependence between observations within the group (Heck, Thomas, & Tabata, 2010). Multilevel modeling, on the other hand, ensures that the predictor variables are analyzed in accordance with the hierarchical structure of the data and obtain unbiased results (Heck et al., 2010).

The studies aiming at determining the variables affecting the students' science achievement are generally performed with single-level analysis for both the national (e.g. high school entrance examinations, etc.) and international (PISA and TIMSS, etc.) exams administered in Turkey (e.g. Acar, 2009; Ötken, 2012; Süer, 2014; Şahin, 2011). Most of these studies were conducted without considering the hierarchical structure of the data. In the TIMSS and TEOG (Transition from Basic Education to Secondary Education) exams, the hierarchical structure of the data necessitates the examination of variables predicting achievement at different levels (individual and school). The use of multi-level analysis in the examination of structures at different levels is more appropriate than the use of single-level models due to the fact that the observations are not independent of each other and the design effect (Hox, 2010). Multilevel analyses are methods of analysis that examine the relationship between variables that characterize individuals and groups. In multilevel analyses, the data structure within the group is hierarchical, and the data should be taken from this hierarchical group (Hox, 2010).

In the literature there are multi-level analysis studies examining students' science achievement in the TIMSS exam (Abazoğlu & Taşar, 2016; Acar & Öğretmen, 2012; Atar, 2014; Atar & Atar, 2012) and in the TEOG exam for subjects such as mathematics and Turkish (Acar, 2013; Doğan & Demir, 2015; Yavuz, Odabaş & Özdemir, 2016). However, in the literature, there are no studies investigating the individual and group level variables affecting the science achievement for the national exams carried out in Turkey by multilevel analysis. In this study, it was aimed to investigate the variables that predict the students' science achievement by multilevel analysis in accordance with the hierarchical structure of the TEOG data. Thus, the extent to which the variables related to individuals and schools related to achievement will be examined in a more unbiased manner. Examination of the students' science achievement by multilevel analysis for a national exam, provides an opportunity to compare the findings of this study with those of single-level analysis and also helps to fill the gap in the literature on this issue. TEOG is a test conducted by the Ministry of National Education (MONE) for the evaluation of student achievement in an integrated manner with the learning process and applied for the evaluation of science achievement. The aim of this study is to examine the science achievement of eighth-grade students who participated in the TEOG science sub-test. By providing scores that are on the same scale, TEOG allows the comparison and inclusion of students (with different characteristics) from different cities and districts of Turkey. Thus, the relationship between the variables included in this study and a national science exam scores can be examined across Turkey. The school-level variables in this study are school region, school type, and teacher experience; and the student-level are the students' socioeconomic level, value given to science, interest in science, self-efficacy and attitude. By using these variables, in this study, the answer to the question To what extent do the school and student level variables predict students' science achievements? is examined. Furthermore, the following research questions guided this study:

1. Do students' science scores show a significant difference among their schools?
2. To what extent do students' science scores are predicted by level-1 (student) variables (interest, value, self-efficacy, attitude, and socioeconomic status)?
3. To what extent do students' science scores are predicted by level-1 and level-2 (school) variables (regional population, type of school, and teacher experience)?

Within the scope of the research, it is assumed that the students answered the questionnaire items in a sincere manner. This research is limited to the answers of the students and teachers to the questionnaire items selected from the TIMSS 2011 measurement tool and the variables determined in the measurement tool.

## METHOD

The related information about the method of the study is presented at the parts below.

### *Participants*

In the study, 1049 8th grade students who took the TEOG exam attending 30 different schools (26 state schools and 4 private schools) in Düzce, Erzurum, Çankırı, Antalya, and Ankara in 2015-2016 school year were participants. 597 of the students were female, and 452 of them were male. In addition, a total number of 41 teachers, 37 of whom were working in a state, and 4 of whom were working in a private school, participated in the study voluntarily. School-level data were collected from the teachers. Participants of the study were selected from conveniently available schools. Therefore, convenience sampling was used in the study.

### *Data Collection Instrument*

Some of the TIMSS 2011 student and teacher questionnaire items were selected and used in the data collection tool of this study. The reasons for using TIMSS items are the research support for items' validity and reliability; comprehensiveness of the items for the related variables and finally comparability property. The relevant TIMSS items were administered to the students and the teachers. Students' TEOG science scores were obtained based on their statements.

The first part of the measurement tool for the students includes 12 demographical items. These are about gender, age, parents' educational level and occupation, home resources (number of books at home, computer, desk, separate room, and internet), and TEOG science score. The second part includes 26 affective items from TIMSS 2011 student questionnaire. The codes for the original TIMSS items were BSBS17A-F, BSBS19A-N, and BSBS18A-E. These items were related to attitude, self-efficacy, interest in science, value given to science. The specific item codes for interest variable are BSBS18A, BSBS18B*, BSBS18C, BSBS18D, BSBS18E; for self-efficacy BSBS19A, BSBS19B*, BSBS19C*, BSBS19D, BSBS19E*, BSBS19F, BSBS19G, BSBS19H, BSBS19I*; for attitude BSBS17A, BSBS17B*, BSBS17D*, BSBS17E, BSBS17F; and for value variable BSBS19J, BSBS19K, BSBS19L, BSBS19M, BSBS19N, BSBS17G'. * items were coded inversely in the study. The measurement tool for the teachers consists of items about teachers' year of experience, regional population of the school, and school type.

### *Data Analysis*

In order to reduce the number of variables to be included in the multi-level regression analysis, the questionnaire items were subjected to exploratory factor analysis, and the obtained variables were used in the regression analysis. The appropriateness of collected data for factor analysis was analyzed by the Kaiser-Meyer-Olkin (KMO) coefficient and Bartlett's sphericity test. In the study, KMO coefficient was calculated as .935, and this value was found to be good (.80 < KMO < .90) in order to continue factor analysis (Büyüköztürk, 2015). In the Bartlett Sphericity Test, the chi-square value ($\chi^2$ = 2067.004; $p$ = .000 < .05) was found to be significant. According to the obtained results, the data showed multivariate normality (Büyüköztürk, Şekercioğlu & Çokluk, 2014). In the factor analysis, the items were analyzed in separate groups for the factors as in the analysis of the 2011 TIMSS measurement tools. Table 1 shows the number of items in each factor, the total explained variance, and KMO. After factor analysis, for interest, value, attitude, self-efficacy, socioeconomic status of the students factor scores were obtained. In addition to these student-level variables, teacher experience, the population in the school region, type of school were considered as independent variables in the regression model. The participant students' TEOG science scores were considered as the dependent variable.

Table 1. Factor Analysis Results for Attitude, Self-Efficacy, Value, Interest and Socioeconomic Status Variables

| Variable | Number of Items | KMO | Total explained variance (%) |
|---|---|---|---|
| Attitude | 6 | .828 | 52.736 |
| Self-efficacy | 9 | .877 | 50.002 |
| Value | 6 | .827 | 53.745 |
| Interest | 5 | .751 | 46.750 |
| Socioeconomic status | 3 | .771 | 39.365 |

In the collected data, there were 35 cases with missing data. In the study, the mean values were assigned for these missing data, and the analyses were performed with 1049 participants. The students' TEOG science scores showed normality. In the analysis, condition indices (CI), variance inflation factor (VIF) and tolerance values were examined for collinearity among the independent variables. The tolerance values of the variables were greater than .20; variance inflation factor (VIF = 1 / (1-R$^2$)) values were less than 10; CI were found to be less than 30. The internal consistency reliability coefficients of each factor were calculated with Cronbach Alpha. The Cronbach Alpha coefficient was α = .83 for the value variable, α = .88 for the self-efficacy variable, α = .81 for the attitude variable, and α = .69 for the interest variable. The Cronbach's Alpha coefficient for the whole measurement tool (α = .921) is over .70, indicating the reliability of the measuring instrument. The data were analyzed with a mixed model (SPSS 20.0). In the following section, multi-level regression analysis and regression models used in this study are explained.

### *Multilevel analysis*

In studies that examine the relationship between individual and society/group, data can be observed at different hierarchical levels, and variables can be defined for each level. Multilevel analyses are methods that examine the relationship between variables that characterize individuals and groups (Hox, 2010). If the data structure is ignored, aggregation and disaggregation problems appear. In the aggregation, researchers are interested in group-level data, so they aggregate the variables that characterize individuals in each group to a higher level (group level). In disaggregation, to analyze data at a single level the variables belonging to the upper level are assigned to the individual level. However, aggregation and disaggregation may cause some errors (Heck et al., 2010). In the hierarchical groups, individual observations are generally not completely independent. Therefore, the mean correlation between the variables measured on students from the same school (so-called intra-class correlations) is higher than the average correlation between the variables measured in different schools. If the sample is not random, participants from the same geographical region will be more similar to each other compared to participants from different geographical regions. Being nonrandom sample (having similar characteristic) leads to standard error estimates that produce incorrect results. To prevent incorrect results design effect has to be considered in analysis. Intra-class correlation (ρ) is used to calculate the design effect. Intra-class correlation is defined as the ratio of variance between the groups compared to the total variance. Intra-class correlation can also be interpreted as the expected correlation between two randomly selected individuals in the same group. Intra-class correlation is calculated by the formula shown in Equation 1.

$$\rho = \sigma_b^2 \Big/ \sigma_b^2 + \sigma_w^2 \qquad (1)$$

The design effect (Deff) depends on both the intra-class correlation and the sample size. Deff for a model with a two-level data structure is shown in Equation 2.

$$\text{Deff}=1+ \rho(n-1) \qquad (2)$$

In this study there are two levels. Level-1 is student-level and level-2 is school-level. The participants' TEOG science scores (Y) were used as the dependent variable. The independent variables at the student level (Level 1) and the variables included in the model at the school level (Level 2) are stated below.

Table 2. Independent Variables of Level- 1 (Student) and Level-2 (School)

| **Level-1 Student level** | **Independent variables** |
|---|---|
| Socioeconomic status | SES ($X_1$) |
| Attitude | TUT($X_2$) |
| Value | DEĞ ($X_3$) |
| Interest | ILG ($X_4$) |
| Self-efficacy | OZY($X_5$) |
| **Level-2 School level** | |
| School region population | BOL($X_6$) |
| School type | TUR($X_7$) |
| Teacher experience | OGR($X_8$) |

The first question of this research is do students' science scores show significant difference among their schools? In order to answer this question, the intra-class correlation and design effect was calculated for the available data. For this purpose, the one-way ANOVA model was established in multilevel analysis.

In the multilevel analysis, the one-way ANOVA model examines the between and within-group components of variances (Heck et al., 2010). This model provides information about intra-class correlation and determines whether a multilevel model is required or not (Tabachnick & Fidell, 2007). One-way ANOVA model is presented in Equation 3.

$$Y_{ij}= \beta_{0j}+ \varepsilon_{ij} \quad (3)$$

The equation of level 2 of the model is given in Equation 4.

$$\beta_{0j}= \gamma_{00}+ u_{0j} \quad (4)$$

Equation 5 is obtained when the Equation 4 is inserted in Equation 3.

$$Y_{ij}= \gamma_{00}+ u_{0j}+ \varepsilon_{ij} \quad (5)$$

This model provides the level of dependence in level 2 through intra-class correlation (ρ). After determining the necessity of multilevel analysis, first level predictor model (level-1 model-random intercepts- constant slope with fixed estimators) was established to answer the second research problem. The model obtained by adding a predictor to the equation used in the estimation of student success is called *the first level predictive model* (Tabachnick & Fidell, 2007). The level-1 estimators are indicated by X. The equation for the student level model is given below in Equation 6. In this equation, the absence of j index in the $\beta_1$ coefficient indicates that the slope is constant for the groups.

$$Y_{ij}= \beta_{0j}+ \beta_1 X_{ij} + \varepsilon_{ij} \quad (6)$$

Equation 7 is used to predict the slope.

$$\beta_1= \gamma_{10} \quad (7)$$

Equation 7 and Equation 4 are inserted in Equation 6, and Equation 8 is obtained. In this equation, when the fixed parameters ($\gamma_{00}$ and $\gamma_{10}$) and random parameters ($u_{0j}$ and $\varepsilon_{ij}$) are edited, Equation 8 is obtained.

$$Y_{ij}= \gamma_{00}+ \gamma_{10}X_{ij}+u_{0j}+ \varepsilon_{ij} \quad (8)$$

By considering student level variables, the Equation 9 is obtained.

$$Y_{ij}= \beta_{0j} + \beta_1(SES)_{ij} + \beta_2(ILG)_{ij}+ \beta_3(DEG)_{ij}+ \beta_4(OZY)_{ij}+ \beta_5(TUT)_{ij}+ \varepsilon_{ij} \quad (9)$$

Through this analysis, β values are determined for the independent variables (SES, ILG, DEG, OZY, and TUT). These values indicate at what level these variables predict the students' science scores. In addition, in order to determine to what extent individual level independent variables added to the model explain the difference between schools, the difference between the variance values for the first level predictive model and the variance values in the one-way ANOVA model are examined. This reduction at variance is calculated by between- and within-group variance estimation ($R^2$). To calculate reduction in variance, Equation 10 is used for between- and within- group variance.

$$\left(\sigma^2_{M1}- \sigma^2_{M2}\right) / \sigma^2_{M1} \quad (10)$$

To answer the third and last research question, school-level variables have been added to the multi-level regression model. Group-level variables are added to the multi-level model (random intercepts fixed slope).

$$\beta_{0j} = \gamma_{00}+ \gamma_{01}W_j+ u_{0j} \quad (11)$$

Adding the independent variables (W and X) at the group level and at the individual level yields the Equation 12. Equation 12 is reached when the terms are arranged.

$$Y_{ij}= \gamma_{00}+ \gamma_{01}W_j + \gamma_{10}X_{ij}+ u_{0j}+ \varepsilon_{ij} \tag{12}$$

Thus, at the school level, variables are added to the equation to explain the variability of the intercepts between schools. Three independent variables in level 2 (school level) have been added to the model. The Equation 13 is obtained when they are placed in Equation 10 at the school level as independent variables.

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(BOL)_j + \gamma_{02}(TUR)_j + \gamma_{03}(OGR)_j + u_{0j} \tag{13}$$

When Equation 13 is combined with the level 1 (student level) variables,

$$Y_{ij}= \gamma_{00} + \beta_1(SES)_{ij} + \beta_2(TUT)_{ij}+ \beta_3(DEG)_{ij}+ \beta_4(ILG)_{ij}+ \beta_5(OZY)_{ij}+ \gamma_{01}(BOL)_j + \gamma_{02}(TUR)_j+ \gamma_{03}(OGR)_j + u_{0j}+ \varepsilon_{ij}$$

is obtained. Through this analysis, the levels of school level (TUR, OGR, BOL) are predicted in terms of predicting student science scores.

## RESULTS

### *Results for the First Research-Problem*

The results of the one-way ANOVA model analysis are given in Table 3. In this model, the average of the students' science scores is determined as 72.76. The standard error of the estimated value is 1.56. In the 95% confidence interval, the real value of the overall science achievement average is in the range of 75.83 - 69.70 points.

Table 3. One-way ANOVA Model Results

| **Fixed effects** | **Coefficient** | **Standard error** | **df** | **t** |
|---|---|---|---|---|
| Average science score | 72.76* | 1.56 | 30.53 | 46.54 |
| **Random effects** | **Variance** | **Standard error** | | **Wald Z** |
| level-1 within-group variation, student level | 308.98* | 13.67 | | 22.60 |
| Level-2 between group variation, school level | 60.37* | 18.56 | | 3.25 |

$^*p < .01$

The variance of the students' science achievement for the school average is estimated as 308.99 (within-group variability), and the variance of the difference of the school means from the general average is 60.37 (between-group variability). Intra-class correlation coefficient is calculated by Equation 1. By using these variance values, it is calculated as 60.37 / (60.37 + 308.98) = 0.163 or 16.3%. When Table 3 is examined, there is a significant difference among TEOG achievement scores (Wald Z = 22.60, $p < .05$). Approximately 16.3% of the differences observed in the students' science scores arise from the differences between schools. Similarly, by using within-group variance: 308.98 / (308.98 + 60.37) = 0.836 or 83.6% is obtained. This value indicates that 83.6% of the total variance stems from the differences among the students. In addition to these values, the design effect (Deff) is calculated in the following way.

$$Deff = 1+ 0.163 ((1049/30) -1) = 5.537$$

Since Deff is 5.537 > 1, it is seen that the data requires multilevel modeling. The results show that, with the average score difference among schools, the development of the model can be continued.

### *Results for the Second Research-Problem*

In the level-1 student model, within- and between-group intercept and slope equations are examined. In order to determine the student characteristics associated with the students' science scores at level 1, some predictive variables are included in the model. These variables are the students' socioeconomic level (SES), attitude (TUT), value (DEG), interest (ILG) and self-efficacy (OZY). Table 4 shows the estimated values of the fixed and random effects of the level 1 model. When the intercept coefficient (208.23) level-1 variables are taken into account in Table 4, it gives the variance value of the differences of the students' science achievement from the school average.

The slope coefficients of independent variables with high t value and statistical significance are socioeconomic level, attitudes, and self-efficacy variables. According to Table 4, the socioeconomic level ($\beta_1$ = 7.36, $p$ < .05) is among the variables affecting student achievement. In addition to this variable, students' attitudes towards science ($\beta_2$ = -3.19, $p$ < .05) affect student achievement at individual level. Self-efficacy perceptions of students ($\beta_5$ = 10.03, $p$ < .05) are also among the variables that affect student achievement. It is concluded that the students' interest in science ($\beta_4$ = -0.32, $p$ > .05) and the value that students give to science ($\beta_3$ = 0.87, $p$ > .05) do not statistically affect student science scores. According to these coefficients, the socioeconomic level ($\beta_1$ = 7.36) and the self-efficacy ($\beta_5$ = 10.03) levels of students affect the science achievement positively. The attitude variable shows a significant negative relationship with the students' TEOG science scores. However, the interest ($p$ =.640 >. 05) and value variables ($p$ = .161 > .05) are not statistically significant. These results show that students with higher socioeconomic levels and higher self-efficacy have higher science scores.

Table 4. Random Intercept Model Results

| **Fixed effect** | **Coefficient** | **Standard error** | **df** | **t** |
|---|---|---|---|---|
| Average science score | 73.10 | 0.84 | 21.94 | 87.35 |
| SES | **7.36*** | 0.63 | 481.11 | 11.63 |
| Attitude | **-3.19*** | 0.76 | 1042.05 | -4.19 |
| Value | 0.87 | 0.62 | 1034.09 | 1.40 |
| Interest | -0.32 | 0.70 | 1035.25 | -0.47 |
| Self-efficacy | **10.03*** | 0.64 | 1038.67 | 15.56 |
| **Random effect** | **Variance** | **Standard error** | | **Wald Z** |
| Within-group variance, student level (Level-1) | 208.23 | 9.27 | | 22.46 |
| Between group variance, school level (Level- 2) | 12.97 | 6.02 | | 2.15 |

*$p$ < .01

In order to examine the influence of socioeconomic status, attitude, self-efficacy, interest, and value variables as within-group variables on the model, the variance between ANOVA and first level predictor model is examined. For this purpose, the estimation of reduction in variance ($R^2$), (308.99-208.23) / 308.99 = 0.326 or 32.6% is obtained.

This result shows that 32.6% of the level-1 variability in student science scores is explained by the variables of student socioeconomic level, attitude, self-efficacy, interest, and value. For the reduction in variance between schools, (60.37-12.97) / 60.37 = 0.785 or 78.5% is obtained.

This result is due to the socioeconomic level, attitude, self-efficacy, interest, and value variables of the students. Between and within-group variance components obtained in the one-way ANOVA model decreased when socioeconomic level, attitude, self-efficacy, interest, and value variables are added to the model. In other words, approximately four-fifths of the variance between schools arises from the differences in the socioeconomic level, attitude, self-efficacy, interest and value status of the students in those schools. Even after socioeconomic level, attitude, self-efficacy, interest, and value variables are included in the model, there is still a significant difference in between- and within-school variability (Wald Z = 2.15, $p$ < .05). In this case, variables at the school level are included in the analysis.

### *Results for the Third Research-Problem*

Level 2 (school level) model is established to determine the predictors of the students' science scores related to school characteristics. In order to explain the difference between school averages in the model, level-1 variables which are socioeconomic level (SES), attitude (TUT), value (DEG), interest (ILG), self-efficacy (OZY) and school-level variables which are school type (TUR (private, state), teacher experience (OGR), and school district (BOL) are included in the model. The results of the analysis are presented in Table 5. Table 5 shows that there is a significant difference between the schools in terms of socioeconomic level, affective characteristics, type of school, teacher experience, and TEOG science achievement scores (Wald Z = 22.46, $p$ < .05). In this case, it is stated that the students' science scores vary between schools. To calculate variance change ($R^2$), between and within-group variances are compared as in the following equation for between groups: (60.37-15.56) / 60.37 = 0.742 or 74.2%. This result indicates that the socioeconomic level, attitude, self-efficacy, interest and value variables of individual level explain 74.2% of the variance between the schools. On the other hand, the coefficient $R^2$ for within-group variances: (308.99-208.01) / 308.99 = 0.327 or 32.7%.

Table 5. Level-2 Random Intercept Model Results

| **Fixed effect** | **Coefficient** | **Standard error** | **df** | **t** |
|---|---|---|---|---|
| Average science score | 7.29 | 4.83 | 23.08 | 15.17 |
| SES | **7.17*** | 0.66 | 723.50 | 10.90 |
| Attitude | **-3.14*** | 0.76 | 1039.39 | -4.11 |
| Value | 0.86 | 0.62 | 1029.22 | 1.38 |
| Interest | -0.33 | 0.70 | 1031.38 | -0.48 |
| Self-efficacy | **10.02*** | 0.65 | 1033.17 | 15.48 |
| School type | -1.25 | 3.67 | 18.63 | -0.34 |
| Teacher experience | -0.01 | 0.57 | 56.86 | -0.01 |
| School region | 0.32 | 0.69 | 42.30 | 0.46 |
| **Random effect** | **Variance** | **Standard error** | | **Wald Z** |
| Level-1 variance | 208.01 | 9.27 | | 22.46 |
| Level-2 variance | 15.56 | 7.15 | | 2.18 |

*$p$ < .01

This result shows that the student socioeconomic level, attitude, self-efficacy, interest and value variables constitute 32.7% of the school variability in the students' science scores. According to Table 5, socioeconomic level ($\beta_1$ = 7.17, $p$ < .05), students' attitudes towards science ($\beta_2$= -3.14, $p$ < .05) and self-efficacy perceptions of students towards science course ($\beta_5$ = 10.02, $p$ < .05) affect the students' science scores. However, the students' interest in science ($\beta_4$ = -0.33, $p$ > .05) and value to science ($\beta_3$ = 0.86, $p$ > .05) do not affect the students' science scores. When Table 5 is examined, it is seen that the school type ($\gamma_{01}$ = -1.25, $p$ > .05), teacher experience ($\gamma_{02}$ = -0.01, $p$ > .05), location of school ($\gamma_{03}$ = -0.32, $p$ > .05) variables do not affect the students' science scores at the school level.

The results of the multilevel analysis can be summarized in the following equation:

Science Scores = 73.29 + 7.17 (SED) – 3.14 (TUT) + 0.86 (DEG) – 0.33 (ILG) + 10.02 (OZY) – 1.25 (BOL) – 0.004 (TUR) + 0.32 (OGR) + $u_{0j} + \varepsilon_{ij}$

In summary, the socioeconomic level, attitude and self-efficacy variables have a significant effect on the students' TEOG science scores. The teacher experience, value, school location, interest, and school type do not have a significant effect on the students' TEOG science scores.

## DISCUSSION and CONCLUSION

In this study, the predictor variables for the 8th grade students' TEOG science scores which are the attitude towards the science, self-efficacy, the value of the science, the students' interest in the science, the student's socioeconomic status, school location, school type, and teacher experience were examined by multi-level regression analysis.

According to the results of the first research problem, there is a significant difference between the average achievement scores of the schools. 16.3% of this difference arises from the schools and 83.7% from the students. This finding aligns with the studies that examined the effect of school and student characteristics variables on student achievement. In these studies, it was expected that most of the variance in achievement will be explained by student characteristics (Odden, Borman & Fermanich, 2009).

In the second research problem, the characteristics of the students were examined to explain the achievement differences among the students and the schools participating in the TEOG exam. The effect of socioeconomic status, attitude, value, interest, self-efficacy variables on science scores were investigated. In the analysis, socioeconomic status, attitude and self-efficacy variables were found to have a statistically significant effect on science achievement, but interest and value variables do not have a statistically significant effect on science achievement. While the socioeconomic status and self-efficacy affected science achievement positively, the attitudes of the students towards science negatively affected the achievement. According to the findings of the analysis, 78.5% of the variance among the schools stems from the students' socioeconomic level, attitude, value, interest, and self-efficacy. In relation to self-efficacy, Atar and Atar (2012) found that students' self-efficacy was a statistical predictor of their science achievement. However, in the study of Akıllı (2015), it was concluded that the students' self-efficacy affected their achievements in a negative way. In another study, it was seen that the socioeconomic status of the students was one of the most important factors affecting the achievement (Öksüzler & Sürekçi, 2010). In addition, in his meta-analysis, Sarıer (2016) found that the most important factors affecting students' achievement were socioeconomic status and self-efficacy. However, Yavuz et al. (2016) stated that the effect of the average socioeconomic status of schools on mathematics achievement was not statistically significant. In our study, the students' socioeconomic status was investigated. The level (individual/group) of the variable included in the analysis also affects the results. The reason for the different findings among the research can stem from the differences between the statistical techniques applied, measurement tools, content, and exam types. In terms of attitude, similar to the results obtained in this study, Kılıç (2016) also concluded that the attitude variable has a negative effect on students' mathematics achievement. On the other hand, Şahin (2011) found that the attitude variable had no significant effect on students' SBS (Achievement level determination exam) science achievement. Regarding attitude, there are also studies showing different results from the findings of this study. For example, in his study, Akıllı (2015) found that the attitudes of 8th grade students predict the TIMSS science scores positively. Pektaş (2010) also stated that attitudes towards science, students' self-efficacy beliefs, the value given to science, and the education level of a family are significant predictors of TIMSS science achievement scores. There are studies in the literature supporting the findings that the value variable does not predict success (Yavuz, Demirtaşlı, Yalçın & Dibek, 2017). Regarding interest in science in some studies in the literature, it has been shown that the interest of students in science significantly predicts success in science (Singh, Mo & Chang, 2006). Obtaining different results from the literature may be due to different analysis methods. In this study, multilevel analysis was used. In multilevel analysis, the problems of aggregation and disaggregation are avoided, and the predictor variables are included in the model at appropriate levels. Therefore, different results may arise from single level analysis methods.

In the third research problem, the student and school characteristics that explain the difference between the students' science scores were examined simultaneously. According to the results, the characteristics of the students and the schools explained 32.7% of the between-school variability. It is found that the school type, the school region, and the teacher experience variables added in Level-2 did not significantly explain the students' science scores. These findings contradict some of the existing research. In one study, it was determined that the less experienced, novice teachers' students had higher scores for application and reasoning questions in TIMSS 2011 (Güner, Sezer & Akkuş-İspir, 2013). In another study, it is stated that teachers with more than five years of experience are more efficient (Greenwald, Hedges & Laine, 1996). While in the literature it was concluded that school type and region variables predicted success (Acar, 2013; Berberoğlu & Kalender, 2005; Karabay, Yıldırım & Güler, 2015), in this study, it was determined that these variables did not predict the students' science scores statistically. However, to investigate this conflicting finding in detail, the school type variable was included in the analysis alone without including the individual level students' characteristics. Then it was found that the school type is the predictor of the students' science scores. In other words, the school type is not the predictor variable of achievement, if it is included in the model with the student characteristics. This finding suggests that it is not the type of schools that matters, but the students who attend those schools. In terms of change in variance with school-level predictors, another interesting result has been observed. The variance between schools increased while it was expected to decrease when level-2 predictor variables are included in the regression model.

According to the findings of this study, the self-efficacy variable has a positive effect on science achievement. For this reason, it is suggested that studies should be conducted to increase the self-efficacy of the students towards the science course. In order to help students to develop self-efficacy, their strengths and positive aspects should be pointed out, emphasized, and supported in the teaching-learning process. In addition, it was determined that the socioeconomic levels of the students had a major significant effect on their achievement. The factors determining the socioeconomic level are parents' education and home resources. In order to increase the achievement of the students, it was determined that the family should be educated first. In Turkey, it may be necessary to follow the innovations in education and to update the education system accordingly to these developments in order to have a positive effect on science achievement. New studies can be done for students to be motivated to learn and understand the importance of science. For example, activities can be planned to show students the relationship of the science courses to real life. Awareness may be raised about the scientific events taking place in Turkey and in the world. Although the experience of the teachers did not have a significant effect on student achievement, there are studies in which teacher experience is determined as an important variable affecting success (Güner et al., 2013). In order to increase the positive effect of teachers on student achievement, new studies should be carried out for teachers who are novice in the profession and competent/experienced teachers in their fields. Teachers may be advised to organize activities for students to love science. The variables that affect the 8th grade students' TEOG science scores were investigated with the items selected from TIMSS 2011 questionnaires. The effect of other variables on achievement can be examined by using other variables from the TIMSS questionnaire. Since the findings of the study were limited to this group of participants, the study could be repeated with participants with different demographic characteristics. In this study, some of the variables that predict achievement differences between schools were determined. From this point of view, the question of what should be emphasized to increase students' science achievement has been answered relatively. However, the undisclosed difference between schools in this study is as high as 20%. In order to explain this ratio, studies that take into account other variables not considered in this study are needed.

## REFERENCES

Abazoğlu, İ., & Taşar, M. F. (2016). Fen bilgisi öğretmen özelliklerinin öğrenci fen başarısı ile ilişkisi: TIMSS 2011 verilerine göre bir durum analizi. *Elementary Education Online, 15*(3), 922-945.

Acar, M. (2013). *Öğrenci başarılarının belirlenmesi sınavında Türkçe dersi başarısının öğrenci ve okul özellikleri ile ilişkisinin hiyerarşik lineer model ile analizi* (Doktora tezi). Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Acar, T. (2009). An application of hierarchical linear modeling: OKS-2006 science test achievement. *Eurasian Journal of Educational Research*, *9*(37), 1-16.

Acar, T., & Öğretmen, T. (2012). Çok düzeyli istatistiksel yöntemler ile 2006 PISA fen bilimleri performansının incelenmesi. *Eğitim ve Bilim*, *37*(163), 178-189.

Akıllı, M. (2015). Regression levels of selected affective factors on science achievement: A structural equation model with TIMSS 2011 data. *Electronic Journal of Science Education*, *19*(1), 1-16.

Akyüz, G. (2006). Türkiye ve Avrupa birliği ülkelerinde öğretmen ve sınıf niteliklerinin matematik başarısına etkisinin incelenmesi. *Elementary Education Online*, *5*(2), 75-86.

Anıl, D. (2009). Factors effecting science achievement of science students in programme for international students' achievement (PISA) in Turkey. *Egitim ve Bilim*, *34*(152), 87-100. Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/view/594/74

Anıl, D. (2011). Türkiye'nin PISA 2006 fen bilimleri başarısını etkileyen faktörlerin yapısal eşitlik modeli ile incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri*, *11*(3), 1253-1266.

Atalmış, E. H., Avgın, S. S., Demir, P., & Yıldırım, B. (2016). Examination of science achievement in the 8th grade level in Turkey in terms of national and international exams depending upon various variables. *Journal of education and Practice, 7*(10), 152-162.

Atar, H. Y. (2014). Öğretmen niteliklerinin TIMSS 2011 fen başarısına çok düzeyli etkileri. *Eğitim ve Bilim*, *39*(172), 121-137.

Atar, H. Y., & Atar, B. (2012). Türk eğitim reformunun öğrencilerin TIMSS 2007 fen başarılarına etkisinin incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri*, *12*(4), 2621-2636.

Berberoğlu, G., Çelebi, Ö., Özdemir, E., Uysal, E., & Yayan, B. (2003). Üçüncü uluslararası matematik ve fen çalışmasında Türk öğrencilerin başarı düzeylerini etkileyen etmenler. *Eğitim Bilimleri ve Uygulama, 2*(3), 3-14. http://ebuline.com/pdfs/3sayi/ebu3_1.pdf adresinden elde edilmiştir.

Berberoğlu, G., & Kalender, İ. (2005). Öğrenci başarısının yıllara, okul türlerine, bölgelere göre incelenmesi. *Eğitim Bilimleri ve Uygulama*, *4*(7), 21-35.

Büyüköztürk, Ş. (2015). *Sosyal bilimler için veri analizi el kitabı*. Ankara: Pegem Akademi.

Büyüköztürk, Ş., Çakan, M., Tan, Ş., & Atar, H. Y. (2014). *TIMSS 2011 ulusal matematik ve fen raporu 8. sınıflar. TIMSS Uluslararası Matematik ve Fen Eğilimleri Araştırması*. Ankara: İşkur Matbaacılık.

Büyüköztürk, Ş., Şekercioğlu, G., & Çokluk, Ö. (2014). *Sosyal bilimler için çok değişkenli istatistik SPSS ve Lisrel uygulamaları.* Ankara: Pegem Akademi.

Doğan, E., & Demir, S. B. (2015). Examination of the relation between TEOG score and school success in terms of various variables. *Journal of Education and Training Studies*, *3*(5), 113-121.

Greenwald, R., Hedges, L., & Laine, R. (1996). The effect of school resources on student achievement. *Review of Educational Research*, *66*(3), 361-396.

Güner, N., Sezer, R., & Akkuş-İspir, O. (2013). İlköğretim ikinci kademe öğretmenlerinin TIMSS hakkındaki görüşleri. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, *1*(33), 11-29.

Heck, R. H., Thomas, S. L., & Tabata, L. N. (2010*). Multilevel and longitudinal modeling with IBM SPSS.* New York, NY: Taylor & Francis Group.

Hox, J. J. (2010). *Multilevel analysis techniques and applications*. Great Britain: Routledge.

Karabay, E., Yıldırım, A., & Güler, G. (2015). Yıllara göre PISA matematik okuryazarlığının öğrenci ve okul özellikleri ile ilişkisinin aşamalı doğrusal modeller ile analizi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, *1*(36), 137-151.

Kılıç, A. (2016). 8*. sınıf öğrencisinin matematik dersine karşı tutumu ile teog sınav sonuçları arasındaki ilişki* (Yüksek lisans tezi). Çağ Üniversitesi, Mersin.

Odden, A., Borman, G., & Fermanich, M. (2009). Assessing teacher, classroom, and school effects, including fiscal effects. *Peabody Journal of Education*, *79*(4), 4-32.

Öksüzler, O., & Sürekçi, D. (2010). Türkiy'de ilköğretimde başarıyı etkileyen faktörler: Bir sıralı lojit yaklaşımı. *Finans Politik & Ekonomik Yorumlar*, *47*(543), 79-90.

Ötken, Ş. (2012). *İlköğretim 7. SINIF SBS başarısını yordayan değişkenlerin belirlenmesi* (Yüksek lisans tezi). Hacettepe Üniversitesi, Ankara.

Pektaş, M. (2010). *Uluslararası matematik ve fen bilimleri eğilimleri çalışması (TIMSS) verilerine göre Türkiye örnekleminde fen bilimleri başarısını etkileyen bazı değişkenlerin incelenmesi* (Yüksek lisans tezi). Hacettepe Üniversitesi, Ankara.

Sarıer, Y. (2016). Türkiye'de öğrencilerin akademik başarısını etkileyen faktörler: Bir meta-analiz çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(3), 609-627. doi: 10.16986/HUJE.2016015868.

Singh, K., Mo, Y., & Chang, M. (2006, November). *Science achievement: Effect of self and engagement variables.* Paper presented at the APERA Conference, Hong Kong. Retrieved from http://edisdat.ied.edu.hk/pubarch/b15907314/full_paper/1672708960.pdf

Süer, N. (2014). *Öz-düzenleme becerilerinin TEOG sınavı üzerinde etkisi* (Yüksek lisans tezi). Yıldız Teknik Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.

Şahin, M. D. (2011). *İlköğretim 7. Sınıf öğrencilerinin seviye belirleme sınavı (SBS) 2010 fen ve teknoloji alt test başarılarına etki eden bazı faktörler* (Yüksek lisans tezi). Hacettepe Üniversitesi, Ankara.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston, MA: Pearson.

TUBİTAK. (2016). *MEB için "fen, teknoloji, mühendislik, matematik-fetemm modeli (STEM) ile eğitim"*. Kocaeli: Tübitak Bilgem TBAE.

Uçar, S., & Öztürk, D. (2010). TIMSS verileri kullanılarak Tayvan ve Türkiye'deki 8. sınıf öğrencilerinin fen başarısına etki eden faktörlerin belirlenmesi ve karşılaştırılması. *Ç.Ü. Sosyal Bilimler Enstitüsü Dergisi, 19*(2), 241-256.

Uzun, N. B., Gelbal, S., & Öğretmen, T. (2010). TIMSS-R fen başarısı ve duyuşsal özellikler arasındaki ilişkinin modellenmesi ve modelin cinsiyetler bakımından karşılaştırılması. *Kastamonu Eğitim Dergisi, 18*(2), 531-544.

Yavuz, H. Ç., Demirtaşlı, N. R., Yalçın, S., & Dibek, M. İ. (2017). Türk öğrencilerin TIMSS 2007 ve 2011 matematik başarısında öğrenci ve öğretmen özelliklerinin etkileri. *Eğitim ve Bilim*, *42*(189), 27-47.

Yavuz, S., Odabaş, M., & Özdemir, A. (2016). Öğrencilerin sosyoekonomik düzeylerinin TEOG matematik başarısına etkisi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 7*(1), 85-95.

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(4); 378-390

# The Effects of Log Data on Students' Performance

Hatice Çiğdem YAVUZ *

**Abstract**
This study aimed to assess the relationships between response times (RTs), the number of actions taken to solve a given item, and student performance. In addition, the interaction between the students' information and communications technology (ICT) competency, reading literacy, and log data (time and number of actions) were examined in order to gain additional insights regarding the relations between student performance and log data. The sample consisted of 2 348 students who participated in the triennial international large-scale assessment of the Programme for International Student Assessment (PISA). For the current study, 18 items in the one cluster of the 91st booklet were chosen. To achieve the aim of the study, explanatory item response modeling (EIRM) framework based on generalized linear mixed modeling (GLMM) was used. The results of this study showed that students who spent more time on items and those that took more actions on items were more likely to answer the items correctly. However, this effect did not have variability across items and students. Moreover, the interaction only with reading and the number of actions was found to have a positive effect on the students' overall performance.

*Key Words:* Test-taking behaviors, explanatory item response modeling, log data, technology-based assessment.

## INTRODUCTION

Depending on the stakes or context of the tests, students adapt different test-taking behaviors. To explore these behaviors, much research has been undertaken in psychometric practice. With the emerging utilization of technology in testing, it has become possible to analyze test-takers' behaviors in detail in relation to many psychometrical aspects. Considering the feasibility of administration of computerized assessments in education, computer-generated log-files are able to provide rich information in this context.

A student log file records all the data produced by the student during testing. Log files make it possible to see beyond students' overall performance by determining, for example, what actions have been undertaken, and how much time has been spent for a specific item. The information gathered in log files reveals a different perspective concerning students' performance and cognitive behaviors (Greiff, Wüstenberg & Avvisati, 2015). Moreover, log files can offer valuable feedback about students' learning and cognitive abilities (Greiff et al., 2014). Many recent studies have shown that students' log files provide validity evidence (e.g., Lee & Jia, 2014; Wise & DeMars, 2005), possible associations with student performance (Goldhammer et al., 2014; Greiff et al., 2015), and a better understanding on non-traditional competences (Azzolini, Bazoli, Lievore, Schizzerotto, & Vergolini, 2019).

In particular, from the students' log data, the response time (RT) has been the subject of many studies within the field of psychology and psychometrics (e.g., Goldhammer, Naumann & Greiff, 2015; Lee & Haberman, 2016). RT has been used to gain a better understanding of mental activity in psychology, and the utilization of RT is also on the rise in testing over the last few decades (Schnipke & Scrams, 2002). This is because time plays an important role in examining the process of answering items in detail. In this sense, RT has been examined as an indicator of test-taking motivation/engagement (Wise & DeMars, 2005), rapid-guessing behavior (Lee & Jia, 2014), or a characteristic of student performance (Goldhammer et al., 2014).

* Ph. D., Cukurova University, Faculty of Education, Adana-Turkey, hcyavuz@cu.edu.tr, ORCID ID: 0000-0003-2585-3686

To cite this article:
Yavuz, H. Ç. (2019). The effects of log data on students' performance. *Journal of Measurement and Evaluation in Education and Psychology, 10*(4), 378-390. doi: 10.21031/epod.564232

*Received: 13.05.2019*
*Accepted: 05.11.2019*

Previous studies in which RT was examined in the context of test-taking engagement have revealed that a lower RT can be interpreted as a validity thread (Wise Kingsbury, Thomason & Kong, 2004; Wise & DeMars, 2005; Rios, Guo, Mao & Liu, 2017). Together with this, most researchers consider RT as being associated with the cognitive ability of individuals (Kyllonen & Zu, 2016). Recent studies in testing propose that the relationship between student performance and RT changes depending on the features of items/tasks and students.

In their study, Goldhammer et al. (2014) examined the time effect in reading and problem solving using the items of the Programme for the International Assessment of Adult Competencies (PIAAC). They found that the time effect depended on item difficulty and test-takers' ability. In this sense, the time had a positive effect on problem-solving items while the opposite relationship was found for reading items. With a similar purpose, item RT was investigated using a computerized version of Raven's Advanced Progressive Matrices (RAPM) test (Goldhammer et al., 2015). According to the findings of the study, item RT had a negative effect on the overall performance of test-takers. However, this effect differed in that it was highly negative for easy items among higher-performing test-takers, but not high enough for difficult items and lower-performing test-takers. In another study (Greiff, Niepel, Scherer & Martin, 2016) using students' RT, it was revealed that spending an extremely low or high level of time led to lower performance in complex problem-solving. Lee and Haberman (2016) used RT to investigate test-taking behaviors in an international language assessment and found that the behaviors and RTs of examinees from different countries did not generally follow a stable trend. On the other hand, in their study, higher-performing examinees showed a more stable trend within each country in terms of RTs. In another study by Dodonova & Dodonov (2013), the relationship between cognitive ability and RT of individuals was examined using the RAPM test. The result of their research showed that higher-performing individuals had lower RTs than lower-performing individuals; however, this association changed in relation to more difficult items.

The aim of the current study was also to model RT as a characteristic of student performance and examine the effect of the number of actions taken to solve a given item using the Programme for International Student Assessment (PISA) 2015 data. In addition, the interaction between the students' information and communications technology (ICT) competency, reading literacy, and log data (time and number of actions) were examined in order to gain additional insights regarding the relations between student performance and log data. An only a limited number of studies considered the investigation of the interactions between log data and other possible indicators, such as reading ability or technological competencies which can have a role in shaping this data. Thus, to provide more information from students' log data, the current study aimed to assess the relationships between RTs, the actions taken to solve a given item, and student performance.

Considering the results of the above-mentioned research studies and the effort required to give correct answers to the items in PISA, it was assumed, in this study, that RT has a positive effect on the overall student performance. Therefore, it was expected that the more students spent time on items, the more their probability of answering items correctly would increase. Since spending less time on items is considered as rapid guessing and having lower levels of test engagement, it was also expected that students with higher ability would spend more time on items. Moreover, it was also assumed in the current study that RT increased depending on item difficulty regardless of students' ability, given the results of various studies (e.g., Goldhammer & Klein-Entink, 2011; Goldhammer et al., 2014; Klein-Entink, Fox & van der Linden, 2009) indicating that the difficulty of items had a moderating effect on performance. Moreover, students' reading ability can affect RT when answering items, since an item needs to be read before giving a response to the item. The interaction between reading performance and time will vary depending on the reading load of the items. However, in the current study, it was assumed that this interaction would have a negative effect on student performance. Apart from their reading ability and understanding, the student's RT also may be affected by the level of their ICT competencies since during the process of solving the item in computerized tests, such as PISA, students need to press buttons, drag and drop, and select lists (Organisation for Economic Cooperation and Development-OECD, 2017a). Thus, it was expected that students having a lower ability on ICT would spend more time on items, and it was assumed that the interaction between ICT competence and time would negatively affect overall student performance.

Although extensive research has been carried out on the relationship between RT and test-takers' ability, a limited number of research (He, von Davier, & Han, 2018; Herborn, Stadler, Mustafić & Greiff, 2018) was found in the literature regarding how the number of actions taken to solve a given item affect student performance. Since these studies were in the context of problem-solving behaviors, additional research can be undertaken to find associations between the number of actions taken by students while answering items during testing and students' overall performance. In this way, it would be possible to compare the effects of log data such as the number of actions in different types of assessments. For instance, unlike paper-pencil assessments, students needed to undertake several actions in order to answer the items in PISA 2015. Hence, it was expected that students engaging in more actions on items would have a positive effect on overall student performance. Moreover, it was also assumed that the number of actions increased depending on item difficulty regardless of the students' ability in this study. Moreover, students' ICT competencies might have affected the number of actions taken when answering items in PISA 2015. Students having higher ICT competence and taking more actions to answer to items might be able to solve problems better, but for those with lower ICT competence undertaking irrelevant actions would make no difference in answering the items correctly. Thus, in this study, it was assumed that this interaction between ICT competence and the number of actions would have positively affected student performance. Likewise, it was expected that the interaction between the number of actions and reading would have a positive effect. In this sense, the following four research questions were addressed:

1. Does time have a significant effect on overall student performance?
2. Does the interaction between reading, ICT competence, and time have a significant effect on overall student performance?
3. Does the number of actions have a significant effect on overall student performance?
4. Does the interaction between reading, ICT competence, and the number of actions have a significant effect on overall student performance?

## METHOD

The aim of this study was to investigate the effects of log-data on students' performance. To achieve this aim, explanatory item response modeling was used. RT and the number of actions were modeled as covariates. Sample, data collection instruments and data analysis are described in the following section.

### *Sample*

The sample consisted of students who participated in the triennial international large-scale assessment of PISA in 2015, which assesses the key knowledge and skills of 15-year-old students, focusing on reading, mathematics, and science literacy. PISA also uses questionnaires in order to obtain information regarding various aspects of students, schools, and countries. In PISA 2015, apart from students from schools in 15 countries unable to fulfill the technological requirements, all participants completed the tests and questionnaires via computer. Thus, students' log files were available in PISA 2015. In each cycle of PISA, one of the core domains is tested in detail, and in 2015, the major domain was science.

In order to avoid item position effects, 2 348 students who answered 27 items in the same order in the one cluster of the 91st test booklet, which was taken by the largest number of students, were chosen for this study. However, some students had to be excluded from the analysis due to not having completed/taken the ICT competency questionnaire (n = 635), having an extremely large number of actions or RTs (n = 147); therefore, the final sample consisted of 1 566 students (51% female; $\bar{X}_{age}$ = 15.78, $SD_{age}$ = 0.29).

### *Data Collection Instruments*

#### *Items*

In PISA 2015, the scientific literacy items focused on three competencies (explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically) (OECD, 2017b). In this cycle of PISA, some items required the completion of interactive tasks, meaning that students need to manipulate, variables in simulation given on items (OECD, 2017a). Each student first received two 30-minute booklets of science tasks and two 30-minute booklets for the other domains (OECD, 2017c).

Since the 91st booklet was taken by the largest number of students in PISA 2015, the items in the one cluster of this booklet were chosen for the current study. Of the items in this cluster, two polytomous items, one item not having the timing data, and six items having low item discrimination values were not included; therefore, only 18 science items were selected for the analyses. In this study, log data regarding response times and the number of actions of those items were included. Response time variable indicates how much time was spent answering each item and the number of actions variable indicates how many actions were taken to answer a given item by students (such as clicks, keypresses, and drag/drop events).

Reading literacy and ICT competence were also utilized as predictors in this study. Reading literacy is defined by OECD (2017b) as “understanding, using, reflecting on and engaging with written texts, in order to achieve one’s goals, develop one’s knowledge and potential, and participate in society” (p. 51). In PISA 2015, three aspects (access and retrieve, integrate and interpret, reflect and evaluate) were defined to assess reading literacy by using mixed response format items. Students’ perceived ICT competence was assessed by asking them several questions regarding their level of comfort in using various digital devices (OECD, 2017b). An index variable was calculated from these responses for each student in PISA 2015, and this index was used in the present study.

### *Data Analysis*

To achieve the aim of the study, explanatory item response modeling (EIRM) framework based on generalized linear mixed modeling (GLMM) (De Boeck et al., 2011; De Boeck & Wilson, 2004) was used. With this framework, properties of items and persons are modeled as explanatory covariates in order to explain individuals’ responses in a broader approach (Wilson, De Boeck & Carstensen, 2008). In the context of EIRM, responses are treated as repeated observations nested within students. Unlike traditional item response theory (IRT) models, EIRM allows including item- and person-level covariates in the measurement model to explain variances in the latent abilities of individuals. In the framework of GLMM, EIRM is the complex extension of the Rasch model (Rasch, 1960), “in which the clustering of item responses within respondents is a function of item-specific fixed effects and one person-specific random effect” (Briggs, 2008, p. 93). More detailed information about how GLMM is formulated as a Rasch model can be found in Rijmen, Tuerlinckx, De Boeck, and Kuppens (2003) and Briggs (2008).

In this study, RT and the number of actions were modeled as covariates separately. For data preparation, time-variable was initially log-transformed as suggested in the literature in order to obtain a better model fit (van der Linden, 2009). The number of actions, reading literacy, and ICT competence variables were also normalized. Outliers (147 students) were excluded from data analysis. After this process, the data was translated into the long format using the “reshape” package (Wickham, 2012) in R (R Development Core Team, 2018).

For the study, first, the data fit was examined for the Rasch model by obtaining related fit indices and checking other required assumptions. Since the Infit and Outfit indices for items ranged between 0.5 and 1.5 (De Ayala, 2009), the item fit was confirmed. For unidimensionality, the average RMSEA value was found to be .03 less than .05, indicating that the data was fitted to a one-factor model. When the local independence assumption was checked with Yen’s Q3 statistics, all residual correlations for all pairs of items were found to be below .20, indicating that item responses are independent in the data. These assumptions were examined using the “sirt” package (Robitzsch, 2019) in R. After the assumptions were met, explanatory IRT models were tested using the “lme4” package (Bates, Maechler & Bolker, 2012) in R. Within the approach of Goldhammer et al. (2014, 2015) and as described by Desjardins and Bulut (2018), all explanatory IRT models tested separately for time and action variables in this study are as follows:

Model 0: response ~ -1 + time/action + (1 | id) + (1 | item)

Model 1: response ~ -1 + time/action + (1 | id) + (1 + time/action | item)

Model 2: response ~ -1 + time/action + (1 + time/action | id) + (1 + time/action | item)

Model 3: response ~ -1 + time/action * reading + (1 | id) + (1 | item)

Model 4: response ~ -1 + time/action * ictcom + (1 | id) + (1 | item)

These models were compared using Akaike’s information criterion (AIC) and Bayesian information criterion (BIC) values.

## RESULTS

According to the results of the initial analysis, all items were fitted to the Rasch model, and the correlation between students’ abilities estimated using the selected items in this study and the performance scores obtained from PISA was found to be .91. The coefficient Alpha value was calculated as .81, meaning that the items had high internal consistency. The item statistics, item parameters, and fit statistics are given in Table 1.

Table 1. The Item Statistics, Item Parameters, And Fit Statistics

| Item | Item Difficulty | Item Discrimination | Item Easiness | Outfit | Infit |
|---|---|---|---|---|---|
| 1 | 0.43 | 0.41 | -0.33 | 1.15 | 1.11 |
| 2 | 0.44 | 0.49 | -0.29 | 1.08 | 0.99 |
| 3 | 0.61 | 0.53 | 0.56 | 0.89 | 0.93 |
| 4 | 0.55 | 0.40 | 0.24 | 1.21 | 1.12 |
| 5 | 0.58 | 0.41 | 0.41 | 1.12 | 1.20 |
| 6 | 0.53 | 0.51 | 0.14 | 0.96 | 0.98 |
| 7 | 0.70 | 0.43 | 1.08 | 0.99 | 1.04 |
| 8 | 0.54 | 0.51 | 0.19 | 0.94 | 0.97 |
| 9 | 0.40 | 0.47 | -0.52 | 1.02 | 1.02 |
| 10 | 0.70 | 0.52 | 1.06 | 0.91 | 0.90 |
| 11 | 0.41 | 0.50 | -0.47 | 0.96 | 0.98 |
| 12 | 0.30 | 0.50 | -1.02 | 0.89 | 0.94 |
| 13 | 0.84 | 0.39 | 2.05 | 0.86 | 0.93 |
| 14 | 0.53 | 0.58 | 0.16 | 0.86 | 0.89 |
| 15 | 0.56 | 0.46 | 0.32 | 1.05 | 1.04 |
| 16 | 0.49 | 0.49 | -0.04 | 1.05 | 1.01 |
| 17 | 0.68 | 0.49 | 0.93 | 0.89 | 0.98 |
| 18 | 0.50 | 0.61 | -0.01 | 0.79 | 0.85 |

Note: Item difficulty and discrimination were calculated based on classical test theory. Item easiness and item fit indices were obtained according to the Rasch model in the framework of GLMM.

As shown in Table 1, the easiness of the items ranged from -1.02 to 2.05, with the average difficulty being 0.24, which means that the items were of moderate difficulty overall. The results from EIRMs about RT are presented in Table 2, and EIRM related to the number of actions are given in Table 3.

Table 2. Results from EIRMs about RT

| **Predictor** | **Model 0** | | **Model 1** | | **Model 2** | | **Model 3** | | **Model 4** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | (SE) | Estimate | (SE) | Estimate | (SE) | Estimate | (SE) | Estimate | (SE) |
| Time | 0.04*** | .01 | 0.02 | .01 | 0.02 | .01 | 0.04*** | .01 | .04*** | .01 |
| Reading | | | | | | | 1.17*** | .17 | | |
| Time*Reading | | | | | | | -0.01 | .02 | | |
| ICT competency | | | | | | | | | 0.15 | .17 |
| Time* ICT competency | | | | | | | | | -0.01 | .02 |
| Var(Id) | 1.15 | | 1.11 | | | | 0.12 | | 1.15 | |
| Var(Item) | 0.56 | | 11.60 | | | | 0.57 | | 0.56 | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3. Results from EIRMs about the Number of Actions

| **Predictor** | **Model 0** | | **Model 1** | | **Model 2** | | **Model 3** | | **Model 4** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | (SE) | Estimate | (SE) | Estimate | (SE) | Estimate | (SE) | Estimate | (SE) |
| Action | 0.33*** | .02 | -0.15 | .26 | -0.18 | .48 | 0.25*** | .02 | 0.32*** | .02 |
| Reading | | | | | | | 0.99*** | .02 | | |
| Action*Reading | | | | | | | 0.07*** | .02 | | |
| ICT competency | | | | | | | | | 0.07* | .03 |
| Action* ICT competency | | | | | | | | | 0.02 | .02 |
| Var(Id) | 1.10 | | 1.09 | | 1.11 | | 0.13 | | 1.10 | |
| Var(Item) | 0.78 | | 0.54 | | 0.55 | | 0.74 | | 0.78 | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As can be seen in the tables given above, the overall effects of RT and the number of actions were statistically significant ($\beta_{time} = 0.04$, $\beta_{action} = 0.33$, $p < .001$). The positive effects indicated that students spending more time on items and those taking more actions on items were more likely to answer the items correctly. However, when RT and the number of actions were included as random effects in addition to being fixed effects, the estimated effects of these variables were not significant ($\beta_{time} = 0.02$, $\beta_{action} = -0.15$, $p > .05$). This finding shows that the effects of RT and the number of actions were not associated linearly with the abilities of students and difficulties of items. Thus, the results indicate that the variation of RT and the number of actions taken by higher performing students on easy or difficult items differed from those of lower-performing students on easy or difficult items. Thus, the variability of RT and the number of actions was unequal across items and students.

The models including interactions between log data and reading literacy and ICT competency showed that all interactions except the interaction between the number of actions taken and reading literacy were found to be a non-significant predictor. This finding shows that students' level of ICT competency did not differ depending on RT and the number of actions taken by students in order to answer the items correctly. However, students with higher reading literacy performance took a greater number of actions.

Table 4. Model Fit Indices of the EIRMs about RT

| **Model** | **AIC** | **BIC** | **Loglik** | **Chisquare** |
|---|---:|---:|---:|---:|
| Model 0 | 33240.0 | 33264.7 | -16617.0 | - |
| Model 1 | 33087.0 | 33128.1 | -16538 | 157.06 *** |
| Model 2 | 33058.6 | 33116.3 | -16522.3 | 32.313 *** |
| Model 3 | 31347.8 | 31388.9 | -15668.9 | 1896.20 *** |
| Model 4 | 33237.8 | 33278.9 | -16613.9 | 6.26 * |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Note: All other models were compared with Model 0

Table 5. Model Fit Indices of the EIRMs about the Number of Actions

| Model | AIC | BIC | Loglik | Chisquare |
|---|---|---|---|---|
| Model 0 | 33011.5 | 33036.2 | -16502.8 | - |
| Model 1 | 32966.3 | 33007.5 | -16478.2 | 49.16 *** |
| Model 2 | 32957.4 | 33015.0 | -16471.7 | 12.98 ** |
| Model 3 | 31169.6 | 31210.7 | -15579.8 | 1845.9 *** |
| Model 4 | 33008.2 | 33049.4 | -16499.1 | 7.31* |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Note: All other models were compared with Model 0

As seen in Tables 4 and 5, Model 3 showed the best fit in terms of AIC and BIC fit statistics. It should be noted that Model 1 having a related variable as a random effect on item level seems to fit the data better than other models.

## DISCUSSION and CONCLUSION

The aim of this study was to assess the relationships between RTs, the number of actions taken to solve a given item, and student performance. In addition, the interaction between the students' ICT competency, reading literacy, and log data (time and number of actions) were examined in order to gain additional insights regarding the relations between student performance and log data. The results of this study showed that students who spent more time on items and those that took more actions on items were more likely to answer the items correctly. However, this effect did not have variability across items and students.

In this study, it was assumed that RT and the number of actions had a positive effect on overall student performance. As hypothesized, the results revealed that students spending more time on items and those taking more actions on items were more likely to answer the items correctly. Moreover, it was also assumed that RTs depended on item difficulty and student ability in the study. Unexpectedly, this effect did not have variability across items and students, and broadly, this finding did not support the findings from other studies (Dodonova & Dodonov, 2013; Goldhammer & Klein-Entink, 2011; Goldhammer et al., 2015; Lasry, Watkins, Mazur & Ibrahim, 2013; Verbić & Tomić, 2009), which found a negative relationship between RT and abilities of individuals on a particular test. Furthermore, they found that RT varied significantly across items and individuals having a different level of abilities; however, since other studies investigated tests measuring cognitive skills, RTs may play a different role in those tests. This inconsistency may be due to the item structure used in PISA. The science items used in PISA have different features in terms of context than cognitive tests. Similarly, Lee and Haberman (2016), investigating RT as a pacing and speediness indicator using PISA data sets, found that the RTs of examinees from different counties were not following a stable trend in general. Similar to items in PISA that measure not a cognitive structure but something more like an achievement in a particular field, some studies (Klein-Entink et al., 2009) did not find a relationship between RT and student performance on Scholastic Aptitude Test (SAT). Hence, it may be concluded that item types and more specifically the aim of the test also affect RT. Another possible explanation for this could be the testing conditions (Lee & Jia, 2014). As Goldhammer et al. (2014) stated, "when collecting time information across tasks and individuals that are heterogeneous in difficulty and skill level, respectively, the role of time and its interpretation may differ" (p. 624) and the same finding occurred in this study. All the discussions undertaken concerning RT can be applied to the number of actions. However, further evidence is certainly needed to understand the effect of the number of actions on answering items. Given that all items were not released in PISA, future studies could use other types of items and tests in which they can examine item features in more detail while looking for an effect on RT and the number of actions.

In the present study, several effects of interactions were examined. It was assumed that the interaction between RT, reading, and ICT competence would have a negative effect on student performance. However, none were found to have a significant effect on student performance, and these results are likely to be related to previous findings. Given the non-uniform distribution of RTs among items and students, RTs of students having a higher reading ability or ICT competency would also have a similarly non-uniform distribution. The finding related to students' reading ability supports the work of Golhammer et al. (2014) and Petscher, Mitchell, and Foorman (2015). In the study by Petscher et al. (2015), the variability of RTs of students having higher reading ability showed more functional information compared to students with lower or moderate ability. On the contrary, Su and Davison (2019) found that students with high reading ability had lower RTs while answering the items correctly. However, since only science literacy items selected for this study, students' abilities could have played a different role in RTs on items. Moreover, this result may be due to the students' test-taking behaviors. Wise (2006) argued that students adopting rapid-guessing behavior spent less time on items, especially those with a high reading load. As Wu, Chen, and Stone (2018) stated, students' test-taking behavior is not a trait, but a reaction to that particular test, and students' RTs and other performances depend on test features. In this sense, non-significant interactions between those variables cannot be ascribed to the other assessments, and PISA can be classified as a low stake assessment. For that, future studies with similar purposes may use high stakes tests in order to explore those interaction effects.

In the current study, it was also expected that the interaction between the number of actions, reading competence, and ICT competence would have a positive effect on student performance. While the interaction with ICT and the number of actions did not have a significant effect on overall student performance, interaction with reading and the number of actions was found to have a positive effect on the students' overall performance. In this sense, it could be argued that ICT competence and the number of actions do not have a relationship in terms of students' likelihood of answering items correctly. The study by Lasry et al. (2013) demonstrated that students with lower confidence spent more time on items. Following the same logic, it was assumed that students' ICT competence could play a role in students' performance together with the number of actions they had taken. This result is likely to be related to the variation of those features among students with different levels of abilities. On the contrary, a positive interaction effect between the number of actions and reading was found in the current study. That is, the effect of the number of actions on the overall performance was higher in students who possessed the higher reading ability. This may be due to students with a high reading ability tending to take more actions by trying harder on items considering the high impact on the overall science performance of the students.

The present study proposes that the effect of time does not have a uniform trend across items and students. However, it should be noted that in this study, only a limited number of items were included in order to avoid possible item position effects; thus, the results and interpretations of this study may not cover all booklets used in PISA. Therefore, other types of research design should be implemented in the future to generalize these findings. Many other interaction effects could be included in order to explain the role of RT and the number of actions on students' performance, as explained variances found in the study suggest that there are further variables having a role in the students' log data and performance. Future studies can include other possible interactions to explain relationships between those variables. Furthermore, it would be interesting to test the role of RT and the number of actions with other IRT-based models. This could provide more detailed information to replicate this study, allowing for not only multiple-choice items but also constructed response items to be included.

## REFERENCES

Azzolini, D., Bazoli, N., Lievore, I., Schizzerotto, A., & Vergolini, L. (2019). *Beyond achievement. A comparative look into 15-year-olds' school engagement, effort and perseverance in the European Union.* Luxembourg: Publications Office of the European Union.

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes, 2011*[R package version 0.999375-42]. Retrieved from http://CRAN.R-project.org/package=lme4.

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education, 21*(2), 89-118. doi: 10.1080/08957340801926086

De Ayala, R. J. (2009). *The theory and practice of Item Response Theory*. New York, NY: The Guilford Press.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer. doi: 10.1007/978-1-4757-3990-9

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuer- linckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1-28. doi: 10.18637/jss.v039.i12

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. New York, NY: Chapman and Hall/CRC.

Dodonova Y. A., & Dodonov Y. S. (2013). Faster on easy items, more accurate on difficult ones: Cognitive ability and performance on a task of varying difficulty. *Intelligence, 41*(1), 1-10. doi: 10.1016/j.intell.2012.10.003

Goldhammer, F., & Klein-Entink, R. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence, 39*(2-3), 108-119. doi: 10.1016/j.intell.2011.02.001

Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence, 3*(1), 21-40. doi: 10.3390/jintelligence3010021

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608-626. doi: 10.1037/a0034716

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior, 61*(2016), 36-46. doi: 10.1016/j.chb.2016.02.095

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*(2015), 92-105. doi: 10.1016/j.compedu.2015.10.018

Greiff, S., Wüstenberg, S., Csapo, B., Demetriou, A., Hautamaki, J., Graesser, A. C., Martin, R. (2014). Domain-general problem-solving skills and education in the 21st century. *Educational Research Review, 13*(2014), 74-83. doi: 10.1016/j.edurev.2014.10.002

He, Q., von Davier, M., & Han, Z. (2018). Exploring process data in problem-solving items in computer-based large-scale assessments. In H. Jiao, W. R. Lissitz, & A. Van Wie (Eds.), *Data analytics and psychometrics: Informing assessment practices* (pp. 53-76). Charlotte, NC: Information Age Publishing.

Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2018). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*. Online first. doi: 10.1016/j.chb.2018.07.035

Klein-Entink, R. H., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika, 74*(1), 21-48. doi: 10.1007/s11336-008-9075-y

Kyllonen, P., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence, 4*(4), 1-29. doi: 10.3390/jintelligence4040014

Lasry, N., Watkins, J., Mazur, E., & Ibrahim, A. (2013). Response times to conceptual questions. *American Journal of Physics, 81*(9), 703-706. doi: 10.1119/1.4812583

Lee, Y. H., & Haberman, S. J. (2016) Investigating test-taking behaviors using timing and process data. *International Journal of Testing, 16*(3), 240-267, doi: 10.1080/15305058.2015.1085385

Lee, Y. H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education, 2*(8), 1-24. doi: 10.1186/s40536-014-0008-1

Organisation for Economic Cooperation and Development. (2017a). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving*. Paris: OECD Publishing. doi: 10.1787/9789264281820-en

Organisation for Economic Cooperation and Development. (2017b). *PISA 2015 results (Volume V): Collaborative problem solving*. Paris: OECD Publishing.

Organisation for Economic Cooperation and Development. (2017c). *PISA 2015 technical report*. Paris: OECD Publishing.

Petscher, Y., Mitchell, A. M., & Foorman, B. R. (2015). Improving the reliability of student scores from speeded assessments: An illustration of conditional item response theory using a computer-administered measure of vocabulary. *Reading and Writing, 28*(1), 31-56. doi: 10.1007/s11145-014-9518-z

R Development Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*(2), 185-205. doi: 10.1037/1082-989X.8.2.185

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing, 17*(1),74-104. doi: 10.1080/15305058.2016.1231193

Robitzsch, A. (2019). *Package "sirt".* Retrieved from https://cran.r-project.org/web/packages/sirt/sirt.pdf

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In N. C. Mills., M. T. Potenza, J. J. Fremer & C. W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237-266). New York, NY: Psychology Press.

Su, S., & Davison, M. L. (2019) Improving the predictive validity of reading comprehension using response times of correct item responses. *Applied Measurement in Education, 32*(2), 166-182. doi: 10.1080/08957347.2019.1577247

van der Linden, W. J. (2009). Conceptual issues in response- time modeling. *Journal of Educational Measurement, 46*(3), 247-272. doi: 10.1111/j.1745-3984.2009.00080.x

Verbić, S., & Tomić, B. (2009). *Test item response time and the response likelihood.* Retrieved from http://arxiv.org/ftp/arxiv/papers/0901/0901.4356.pdf

Wickham, H. (2012). *reshape2: Flexibly reshape data: a reboot of the reshape package*. Retrieved from http://cran.ms.unimelb.edu.au/web/packages/reshape2/

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J Hartig, E Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 91-120). Cambridge, MA: Hogrefe.

Wise, S. L. (2006). An Investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95-114. doi: 10.1207/s15324818ame1902_2

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1-17. doi: 10.1207/s15326977ea1001_1

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Wu, A. D., Chen, M. Y., & Stone, J. E. (2018). Investigating how test-takers change their strategies to handle difficulty in taking a reading comprehension test: Implications for score validation. *International Journal of Testing, 18*(3), 253-275. doi: 10.1080/15305058.2017.1396464

# Bilgisayar Ortamında Kaydedilen Madde Yanıtlama Verilerinin Öğrenci Performansına Etkisi

### *Giriş*

Testlerin amacına veya içeriğine bağlı olarak, öğrenciler farklı test yanıtlama davranışları benimsemektedirler. Bu davranışları incelemek adına, psikometri alanında birçok araştırma yürütülmüştür. Testlerin geliştirilmesinde ve uygulanmasında teknoloji kullanımımın artmasıyla birlikte öğrencilerin test yanıtlama davranışlarını daha detaylı bir şekilde incelemek mümkün olmuştur. Bu bağlamda, eğitimde bilgisayara dayalı ölçme uygulamalarının artmasıyla bilgisayar ortamında kaydedilen log dosyaları[1] (log files) zengin bilgi sağlamaktadır.

Bir öğrenciye ilişkin log dosyasına, öğrencinin bilgisayar ortamında testi yanıtlarken yaptığı tüm işlemler kaydedilmektedir. Log dosyalarında kaydedilmiş veriler *log verileri* adını almaktadır. Eğitim alanındaki log verileri de genellikle madde yanıtlama verilerini içermektedir. Log verileri öğrencilerin performansına ve bilişsel davranışlarına ilişkin farklı bakış açısı sunmaktadır (Greiff, Wüstenberg, & Avvisati, 2015). Yapılan çalışmalarda öğrenci log verileri, geçerlik kanıtı elde etme (Lee & Jia, 2014; Wise & DeMars, 2005), öğrenci performansıyla ilgili olası ilişkileri ortaya koyma (Goldhammer ve diğerleri, 2014; Greiff ve diğerleri, 2015) ve öğrencinin bilişsel olmayan yeterliklerini daha detaylı anlama (Azzolini, Bazoli, Lievore, Schizzerotto, & Vergolini, 2019) amacıyla kullanılmıştır.

[1] Çalışmada *log* olarak ifade edilen terimin Türkçe karşılığı olarak *günlük*, *kütük* veya *kayıt* terimlerine rastlanılmıştır. Bu terimler eğitim dışında diğer alanlara (örn., bilgisayar, yazılım) özgü olduğundan dolayı, bu çalışmada bu terimin söz konusu Türkçe karşılıkları kullanılmamıştır. Bu nedenle, Türkçe metinde *log files*, *log dosyaları* ve *log data*, *log veri* olarak kullanılmıştır. Ayrıca, çalışmada *log veri*, bilgisayar ortamında kaydedilen madde yanıtlama verileri olarak tanımlanmıştır.

Psikoloji ve psikometri alanında, öğrencilerin log verileri arasında en çok yanıtlama süresi odak noktası olmuştur (Goldhammer, Naumann, & Greiff, 2015; Lee & Haberman, 2016). Yanıtlama süresiyle ilgili olarak psikolojide bireylerin zihinsel aktivitelerini daha iyi anlama amacıyla araştırmalar yapılmıştır. Ayrıca psikometri alanında da yanıtlama süresinin kullanımı giderek önem kazanmaktadır (Schnipke & Scrams, 2002). Bunun nedeni, madde yanıtlama süresi, bireylerin maddeyi yanıtlama sürecine ilişkin detaylı bilgi sağlamaktadır. Bu bağlamda, yanıtlama süresi test yanıtlama motivasyonuna/bağlılığına (Wise & DeMars, 2005), hızlı-tahmin davranışına (Lee & Jia, 2014) ilişkin bir gösterge ya da öğrenci performansının karakteristik bir özelliği olarak incelenmiştir.

Bireylerin belirli bir alandaki performansları ile yanıtlama süresi arasındaki ilişkiyi inceleyen zengin bir alanyazın olmasına rağmen, bir maddeyi cevaplarken yapılan toplam eylem sayısının öğrenci performansını nasıl etkilediğine ilişkin sınırlı sayıda çalışmaya rastlanmıştır (He, von Davier, & Han, 2018; Herborn, Stadler, Mustafić, & Greiff, 2018). Söz konusu çalışmalar problem çözme alanında gerçekleştirildiğinden dolayı, yanıtlama süresi dışında madde düzeyinde tutulan diğer log verilerinin öğrenci performansıyla ilişkisini inceleyen araştırmalara ihtiyaç duyulmamaktadır. Böylelikle, farklı yapılardaki testlerde ne şekilde ve nasıl log veri toplanılması gerektiğine ilişkin bulgular elde edilebilir. Bununla birlikte, alanyazında log verileri ile diğer ilgili olabilecek değişkenlerin etkileşimlerinin araştırıldığı sınırlı sayıda araştırma bulunmaktadır. Bu nedenle, madde düzeyinde tutulan log verilerinden daha fazla bilgi edinmek amacıyla, bu çalışmada öğrencilerin performansıyla maddeyi yanıtlama süreleri ve maddeyi yanıtlarken yaptıkları eylem sayıları arasındaki ilişkinin Uluslararası Öğrenci Değerlendirme Programının (Programme for International Student Assessment-PISA) 2015 verileri kullanılarak incelenmesi amaçlanmıştır. Buna ek olarak, öğrencilerin ilgili log verileriyle okuduğunu anlama ve bilgi iletişim teknolojileri (BİT) yeterlikleri arasındaki etkileşim etkileri de incelenmiştir. Bu kapsamda, çalışmada şu sorulara yanıt aranmıştır:

1. Madde yanıt süresi öğrencinin genel performansı üzerinde manidar etkiye sahip midir?
2. Okuduğunu anlama, BİT yeterlikleri ile yanıtlama süresi arasındaki etkileşimler öğrencinin genel performansı üzerinde manidar etkiye sahip midir?
3. Yapılan eylem sayısı öğrencinin genel performansı üzerinde manidar etkiye sahip midir?
4. Okuduğunu anlama, BİT yeterlikleri ile eylem sayısı arasındaki etkileşimler öğrencinin genel performansı üzerinde manidar etkiye sahip midir?

## *Yöntem*

### *Örneklem*

Bu çalışmanın katılımcılarını her üç yılda gerçekleşen PISA 2015'teki katılımcıları oluşturmaktadır. Çalışmada, madde konum etkilerini (item position effects) önlemek için PISA'da fen okuryazarlığıyla ilgili olan 27 maddeyi aynı sırada yanıtlamış 2348 öğrenci seçilmiştir. Söz konusu maddeler en fazla öğrenci tarafından cevaplanan 91. test kitapçığının bir formundan seçilmiştir. Çalışmaya dâhil edilen öğrencilerden 635'i BİT yeterlik anketini almadığından, 147'si de log verilerinin uç değerlerde olması sebebiyle veri setinden çıkarılmıştır. Bu nedenle, çalışma 1566 (%51 kız, $\bar{X}_{yaş}$ = 15.78, $SS_{yaş}$ = 0.29) öğrenci verisi üzerinde gerçekleştirilmiştir.

### *Veri toplama araçları*

*Maddeler:* PISA 2015'te en fazla öğrenci tarafından cevaplanan kitapçık 91. test kitapçığı olduğundan dolayı, bu test kitapçığındaki bir formda yer alan fen okuryazarlığına ilişkin 27 madde seçilmiştir. Bu maddelerden iki tanesi çoklu puanlanan madde, altı tanesi düşük madde ayırt ediciliğine sahip olduğu ve bir tanesi de yanıtlama süresine ilişkin veriye sahip olmadığı için çalışmaya dâhil edilmemiştir. Böylelikle, çalışmadaki analizler toplam 18 madde kullanılarak yapılmıştır. Çalışmada değişken olarak, seçilen 18 maddeye ait yanıtlanma süreleri ve bu maddeler üzerinde yapılan eylem sayıları da seçilmiştir. Yanıtlama süresi, öğrencinin her bir madde üzerinde ne kadar süre harcadığını göstermektedir. Eylem sayısı ise öğrencinin her bir madde üzerinde ne kadar sayıda eylem gerçekleştirdiğini göstermektedir. PISA 2015'te tıklama, tuşlama, ekran üzerinde tutma veya çekme işlemlerinin tümü eylem sayısı olarak kaydedilmiştir.

*Okuduğunu anlama becerisi:* Öğrencilerin PISA 2015'teki okuduğunu anlama alanında yanıtladıkları maddelerden elde ettikleri başarı puanlarıdır. Ekonomik Kalkınma ve İşbirliği Örgütü bazen de İktisadi İşbirliği ve Gelişme Teşkilatı (Organisation for Economic Co-operation and Development-OECD) (2017) tarafından okuduğunu anlama becerileri bireyin yazılı metinleri kullanarak, üzerinde düşünerek, anlayarak amaçlarını gerçekleştirme, bilgisini ve potansiyelini geliştirme ve toplum içerisinde katılımına yönelik beceriler olarak tanımlamaktadır.

*BİT yeterliliği:* Öğrencilerin BİT yeterliliği, onların çok çeşitli dijital aletleri kullanım yeterliklerine ilişkin maddelerden alınan yanıtlarla ölçülmüştür (OECD, 2017b). PISA 2015'te öğrencilerin bu maddelere verdiği yanıtlardan indeks değişkeni geliştirilmiştir. Bu çalışmada da bu indeks değişkeni kullanılmıştır.

### *Verilerin analizi*

Verilerin analizinde genelleştirilmiş doğrusal karma model (generalized linear mixed modelling-GLMM) yöntemi kapsamındaki açımlayıcı madde tepki modeli (explanatory item response modelling-EIRM) (De Boeck ve diğerleri, 2011; De Boeck & Wilson, 2004) kullanılmıştır. Bu yöntem çerçevesinde, madde ve birey özellikleri, bireylerin yeteneklerini daha detaylı açıklama amacıyla açımlayıcı değişkenler (explanatory covariates) olarak modele alınabilmektedir (Wilson, De Boeck & Carstensen, 2008). EIRM'de maddeler bireylerden elde edilmiş tekrarlı ölçümler olarak modellenmektedir. EIRM'de, geleneksel madde tepki kuramı (MTK) modellerinin aksine madde ve birey düzeyinde yordayıcı değişkenler de eklenerek bireylerin örtük yeteneklerindeki varyans belirlenebilmektedir.

Bu çalışmada, öncelikli olarak söz konusu modeller için varsayımlar test edilmiştir. Verilerin Rasch modeline uyması için gereken uyum istatistik değerleri hesaplanmıştır. Uyum istatistiklerinin ve varsayımların gereken koşulları sağlamasından sonra, açımlayıcı madde tepki modelleri R programında "lme4" paketi (Bates, Maechler & Bolker, 2012) kullanılarak test edilmiştir. Goldhammer ve diğerlerinin (2014, 2015) yaklaşımı çerçevesinde, Desjardins ve Bulut'ta (2018) açıklandığı şekliyle, tüm açımlayıcı madde tepki modelleri yanıtlama süresi ve eylem sayısı için ayrı ayrı şu modeller kullanılmıştır:

Model 0: yanıt ~ -1 + zaman/eylem + (1 | birey) + (1 | madde)

Model 1: yanıt ~ -1 + zaman /eylem + (1 | birey) + (1 + zaman / eylem | madde)

Model 2: yanıt ~ -1 + zaman /eylem + (1 + zaman /eylem | birey) + (1 + zaman /eylem | madde)

Model 3: yanıt ~ -1 + zaman /eylem * okuma+ (1 | birey) + (1 | madde)

Model 4: yanıt ~ -1 + zaman /eylem * bit + (1 | birey) + (1 | madde)

## ***Sonuç ve Tartışma***

Çalışma kapsamında kurulan ilk modelin sonuçlarına göre madde üzerindeki yanıtlama süresinin ve yapılan eylem sayısının öğrencinin genel performansı üzerinde pozitif ve manidar bir etkiye sahip olduğu belirtilebilir ($\beta_{time}$ = 0.04, $\beta_{action}$ = 0.33, $p$ < .001). Bu bulgulara göre öğrencilerin bir madde üzerinde daha fazla zaman harcaması veya daha fazla eylemde bulunması onların maddeleri doğru yanıtlama olasılıklarını arttırmaktadır. Bununla birlikte, yanıtlama süresi ve eylem sayısının sabit etkilerine ek olarak tesadüfi etkilerine bakıldığında, kestirilen etkilerin manidar olmadığı tespit edilmiştir ($\beta_{time}$ = 0.02, $\beta_{action}$ = -0.15, $p$ > .05). Bu bulgu maddeyi cevaplamada uzun ya da kısa süre geçiren ve maddeyi cevaplarken daha fazla sayıda ya da az sayıda eylem yapan öğrencilerin yetenek düzeyleri arasında ilişkinin sabit olmadığını göstermektedir. Başka bir ifadeyle, örneğin yüksek yetenek düzeyinde olan öğrencilerin kolay ya da zor maddeler üzerinde geçirdiği sürenin değişkenliğinin diğer öğrencilere göre daha farklı olduğu belirtilebilir. Benzer şekilde, yüksek yetenek düzeyinde olan öğrencilerin kolay ya da zor maddeler üzerinde yaptığı toplam eylem sayısının değişkenliğinin diğer öğrencilere göre daha farklı olduğu belirtilebilir. Özetle, öğrencilerin madde üzerindeki yanıtlama süresi ve eylem sayısıyla, öğrenci performansları ve maddeler arasında doğrusal bir ilişki bulunmaktadır. Başka bir ifadeyle, yanıtlama süresi ve eylem sayısının değişkenliği maddeler ve öğrenciler arasında benzerlik göstermemektedir. Yanıtlama süresiyle ilgili olan bulgular, alanyazındaki bazı çalışmalarla paralellik göstermemektedir (Dodonova & Dodonov, 2013; Goldhammer ve diğerleri, 2015; Goldhammer & Klein-Entink, 2011; Lasry, Watkins, Mazur & Ibrahim, 2013; Verbić & Tomić, 2009). Bunun nedeni, bu çalışmalarda kullanılan testlerin daha çok bireylerin zihinsel becerilerini ölçmeye yönelik olması olabilir. Çünkü PISA testlerinin kullanıldığı bir başka çalışmada bu çalışmanın bulgusuna benzer bir bulguya ulaşılmıştır (Lee & Haberman 2016). Benzer şekilde, öğrencilerin belirli alanlardaki başarılarına odaklanan bir başka çalışmada da benzer sonuçlar elde edilmiştir (Klein-Entink, Fox & van der Linden, 2009). Buradan hareketle, yanıtlama süresinin farklı testlerde farklı rolleri üstlendiği ve farklı yorumlandığı belirtilebilir (Goldhammer ve diğerleri, 2014).

Çalışmada incelenen etkileşim etkilerinden sadece eylem sayısı ve okuduğunu anlama arasındaki etkileşimin pozitif yönde manidar etkisinin olduğu bulunmuştur. Okuduğunu anlama becerileri yüksek olan öğrencilerin madde üzerinde daha fazla eylemde bulunduğu belirtilebilir. Bu durum, okuduğunu anlama becerileri yüksek olan öğrencilerin genel olarak fen okuryazarlığında da başarılı olması ile açıklanabilir. Aynı zamanda, bu çalışmada öğrencilerin BİT yeterlikleriyle madde yanıtlama süreleri veya maddede yapılan eylem sayıları arasında herhangi bir ilişkinin olmadığı belirlenmiştir.

Bu çalışmanın bulguları sınırlılıkları çerçevesinde değerlendirilmelidir. Çalışmada, PISA 2015'te yer alan tek bir kitapçığın bir formundaki sınırlı sayıda madde ele alınmıştır. İleride yapılacak çalışmalar, madde konum etkilerini de göze alarak daha fazla sayıda kitapçık ve madde üzerinde yürütülebilir. Ayrıca bu çalışmada ele alınmayan diğer log verileri ile ilgili olabilecek değişkenlerle etkileşim etkileri incelebilir. Bunun yanında, sadece çoktan seçmeli maddeler yerine açık uçlu maddeler üzerinde de log verilerinin etkileri araştırılabilir.

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(4); 391-405

# The Effects of Student and School Level Characteristics on Academic Achievement of Middle School Students in Turkey *

Pınar KARAMAN ** Burcu ATAR ***

**Abstract**
The purpose of the study was to examine the student-level and school-level variability that affect middle school students' academic achievement. Student background and school context on student academic achievement were examined. Participants of the study consisted of 1053 seventh and eighth grade middle school students from 10 schools in the cities of Ankara and Sinop, Turkey. The research study analysed using two-level hierarchical linear modeling (HLM). Data were analysed with three HLM models: (1) random effects one-way ANOVA model, (2) random coefficients regression model, (3) intercepts and slopes-as outcomes model. The results of the analyses showed that at the student level, gender, SES, and number of siblings were found to have statistically significant effects on student GPA. When considering the practical importance of student level variables, SES, and number of siblings have small effects, but gender has a moderate effect on students' school achievements. On average, female students perform higher than male students in terms of their GPA scores. At the school level, educational school resources have a significant effect on predicting academic achievement. It has been shown that school resources have a moderate effect on students' academic achievements.

*Key Words:* Hierarchical linear modeling, academic achievement, student GPA, gender, SES, school resources.

## INTRODUCTION

Academic achievement is one of the most important determinants of education quality. Educational researchers agree that many factors have an impact on students' achievements (Börkan & Bakış, 2016; Coleman et al., 1966; Engin-Demir, 2009; Gelbal, 2008). To monitor the quality of education, educational assessment studies associated with academic achievement are taken into consideration in many countries. Therefore, studies related to the determinants of student achievement are dramatically increased over several decades. Student achievement depends on several factors, such as individual factors, family factors, school factors.

The research studies have shown that student characteristics such as gender, age, motivation, attitudes towards courses, self-efficacy, students' efforts, being bullied at school have significant impacts on academic achievement (Engin-Demir, 2009; Gevrek & Sieberlich, 2014; Ma, 2001; Özberk, Atalay-Kabasakal & Boztunç-Öztürk, 2017, Yavuz, Demirtaşlı, Yalçın, & İlgün-Dibek, 2017). Family background characteristics such as family socioeconomic status (SES), family size or number of children in the family, and parental education are related to educational achievement (Alacacı & Erbaş, 2010; Börkan & Balkış, 2016; Downey, 2001; Engin-Demir, 2009; Kalender & Berberoglu, 2009; Ministry of National Education-MoNE, 2007). The students whose families have a lower status, a lower level of education, and a bigger size are more likely to have lower academic performance in schools (Gamboa & Waltenberg, 2012; Willms, 1996). On the other hand, some students with low SES are able to show much higher academic performance than their peers with high SES (Erberber et al., 2015; Organisation for Economic Co-operation and Development-OECD, 2011; Özberk et al., 2017). These students are

* This study was presented at the International Congress of Eurasian Educational Researches held in Ankara on June 19–22, 2019.

** Assist. Prof., Sinop University, Faculty of Education, Sinop-Turkey, pkaraman1626@gmail.com, ORCID ID: 0000-0002-2218-2701
*** Assoc. Prof., Hacettepe University, Faculty of Education, Ankara-Turkey, burcugulum@gmail.com, ORCID ID: 0000-0003-3527-686X

To cite this article:
Karaman, P. & Atar, B. (2019). The effects of student and school level characteristics on academic achievement of middle school students in Turkey. *Journal of Measurement and Evaluation in Education and Psychology, 10*(4), 391-405. doi: 10.21031/epod.564819

*Received: 14.05.2019*
*Accepted: 28.10.2019*

called as academically resilient students. Research studies have shown that family characteristics are strong effects on student achievement whereas school characteristics have weak effects (Baker, Goesling, & Letendre, 2002; Brooks-Gunn and Duncan, 1997; Coleman et al., 1966; Heyneman & Loxley, 1983). However, there has been considerable debate on whether school characteristics have a significant effect on student outcomes (Chevalier & Lanot, 2002; Hanushek, 1997). Several research implied that in some contexts, school resources and teacher characteristics have a significant impact on student achievement (Atar, 2014; Bilican-Demir, 2018; Darling-Hammond, 2000; Glewwe, Kremer, Moulin & Zitzewitz, 2004; Leon & Valdivia, 2015; Phan, 2008; Sweetland & Hoy, 2000; Tavşancıl & Yalçın, 2015; Yavuz et al., 2017). School characteristics, especially in developing countries, determine the school quality. To examine school effects, different strategies can be used in the studies such as student-teacher ratio, school size, class size, instructional materials, teacher quality, school resources (libraries, labs, computers, etc.) (Leon & Valdivia, 2015; Willms & Somers, 2001). The results indicated that schools with better physical facilities (e.g., libraries, labs, textbooks) and qualified teachers, especially for developing countries, contribute positively to increase student achievement (Alacacı & Erbaş, 2010; Baker et al., 2002).

### *Assessment of Student Achievement*

Several methods can be used to assess student achievement. Final grades or grade point average (GPA) are generally used for students' achievements at school. On the other hand, standardized achievement tests are also used to assess student achievement (Petrill & Wilkerson, 2000). International educational large-scale assessments such as The Trends in International Mathematics and Science Study (TIMSS), Programme for International Student Assessment (PISA), and Progress in International Reading Literacy Study (PIRLS) and national large-scale assessments are generally used to evaluate student achievement. Numerous studies have been conducted in Turkey to examine student achievement on TIMSS, PISA, or PIRLS data (Akyüz, 2014; Alacacı & Erbaş, 2010; Anıl, 2009; Atar, 2014; Atar & Atar, 2012; Dincer & Uysal, 2010; Özberk et al., 2017; Özdemir, 2016; Yalçın, Demirtaşlı, İlgün-Dibek, & Yavuz, 2017). However, a few studies conducted in Turkey to examine student academic achievement on national large scale assessment such as Placement Test Results (SBS), Student Achievement Determination Exam (ÖBBS), Transition from Primary to Secondary education (TEOG) or on students' GPA in schools (Börkan & Bakış, 2016; Çiftçi, 2015; Engin-Demir, 2009; Gelbal, 2008; Yavuz, Tan & Atar, 2019).

The literature showed that academic achievement and its relationship with student characteristics and school characteristics is one of the enduring issues. Student characteristics such as gender, SES, number of siblings were examined in the study since these variables are mostly used contextual variables and likely to influence educational achievement. To determine whether school characteristics make a difference in student achievement, three categories (school size, student-teacher ratio, school resources) were measured. Therefore, the aim of the study was to provide empirical evidence on the relationship between student and school characteristics and student GPA in Turkey. Multilevel modeling was used to assess these factors on student achievement. Four research questions were investigated in the study:

1. How much do schools differ in their mean academic achievements?
2. How much do schools differ regarding the association between student level variables (i.e., gender, SES, number of siblings) and academic achievement?
3. Are school level variables (school size, student-teacher ratio, school resources) significant predictors of mean academic achievement?
4. Are school level variables (school size, student-teacher ratio, school resources) significant predictors of within school associations?

## METHOD

### *Sample and Data*

The study group included 1053 Grade 7 and Grade 8 students from 10 public middle schools in the cities of Ankara and Sinop, Turkey. A typical case sampling method was used to represent the average of middle school students in the province of Ankara and Sinop (Büyüköztürk, Çakmak, Akgün, Karadeniz & Demirel, 2008). The participants consisted of 512 females (48.6%) and 541 males (51.4%). The average age was 13.46 years, and age range was between 12-15.

### *Data Collection Instrument*

1053 middle school students in 10 schools have completed survey questions which including only demographic questions. Several demographic questions (gender, SES, number of siblings) were asked to the students in the survey. While some of the variables were categorical, some others were continuous. Variables that are thought to affect student achievement were determined. Gender, SES, and the number of siblings were assigned as student level variables. School size, student-teacher ratio, and educational resources were assigned as school level variables. School level variables were obtained from the Ministry of National Education (MEB) e-school system. Students' GPA as composite achievement scores were obtained from school administrative records. In schools, teacher-based exams are applied to students and GPA affects students' high school placement results.

Students' GPA scores were included as a continuous dependent variable in the HLM analyses. Since gender is a dummy variable, female students were coded as 1, and male students were coded as 2. SES was measured with parental income. Students were asked to provide information about their family's SES in the survey. SES was ranged from lower to upper as low SES, lower-middle SES, middle SES, upper-middle SES, and high SES. This variable was coded as *low* = 1, *lower-middle* = 2, *middle* = 3, *upper-middle* = 4, and *high* = 5. Educational resources (e.g. music room, art room, computer lab, science lab, library, conference room, atelier, sports room) in schools were examined. Scoring school resources was ranged from the highest score (8) to the lowest score (1). Schools' scores between 7-8 score, 5-6 score, 3-4 score, and 1-2 score were categorized as *a lot* (4), *some* (3), *little* (2), and *very little* (1), respectively. Therefore, SES and educational resources have been considered as ordinal variables. The number of siblings, school size, and student-teacher ratio were continuous variables in the study. School size was measured by the number of students per school. The student level and school level variables have shown in Table 1. The mean values of categorical variables such as gender, SES, and educational resources represent the proportion of frequency of these variables in Table 1.

Table 1. Descriptive Statistics for Variables

| Variables | N | Mean | Sd |
|---|---:|---:|---:|
| **Student level** | | | |
| Gender | 1053 | 1.51 | 0.50 |
| SES | 1053 | 3.36 | 0.76 |
| Number of Siblings | 1053 | 2.34 | 0.96 |
| **School level** | | | |
| School Size | 10 | 492.30 | 181.37 |
| Student-teacher ratio | 10 | 13.40 | 1.77 |
| Educational resources | 10 | 2.70 | 0.82 |
| Outcome variable (GPA) | 1053 | 83.94 | 12.10 |

### *Design of the Study*

This study aimed to examine the effects of variables at the student level and school level on middle school students' academic achievement in Turkish public schools. Due to the nested nature of data, the

Hierarchical Linear Modeling methodology was used in the present study. Conducting HLM analysis for nested structure of data helps to prevent making a Type I error and biased results (Gill, 2003; Osborne, 2000; Raudenbush & Bryk, 2002). HLM helps to determine the direct effects of variables at individual level and student level (Hox, 1995). For HLM analysis, adequate sample sizes must be obtained. There are several suggestions about the number of groups required for multilevel model (MLM) studies. The minimum cluster size of 20 (Tabachnick & Fidell, 2014), cluster size of 30 (Kreft, 1996), or even cluster size of 50 (Hox, 1998, 2010) is recommended in MLM studies. Moreover, the simulation studies advise that multilevel model should not be used if the number of clusters less than 10 (McNeish & Stapleton, 2016; Snijders & Bosker, 1993). When using small sample size for MLM studies, restricted maximum likelihood or Kenward-Roger adjustment is recommended to reduce biased estimates (Boedeker, 2017; McNeish & Stapleton, 2016). In this study, maximum and minimum number of students in schools was 235 and 68, respectively. Two-level models are analyzed using restricted maximum likelihood estimation by default in HLM 7 software (Raudenbush, Bryk, Cheong, Congdon & du Toit, 2011).

### *Data Analysis*

For HLM analysis, the two-level model was applied that student level was at the first level, and school level was at the second level. Student variables as the lowest level of the hierarchy are nested within schools (level 2). Analyzing the level 1 (student level) and level 2 (school level) regression relationship helps to determine the relationship between the predictors and outcome variables (Woltman, Feldstain, MacKay & Rocchi, 2012). Each level in the hierarchical structure has its own sub-model that explains the relationships among the variables. The student level factors in the HLM analyses included gender, SES, and family size (number of siblings). School level factors were school size, student-teacher ratio, and educational recourses. Before the analysis, the assumptions of HLM were checked. The normality of error terms (level 1 residuals and level 2 residuals) was assessed (Raudenbush et al., 2011). QQ plots showed that the residuals are normally distributed.

The HLM modelling consisted of three steps. In the first step, null (unconditional) model with random effects ANOVA model was created with only student level outcome variable but not included predictors at student level and school level. It gives the proportion of variance in middle school students' academic achievement among schools. The variance of students' GPA scores was analyzed at the individual level and also at school level. Student level variables were centered around their group means, and school level variables were centered around their grand means in the HLM analysis. Centering can help the interpretation of the model intercepts easily by transforming these scores (Enders & Tofighi, 2007; Raudenbush & Bryk, 2002).

#### *Random effects one-way anova model*

Equations for random effects Anova model regarding this study are as follows:

Level 1 Model (Student Level): $Y_{ij} = \beta_{0j} + r_{ij}$

Level 2 Model (School Level): $\beta_{0j} = \gamma_{00} + u_{0j}$

In student level model, $Y_{ij}$ refers to GPA of student *i* in school *j*. $\beta_{0j}$ refers to the mean of student GPA in school $j$, and $r_{ij}$ refers to deviation of student GPA in school $j$ from mean student GPA of school $j$. $\gamma_{00}$ is the grand mean of student *GPA of* $j$ schools, *and* $u_{0j}$ is the deviation of the mean of student GPA of school $j$ from grand mean of student GPA.

#### *Random coefficient regression model*

In the model, the independent variables (gender, SES, number of siblings) were examined to determine whether they have a significant effect on students' GPA, on average. Equations for random coefficient regression model are as follows:

Level 1 model:

$Y_{ij} = \beta_{0j} + \beta_{1j}(gender_{ij}) + \beta_{2j}(SES_{ij}) + \beta_{3j}(number\ of\ sibling_{ij}) + r_{ij}$

Level 2 model*:*

$\beta_{0j} = \gamma_{00} + u_{0j}$

$\beta_{1j} = \gamma_{10} + u_{1j}$

$\beta_{2j} = \gamma_{20} + u_{2j}$

$\beta_{3j} = \gamma_{30} + u_{3j}$

*Intercepts and slopes-as outcomes model*

Intercept and slope coefficients are outcomes in the model. This model also called as full model since both student level and school level variables were included. Equations for intercepts and slopes-as outcomes model regarding this study are as follows:

Level 1 model:

$Y_{ij} = \beta_{0j} + \beta_{1j}(gender_{ij}) + \beta_{2j}(SES_{ij}) + \beta_{3j}(number\ of\ sibling_{ij}) + r_{ij}$

Level 2 model:

$\beta_{0j} = \gamma_{00} + \gamma_{01}(schoolsize) + \gamma_{02}(student - teacher\ ratio) + \gamma_{03}(school\ recources\ ) + u_{0j}$

$\beta_{1j} = \gamma_{10} + u_{1j}$

$\beta_{2j} = \gamma_{20} + u_{2j}$

$\beta_{3j} = \gamma_{30} + u_{3j}$

## RESULTS

### *Results of The First Research Question (How much do schools differ in their mean academic achievements?):*

The random-effects Anova model determines whether there is enough school variance to justify the use of multilevel analysis for data set. None of the predictors at level 1 and level 2 here are included in the null (unconditional) model. The result of the one way ANOVA with random effects were presented in Table 2.

Table 2. Estimation of Fixed Effect on Anova Model

| Fixed Effect | Coefficient | Standard Error | t ratio | df |
|---|---|---|---|---|
| Average GPA,, $\gamma_{00}$ | 83.07 | 1.52 | 57.59** | 9 |

** *p* < .001

Table 3. Estimation of Random Effects Anova Model

| Random effect | Variance | χ2 | df |
|---|---|---|---|
| School level, $u_{0j}$ | 21.54 | 116.07** | 9 |
| Level 1 effect, $r_{ij}$ | 133.67 | | |

** *p* < .001

According to Table 2, overall school mean was 83.07 with 1.52 standard error. And in Table 3, the within-school variance was estimated as 133.67. The between-school variance was estimated as 21.54. The results showed that school level variance was statistically significant ($\chi^2_{(9)}$ = 116.07, $p$ < .001). Indicating that mean student GPA was significantly varied among schools. The null model also provides the estimate of the intraclass correlation coefficient. The intraclass correlation coefficient (ICC) was calculated to indicate the proportion of variance in student GPA among schools. The intraclass correlation was calculated as $\rho\rho = \tau_{00} \ / \ (\tau_{00} \ + \sigma^2)$ = 21.54 / (21.54 + 133.66) = .14 which indicated that 14% of total variance in student GPA was accounted for by differences among schools. 86% of the variability in student GPA resulted from the within-school variance. It has been found that estimated ICC value was larger than threshold of 5% (Bliese, 2000). The result suggested that HLM analysis is necessary for the nested data.

***Results of the Second Research Question (How much do schools differ regarding the association between student level variables (i.e., gender, SES, number of siblings) and academic achievement?):***

Table 4 and Table 5 showed that the results obtained from the random coefficient model analysis.

Table 4. Estimation of Fixed Effects on Random Coefficient Model

| Fixed effect | Coefficient | Standard Error | t-ratio | df | Effect size |
|---|---|---|---|---|---|
| Average GPA, $\gamma_{00}$ | 83.07 | 1.43 | 57.84** | 9 | |
| Gender, $\gamma_{10}$ | -4.82 | 1.17 | -4.09* | 9 | .43 |
| SES, $\gamma_{20}$ | 1.08 | 0.44 | 2.42* | 9 | .10 |
| Number of Sibling, $\gamma_{30}$ | -1.28 | 0.47 | -2.74* | 9 | .11 |

**$p$ < .001; *$p$<.05

Table 5. Estimation of Variance Components on The Random Coefficient Model

| Random effect | Variance | Standard Deviation | χ2 | df |
|---|---|---|---|---|
| School level, $u_{0j}$ | 21.46 | 4.63 | 121.48** | 9 |
| Level 1 effect, $r_{ij}$ | 124.94 | 11.17 | | |

** $p$ < .001

The findings indicated that the mean effects of the gender, SES, and number of siblings on student GPA were statistically significant. The independent variables had a significant effect on students' GPA scores at the student level. The mean slope values associated with the independent variables were estimated as -4.82, 1.08, -1.28, respectively. Negative coefficient value for gender suggests that on average, female students' GPA scores were about five points higher than male students when holding other variables constant ($\gamma_{10}$= -4.82). And also on average, one unit increase in number of siblings, student GPA score decreased one point when controlling all other variables ($\gamma_{30}$ = -1.28). It indicated that number of siblings was negatively correlated with student GPA score. On the other hand, SES positively contributed to students' GPA scores ($\gamma_{20}$ = 1.08). The effect size of each variable was also estimated to interpret the practical significance of variables (Kelley & Preacher, 2012). The effect size of each variable was estimated as .43, .10, and .11, respectively. Female students' GPA on average is 0.43 standard deviation higher than that of male students. It means that gender variable has moderate effect on student GPA. On the other hand, SES and number of siblings variables on academic achievement have a small effect (Cohen, 1992).

After student level variables were added to the model, within-school variance was reduced from 133.67 to 124.94. The results suggested that these variables in students' GPA scores explain only 7% of within-school variability ($r^2$= .07).

### *Results of the Third Research Question (Are school level variables (school size, student-teacher ratio, school resources) significant predictors of mean academic achievement?)*

The results of the intercepts and slopes as outcomes model for fixed effects were presented in Table 6.

Table 6. Results of The Fixed Effect in the Full Model

| **Fixed effect** | **Coefficient** | **Standard Error** | **t-ratio** | **df** | **Effect Size** |
|---|---|---|---|---|---|
| Intercept (GPA), $\gamma_{00}$ | 83.03 | 1.24 | 66.65** | 9 | |
| **Student level** | | | | | |
| Gender, $\gamma_{10}$ | -4.66 | 1.04 | -4.44* | 9 | -.40 |
| SES, $\gamma_{20}$ | 1.07 | 0.44 | 2.39* | 9 | .09 |
| Number of Sibling, $\gamma_{30}$ | -1.25 | 0.48 | -2.60* | 9 | -.10 |
| **School level** | | | | | |
| School size, $\gamma_{01}$ | 0.003 | 0.005 | 0.67 | 6 | |
| Student-teacher ratio, $\gamma_{02}$ | -0.79 | 0.36 | -2.19 | 6 | |
| School resources, $\gamma_{03}$ | 3.11 | 1.00 | 3.09* | 6 | .27 |

** $p < .001$; *$p < .05$

At the student level, gender, SES, and the number of siblings were found to have a significant impact on student GPA. The coefficient values of independent variables were estimated to be -4.66, 1.07, and -1.25, respectively. Negative coefficient value for gender suggests that on average, female students' GPA scores were about five points higher than male students when holding other variables constant ($\gamma_{10}$= -4.66). And also on average, one unit increase in number of siblings, student GPA score decreased one point when controlling all other variables ($\gamma_{30}$= -1.25). It indicated that number of siblings was negatively correlated with student GPA score. On the other hand, SES positively contributed to students' GPA scores. At the school level, only school resources found to have statistically significant effect on mean academic achievement (p = 0.021). It suggested that school educational resources were positively related to students' academic performance. And also the effect sizes of the variables at student level and school level were estimated. Effect sizes for student variables were found -0.40, 0.09, and -0.10, respectively. While gender variable had medium effect on student GPA, SES and number of siblings variables had small effect on student GPA. At the school level, effect size of school resources indicated that an increase of one standard deviation in school resources would result in an increase of 0.27 standard deviation in the school mean student GPA. It showed that school resources had approximately medium effect on academic achievement.

### *Results of the Fourth Research Question (Are school level variables (school size, student-teacher ratio, school resources) significant predictors of within school associations?)*

The results of the intercepts and slopes as outcomes model for random effects were presented in Table 7.

Table 7. Estimation of Variance Components on the Full Model

| **Random effect** | **Variance** | **Standard Deviation** | $\chi^2$ | **df** |
|---|---|---|---|---|
| School level, $u_{0j}$ | 19.23 | 4.38 | 122.92** | 6 |
| Level 1 effect, $r_{ij}$ | 124.96 | 11.17 | | 9 |

**$p < .001$

According to Table 7, adding student level and school level variables to the null model decreased school variability from 21.54 to 19.23. This finding indicated that school level variables explained 11% of the between-school variability in students' GPA scores. And also student variance in the full model decreased between from 133.67 to 124.96. It showed that student level variables explained 7% of the within-school variability in students' GPA scores. In comparison with the null model, final model explained approximately 7% of the variance at the student level, and 11% of the variance at the school level and remaining variability is still statistically significant (*p* < .001).

## DISCUSSION and CONCLUSION

This study empirically investigated the effects of student characteristics and school characteristics on the academic achievement of middle school students in Turkey. The findings indicated that student characteristics including gender, SES, and the number of siblings have significant effects on academic achievement. Student variables explained 7% variance in academic achievement. Gender has strongly significant effect on student academic achievement. Female students had higher average GPA scores than male students after controlling other variables. This finding is consistent with several studies (Börkan & Bakış, 2016; Dayioğlu & Türüt-Aşık, 2007; Engin-Demir, 2009; Ferreira & Gignoux, 2010; Gevrek & Seiberlich, 2014; Güvendir, 2014; Van Houtte, 2004). For example, Engin-Demir (2009) studied with sixth, seventh, and eighth grade students to investigate factors influencing their academic success by using their GPA. This study found that gender is the most important factor among student characteristics. On average, female students had higher achievement scores than male students in that study. Dayioğlu and Türüt-Asık (2007) examined the gender gap in academic performance for undergraduate students. They found that female students outperform male students in cumulative GPA, but the gender gap in university entrance exam scores was in reverse. Several reasons may explain why female students outperform male students in schools. Their attitudes and self-efficacy toward school, sense of school belongings, academic motivation, their efforts toward courses influence female and male students' academic achievement differently (Batyra, 2017; Engin-Demir, 2009; Gevrek & Seiberlich, 2014; Johnson, Crosnoe & Elder 2001; OECD, 2016; Van Houtte, 2004; Veenstra & Kuyper, 2004). Besides, gender equity for school achievement is very important. Turkey has made great efforts to advance gender equity since 2000. Since school enrollment, especially for females, has increased in primary and secondary education, gender differences in academic achievement are disappearing progressively in Turkey. The result of the present study may also show the positive effects of projects related to gender equity in schools throughout Turkey (The United Nations Children's Fund-UNICEF,2016). On the other hand, female students tend to show lower performance than male students in some subjects, especially in science and maths (Atar & Atar, 2012; Berberoğlu, 2004; Chiu & Xihua, 2008; Farkas, Sheehan, & Grobe, 1990; Wößmann, 2003). Literature generally showed that gender differences exist in academic performance of students all around the world. Therefore, more research is needed to examine gender gap in academic achievement for gender equity in education.

Although effect sizes are small, the effects of the number of siblings and SES on academic achievement were significant. It was found that low SES students are more likely to get a lower GPA. Similarly, vast majority of research revealed that the students living in a low socio-economic status family show poorly performance in schools (Alacacı & Erbaş, 2010; Atar & Atar, 2012; Aypay, Erdogan, & Sozer, 2007; Bellibas, 2016; Dincer & Uysal, 2010; Flores, 2007; Gelbal, 2008; Kalaycıoğlu, 2015; Ma & Klinger, 2000; Perry & McConney, 2010; Sirin, 2005; Smits & Hosgör, 2006). Sirin (2005) used meta-analysis to examine the family effects on academic achievement. The results showed that socioeconomic structure has a medium to strong impact on academic achievement. The author suggested that to prevent overestimating the effects of SES using multiple components of SES (e.g. income, education, and occupation) is important. The present study also showed the negative siblings effects on academic achievement. Especially in developing countries and western countries, a negative relationship exists between large number of siblings and educational outcomes (Buchmann & Hannum, 2001; Downey, 2001; Gelbal, 2008).

The impacts of school variables on academic achievement were examined. The findings revealed that approximately 11% of the variation in student GPA was explained by differences among schools. School quality was measured with school size, teacher-student ratio, and school resources. The effect of educational resources of schools (e.g., library, computer labs, science labs, music room) on academic achievement was moderate. School size and teacher-student ratio had no statistically significant effect on student achievement. The research findings showed that the effect of school resources on academic achievement was significant. However, there is no consensus about the effect of school resources on academic achievement. While most of the research found that school characteristics do not have significant effect on educational achievement research in developed and developing countries (Coleman et al.,1966; Hanushek, 1997; Hanushek & Luque, 2003), some research emphasized that school resources are associated with student outcomes especially in developed countries (Card & Krueger, 1996; Fuller & Clarke, 1994; Glewwe et al.,2004; Leon & Valdivia, 2015; Özberk et al., 2017). Leon and Valdivia (2015) concluded that when the distribution of schools was unequal, the influence of school characteristics on academic achievement was significant in developing countries. The authors suggested that improving school quality especially in poorer areas can help to close gender gap and socioeconomic gap in student achievement. The school with better physical environment is positively related to student outcomes (Adeogun & Osifila, 2008; Krueger, 2003; Parcel & Dufur, 2001). The present study showed that increases in educational resources in schools have a significant impact on student academic achievement. Therefore, this study suggests that investigating the determinants of student achievement is crucial to increase quality of education. More progress should be made to decrease the achievement gap in schools with educational policy movements in Turkey.

The study has also some limitations. Not many variables at student level and school level that effect student GPA were examined in this study. Student characteristics were measured with middle school students' background (demographic variables). However, it is also useful to examine the effect of other student variables on academic achievement (e.g. personality, intelligence). To determine the quality of schools, numerous resources can be considered such as teacher quality, institutional quality, physical resources, etc. School characteristics were measured into three categories in the present study. More variables should also be considered to measure school quality in further studies. School SES, geographical distribution of schools, school types, which may also potentially impact educational attainment, can also be considered in further studies. More research is needed to investigate the determinants of student achievement. Another limitation of this study was using self-reported data except students' GPAs. And also in the study, acceptable low limit to sample size at group level was used. Since getting larger groups is difficult for several reasons, the number of groups is usually a methodological concern in multilevel studies (Maas & Hox, 2005). Therefore, further studies should be conducted to larger number of schools.

## REFERENCES

Adeogun, A. A., & Osifila, G. I. (2008). Relationship between educational resources and students' academic performance in Lagos State Nigeria. *International Journal of Educational Management, 5*(6), 144-153. Retrieved from http://www.unilorin.edu.ng/ejournals/ijern

Akyüz, G. (2014). TIMSS 2011'de öğrenci ve okul faktörlerinin matematik başarısına etkisi. *Eğitim ve Bilim, 39*(172), 150-162. Retrieved from: http://egitimvebilim.ted.org.tr/index.php/EB/article/view/2867

Alacacı, C., & Erbaş, A. K. (2010). Unpacking the inequality among Turkish schools: Findings from PISA 2009. *International Journal of Educational Development, 30*(2), 182-192. doi: 10.1016/j.ijedudev.2009.03.006.

Anıl, D. (2009). Uluslararası öğrenci başarılarını değerlendirme programı (PISA)'nda Türkiye'deki öğrencilerin fen bilimleri başarılarını etkileyen faktörler. *Eğitim ve Bilim, Eğitim ve Bilim, 34(152 ),* 87-100. Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/view/594/74

Atar, H. Y. (2014). Öğretmen niteliklerinin TIMSS 2011 fen başarısına çok düzeyli etkileri. *Eğitim ve Bilim, 39*(172), 121-137. Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/view/2894

Atar, H. Y., & Atar, B. (2012). Examining the effects of Turkish Education reform on students' TIMSS 2007 science achievements. *Educational Sciences: Theory and Practice, 12*(4), 2632-2636. Retrieved from https://files.eric.ed.gov/fulltext/EJ1002867.pdf

Aypay, A., Erdogan, M., & Sozer, M. A. (2007). Variation among schools on classroom practices in science based on TIMSS-1999 in Turkey. *Journal of Research in Science Teaching 44*(10), 1417-1435. doi: 10.1002/tea.20202

Baker, D. P., Goesling, B., & Letendre, G. K. (2002). Socioeconomic status, school quality, and national economic development: A cross-national analysis of the "Heyneman Loxley Effect" on Mathematics and Science achievement. *Comparative Education Review*, *46*(3), 291-312. doi: 10.1086/341159

Batyra, A. (2017). *Gender gaps in student achievement in Turkey: Evidence from the programme for international student assessment (PISA) 2015*. Istanbul: Education Reform Initiative.

Bellibaş, M. Ş. (2016). Who are the most disadvantaged? Factors associated with the achievement of students with low socio-economic backgrounds. *Educational Sciences: Theory &Practice, 16*(2), 691-710. doi: 10.12738/estp.2016.2.0257

Berberoğlu, G., (2004). *Student learning achievement*. Paper Commissioned for the Turkey ESS. World Bank, Washington, DC.

Bilican-Demir, S. (2018). The effect of teaching quality and teaching practices on PISA2012 mathematics achievement of Turkish students. *International Journal of Assessment Tools in Education, 5*(4), 645-658. Retrieved from https://doi.org/10.21449/ijate.463409

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multi-level theory, research and methods in organizations: Foundations, extensions, and new directions* (pp. 349-381). San Francisco, CA: Jossey-Bass.

Boedeker, P. (2017). Hierarchical linear modeling with maximum likelihood, restricted maximum likelihood, and fully bayesian estimation. *Practical Assessment, Research & Evaluation, 22*(2), 1-19.

Börkan, B., & Bakış, O. (2016). Determinants of academic achievement of middle schoolers in Turkey. *Educational Sciences: Theory & Practice, 16*(6), 2193-2217. doi: 10.12738/estp.2016.6.0227

Brooks-Gunn, J., & Duncan, G.J., (1997). The effects of poverty on children. *The Future of Children, 7*(2), 55-71. doi: 10.2307/1602387

Buchmann, C., & Hannum, E. (2001). Education and stratification in developing countries: A review of theories and research. *Annual review of Sociology, 27*, 77-103. doi: 10.1146/annurev.soc.27.1.77

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2008). *Bilimsel araştırma yöntemleri* (4. Baskı). Ankara: Pegem A Yayıncılık.

Card, D., & Krueger, A. (1996). School resources and student outcomes: an overview of the literature and new evidence from North and South Carolina. *Journal of Economic Perspectives 10*(4), 31-40. doi: 10.3386/w5708

Chevalier, A., & Lanot, G., (2002). The relative effect of family characteristics and financial situation on educational achievement. *Education Economics 10*(2), 165-181. doi: 10.1080/09645290210126904

Chiu, M. M., & Xihua, Z. (2008). Family and motivation effects on mathematics achievement: Analyses of students in 41 countries. *Learning and Instruction, 18*(4), 321-336. doi: 10.1016/j.learninstruc.2007.06.003

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. doi: 10.1037/0033-2909.112.1.155

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartlant, J., Mood, A.M., Weinfall, F.D., & York, R.L. (1966). Equality of Educational Opportunity. Department of Health, Education and Welfare, Washington, DC. doi:10.3886/ICPSR06389.v3

Çiftçi, Ş. K. (2015). Effects of secondary school student' perceptions of mathematics education quality on mathematics anxiety and achievement. *Educational Sciences: Theory & Practice, 15*(6), 1487-1502. doi: 10.12738/estp.2015.6.2829

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of the state policy evidence. *Education Policy Analysis Archives 8*(1), 1-30. doi: 10.14507/epaa.v8n1.2000

Dayioğlu, M., & Türüt-Aşik, S. (2007). Gender differences in academic performance in a large public university in Turkey. *Higher Education, 53*(2), 255-77. doi: 10.2307/29735052

Dincer, M. A., & Uysal, G. (2010). The determinants of student achievement in Turkey. *International Journal of Educational Development, 30*(6), 592-598. doi: 10.1016/j.ijedudev.2010.05.005

Downey, D. B. (2001). Number of siblings and intellectual development. The resource dilution explanation. *American Psychologist, 56*(6), 497-504. Retrieved from http://dx.doi.org/10.1037/0003-066X.56.6-7.497

Enders, C. K. & Tofighi, D. D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue, *Psychol. Methods, 12*(2), 121-138. doi: 10.1037/1082-989X.12.2.121

Engin-Demir, C. (2009). Factors influencing the academic achievement of the Turkish urban poor. *International Journal of Educational Development, 29*, 17-29. doi: 10.1016/j.ijedudev.2008.03.003

Erberber, E., Stephens, M., Mamedova, S., Ferguson, S., & Kroeger, T. (2015). *Socioeconomically disadvantaged students who are academically successful: Examining academic resilience crossnationally*. IEA's Policy Brief Series, No. 5, Amsterdam, IEA.

Farkas, G., Sheehan, D., & Grobe, R.P. (1990). Coursework mastery and school success: gender, ethnicity, and poverty groups within an urban school district. *American Educational Research Journal 27*(4), 807-827. Retrieved from https://doi.org/10.3102/00028312027004807

Ferreira, F. H., & Gignoux, J., (2010). Inequality of opportunity for education: Turkey. In: R. Kanbur & S. Michael (Eds.), *Equity and growth in a globalizing world. Commission on growth and development* (pp. 131-156). Washington DC: The World Bank.

Flores, A. (2007). Examining disparities in mathematics education: Achievement gap or opportunity gap? *The High School Journal, 91*(1), 29-42. doi: 10.1353/hsj.2007.0022

Fuller, B., & Clarke, P. (1994). Raising school effects while ignoring culture? Local conditions and the influence of classroom tools, rules and pedagogy. *Review of Educational Research 64*(1), 122-131. Retrieved from https://doi.org/10.3102/00346543064001119

Gamboa, L. F., & Waltenberg, F. D. (2012). Inequality of opportunity in educational achievement in Latin America: Evidence from PISA 2006–2009. *Economics of Education Review, 31*(5), 694-708. doi: 10.1016/j.econedurev.2012.05.002

Gelbal, S. (2008). Sekizinci sınıf öğrencilerinin sosyoekonomik özelliklerinin türkçe başarısı üzerinde etkisi. *Egitim ve Bilim, 33*(150), 1-13. Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/view/626/96

Gevrek, Z. E., & Seiberlich, R. R. (2014). Semiparametric decomposition of the gender achievement gap: An application for Turkey. *Labour Economics, 31*, 27-44. Retrievd from https://doi.org/10.1016/j.labeco.2014.08.002

Gill, J. (2003). Hierarchical linear models. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement*. New York, NY: Academic.

Glewwe, P., Kremer, M., Moulin, S., & Zitzewitz, E. (2004). Retrospective vs. prospective analyses of school inputs: The case of flip charts in Kenya. *J. Dev. Econ., 74*(1), 251-268. doi: 10.1016/j.jdeveco.2003.12.010

Güvendir, M. A. (2014). Öğrenci başarılarının belirlenmesi sınavında öğrenci ve okul özelliklerinin Türkçe başarısı ile ilişkisi. *Eğitim ve Bilim, 39*(172), 163-180.

Hanushek, E.A. (1997). Assessing the effects of school resources on student performance: an update. *Educational Evaluation and Policy Analysis, 19*(2), 141-164. doi: 10.3102/01623737019002141

Hanushek, E. A., & Luque, J. A. (2003). Efficiency and equity in schools around the world. *Economics of Education Review, 22*(5), 481-502. Retrieved from https://ideas.repec.org/a/eee/ecoedu/v22y2003i5p481-502.html

Heyneman, S. P., & Loxley, W. A. (1983). The effect of primary school quality on academic achievement across 29 high- and low-income countries. *American Journal of Sociology 88*(6), 1162-1194.

Hox J. J. (1995). *Applied multilevel analysis*. Amsterdam: TT-Publicaties.

Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). Berlin: Springer.

Hox, J. (2010). *Multilevel analyses: Techniques and applications* (2nd Ed.). Mahwah, NJ: Erlbaum.

Johnson, M. K., Crosnoe, R., & Elder, G. H. (2001). Students' attachment and academic engagement: The role of race and ethnicity. *Sociol. Educ. 74*(4), 318-340. Retrieved from http://dx.doi.org/10.2307/2673138

Kalaycıoglu, D. B. (2015). The influence of socioeconomic status, self-efficacy, and anxiety on mathematics achievement in England, Greece, Hong Kong, the Netherlands, Turkey, and the USA. *Educ Sci: Theory Pract., 15*(5), 1-11. doi: 10.12738/estp.2015.5.2731

Kalender, I., & Berberoglu, G. (2009). An assessment of factors related to science achievement of Turkish students. *International Journal of Science Education, 31*(10), 1379-1394. doi: 10.1080/09500690801992888

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*, 137-152. doi: 10.1037/a0028086

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies* (Unpublished manuscript). California State University, Los Angeles.

Krueger, A.B. (2003). Economic considerations and class size. *The Economic Journal, 113*(485), F34-F63. doi: 10.1111/1468-0297.00098

Leon, G., & Valdivia, M. (2015). Inequality in school resources and academic achievement: Evidence from Peru. *International Journal of Educational Development 40,* 71–84. Retrieved from http://dx.doi.org/10.1016/j.ijedudev.2014.11.015

Ma, X., (2001). Stability of socioeconomic gaps in mathematics and science achievement among Canadian schools. *Canadian Journal of Education 26 (1),* 97-118. doi: 10.2307/1602147

Ma, X., & Klinger, D. A. (2000). Hierarchical linear modeling of student and school effects on academic achievement. *Canadian Journal of Education, 25*(1), 41-55.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92. doi: 10.1027/1614-1881.1.3.86

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295-314. Retrieved from http://dx.doi.org/10.1007/s10648-014-9287-x

National Ministry of Education (MoNE), (2007). *Report on student assessment program (SAP) 2005: Mathematics*. Ankara: MoNE Directorate of Education and Instruction.

Organisation for Economic Co-operation and Development. (2011). *Lessons from PISA for the United States, strong performers and successful reformers in education*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2016). *PISA 2015 Results: Excellence and Equity in Education (Vol. I)*. Paris: OECD.

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research, and Evaluation, 7*(1), 1-3.

Özberk, E. H., Atalay- Kabasakal, K., & Boztunç-Öztürk, N. (2017). Investigating the factors affecting Turkish students' PISA 2012 mathematics achievement using hierarchical linear modeling. *Hacettepe University Journal of Education, 32(3),* 544-559. doi: 10.16986/HUJE.2017026950

Özdemir, C. (2016). Equity in the Turkish education system: A multilevel analysis of social background influences on the mathematics performance of 15-year-old students. *European Educational Research Journal, 15*(2), 193-217. doi: 10.1177/1474904115627159

Parcel, T.L., & Dufur, J.M., (2001). Capital at home and at school: effects on student achievement. *Social Forces 79* (3), 881-911. doi: 10.1353/sof.2001.0021

Perry, L., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record, 112*(4), 1137-1162.

Petrill, S. A., & Wilkerson, B. (2000). Intelligence and achievement: A behavioral genetic perspective. *Educational Psychology Review, 12*(2), 185-199. Retrieved from https://doi.org/10.1023/A:1009023415516

Phan, H. T. (2008). *Correlates of mathematics achievement in developed and developing countries: an hlm analysis of timss 2003 eighth-grade mathematics scores* (Doctoral dissertation).University of South Florida.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods,* (2nd Ed.). Thousand Oaks, CA: Sage Publications.

Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., & du Toit, M. (2011). HLM 7: Hierarchical Linear and Nonlinear Modeling [Computer Software]. Lincolnwood, IL: Scientific Software International, Inc.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta- analytic review of research. *Review of Educational Research, 75*(3), 417-453. doi: 10.3102/00346543075003417

Smits, J., & Gündüz-Hoşgör, A. (2006). Effects of family background characteristics on educational participation in Turkey. *International Journal of Educational Research, 26*(5), 545-560. Retrieved from http://dx.doi.org/10.1016/j.ijedudev.2006.02.002

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling* (2nd Ed.). London: Sage.

Sweetland, S. R., & Hoy, W. K. (2000). School characteristics and educational outcomes: toward an organizational model of student achievement in middle schools. *Educational Administration Quarterly, 36(*5), 703-729. doi: 10.1177/00131610021969173

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6th Ed). New York, NY: Pearson Education

Tavşancıl, E., & Yalçın, S. (2015). A determination of Turkish student's achievement using hierarchical linear models in trends in ınternational mathematics-science study (TIMSS) 2011. *Anthropologist, 22*(2), 390-396. doi: 10.1080/09720073.2015.11891891

The United Nations Children's Fund, (2016). Gender equality in secondary education. a literature review for UNICEF, NATCOM and Aydın Doğan Foundation.

Van Houtte, M. (2004). Why boys achieve less at school than girls: The difference between boys' and girls' academic culture. *Educational Studies, 30*(2), 159-173. doi: 10.1080/0305569032000159804

Veenstra, R., & Kuyper, H. (2004). Effective students and families: The importance of individual characteristics for achievement in high school. *Educational Research and Evaluation, 10*(1), 41-70. doi: 10.1076/edre.10.1.41.26302

Willms, J.D., (1996). Indicators of mathematics achievement in Canadian elementary schools. In: HRDC (Eds.), *Growing up in Canada: National longitudinal study of children and youth* (pp. 69-82). Ottawa, Ontario: Human Resources Development Canada and Statistics.

Willms, D. J., Somers, M. A. (2001). Family, classroom, and school effects on children's educational outcomes in Latin America. *School Effectiveness and School Improvement 12*(4), 409-445.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology, 8*(1), 52-69. doi: 10.20982/tqmp.08.1.p052

Wößmann, L. (2003). Schooling resources, educational institutions and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics, 65*(2), 117-170. doi: 10.1111/1468-0084.00045

Yalçın, S., Demirtaşlı, R. N., İlgün-Dibek, M., & Yavuz, H. Ç. (2017). The effect of teacher and student characteristics on TIMSS 2011 mathematics achievement of fourth-and eighth-grade students in Turkey. *International Journal of Progressive Education, 13*(3), 79-94.

Yavuz, H. Ç., Demirtaşlı, R. N., Yalçın, S., & İlgün-Dibek, M. (2017). The effects of student and teacher level variables on TIMSS 2007 and 2011 mathematics achievement of Turkish students. *Education and Science, 42(189),* 27-47. doi: 10.15390/EB.2017.6885

Yavuz, E., Tan, Ş., & Atar, H., Y. (2019). Effects of students and school variables on SBS achievements and growth in mathematic. *Journal of Measurement and Evaluation in Education and Psychology, 10*(1), 96-116. doi: 10.21031/epod.493297

# Türkiye'de Öğrenci ve Okul Özelliklerinin Ortaokul Öğrencilerinin Akademik Başarılarına Etkileri

## *Giriş*

Akademik başarı eğitim sisteminin niteliğine yönelik en önemli belirleyicilerden biridir. Birçok faktörün akademik başarıyı etkilediği görülmektedir (Börkan & Bakış, 2016; Coleman ve diğerleri, 1966; Engin-Demir, 2009; Gelbal, 2008). Araştırmalar sadece aile özelliklerinin değil aynı zamanda okul ve öğrenci özelliklerinin de öğrenci başarısını etkileyen önemli faktörler olduğunu göstermektedir (Alacacı & Erbaş, 2010; Bellibaş, 2016; Börkan & Bakış, 2016; Engin-Demir, 2009; Kalender & Berberoglu, 2009; MEB, 2007).

Cinsiyet, yaş, motivasyon, derslere yönelik tutumlar, öz-yeterlik, öğrencilerin çabaları, okulda zorbalığa uğramak gibi birçok öğrenciye ait bireysel özellikler olup akademik başarı üzerinde anlamlı etkilere sahiptir (Engin-Demir, 2009; Ma, 2001; Özberk, Atalay-Kabasakal & Boztunç-Öztürk, 2017; Yavuz, Demirtaşlı, Yalçın, & İlgün-Dibek, 2017). Ailenin sosyo ekonomik özellikleri, aile büyüklüğü ya da ailedeki kardeş sayısı, ebeveynlerin eğitim düzeyi öğrenci başarısında etkili olabilmektedir (Alacacı & Erbaş, 2010; Börkan & Balkış, 2016; Downey, 2001; Engin-Demir, 2009; Kalender & Berberoglu, 2009; MEB, 2007). Okul ve öğretmen özellikleri de öğrenci başarısında etkili faktörlerdir (Atar, 2014; Bilican-Demir, 2018; Darling-Hammond, 2000; Phan, 2008; Tavşancıl & Yalçın, 2015; Yavuz ve diğerleri, 2017). Öğrenci başarısı üzerinde sınıf büyüklüğü, okul büyüklüğü, okulun bulunduğu bölge, ortalama SES (Sosyo-Ekonomik Statü), öğretmen öğrenci oranı, öğretmen niteliği, eğitim kaynakları, çevre gibi faktörler okullar arasında farklılık oluşturabilmektedir (Leon & Valdivia, 2015; Willms & Somers, 2001).

Öğrenci başarısı değerlendirilirken birkaç yöntem kullanılmaktadır. Genel olarak final notları ya da not ortalamaları dikkate alınmaktadır. Standartlaştırılmış başarı testleri de öğrenci başarısı değerlendirilirken kullanılabilmektedir (Petrill & Wilkerson, 2000). Uluslararası geniş ölçekli değerlendirme (örneğin; The Trends in International Mathematics and Science Study-TIMSS, Programme for International Student Assessment-PISA, and Progress in International Reading Literacy Study-PIRLS) ve ulusal geniş ölçekli değerlendirme ile öğrenci başarısı değerlendirilmektedir. Türkiye'de öğrenci başarısı üzerine birçok çalışmanın uluslararası TIMSS, PISA veya PIRLS veri setleri kullanılarak gerçekleştiği görülmektedir (Akyüz, 2014; Alacacı & Erbaş, 2010; Anıl, 2009; Atar, 2014; Atar & Atar, 2012; Dincer & Uysal, 2010; Özdemir, 2016; Özberk ve diğerleri, 2017; Yalçın ve diğerleri, 2017). Ancak Türkiye'de akademik başarıya yönelik sadece birkaç çalışmada ulusal geniş ölçekli değerlendirmenin (örneğin; SBS, ÖBBS, TEOG) ya da başarı ortalamalarının kullanılarak gerçekleştiği görülmektedir (Börkan & Bakış, 2016; Çiftçi, 2015; Engin-Demir, 2009; Gelbal, 2008; Yavuz, Tan & Atar, 2019). Bu çalışma ortaokul öğrencilerinin akademik başarılarını etkileyen öğrenci ve okul özelliklerinin incelenmesini amaçlamaktır. Akademik başarı öğrencilerin genel not ortalamaları ile ölçülmüştür. Bu çalışmada dört araştırma sorusuna yanıt aranmıştır.

1. Okullar öğrencilerin ortalama akademik başarılarında ne kadar farklılık oluşturmaktadır?
2. Okullar öğrenci düzeyindeki değişkenler (örneğin, cinsiyet, SES, kardeş sayısı) ve akademik başarı arasındaki ilişkiye bağlı olarak ne kadar farklılık oluşturmaktadır?

3. Okul düzeyinde değişkenler (okul büyüklüğü, öğrenci-öğretmen oranı, okul kaynakları) ortalama akademik başarının anlamlı yordayıcıları mıdır?
4. Okul düzeyinde değişkenler (okul büyüklüğü, öğrenci-öğretmen oranı, okul kaynakları) okullar arası ilişkide anlamlı yordayıcılar mıdır?

### *Yöntem*

Bu çalışmada öğrenci düzeyinde ve okul düzeyinde değişkenlerin öğrenci başarısı üzerindeki etkilerini incelemek için hiyerarşik linear modelleme (HLM) yöntemi kullanılmıştır. İç içe gruplanmış yapıdaki veriler için HLM analizi kulllanılması Tip I hata yapmayı ve yanlı sonuçların önlenmesini sağlamaktadır (Gill, 2003; Osborne, 2000; Raudenbush & Bryk, 2002). Çalışma grubunu, Ankara ve Sinop il merkezlerinde 10 ortaokula devam eden toplam 1053 yedinci sınıf ve sekizinci sınıf öğrencisi oluşturmuştur. Katılımcıların 512'sini (% 48.6) kız öğrenciler, 541'ini (% 51.4) ise erkek öğrenciler oluşturmuştur. Ortalama yaş 13.46 olup yaş aralığı 12 ile 15 arasında değişmektedir. Ortaokul öğrencilerine anket aracılığı ile çeşitli demografik sorular (cinsiyet, yaş, SES, kardeş sayısı) ve akademik başarı ortalamaları sorulmuştur. Veri analizi için HLM 7 kullanılmıştır. İki düzeyli HLM modeli kullanılarak öğrenci düzeyindeki ve okul düzeyindeki değişkenlerin akademik başarı üzerindeki etkileri incelenmiştir. Cinsiyet, SES ve kardeş sayısı öğrenci düzeyindeki değişkenleri oluştururken okul büyüklüğü, öğrenci-öğretmen oranı ve okul kaynakları okul düzeyindeki değişkenleri oluşturmuştur. Çalışmada öğrencilerin okullardaki dağılımı incelendiğinde, en yüksek öğrenci sayısının 235 ve en düşük öğrenci sayısının 68'dir. Çalışmada iki düzeyli model, HLM 7'nin hesapladığı sınırlandırılmış maximum olabilirlik ölçümü kullanılarak analiz edilmiştir (Raudenbush, Bryk, Cheong, Congdon & du Toit, 2011).

### *Sonuç ve Tartışma*

Bu çalışmada Türkiye'deki ortaokul öğrencilerinin akademik başarılarını etkileyen öğrenci ve okul özellikleri incelenmiştir. Araştırma bulguları, öğrenci özelliklerinin (cinsiyet, SES ve kardeş sayısı) ortaokul öğrencilerinin akademik başarıları üzerinde istatistiksel olarak anlamlı etkiye sahip olduğunu göstermiştir. Öğrenci değişkenlerinin akademik başarı üzerinde açıkladığı varyans oranı %7'dir. Cinsiyetin öğrenci başarısı üzerinde güçlü bir etkiye sahip olduğu ortaya çıkmıştır. Diğer değişkenler kontrol edildiğinde, kız öğrenciler erkek öğrencilere göre daha yüksek başarı ortalamasına sahiptir. Bu araştırma sonucu diğer araştırma sonuçları ile benzerlik göstermektedir (Börkan & Bakış, 2016; Güvendir, 2014; Engin-Demir, 2009; Van Houtte, 2004). Araştırmalar bazı sebeplerden dolayı kız öğrencilerin erkek öğrencilere göre daha iyi performans gösterdiklerini ortaya koymaktadır. Öğrencilerin tutumları, öz-yeterlikleri, okula bağlılıkları, akademik motivasyonları, derslerdeki çabaları kız ve erkek öğrencilerin akademik başarılarını farklı şekilde etkilemektedir (Batyra, 2017; Engin-Demir, 2009; Gevrek & Seiberlich, 2014; Van Houtte, 2004; Veenstra & Kuyper, 2004). Ayrıca, cinsiyet eşitliği okul başarısı için çok önemlidir. Türkiye'de 2000 yılından itibaren cinsiyet eşitliğini arttırmak adına önemli çalışmalar yapılmıştır. İlkokul ve ortaokulda özellikle kız öğrencilerin okullaşma oranları arttırılarak kız ve erkek öğrencilerin akademik başarıları arasındaki farklılık önemli ölçüde azalmıştır. Bu araştırma sonucunun da Türkiye'de okullarda cinsiyet eşitliğine yönelik yapılan projelerin olumlu etkilerini gösterdiği söylenebilir(The United Nations Children's Fund-UNICEF, 2016). Diğer taraftan kız öğrencilerin bazı alanlarda özellikle fen ve matematikte erkek öğrencilere göre daha düşük performans gösterme eğiliminde oldukları görülmektedir (Berberoğlu, 2004; Chiu & Xihua, 2008; Farkas, Sheehan & Grobe, 1990; Wößmann, 2003). Alan yazın genel olarak öğrencilerin akademik performanslarının cinsiyetlerine göre farklılık gösterdiğini ortaya koymaktadır. Bu nedenle, bu alana yönelik daha fazla çalışma yapılması oldukça önemlidir.

Etki büyüklüğü düşük olmasına rağmen, kardeş sayısı ve SES değişkenlerinin akademik başarı üzerinde anlamlı etkiye sahip olduğu ortaya çıkmıştır. Düşük SES'e sahip öğrencilerin daha düşük akademik ortalamaya sahip olma ihtimalinin daha yüksek olduğu bulunmuştur. Benzer şekilde, birçok araştırma düşük sosyo ekonomik statüye sahip aile ile yaşayan öğrencilerin okullarda düşük performans gösterdiklerini açığa çıkarmıştır (Alacacı & Erbaş, 2010; Atar & Atar, 2012; Aypay, Erdogan, & Sozer, 2007; Bellibas, 2016; Dincer & Uysal, 2010; Flores, 2007; Gelbal, 2008; Kalaycıoğlu, 2015; Perry & McConney, 2010). Aynı zamanda bu araştırmada, kardeş sayısının akademik başarı üzerindeki negatif etkisi ortaya çıkmıştır. Özellikle gelişen ülkeler ve batı ülkelerinde, çok sayıda kardeş ve eğitim çıktıları arasında negatif yönde ilişki bulunmaktadır (Buchmann & Hannum, 2001; Downey, 2001; Gelbal, 2008).

Araştırmada okul değişkenlerinin akademik başarı üzerindeki etkileri incelenmiştir. Öğrenci başarı ortalaması üzerinde yaklaşık %11 oranında varyans, okullar arasındaki farklılıklar aracılığı ile açıklanmaktadır. Okulun niteliği, okul büyüklüğü, öğretmen-öğrenci oranı ve okul kaynakları gibi değişkenler ile ölçülmüştür. Okul kaynaklarının (örneğin, kütüphane, bilgisayar laboratuvarı, fen laboratuvarı, müzik odası gibi) öğrenci başarısı üzerinde etkisi orta düzeydedir. Ancak okul büyüklüğü ve öğretmen-öğrenci oranının öğrenci başarısı üzerinde istatistiksel olarak anlamlı bir etkiye sahip olmadığı ortaya çıkmıştır. Alan yazın incelendiğinde okul kaynaklarının akademik başarı üzerindeki etkisine yönelik ortak bir görüş olmadığı görülmektedir. Bazı çalışmalar okul kaynaklarının akademik başarı üzerinde etkisinin olmadığını göstermektedir (Coleman ve diğerleri,1966; Hanushek, 1997; Hanushek & Luque, 2003). Diğer taraftan bazı çalışmalar, okul kaynaklarının öğrenci çıktıları ile ilişkili olduğunu ortaya koymuştur (Card & Krueger, 1996; Fuller & Clarke, 1994; Özberk ve diğerleri, 2017). Daha iyi fiziksel ortama sahip bir okul, öğrenci başarısını pozitif yönde etkileyebilmektedir (Adeogun & Osifila, 2008; Krueger, 2003; Parcel & Dufur, 2001). Bu çalışmada da, bu araştırmaları destekleyen bulgulara ulaşılmıştır.

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(4); 406-421

# Investigation of the Reliability of Teachers, Self and Peer Assessments at Primary School Level with Generalizability Theory *

Eda GÜRLEN **　　Nagihan BOZTUNÇ ÖZTÜRK ***　　Emel EMİNOĞLU ****

**Abstract**

This study aims at determining the reliability coefficients of teacher, self and peer assessments carried out at primary school level. In line with this aim, an interdisciplinary approach is adopted, and the notion of helpfulness included within the scope of values education is addressed in connection with the practices followed in Turkish, social studies and music lessons. The study group consists of 30 students of the third graders from a public school in the city of Ankara. In the light of the aim of the study, the Generalizability Theory is used for the data analysis. It is found out at the end of the study that the variance component estimated for the main effect of the student is the largest component of the total variance in all three lessons. When G and Φ coefficients are examined, reliability coefficients are found to be over .80 in music, and over .90 in Turkish and social studies. According to G-Facet analysis results, when teacher and peer assessments are excluded from the analysis, respectively, G and Φ coefficients have a decreasing tendency whereas these coefficients increase when self-assessment is excluded from the analysis. Especially in the music lesson, the reliability coefficients obtained by excluding teacher and peer assessments from the analysis are found to be around .60, which is a remarkable result.

*Key Words:* Teacher assessment, self-assessment, peer assessment, Generalizability Theory.

## INTRODUCTION

Evaluation, which is an important element of the education system, has important functions such as providing information about the effectiveness and efficiency level of the teaching process, determining the degree of achievement of the previously-set goals and revealing the strengths and weaknesses of the practices followed during lessons. Implementing the evaluation activities thoroughly ensures the continuous control of the education and thus makes it possible to find a quick remedy for the troubles that come out at any stage of education and produce robust solutions for problems. Moreover, it enables the identification and then the elimination of learning difficulties and deficiencies by monitoring student development. It also identifies sources of success and failure and helps to uncover elements that affect education positively and negatively. Thus, it becomes possible to support the practices that improve the quality of education and to take timely measures against obstacles and threats. By also shedding light on planning and orientation studies for the future, education can be improved efficiently and quickly (Çeçen, 2011; İşman & Eskicumalı, 2003; Kurudayıoğlu, Şahin & Çelik, 2008; Turgut & Baykul, 2015; Yaşar, 2017).

Teachers use different methods in order to make an assessment that can reveal every aspect of the change created by all educational activities. As a result of these methods, students are assessed from the perspectives of teachers and experts according to the criteria prepared by them. However, education and training are processes that come to life with interaction. The fact that the point of view of the students actively participating in this interaction is not included in the assessment activities constitutes

* A part of this article was presented as an oral presentation at the 17th International Primary Teacher Education Symposium.

** Assoc. PhD., Hacettepe University, Faculty of Education, Ankara-Turkey, edaerdem@hacettepe.edu.tr, ORCID ID: 0000-0002-1719-9840

*** PhD., Hacettepe University, Lifelong Learning Center, Ankara-Turkey, nagihanboztunc@hacettepe.edu.tr, ORCID ID: 0000-0002-2777-5311

**** Teacher, Talatpaşa Secondary School, Ankara-Turkey, emel.eminoglu@gmail.com, ORCID ID: 0000-0003-1712-8652

To cite this article:
Gürlen, E., Boztunç-Öztürk, N., & Eminoğlu, E. (2019). Investigation of the reliability of teachers, self and peer assessments at primary school level with generalizability theory. *Journal of Measurement and Evaluation in Education and Psychology, 10*(4), 406-421. doi: 10.21031/epod.583891

*Received: 28.06.2019*
*Accepted: 21.08.2019*

an important deficiency. Assessment activities conducted in this way will not become meaningful enough for the students who do not participate in the process and therefore will not perform their intended functions fully. The assessment activities can be meaningful only if the students use the assessment criteria for their own studies as well as other studies. In this way, students can realize that the assessment process is a deep learning experience. By comparing their own work included in an activity with other students' work related to the same activity, they can reach a more in-depth learning level and understand the working principles of the mind during the assessment process. Thus, they can also have an idea about how the teacher performs the assessment process. This can open the way for the teacher-student dialogue and enable the students to think about the arrangements to be made after the assessment and to take responsibility. Students who take responsibility for their own learning processes have the opportunity to become independent learners who think, direct, realize their own development, organize their own work, criticize themselves, and learn. An individual who has the ability to decide whether a behavior they exhibit meets the criteria related to that behavior will also have the ability to control their own behavior independently of any authority during their life. Therefore, assessment activities will contribute to the education of individuals who have gained autonomy for lifelong learning (Race, 2001; Sünbül, 2007; Wilson & Jan, 1993).

The world of education, which has discovered this aspect of assessment activities, tends to assess individual's learning through methods in which the individual is at the center. Such assessments, although they are more costly and time-consuming, contribute to the learning of the students and the professional development of the teachers by integrating learning and assessment. These activities represent not only a scoring exercise but also a dynamic process in which learning skills develop through active participation whereas in-depth learning turns out to be a possible phenomenon. Students' involvement in this process allows them to understand that assessment is not just a grading process. Students who are not adequately informed about the objectives and functions of the assessment may not be able to fully understand the points that their teaching activities are intended to achieve. When students are not fully aware of what is expected from them, their motivation to learn can be affected adversely. This may lead them to develop negative attitudes towards learning. Students who understand the purpose and necessity of the assessment activity can explore their strengths and weaknesses by approaching the assessment criteria more realistically. Self-discovering students focus directly on learning by taking responsibility for their own learning, and they turn out to be self-confident, critical and independent learners (Ballantyne, Huges & Mylonas, 2002; Boud, 1986; Cihanoğlu, 2008; Cram, 1995; Falchikov, 2001; Tekindal, 2014; Topping, Smith, Swanson & Elliot, 2000).

In order for such assessment activities to be carried out objectively, students should be included in the process from the first stage of assessment. Students should actively participate in the process of deciding on the type of assessment, determining which learning outcomes will be assessed, and establishing the criteria to be used. Teachers and students should discuss and agree on these issues. There should be a harmonious relationship among those who are involved in the assessment. Thus, students can realize the ideal behaviors expected from them, the reasons why they are expected to display these behaviours and the necessity of learning. Therefore, it will be possible to develop the skills to establish a criterion for a specific behavior and grading the quality of that behavior. The participation of students in these discussions will also be beneficial in terms of communication and self-expression skills. With all this learned, students can manage their own learning processes from the beginning till the end. They can decide on what is needed to raise their learning levels (Alıcı, 2010; Stiggins and Chappius, 2005; Woolfolk 2002).

The participation of students in assessment activities also contributes to the creation of a healthy teaching-learning environment. These activities give the teacher information about how the student thinks and, therefore, can learn. They enable teachers to recognize students in different aspects including affective characteristics. Thus, they guide the teacher in organizing teaching activities. They also help the student to understand how the teacher thinks. When students get involved in the process using similar ways of thinking, they feel that they become part of the learning environment. When students fulfil their potentials, their academic self-concept develops in a positive way; and they become more self-confident. They get proud of what they have learned by seeing their achievements and their level of development over time. Thus, they get happy and turn out to be willing to learn (Bahar, 2006; Stiggins and Chappius, 2005).

Assessment activities carried out with the participation of students allow students to become aware of other students' learning after they become aware of their own learning. Starting from themselves, the students take responsibility for each other's learning and gain the ability to assess other individuals. That's why self-assessment and peer-assessment are the two most important types of assessment that enable students to improve in this way.

Self-assessment means that students make judgments about the extent to which they fulfill these criteria by applying the assessment criteria for their own studies. Thus, students discover what they know, what they can do, how they feel, and how they learn. In this discovery process, students who have the opportunity to use their high-level thinking skills from a critical point of view are provided with the skill to make sense of themselves objectively. By becoming familiar with their strengths and weaknesses, students become aware of their learning problems. They can produce solutions to their own learning problems by using the detailed information they have acquired about their own learning paths. They develop an ability to plan their future studies and work by judging their learning experiences. Therefore, it can be said that a student who has the ability to evaluate his/her achievement will reach the competency level necessary to achieve greater success. Thus, students should be supported to form a set of productive and realistic objectives with an action plan based on the feedback resulting from the self-assessment (Alıcı, 2010; Boud, 1986; Kutlu, Doğan & Karakaya, 2008; Mistar, 2011; Stiggins, 1997; Tekindal, 2014).

From the perspective of cognitive and constructivist learning theories, it is seen that self-assessment helps the learner to structure the knowledge. According to these theories, newly-acquired information can be meaningful for students only when they associate the new pieces of information with the already existing ones. Self-assessment contributes to establishing a link between the existing knowledge and understanding and the new ones by giving meaningful feedback to students based on the criteria they have internalized before. In this way, students learn by constantly comparing their knowledge and understanding with their learning objectives. This shows that self-assessment is also effective in establishing learning goal orientation. Learning objectives require a certain degree of internal processing of information. Self-assessment contributes to the motivation of the learning type as it improves internal control, knowledge, understanding, and skills so that students can be aware of their progress towards understanding the information fully. (McMillan, trans. 2015).

Self-evaluation is closely related to the development of an individual's reflection ability. Reflection involves one's self-monitoring as an external observer and the development of decision-making skills for better action in the future (Osterman & Kottkamp, 1993). Students' developments in reflective behaviors and skills constitute the most important point in self-assessment. In order to make progress in this regard, it is necessary to clearly define which behaviors and skills will be assessed and the corresponding trends. In order to obtain reflective comments about students' work, what is expected from them should be clearly stated. Simple examples can be used to visualize trends in this field. It can be started by questioning the accuracy of the answers given by the students to the questions about the lesson. Afterwards, questions such as why the answer is not correct, what the wrong answer exactly tells the student, and what needs to be done in order to give the correct answer can be asked (McMillan, trans. 2015).

Self-assessment tools can be prepared in different ways. They may range from a format that is prepared in a draft form of checklists and questions to a format that questions the reflections they have produced from a composition before; however, what is important is that students should take responsibility for their learning by determining what they have learned and in which areas they have problems whatever the chosen self-assessment tool is (Bahar, Nartgün, Durmuş & Bıçak, 2008). Also, students' self-assessments should be kept in students' personal development files (Woolfolk, 2002).

There are a number of factors that prevent self-assessment from being performed in a healthy way. Such factors include students who are biased about assessing their own learning because of having difficulty in making objective interpretations, who overestimate or underestimate their own abilities, who are able to make self-evaluation because of being unaware of their own abilities, who do not consider themselves sufficient to perform self-assessment or who believe that assessment should only be done by the teacher. In this case, continuous self-assessment, clarification of how students can make self-assessment and encouraging students to self-assessment will be effective in eliminating these factors (Alıcı, 2010; Tekindal, 2014).

On the other hand, when peer assessment is in question, students evaluate the performances or quality of the products belonging to others by applying the relevant criteria to the work of other students of similar status. Thus, they learn new pieces of information together and from each other via examining and criticizing different works. Peer review involves providing students with feedback from their peers about the quality of their work. Peer feedback encourages working together and learning together. Students increase awareness about their own learning needs by seeing their strengths and weaknesses. They can even get to know each other better than their teachers and give more detailed feedback. In this respect, peers can provide feedback to a greater number of students than the teacher in crowded classrooms. Thus, they can develop each other's talents and skills. However, students' mastering in performing an effective peer review requires a lot of practice. The assessment criteria should also be clear, appropriate, and discussed with the students. (Ballantyne et al., 2002; Falchikov, 1986, 2001; Topping et al., 2000; Tekindal, 2014).

Peer review has turned out to be a part of our success development since the first years of our lives. When children get informal feedback from their peers, this contributes to their social development to a great extent. The social development of the students can be accelerated significantly when the power of peer feedback is included in the planned assessment activities. Students have the opportunity to improve the quality of their products through teamwork. They can see the mistakes and deficiencies in their studies from the point of view of their friends although they do not realize these mistakes and deficiencies on their own. Thus, the defects can be corrected, and the works can be carried to higher levels. It is no doubt that students also develop a number of social skills such as communication, cooperation, discussion besides improving their products of studies in such a process. Students learn to criticize each other constructively and accept criticism with tolerance. When they work together in this way, they can see themselves as a member of the community and develop a sense of belonging. They grow up as individuals who can use what they learn from their peers both in their own personal development and in the development of the society as a whole (Alıcı, 2010; Tekindal, 2014).

Initially, peer assessment, as well as self-assessment, may be difficult to perform objectively. Students are more likely to behave subjective when evaluating their peers whom they like and who are more popular than others in the class. However, when these studies are carried out routinely at regular intervals, students will start to carry out better assessments. The purpose, importance and implementation steps of peer assessment should be clearly explained to the students in order to improve peer-assessment process. It should be emphasized that it is necessary to make a distinction between the students' features to be assessed and other qualities of these students that will be excluded from the peer-assessment process. Peer-assessment will be more objective when students start to feel that they are working together and not competing. Moreover, it is possible to carry out the peer-assessment process more objectively when students do not know whose product is being assessed, and assessment is done by more than one student or the students to assess a product are chosen randomly (Bahar et al., 2008; Alıcı, 2010).

When self-assessment and peer-assessment are used together, they help and develop each other. However, students should be able to use their assessment skills actively and correctly in order to achieve this development. This is closely related to providing students with the opportunity to grow up in a culture of assessment and evaluation. Researches show that performing such activities routinely from the first year of primary education contributes significantly to critical thinking skills (Alıcı, 2010). In addition to this, when the related literature is examined, it is seen that such assessment should be carried out continuously in order to handle this process in a healthy way. Therefore, students' participation in assessment activities should be ensured from the first stages of education. Assessment activities, which include both students' and teachers' perspectives, will provide more detailed data and develop more effective solutions. Examining the students' point of view by comparing them with the teachers' point of view will guide the development of assessment activities in this field. In this case, it is important to determine whether the primary school students in the first stages of education differ from the teachers who are experts in the field in terms of evaluating their own and their peers' work according to certain criteria. If so, identifying the scope of this difference is important to determine where we are in the field of assessment. When the related literature is reviewed, it is clear that there are numerous studies on self- and peer assessment in Turkey, but there are a limited number of studies that examined self- and peer assessment through comparing the reliability of these assessments. Considering that reliability is one of the significant limitations of such assessments, it is thought that addressing this issue is important in terms of revealing the level reached in studies that are have been carried out about assessment involving students' participation. Therefore; this study aims at determining the reliability of the scores obtained from teacher and student (self and peer) assessments in primary school level. For this purpose, the researchers examined the change of reliability of scores obtained via self- and peer-assessment while evaluating the exemplar event-driven performance works that were done in Turkish, social sciences and music lessons at third grade of a primary school. It is thought that the study will contribute to more efficient assessment studies by examining the self-assessment and peer-assessment skills of primary school students.

## METHOD

This study is a descriptive study since it is aimed to determine the reliability of teacher, self and peer assessments performed in the performance works done in Turkish, social sciences and music lessons at the third grade of primary school.

### *Study Group*

The study group consists of 30 third grade students (14 boys and 16 girls) studying at a primary school in the city of Ankara. It was decided during the study that five students to be selected randomly among 30 students would score for peer assessment. As a result, the remaining 25 students were included in the study as the measurement object.

### *Data Collection Tools*

The assessment, self-assessment and peer-assessment scales prepared by the Ministry of National Education and included in the teacher's guide books were used as data collection tools after being simplified in accordance with performance works that had been prepared in line with the expert opinions (5 classroom teachers, 2 Turkish teachers, and 1 music teacher).

Writing Skills Assessment Scale included in the teacher's guide book which has been used since 2013-2014 academic year upon the approval of the Ministry of National Education was used to assess the writing skills of the students in the Turkish lesson (Milli Eğitim Bakanlığı-MEB, 2013). Taking into consideration the length of time of the implementation and the performance task, a grading scale that consists of four criteria was created by selecting and arranging critical criteria among the ten measures included in the scale in accordance with the opinions of experts. The generated grading scale is given in Table 1.

Table 1. Grading Scale Used in Turkish Performance Task

| **Criterion** | **Rating** | | |
|---|---|---|---|
| | **1** | **2** | **3** |
| Following spelling rules | | | |
| Writing meaningful and normative sentences | | | |
| Writing events in order of occurrence | | | |
| Including the main idea in writings | | | |

Discussion Scale included in the teacher's guide book which has been used since the 2015-2016 academic year upon the approval of Ministry of National Education was used to assess the discussion skills of the students in the social sciences lesson (MEB, 2017a). Taking into consideration the length of the time of the implementation and the prepared performance task, a grading scale consisting of four criteria was created by selecting and arranging critical criteria among the ten criteria in the scale in accordance with expert opinions. The generated grading scale is given in Table 2.

Table 2. Grading Scale Used in Social Sciences Performance Task

| **Criterion** | **Rating** | | |
|---|---|---|---|
| | **1** | **2** | **3** |
| Being able to express his/her idea clearly | | | |
| Interpreting the questions correctly and giving appropriate answers to the questions | | | |
| Following the rules of discussion | | | |
| Controlling the tone of voice and gestures | | | |

Analytical-Rate Grading Scale for Song/Folk/March Performances included in the teacher's guide book which has been used since the 2017-2018 academic year upon the approval of Ministry of National Education was used to assess the singing performances of the students in the Music lesson (MEB, 2017b). Taking into consideration the length of time of the implementation and the performance task, the grading scale consisting of four criteria was created by selecting and arranging critical criteria among the six criteria included in the grading scale in accordance with expert opinions. The generated grading scale is given in Table 3.

Table 3. Grading Scale Used in Music Performance Task

| **Criterion** | **Rating** | | |
|---|---|---|---|
| | **1** | **2** | **3** |
| Singing the lyrics of the song correctly | | | |
| Singing the tune of the song correctly | | | |
| Paying attention to the rhythm of the song | | | |
| Using his/her voice correctly and effectively | | | |

As a result, grading scales, which consist of four criteria and each of which is specific to the course, were used in each lesson. Grading scales are rated with three different smiley icons in accordance with the age group. The data set was prepared by the researchers as 1-2-3, which is the scoring equivalence of smiley icons.

### *Data Collection Procedure*

An interdisciplinary approach has been used in this study. The issue of helpfulness within the scope of values education has been addressed in relation to the practices in three lessons. It is thought that it will be possible to examine the same subject from the angels of different methods in different disciplines by means of adopting such an approach, and in this way, it will be possible to obtain more detailed data. Moreover, it is aimed to help the students make a healthier assessment by organizing different knowledge and skills to form a meaningful whole and get students gain this meaningful whole. At the same time, it is thought that it will be possible to examine the differences in the perspectives of teachers and students about the assessment of different disciplines.

In the research, the same students' group was asked to do both peer and self-assessment in three different lessons. While the students' group remained the same, it was ensured that different teachers made the assessment in different lessons. In this case, firstly, the teacher and the students were informed about the type of assessment before the research started. Teacher assessments were conducted by two classroom teachers and one music teacher, each working in a public school with expertise and experience in the field. While the teacher of the class in which this study was being conducted made the assessment in social sciences lesson, the teacher of a different class made the assessment in Turkish lesson. In music lesson, the music teacher, who is also one of the researchers of this study, made the assessment. It was decided that music lesson should be conducted by a music teacher who had received a music education as Music lesson requires special skills and the scoring should be done as neutrally as possible. Since the music teacher is one of the researchers of this study, she already has detailed information about the grading scale and the scoring process. On the other hand, the classroom teachers that were to do scoring in Turkish and social sciences lessons were informed about the types of assessment and the grading scales in advance. For this purpose, classroom teachers were given training on how to do assessments using a grading scale, and they were provided with the opportunity to examine exemplary implementations with the researchers. In the process of informing students, short training was given on teacher assessment, self-assessment, peer assessment, and grading scales.

In the Turkish lesson, students were allowed to watch a cartoon film that was telling a fairytale based on the importance of helpfulness. The film was stopped at half, and the students were asked to write the end of the fairytale. After all the students completed their studies, they went to the blackboard one by one and read the rest of the fairytale as they had completed. Since writing rules were also included in the assessment, students' papers were examined by the peer students and the teacher immediately after each student finished reading. In this way, the writing skills of the students were assessed by the teacher and the students.

In the social sciences lesson, students were allowed to watch a short film that was explaining how charity can create a cycle by awakening the sense of helpfulness in people. Then, the students were asked to discuss in groups the positive and negative results that charity could produce based on the events they had watched. Groups of four students were established as they wished and the two groups mutually had the opportunity to discuss the topic. After each group finished discussion, the students' ability to discuss within the group was assessed by using grading scales prepared by teachers and students.

In the music lesson, a song that teaches the importance of helpfulness was taught to students by using ear-to-ear teaching method. Then, the students were asked to sing the song individually. The song performance of the students was assessed by the teacher and the students.

### *Data Analysis*

In this study, it is aimed to determine the reliability of the scores obtained as a result of teacher, self and peer assessment. When the literature is examined; it is clear that Classical Test Theory (CTT), Generalizability (G) Theory and Item Response Theory (IRT) are employed to identify the reliability of the measurement results (Güler, 2011). Especially when it is focused on the studies that try to determine the reliability between different raters, it is seen that G Theory or IRT-based methods have been preferred more frequently compared to CTT (Atılgan, 2005; Börkan, 2017; Büyükkıdık & Anıl, 2015; Farrokhi, Esfandiari & Dalili, 2011; Farrokhi, Esfandiari & Schaefer, 2012; Karakaya, 2015; Matsuno, 2009; Nalbantoğlu-Yılmaz, 2017; Taşdelen-Teker, Şahin & Baytemir, 2016; Yıldıztekin, 2014). If a comparison is made on the basis of CTT and G-theory, it is seen that only one error source is allowed to be estimated in the reliability determination studies based on the CTT, while all error sources can be included in the analysis in the reliability analysis based on G theory. In addition to his; in G theory, the sources of error can be addressed separately and the interactions of error sources can be determined as a result of the analysis (Brennan, 2001; Güler, 2009; Güler, Kaya-Uyanık & Taşdelen-Teker, 2012; Shavelson & Webb, 1991). The study carried out by Taşdelen-Teker and Güler (2019) shows that G Theory is frequently used especially in inter-rater reliability and standard-setting studies. Due to these advantages and application areas of G Theory, in this study, G Theory was preferred in order to determine the reliability between different types of raters (teacher-self-peer).

In accordance with the purpose of this study, by using the students (s) who are the measurement objects and the rater type (r) and criterion (c) variability sources, the analysis was conducted on full crossed random two-facet design (sxrxc). The peer assessment, which is one of the rater types, was included in the analysis and the average score of those given to 25 students by the chosen 5 was taken. For the three courses covered in the study, the predicted variance components, G and Phi coefficients were calculated to determine the main and common effects of the variables that constitute the sources of variability. In addition, G and Phi coefficients were also calculated by using G-facets analysis when the rater types were excluded from the analysis respectively. The analyses were performed using the EduG 6.1 package program.

## RESULTS

In this section, estimated variance components, reliability values and G-Facet components done according to rater type of teacher, self and peer assessment scores are given under separate titles for Turkish, social sciences and music lessons respectively.

### *1. Turkish Lesson*

For the G study of sxrxc pattern which is completely crossed in Turkish lesson; the estimated variance components and percentages of total variance explanation are given as the main effects of s, r and c, and the common effects of sr, sc, rc, and src in Table 4.

Table 4. Estimated Variance Components for the Turkish Lesson

| **Sources of Variance** | **Sum of Squares** | ***df*** | **Mean of Squares** | **Variance ($\sigma^2$)** | **%** |
|---|---|---|---|---|---|
| Student (s) | 77.16853 | 24 | 3.21536 | 0.25745 | 64.0 |
| Rater Type (r) | 4.10027 | 2 | 2.05013 | 0.01712 | 4.3 |
| Criterion (c) | 3.32627 | 3 | 1.10876 | 0.01068 | 2.7 |
| sr | 6.17307 | 48 | 0.12861 | 0.00674 | 1.7 |
| sc | 7.12373 | 72 | 0.09894 | -0.00090 | 0.0 |
| rc | 1.86453 | 6 | 0.31076 | 0.00836 | 2.1 |
| src,e | 14.63547 | 144 | 0.10164 | 0.10164 | 25.3 |
| Total | 114.39187 | 299 | | | 100% |

It is seen in Table 4 that the estimated variance component (0.258) explains the 64.0% of the total variance for the main effect of student (s) in Turkish lesson. The main effect of the student has the biggest share in the total variance. Therefore, it can be concluded that the assessment process can determine the differences between students.

It is also clear that the estimated variance component (0.017) for the main effect of rater type (r) explains 4.3% of the total variance. The main effect of the rater type is the variance component which has the third-largest share in the total variance. According to this, it can be said that the scores given by the teacher, self and peers differ slightly.

It is seen that the estimated variance component (0.011) for the main effect of criterion (c) explains 2.7% of the total variance. In this case, it can be said that the given scores differ slightly from one criterion to another.

When the common effect values are examined, it is seen that the estimated variance component (0.007) for the common effect of student-rater type (sr) explains 1.7% of the total variance. The common effect of the student-rater type (sr) has the second-lowest variance of the total variance. In this case, it can be said that the scores given to students by different types of raters do not change much.

It is seen that the estimated variance component (-0.001) for the common effect of student-criterion (sc) explains 0.0% of the total variance. Student-criterion (sc) common effect has the lowest variance in the total variance, having a negative value. In cases where variance is negative, Cronbach et al. (1972) suggested that the variance value be zero (as cited in Doğan & Anadol, 2017). The reason for the variance being negative may be that the study group is small, or the measurement pattern is not suitable (Taşdelen-Teker et al., 2016). In this study, since there was no problem with the pattern, the finding is thought to be related to the size of the study group. Based on that, when the total variance of the student-criterion (sc) common effect is considered to be zero, it can be said that this effect does not contribute to the total variance. In short, students' performances do not differ according to criteria.

It is seen that the estimated variance component (0.008) for the common effect of rater type-criterion (rc) explains 2.1% of the total variance. This finding shows that there is a slight difference in the scoring from one criterion to the other according to the rater type.

As is seen, student-rater type-criterion (residual) common effect variance component (0.102) explains 25.3% of the total variance. This ratio is the second-largest value in the total variance. However, the share of the student-rater type-criterion (residual) common effect variance component in the total variance is expected to be small (Shavelson & Webb, 1991). As a result, this situation may indicate that the student-rater type-criterion common effect and/or the random error in the measurement can be large.

When G and Phi coefficients are examined, G coefficient is found to be .96 based on relative error variance, and Phi coefficient is found to be .93 based on absolute error variance. It can be said that these values are quite high values within the acceptable limits of the reliability coefficient (Brennan, 2001).

As a result of the G-facets analysis, the reliability coefficients obtained when each of the rater types is not included in the analysis respectively are given in Table 5.

Table 5. G-Facets Analysis of Rater Types

| Facet | Level | G | Φ |
|---|---|---|---|
| Rater Types ($n_r$ = 3) | Teacher Assessment | .92 | .88 |
| | Self Assessment | .97 | .94 |
| | Peer Assessment | .92 | .86 |

As is clear in Table 5, the G and Φ coefficients decrease slightly when the teacher or peer assessments are excluded from the analysis. However, the obtained reliability coefficients are quite high. As a result of excluding the self-assessment from the analysis, both G and Φ coefficients increase.

## *2. Social Sciences Lesson*

For the G study of sxrxc pattern, which is completely crossed in the Social Sciences lesson, the estimated variance components and total variance explanation percentages are given as s, r and main effects and sr, sc, rc, and src common effects in Table 6.

Table 6. Estimated Variance Components for Social Sciences Lesson

| Sources of Variance | Sum of Squares | *df* | Mean of Squares | Variance ($\sigma^2$) | % |
|---|---|---|---|---|---|
| Student (s) | 109.41813 | 24 | 4.55909 | 0.35791 | 74.6 |
| Rater Type (r) | 1.01840 | 2 | 0.50920 | 0.00034 | 0.1 |
| Criterion (c) | 0.59987 | 3 | 0.19996 | -0.00254 | 0.0 |
| sr | 9.62827 | 48 | 0.20059 | 0.03713 | 7.7 |
| sc | 8.33013 | 72 | 0.11570 | 0.02120 | 4.4 |
| rc | 1.95973 | 6 | 0.32662 | 0.01098 | 2.3 |
| src,e | 7.50027 | 144 | 0.05209 | 0.05209 | 10.9 |
| Total | 138.45480 | 299 | | | 100% |

It is seen in Table 6 that the estimated variance component (0.358) explains the 74.6% of the total variance for the main effect of student (s) in social sciences lesson. As a result of obtaining the highest variance ratio from the student variable, it can be concluded that the assessment process can identify the differences between students.

It is seen that the estimated variance component (0.000) for the main effect of the rater type (r) explains 0.1% of the total variance. The main effect of the rater type is the variance component which has the second smallest share in the total variance. According to this, it can be said that the scores given by the teacher, self, and peer show almost no significant difference.

It is observed that the estimated variance component (-0.003) for the main effect of criterion (c) explains 0.0% of the total variance. The main effect of the criterion has the lowest variance in the total variance while it gets a negative value. If the total variance of this variable is considered as zero, it can be said that this effect does not contribute to the total variance. In short, the scoring does not differ according to the criteria.

When the common effect values are examined, it is seen that the estimated variance component (0.037) for the common effect of student-rater type (sr) explains 7.7% of the total variance. The student-rater type (sr) of the common effect has the third-highest variance in the total variance. In this case, it can be said that the scores given to students by the different rater types vary.

It is clear in Table 6 that the estimated variance component (0.021) for the common effect of student-criterion (sc) explains the 4.4% of the total variance. Student-criterion (sc) common effect has the lowest third variance in total variance. As a result, students' performances differ slightly according to the criteria.

It is seen that the estimated variance component (0.011) for the common effect of rater type-criterion (rc) explains 2.3% of the total variance. While this indicates that the rater-criterion (rc) common effect has the lowest third variance value, it can be said that the rater type may differ slightly from criterion to criterion.

Student-rater type-criterion (residual) common effect variance component (0.053) appears to explain 10.9% of the total variance. While this ratio appears to have the second largest value in the total variance, it may be an indicator that the student-rater type-criterion common effect and/or random errors in measurement may be large.

When G and Phi coefficients are examined; both G coefficient and Phi coefficient are found to be .94. It can be said that the obtained relative and absolute reliability coefficients are quite high within the acceptable limits.

As a result of the G-Facets analysis, the reliability coefficients obtained when each of the rater types is not included in the analysis respectively, are given in Table 7.

Table 7. G-Facets Analysis of Rater Types

| **Facet** | **Level** | **G** | **Φ** |
|---|---|---|---|
| Rater Types ($n_r$ = 3) | Teacher Assessment | .91 | .90 |
| | Self Assessment | .97 | .96 |
| | Peer Assessment | .89 | .88 |

As is clear in Table 7, the G and Φ coefficients decrease slightly when the teacher or peer assessments are excluded from the analysis. This decrease was found to be slightly higher in peer assessment, but the obtained reliability coefficients are still quite high. It is seen that both reliability coefficients increased slightly when G and Φ coefficients are excluded from the analysis.

### *3. Music Course*

For the G study of sxrxc pattern which is completely crossed in Music lesson, the estimated variance components and total variance explanation percentages are given as the main effects of s, r; and the common effects of sr, sc, rc, and src in Table 8.

Table 8. Estimated Variance Components for Music Lesson

| Sources of Variance | Sum of Squares | *df* | Mean of Squares | Variance ($\sigma^2$) | % |
|---|---|---|---|---|---|
| Student (s) | 71.25013 | 24 | 2.96876 | 0.21139 | 47.3 |
| Rater Type (r) | 5.35707 | 2 | 2.67853 | 0.02104 | 4.7 |
| Criterion (c) | 0.18600 | 3 | 0.06200 | -0.00453 | 0.0 |
| sr | 17.49627 | 48 | 0.36451 | 0.06021 | 13.5 |
| sc | 13.77067 | 72 | 0.19126 | 0.02253 | 5.0 |
| rc | 2.00400 | 6 | 0.33400 | 0.00841 | 1.9 |
| src,e | 17.80933 | 144 | 0.12368 | 0.12368 | 27.7 |
| Total | 127.87347 | 299 | | | 100% |

It is seen in Table 8 that the estimated variance component (0.211) for the main effect of the student (s) explains 47.3% of the total variance in music lesson. As a result of the scoring performed within the scope of music lesson, it can be concluded that differences between students can be identified.

It is seen that the estimated variance component (0.021) for the main effect of rater type (r) explains 4.7% of the total variance. Considering the main Moreover of the rater type; it can be said that the scores given by the teacher, self and peer vary.

It is observed that the estimated variance component (-0.05) for the main effect of criterion (c) explains 0.0% of the total variance. The main effect of the criterion has the lowest variance in the total variance while it gets a negative value. If the total variance of this variable is considered as zero, it can be said that this effect does not contribute to the total variance. In short, the scoring does not differ according to the criteria.

When the common effect values are examined, it is seen that the estimated variance component (0.060) for the common effect of student-rater type (sr) explains 13.5% of the total variance. In this case; while the student-rater type (sr) has the third-highest variance in the total variance, it can be said that with this finding, the scores given to students by different rater types differ.

While the student-criterion (sc) explains 5.0% of the total variance of the estimated variance component (0.023) for the common effect; it can be said that the scores given to the students differ according to the criteria. Considering the estimated variance component for the rater type-criterion (rc) common effect; it explains 1.9% of the total variance. According to this result; the scores obtained by the rater type according to the criteria differ slightly.

It is seen that the estimated variance component (0.008) for the common effect of rater type-criterion (rc) explains 1.9% of the total variance.

While the student-rater type-criterion (residual) common effect variance component (0.124) explains 27.7% of the total variance, this value is the second largest value in the total variance. Therefore, it can be said that the common effect of student-rater type-criterion and/or random errors in measurement may be large.

When the G and Phi coefficients obtained in the analysis are examined; the G coefficient is found to be .85 and the Phi coefficient is .83. It is seen that the obtained reliability coefficients are within the accepted limits according to the literature (Brennan, 2001).

The reliability coefficients obtained when each of the rater types in G-facets analysis is not included in the analysis respectively, are given in Table 9.

Table 9. G-Facets Analysis of Rater Types

| Facet | Level | G | Φ |
|---|---|---|---|
| Rater Types ($n_r$ = 3) | Teacher Assessment | .63 | .60 |
| | Self Assessment | .97 | .96 |
| | Peer Assessment | .68 | .62 |

In Table 9; there is a significant decrease in the G and Φ coefficients obtained by excluding teacher or peer assessments from the analysis. The obtained reliability coefficients are lower than the acceptable reliability coefficient limit in the literature. There is an increase in both G and Φ coefficients as a result of excluding the self-assessment type from the analysis.

## DISCUSSION and CONCLUSION

This study aims at identifying the reliability of the scores given by third-grade elementary school students through self and peer assessment methods and that of the scores obtained as a result of teacher assessment. An interdisciplinary approach has been adopted for that purpose, and the notion of helpfulness has been associated with Turkish, Social Sciences and Music lessons within the scope of values education.

G-theory was used in the study as it was aimed to include more than one source of error in the analysis and to examine the sources of variance in detail. Thanks to the advantages of the relevant theory, both main and interactive effects of variance sources were examined, and relative as well as absolute reliability coefficients were estimated.

When Turkish lesson is in question, it is seen that the component explaining the total variance is the main effect of the student (s). The fact that the main effect of the student (s) has the largest percentage of explanation is desirable during the assessment process, because it is obtained that the differences between students can be revealed by the assessment process (Atılgan, 2005; Doğan & Anadol, 2017; Taşdelen-Teker et al., 2016). It is seen that the total variance is the second mostly explained component by the residues (src,e) following the main effect of the student (s). This result may be an indicator that the common effect of student-rater type-criterion (src,e) and/or random errors may be large. The cause of random errors in this lesson can be that students who do not encounter such practices frequently experience a lack of excitement and motivation. Considering the main effect of the criterion (c) variable; it is seen that it explains 2.7% of the total variance. When evaluated in terms of criteria, it can be said that student and teacher perspectives differ in some of the criteria within the scope of writing skills. Another noteworthy finding obtained in the context of the Turkish lesson is that the common effect of student-criterion (sc) does not contribute to the total variance. In short, students' performances do not differ according to the criteria included in the grading scale. In this case, it can be said that these criteria assess the same skills.

When the results related to the social sciences lesson are considered, the main component explaining the total variance was the student (s) main effect, and after that, the largest share in explaining the total variance belongs to residues (src,e). It can be said that differences between the students can be revealed in the assessments made within the scope of social sciences lesson with the biggest share of the main effect of the students. The sources of random errors that may occur in this lesson are thought to be that there might be distractions and noise generated in the classroom by the students who did not participate in the activity. The main effect of rater type (r) on estimated variance values of social sciences lesson has a relatively small share in total variance. In other words, it can be said that the scores given in teacher, self and peer assessments show almost no significant difference. In the research, considering that the teacher's assessment of the social sciences lesson is done by that classroom's teacher, the result obtained is thought to be based on the fact that the teacher knows the students in the classroom better and that the students can score more easily in an assessment environment made by the classroom teachers.

When the results related to the estimated variance components within the scope of music lesson are considered, the main effect of the student (s) is the component that explains most of total variance in music lesson as is the case in other lessons. In this respect, differences among the students have been revealed in the assessment made in Music lesson. In addition, the effect of residues (src,e) in explaining the total variance has the largest share following the main effect of the student (s). Among the reasons why residues in music lessons have a high share, the reaction from the class during the individual performance of some students, and the excitement of students unfamiliar with individual performance can be included as the sources of random errors. Another remarkable finding obtained in the context of music lesson is that the main effect of the measure (c) does not contribute to the total variance; in other words, the scoring does not differ according to the criteria. This situation can be explained by the fact that all of the criteria are directed towards singing skills and the level of musical ability of the students has the same effect on the skills related to the criteria.

When G and Φ coefficients are examined, it is seen that G and Φ coefficients obtained for all three lessons are considerably higher than the acceptable value of .80 in the literature (Brennan, 2001). When the G and Φ coefficients are handled on lesson base, it is seen that the coefficients obtained in music lessons are lower than the coefficients obtained in Turkish and Social Sciences lessons. The coefficients obtained in the Turkish and Social Sciences lessons are above .90, and they are very close to each other. In the study, the fact that the teacher assessments in Turkish and social sciences were made by the classroom teachers and the assessment in music lesson was conducted by the music teacher can be considered as a factor affecting the reliability.

When the values obtained as a result of G-facet analysis of rater types are evaluated, if teacher and peer assessments are not included in the analysis for all three lessons, G and Φ coefficients decrease. While the new G and Φ coefficients obtained as a result of these decreases are still higher than the acceptable reliability coefficient for Turkish and social sciences lessons, they are below the acceptable limits for music lesson. When the scores obtained at the end of self-assessment were not added to the analysis, G and Φ coefficients obtained in all three lessons increased. This increase was more in music lesson than it was in other lessons. In the inclusion of peer assessment scores in analysis, the scores of the five raters were averaged. In short, the five raters acted as if they were one rater. In this case, even if one of the peers had not scored very accurately, it may have increased the reliability with the average of the others. But in self-assessment, students may have scored in favor of themselves because they only scored for themselves. When the age characteristics of the students are taken into consideration, instead of exhibiting a biased behavior by giving higher scores to their friends, they are thought to be as careful as possible. In this case, peer assessment and teacher assessment can be expected to be close to each other while self-assessment can be expected to be different from them. A similar result was observed in Salmaner's (2015) study, which examined self, teacher, and peer scores with the multi-surface Rasch measurement model. In this study, Salmaner worked with 5th grade students, and as a result of the analysis, he found out that the most generous raters were self-raters and the strictest raters were teachers or peer raters. When the age group is taken into consideration, it can be said that students' desire to succeed or the anxiety of failure might have created a tendency to give themselves higher scores.

When the literature is examined, it is observed in the studies carried out on the comparison of teacher, peer, and self-assessment at primary school level that students cannot make fully objective assessments; it is generally seen that self-assessments give the most generous scores (Salmaner, 2015; Sarıtaş, 2015). Börkan (2017) also scored the presentation performance of the students by using a grading key in a four-day peer review study with university students. As a result of the study, it was concluded that peer raters generally rated their friends in a very generous manner; and the strictness/generosity levels differed from each other when the raters were compared among themselves. Matsuno (2009) conducted another study in which peer assessment and self-assessment were handled together with teacher assessment. Matsuno (2009) conducted this study with 91 Japanese students between the ages of 19-21 and four teachers. In this study, especially high-performing students gave lower scores than estimated in the self-assessment process, whereas the raters were more tolerant and consistent in the peer assessment process. Regardless of their writing skills, they scored low on high-performing students and high on low-performing students. It was seen that most of the peer raters were consistent and showed less biased interactions than the self-assessment and teacher raters. Farrokhi et al. (2011) conducted a study to determine the tendency of centralism in self-assessment, peer assessment and teacher assessment using the multi-faceted Rasch model. 194 evaluators assessed 188 written compositions with a six-analytical scale and concluded that there was a centrality among peers and self-evaluation. In 2015, Karakaya made a comparison between self-evaluation, peer evaluation and teacher evaluations in evaluating portfolio files of teacher candidates. The findings of the study indicated that the raters were more tolerant in the self-assessment and more rigid in the peer assessment, and it generally found a statistically significant difference between the evaluators. In another study conducted by Nalbantoğlu-Yılmaz (2017) with 56 teacher candidates, it was aimed to determine whether there were differences in self-evaluation and peer evaluations related to a project and to reveal the reliability of the grades given by the teacher candidates and their peers and the scores given by their teachers. As a result of the study, no significant difference was found between the evaluators. It showed that the reliability of self-assessment, peer-assessment and teacher assessments were within acceptable limits.

As is seen in the study results, students should be provided with more opportunities to assess their own works and works of their peers; thus, they should be encouraged to make use of high-level thinking skills such as critical thinking and problem-solving. In this study, a different teaching method called case method was used in order to provide the students with the opportunity to use what they have learned in their daily life, and hence, help them internalize what they have learnt and turned them into a part of permanent learning. Also, the students were asked to make use of alternative assessment skills such as self-assessment and peer-assessment, and thus, the effort made by the students to understand the learning processes deeply was revealed at the end of the study.

The findings of the study show that there should be more space for activities to develop high-level thinking skills such as discussion, critical thinking, and problem-solving which support students' self-assessment and peer-assessment skills. It should be given importance to provide the students with these skills at an early age and to educate individuals who can think scientifically. In addition to the case studies conducted to improve self-assessment and peer-assessment skills, different practices such as problem-based learning and project-based learning should be included more in the curriculum. The interdisciplinary link should be established to contribute to the more effective implementation of curricula.

Choosing the teaching methods appropriate to the level of the students can enable the students to use their self-assessment and peer assessment skills more efficiently. By taking into account the characteristics of student development, appropriate assessment criteria should be determined together with the students to learn the subject. For this purpose, students should have more information about alternative assessment methods. Students should be given performance tasks for self-assessment and peer-assessment, and they should take responsibility for and develop an awareness of their learning.

In this study, the reliability of the rater types in Turkish, social sciences and music lessons was investigated based on an interdisciplinary approach. In different studies, course types and grade levels can be changed, and all teacher assessments can be made by the same teacher as well. The results of such a study can reduce the sources of error that would interfere with the comparison between lessons. In the study, the size of the study group was determined to be 30, but similar studies can be repeated on larger groups of students. The reasons for the low reliability values obtained in music lesson in this study can be examined in detail in different studies.

## REFERENCES

Alıcı, D. (2010). Öğrenci performansının değerlendirilmesinde kullanılan diğer ölçme araç ve yöntemleri. S. Tekindal (Ed.), *Eğitimde ölçme ve değerlendirme* (2. Baskı) içinde (ss. 127-168). Ankara: Pegem Akademi.

Atılgan, H. (2005). Genellenebilirlik kuramı ve puanlayıcılar arası güvenirlik için örnek bir uygulama. *Eğitim Bilimleri ve Uygulama Dergisi*, 4(7), 95-108. Retrieved from http://www.ebuline.com/pdfs/7Sayi/7_6.pdf

Bahar, M. (2006). *Fen ve teknoloji öğretimi*. Ankara: Pegem A Yayıncılık.

Bahar, M., Nartgün, Z., Durmuş, S., & Bıçak, B. (2008). *Geleneksel-alternatif ölçme ve değerlendirme öğretmen el kitabı*. Ankara: Pegem A Yayıncılık.

Ballantyne, R., Huges, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment and Evaluation in Higher Education*, *27*(5), 427-441. doi: 10.1080/0260293022000009302

Börkan, B. (2017). Akran değerlendirmesinde puanlayıcı katılığı kayması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *8*(4), 469-489. doi: 10.21031/epod.328119

Boud, D. (1986). *Implementing student self-assessment*. Sydney: Higher Education Research and Development Society of Australasia.

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

Büyükkıdık, S., & Anıl, D. (2015). Performansa dayalı durum belirlemede güvenirliğin genellenebilirk kuramında farklı desenlerle incelenmesi. *Eğitim ve Bilim*, *40*(177), 285-296. doi: 10.15390/EB.2015.2454

Çeçen, M. A. (2011). Türkçe öğretmenlerinin seviye belirleme sınavı ve Türkçe sorularına ilişkin görüşleri. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 8*(15), 201-211. Retrieved from http://www.acarindex.com/dosyalar/makale/acarindex-1423909379.pdf

Cihanoğlu, M. O. (2008). *Alternatif değerlendirme yaklaşımlarından öz ve akran değerlendirmenin işbirlikli öğrenme ortamlarında akademik başarı, tutum ve kalıcılığa etkileri* (Yayımlanmamış doktora tezi). Dokuz Eylül Üniversitesi Eğitim Bilimleri Enstitüsü, İzmir.

Cram, B. (1995). Self-assessment: From theory to practice. Developing a workshop guide for teachers. In G. Brindley (Ed.), *Language assessment in action* (pp. 271-350). Sydney: National Centre for English Language Teaching and Research, Macquerie University.

Doğan, C. D., & Anadol, H. Ö. (2017). Genellenebilirlik kuramında tümüyle çaprazlanmış ve maddelerin puanlayıcılara yuvalandığı desenlerin karşılaştırılması. *Kastamonu Eğitim Dergisi*, *25*(1), 361-372. Retrieved from https://dergipark.org.tr/tr/pub/kefdergi/issue/27737/309180

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self assesments. *Assesment and Evaluation in Higher Education, 11*(2), 146-166. doi: 10.1080/0260293860110206

Falchikov, N. (2001). *Learning together; Peer tutoring in higher education*. London: Routledge-Falmer.

Farrokhi, F., Esfandiari R., & Dalili, M. V. (2011). Applying the many-facet rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal*, *15*(Innovation and Pedagogy for Lifelong Learning), 70-77. Retrieved from https://pdfs.semanticscholar.org/dd21/ba5683dde8b616374876b0c53da376c10ca9.pdf

Farrokhi, F., Esfandiari R., & Schaefer, E. (2012). A many-facet rasch measurement of differential rates severity/leniency in three types of assessment. *JALT Journal*, *34*(1), 79-102. Retrieved from https://pdfs.semanticscholar.org/d79d/75e55050f9b977ffecd079ba5aadcdc10443.pdf?_ga=2.184016357.2134357192.1569916691-1527179006.1569916691

Güler, N. (2009). Generalizability theory and comparison of the results of g and d studies computed by SPSS and Genova packet programs. *Education and Science*, *34*(154), 93-103.

Güler, N. (2011). Rasgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramı'na göre güvenirliğin karşılaştırılması. *Education and Science*, *36*(162), 225-234.

Güler, N., Kaya-Uyanık, G., & Taşdelen-Teker, G. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi.

İşman, A., & Eskicumalı, A. (2003). *Eğitimde planlama ve değerlendirme* (4. Baskı). İstanbul: Değişim Yayınları.

Karakaya, İ. (2015). Comparison of self, peer and instructor assessments in the portfolio assessment by using many facet rasch model. *Journal of Education and Human Development, 4*(2), 182-192. doi: 10.15640/jehd.v4n2a22

Kurudayıoğlu, M., Şahin Ç., & Çelik, G. (2008). Türkiye'de uygulanan Türk edebiyatı programındaki ölçme ve değerlendirme boyutu uygulamasının değerlendirilmesi: Bir durum çalışması. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 9*(2), 91-101. Retrieved from http://kefad.ahievran.edu.tr/InstitutionArchiveFiles/f44778c7-ad4a-e711-80ef-00224d68272d/d1a3a581-af4a-e711-80ef-00224d68272d/Cilt9Sayi2/JKEF_9_2_2008_91_101.pdf

Kutlu, Ö., Doğan, D., & Karakaya, İ. (2008). *Öğrenci başarısının belirlenmesi, (performansa ve portfolyoya dayalı durum belirleme)*. Ankara: Pegem Akademi.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75-100. doi: 10.1177/0265532208097337

McMillan, H. J. (2015). *Sınıf içi değerlendirme.* (Çev: A. Arı). Ankara: Pegem A Yayıncılık.

Milli Eğitim Bakanlığı. (2013). *İlköğretim Türkçe 3 öğretmen kılavuz kitabı*. Ankara: Milli Eğitim Bakanlığı.

Milli Eğitim Bakanlığı. (2017a). *İlkokul hayat bilgisi öğretmen kılavuz kitabı 3. sınıf*. Ankara: Milli Eğitim Bakanlığı.

Milli Eğitim Bakanlığı. (2017b). *İlköğretim müzik 4 öğretmen kılavuz kitabı*. Ankara: Milli Eğitim Bakanlığı.

Mistar, J. (2011). A study of the valıdıty and relıabılıty of self-assessment. *Teflin Journal*, *22*(1), 45-58. Retrieved from http://journal.teflin.org/index.php/journal/article/viewFile/18/20

Nalbantoğlu-Yılmaz, F. (2017). Reliability of scores obtained from self-, peer-, and teacher-assessments on teaching materials prepared by teacher candidates. *Educational Sciences: Theory & Practice*, *17*(2), 395-409. doi: 10.12738/estp.2017.2.0098

Osterman, K. F., & Kottkamp, R. B. (1993). *Reflective practice for educators: Improving schooling through professional development*. Newbury Park, CA: Corwin Press.

Race, P. (2001). *A briefing on self, peer and group assessment*, Retrieved from https://blogs.shu.ac.uk/teaching/files/2016/09/id9_briefing_on_self_peers_and_group_assessment_snas_901.pdf

Salmaner, R. (2015). *Yazma becerilerinin değerlendirilmesinde öz akran ve öğretmen puanlarının çok yüzeyli rasch ölçme modeliyle incelenmesi* (Yayımlanmamış yüksek lisans tezi). Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Sarıtaş, S. (2015). *Problem çözme becerilerinin değerlendirilmesinde öz, akran ve öğretmen puanlarının çok yüzeyli Rasch ölçme modeli ile incelenmesi* (Yayımlanmamış yüksek lisans tezi). Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. USA: Sage Publications.

Stiggins, J. R. (1997). *Student-centered classroom assessment*. New Jersey, NJ: Merrill, Prentice Hall, Inc.

Stiggins, R., & Chappius, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory into Practıce,* *44*(1), 11–18. Retrieved from https://www.jstor.org/stable/3496986?seq=1#metadata_info_tab_contents

Sünbül, A. M. (2007). *Öğretim ilke ve yöntemleri*. Konya: Çizgi Kitabevi.

Taşdelen-Teker, G., & Güler, N. (2019). Thematic content analysis of studies using generalizability theory. *International Journal of Assessment Tools in Education*, *6*(2), 279-299. doi: 10.21449/ijate.569996

Taşdelen-Teker, G., Şahin, M. G., & Baytemir, K. (2016). Using generalizability theory to investigate the reliability of peer assessment. *Journal of Human Sciences*, *13*(3). 5574-5586. Retrieved from https://j-humansciences.com/ojs/index.php/IJHS/article/view/4155/2035

Tekindal, S. (2014). *Okullarda ölçme ve değerlendirme yöntemleri* (4. Basım). Ankara: Nobel Akademik Yayıncılık.

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment and Evaluation in Higher Education*, *25*(2), 149-169. doi: 10.1080/713611428

Turgut, M. F., & Baykul, Y. (2015). *Eğitimde ölçme ve değerlendirme* (7. Baskı). Ankara: Pegem Akademi.

Wilson, J., & Jan, W. L. (1993). *Thinking for themselves: Developing strategies for reflective learning*. Australia: Eleanor Curtain Publishing.

Woolfolk, A. (2002). *Educational psychology*. New York, NY: Pearson.

Yaşar, M. (2017). Ölçme ve değerlendirmenin önemi. S. Tekindal (Ed.), *Eğitimde ölçme ve değerlendirme* (5. Baskı) içinde (ss. 2-8). Ankara: Pegem Akademi.

Yıldıztekin, B. (2014). *Klasik test kuramı ve genellenebilirlik kuramından puanlayıcılar arası tutarlılığın farklı yöntemlere göre karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(3); 422-434

# Adaptation of Work-Related Rumination Scale into Turkish *

Bilge SULAK AKYÜZ **    Sema SULAK ***

**Abstract**
The aim of the present study was to investigate the validity and reliability of the Turkish version of Work-Related Rumination Scale (T-WRRS). The study was conducted sampling 582 white-collar workers from various fields. In order to determine the construct validity, confirmatory factor analysis was conducted. Additionally, Cronbach Alpha values as an indicator of internal consistency and item-total correlations were utilized for reliability analysis. The results yielded that the Turkish version of WRRS is a reliable scale with three-factor, and it can be used to measure work-related rumination among Turkish workers.

*Key Words:* Work-related rumination, rumination, validation study, CFA.

## INTRODUCTION

Throughout a workday, individuals encounter various emotional, cognitive, and physical demands. At the end of a workday, individuals might feel emotional fatigue due to consuming all the energy levels. In order to reoperate the next day, individuals need to rest and replenish their energy level. After work, time needs to be for individuals to disengage from duties related to work. However, for some individuals, this activity cannot be accomplished as a result of high demands. The process to interfere with successful disengagement from work is called rumination (Cropley, Dijk, & Stanley, 2006; Roger & Jamieson 1988). Previous research in relation to rumination has mainly derived from clinical psychology, and the focus was predominantly on the emotional feature of rumination. Nolen-Hoeksema, Wisco, and Lyubomirsky (2008) defined *rumination* as a recurring thinking process that focuses on distress symptoms and attention is given to the feelings related to the issues. In addition, Martin and Tesser (1996) defined *rumination* as “a class of conscious thoughts that revolve around a common instrumental theme, and that recur in the absence of immediate environmental demands requiring the thoughts” (as cited in Cropley & Zijlstra, 2011, p. 6). Taken together it can be said that *rumination* can be mainly about issues related to self, stressful events, or psychological symptoms one has. *Rumination* is giving attention to the symptoms/stressors, focusing on the possible reasons and outcomes of these symptoms / stressors. Previous studies indicated that rumination was related to several psychological problem such as depression (Lyubomirsky, Caldwall, & Nolen-Hoeksema, 1998; Thomsen, Mehlsen, Christensen & Zachariae, 2003), anxiety (Mellings & Alden, 2000), anger (Hogan & Linden, 2004), poor sleep quality (Thomsen et al., 2003), and somatic symptoms (Brosschot & Van Der Doef, 2006).

Although research in relation to how individuals ruminate about work has not been studied until recently, occupational psychology has given attention to this phenomenon. Sonnentag and Bayer (2005) said occupational psychology focused on thinking about work during leisure time and assessed the detachment from work. Cropley and Zijlstra (2011) speculated that unlike traditional rumination, which was mainly about emotional aspects, work-related *rumination* includes both affective and cognitive aspects. In general, when individuals ruminate, they tend not to have solutions for the problems they have (Nolen-Hoeksema,1987); however, Cropley and Zijlstra opposed to this indicating ruminating about problem(s) can be helpful for individuals. In line with growing interest on this topic, Cropley and Zijlstra (2011) defined *work-related rumination* as “Work-related rumination may be

* This study was funded by Bartin University Coordination of Scientific Research Projects. Project no: 2016-SOS-A-018.
** Assist. Prof. PhD., Bartın University, Faculty of Education, Bartın-Turkey, bsakyuz@bartin.edu.tr, ORCID ID: 0000-0003-3044-3622
*** Assist. Prof. PhD., Bartın University, Faculty of Education, Bartın-Turkey, semasulak@gmail.com, ORCID ID: 0000-0002-2849-321X

To cite this article:
Sulak-Akyüz, B., & Sulak, S. (2019). Adaptation of work-related rumination scale into Turkish. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 422-434. doi: 10.21031/epod.542872
*Received: 21.03.2019*
*Accepted: 06.08.2019*

considered as a thought or thoughts directed to issues relating to work, that is / are repetitive in nature" (p. 6). Individuals ruminate about work in relation to tasks that were not completed, problems that were not solved, and issues that were not clarified with colleagues (Querstret & Cropley, 2012). Thus, *work-related rumination* is not only related to past related issues but also related to future-oriented demands / issues. Considering the fact that work and work-related tasks take more than one-third of a day (Cropley & Zijlstra, 2011), it is expected for individuals to ruminate about work and work-related issues. Hence *work-related rumination* has traits of both traditional rumination due to focusing on past issues as well as traits of worry due to focusing on futuristic events / issues (Flaxman, Menard, Bond & Kinman, 2012).

Over the years, researchers attempted to explore work-related rumination via various instruments. In an instrument developed by Warr (1990), there is a subscale aiming at investigating work strain. After more than a decade, Cropley and Millward-Purvis (2003) developed a three items measure that explores the switching off from work process. In the following years, Sonnentag and Fritz (2007) constructed and proposed an instrument, and one of the sub-scales of the instrument addressed detachment from work. Even though previous research supported the idea that *work-related rumination* has negative consequences, Cropley and Zijlstra (2011) argued otherwise indicating "However thinking and reflecting about work issues can also have beneficial effects and can be associated with positive connotations" (p. 10). As a result, the authors further proposed three distinct types of work-related rumination, which are affective rumination, problem-solving pondering, and detachment. Affective rumination is described as thinking negatively, disturbingly, and persistently about work, which manifests unwanted emotions (Pravettoni, Cropley, Leotta & Bagnara, 2007). Problem-solving pondering, on the other hand, is prolonged thinking about a work-related problem or evaluating solutions on how it can be improved that does not evoke emotional arousal. Finally, detachment is the ease to leave work behind (Cropley & Zijlstra, 2011). In 2012, Cropley, Michalianou, Pravettoni, and Millward utilized this three-factor conceptualization and developed a work-related rumination questionnaire. The aim of the questionnaire is to investigate how people think about work-related issues (Cropley & Zijlstra, 2011).

The aforementioned questionnaire was utilized in several researches. In a study aiming at investigating the relationship between work-related rumination, sleep quality, and work-related fatigue, the three factors structure of the instrument was supported (Querstret & Cropley, 2012). Moreover, affective rumination factor was confirmed via a study investigating the impacts of work-related rumination and recovery on sleep and workplace incivility (Demsky, Fritz, Hammer & Black, 2018). While work-related rumination questionnaire was widely utilized in English, it was translated into other languages. Syrek Weigelt, Peifer and Antoni (2017) conducted a study using the German translation of work-related rumination questionnaire that examined the indirect link between unfinished tasks and sleep by affective rumination and problem-solving pondering. Moreover, in another study aiming at investigating how affective rumination and problem-solving pondering impact overall wellbeing, the Persian translation of work-related rumination questionnaire was utilized (Firoozabadi, Uitdewilligen, & Zijlstra, 2018). According to the results of these two studies, affective rumination and problem-solving are two distinct factors.

### *Purpose of the Study*

Several rumination instruments have been translated into Turkish (Erdur-Baker & Bugay, 2010; Erdur-Baker & Bugay, 2012; Karatepe, Yavuz & Türkcan, 2013); however, these translated instruments mainly focused on traditional rumination that focuses on experiences happened in the past and mostly on distress symptoms of individuals, namely emotional aspects of rumination. However, *work-related rumination* is a combination of both past and future-oriented rumination. As a result, utilizing these instruments to assess work-related rumination can be detrimental. There might be several triggers in relation to work-related rumination. Querstret and Cropley (2012) indicated that some individuals think about unfinished tasks while others ponder about a problem that needs to be addressed, and others might evaluate unwanted issues at work or their relationship with their colleagues. Previous research has been conducted in relation to work-related rumination and various other variables; such as sleep disturbances (Cropley et al., 2006; Querstret, Cropley, & Fife-Schaw, 2016; Querstret, Cropley, Kruger & Heron, 2015; Syrek et al., 2017), fatigue (Querstret & Cropley, 2012; Querstret et al., 2015; Querstret et al., 2016), exhaustion (Donahue et al., 2012; Firoozabadi et al., 2018), depression (Hamesch, Cropley & Lang, 2014), cortisol level (Cropley Rydstedt, Devereux, and Middleton, 2013; Rydstedt, Cropley, Devereux & Michalianou, 2009), well-being (Firoozabadi et al., 2018; Hamesch et al., 2014; Querstret & Cropley, 2012; Syrek et al., 2017), work stressors (Hamesch et al., 2014), work beliefs (Zoupanou, Cropley, & Rydstedt, 2013), unwinding process (Cropley & Millward, 2009), and job strain (Cropley et al., 2006; Cropley & Millward-Purvis, 2003). Thus, in the absence of a Turkish Work-Related Rumination Scale (T-WRRS), it is not possible to garner further information about Turkish workers' rumination traits. Moreover, work-related rumination is a recent phenomenon in literature, and there is no known study in Turkish literature in relation to work-related rumination. Hence, it is crucial to translate and adapt the WRRS into Turkish in order to explore possible underlying and associated factors that are related to work-related rumination. Therefore, the aim of the current study is to translate and adapt work-related scale as well as to examine the factor structure of the scale with Turkish sample. Additionally, this study will contribute to the body of research by adding an instrument that can be utilized by researchers in this field.

## METHOD

This study aimed at translating work-reated rumination scale into Turkish. In this section the participants, data collection procedure, data collection tool, and the data analysis were described.

### *Participants*

A total of 582 while-collar workers were included in the study. The demographics of participants were shown in Table 1.

Table 1. Demographic Properties of Participants

| | | N | % |
|---|---|---|---|
| Gender | Female | 262 | 45.0 |
| | Male | 320 | 55.0 |
| Organization | Public | 294 | 50.5 |
| | Private | 288 | 49.5 |
| Age (M ± S.D.) | | 35.64 ± 9.995 | |
| Daily working hour (M ± S.D.) | | 9.10 ± 2.721 | |
| Year of work (M ± S.D.) | | 10.45 ± 9.392 | |

Cropley et al. (2012) specified white-collar workers as full-time employees from administration, banking, education, health, information technology, marketing, research/science, retail, human resources, insurance, and consultancy. Current study followed similar path, and the occupation composition of the participants was teacher (17.4%), retail (7.6%), administrator (6.9%), soldier / policeman (6.9%), engineer (6.4%), nurse (5.8%), medical professionals (5.7%), human resources (5.5%), officer (4.8%), doctor (4%), accountant (3.6%), businessman (3.4%), pharmacist (2.7%), information technology specialist (2.6%), attorney (2.2%), banking/finance (2.1%), social worker (1.5%), architect (1.4%), veterinarian (1.2%), faculty (1%), and other (7.4%, i.e. insurance agent, technician, journalist, author, cosmetician, secretary and operator). Participants were predominantly from Bartın. Remaining participants were from other cities of Turkey (İstanbul, Ankara, Amasya, Düzce, Kütahya, Isparta, Samsun, Antalya) and reached out through personal communications via snowballing effect.

## *Data Collection Instrument*

### *Work-related rumination scale*

The scale was developed by Cropley et al. (2012). The factor structure of the work-related scale was tested in a study aiming at investigating the relationship between work-related rumination and food choice. In this study, a total number of 268 participants from administration, banking / finance, consultancy, education, health, human resources, insurance, information technology, marketing, retail, and research / science were sampled. The age of the participants ranged from 19 to 63. The scale has twenty-five questions using a 5-point Likert scale (1 = *very seldom* or *never,* 2 = *seldom,* 3 = *sometimes,* 4 = *often* and 5 = *very often* or *always*). According to the factor analysis, three factors emerged accounting for nearly 70% of the variance with eigenvalues greater than one. Concerning oblimin rotation, the variables having .40 or higher loads were retained; this resulted variables on a single factor (Cropley et al., 2012). The results are presented in Table 2.

Table 2. Work-Related Rumination Scale Factor Loadings

| | **Factor 1** | **Factor 2** | **Factor 3** |
|---|---|---|---|
| Affective Rumination | | | |
| Q1 | **.75** | .12 | -.10 |
| Q15 | **.93** | -.15 | .14 |
| Q9 | **.78** | .05 | -.11 |
| Q7 | **.68** | .06 | -.20 |
| Q5 | **.67** | .19 | -.21 |
| Problem-Solving Pondering | | | |
| Q8 | .26 | **.60** | -.17 |
| Q4 | .40 | **.62** | -.03 |
| Q13 | .29 | **.62** | -.08 |
| Q11 | -.34 | **.86** | .04 |
| Q2 | .06 | **.79** | .02 |
| Detachment | | | |
| Q6 | -.37 | -.20 | **.41** |
| Q10 | .10 | .01 | **.78** |
| Q14 | -.02 | .13 | **.88** |
| Q3 | -.03 | -.01 | **.83** |
| Q12 | -.08 | -.10 | **.78** |
| Eigenvalues | 7.30 | 1.79 | 1.32 |
| % of Explained Variance | 48.72 | 11.97 | 8.82 |
| Cronbach’s Alpha | .90 | .82 | .86 |

Note: Factor loadings > .40 are in boldface. (M. Cropley, personal communication, January 25, 2016)

The final scale had 15 items with three factors each of which had five questions. Among all items only item 6 is reverse coded. The first factor was called “affective rumination” that is defined as emotional experiences of work-related thoughts (e.g. “Do you become tense when you think about work-related issues during your free time?”; “Are you troubled by work-related issues when not at work?”). The second factor was called “problem-solving pondering” which was defined as thinking and reflecting about work-related issues (e.g. “In my free time I find myself reevaluating something I have done at work”, “I find solutions to work-related problems in my free time”). Finally, the third factor was called “detachment” that was defined as the ability to switch off from work (e.g. “Do you find it easy to unwind after work?”, “Do you leave work issues behind when you leave work?’). Cronbach’s Alphas were reported .90 for affective rumination, .82 for problem-solving pondering, and .86 for detachment, respectively (Cropley et al., 2012). Querstret and Cropley (2012) confirmed three factors for the scale, indicating nearly 70% of the variance was explained by three factors. They reported Cronbach’s Alpha .90 for affective rumination, .81 for problem-solving pondering, and .88 for detachment. In a study utilizing German translation of the scale, Syrek et al. (2017) reported Cronbach’s Alphas .91 for affective rumination and .84 for problem-solving pondering. They further indicated two-factor model was better in comparison to one-factor model. According to the results of a study using Persian translation of the scale, Firoozabadi et al. (2018) reported Cronbach's Alphas as .91 and .89 for affective rumination and problem-solving pondering, respectively. The authors further indicated in comparison to one-factor model two-factor model was a better fit.

### *Data Collection Procedure*

Prior to translating the instrument, the required permission was taken from the original author of the scale via e-mail. The original scale was translated into Turkish by three experts. Of the experts one of them is specialized in translation and interpretation, the other one is specialized in English literacy, and the last one is specialized in clinical counseling with good command of English. After the translation was completed, the researchers finalized the Turkish version of the scale. In the next step, back translation into English was conducted by an expert in the field of teaching English as a second language. In order to assess the language compatibility, comprehensibility, and clarity of the items, expert consultation was utilized. Experts recommended using *my work* instead of *work* due to language connotations because in Turkish the word *work* cannot be interpreted as a profession. Another recommendation was to use *thinking on / about* instead of *reevaluating* in order to provide better comprehensibility. Taken into consideration all the recommendations, the scale was finalized, and the pilot study was conducted for reliability and validity.

### *Data Analysis*

In order to test the language validity of the scale, English and Turkish versions were administered to the same participants. As a result, Spearman-Brown correlation coefficient was calculated. Furthermore, construct validity was tested utilizing Confirmatory Factor Analysis (CFA). Finally, for internal consistency Cronbach Alpha was used.

## RESULTS

### *Validity Results*

#### *Language validity*

The original and the Turkish version of the WRRS were administered in three weeks intervals to the same participants (N = 16) who were faculty members and had good English proficiency. Spearman Brown correlation coefficient results yielded that these two administrations were correlated for affective rumination (r = .85; $p < .05$), problem solving pondering (r = .73; $p < .05$) and detachment (r = .62, $p < .05$). This result indicated that the T-WRRS had language validity.

#### *Confirmatory factor analysis*

In order to evaluate whether the statistical analysis met the criteria, confirmatory factor analysis assumptions were tested which were determining missing data and outliers, sample size, multicollinearity, and examining univariate as well as multivariate normality (Tabachnick & Fidell, 2001; Ullman, 2012).

The data was collected from 607 participants, and it was screened for possible coding errors and missing values for the analysis. Of the participants, eleven of them were excluded from the analysis due to having inaccurate information. Moreover, fourteen outliers were detected and removed from the data set utilizing box plots. Hence, a total of 582 participants were included in the analysis.

Despite there is no consensus regarding what constitutes adequate sample size for CFA; Klein (2005) said that the parameter and observation ratio needs to be at least 10:1, and Worthington and Whittaker (2006) said that sample size 300 ≥ is acceptable. Thus, sample size (N = 582) is adequate for conducting CFA.

In order to test multicollinearity assumption, VIF and tolerance (T) indices were utilized. In the data set VIF value was found to be lower than 10, and T value was different than zero. This result was indicative of no multicollinearity (Hair, Black, Babin, Anderson & Tatham, 2014).

Concerning normality, the univariate normality assumption was tested utilizing skewness and kurtosis values as well as their critical ratios. According to the results, skewness values ranged from -0.569 to 0.498 and kurtosis values ranged from -1.111 to -0.363. Schumacker and Lomax (2004) indicated that if skewness and kurtosis values are between ± 1.5, the data is distributed normally. This result indicated a normal distribution. Furthermore, maximum likelihood estimation method requires multivariate normally distributed data (Bollen, 1989 as cited in Byrne, 2010; Brown & Moore, 2012; Byrne, 2010). Although there are various measures to test multivariate normality, Mardia's (1970) measure is the widely utilized one. According to Mardia if p values for skewness and kurtosis are greater than .05, multivariate normality is met (Cain, Zhang, & Yuan, 2016). In current study p values were found to be greater than .05, so it can be said the data was clearly multivariate normal.

CFA was conducted sampling 582 participants using IBM SPSS and AMOS 23 software. Firstly, CFA model was created using three factors as latent traits as well as items as observed variables. This model was shown in Figure 1.

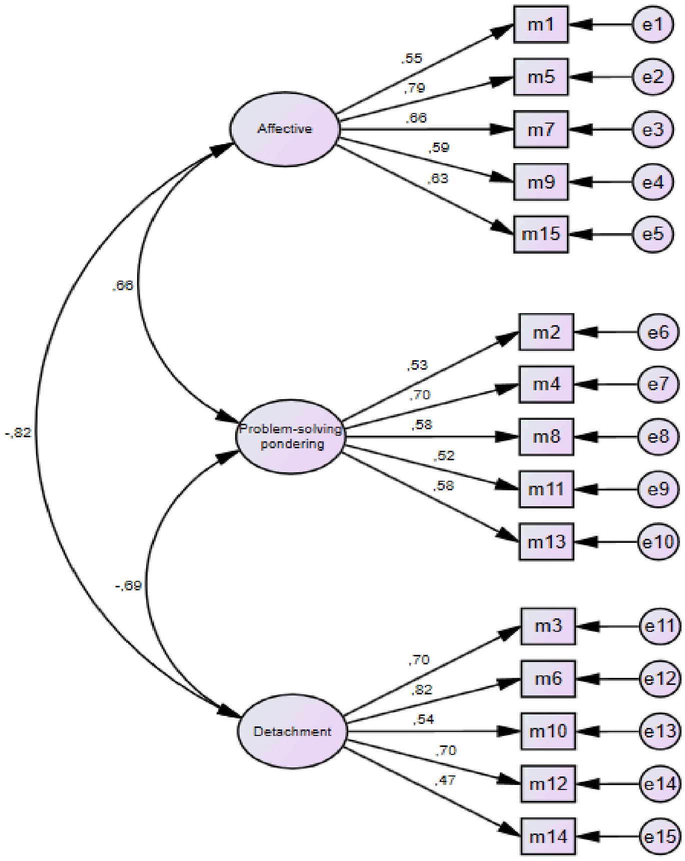


Figure 1. T-WRRS CFA Model

In the second stage, the maximum likelihood method was used in estimating the model. It was aimed to estimate the parameters including the errors of the observed variables, the variances of latent variables, and the regression coefficients related to the paths drawn from the latent variables to the observed variables. Parameter's estimated value, standard error, and critical ratio are given in Appendix A.

Lastly, in order to test the adequacy of model fit, a number of fit indices were used. Several researchers reported good and acceptable fit indices for the adequacy of model fit (Hu & Bentler, 1999; Kline, 2005; Meydan & Sesen, 2011; Tabachnick & Fidell, 2001). These aforementioned fit indices as well as present study's fit indices were presented in Table 3.

Table 3. T-WRRS CFA Model Fit Indices and Criterion Values for Good and Acceptable Fit

| Indices | T-WRRS fit indices | Noble Fit | Acceptable Fit |
|---|---|---|---|
| $\chi^2$/df | 4.04 | $0 \leq \chi^2/df \leq 3$ | $3 < \chi^2/df \leq 5$ |
| GFI | 0.92 | .95 ≤ GFI ≤ 1 | .90 ≤ GFI<.95 |
| IFI | 0.91 | .95 ≤ IFI ≤ 1 | .90 ≤ IFI <.95 |
| TLI | 0.91 | .95 ≤ TLI ≤1 | .90 ≤ TLI <.95 |
| CFI | 0.91 | .95 ≤ CFI ≤ 1 | .90 ≤ CFI < .95 |
| RMSEA | 0.072 | .00 ≤ RMSEA ≤ .05 | .05< RMSEA ≤ .08 |
| SRMR | 0.059 | .00 ≤ SRMR ≤ .05 | .05< SRMR ≤ .10 |

Note: GFI = Goodness of Fit Index, IFI = Incremental Fit Index, TLI = Tucker-Lewis Index, CFI = Comparative Fit Index, RMSEA = Root Mean Square Error of Approximation, SRMR = Standardized Root Mean Square Residual.

When the fit indices for the present study were compared to good fit and acceptable fit indices criterion, it was concluded that the values $\chi^2$/df, GFI, IFI, TLI, CFI, RMSEA, and SRMR met the criterion for acceptable fit.

### *Reliability Analysis*

Reliability of the T-WRRS was examined by assessing the internal consistency coefficient Cronbach's Alpha. The reliability results are shown in Table 4.

Table 4. Reliability Analysis Results for T-WRRS

| Sub-Scale | Item No | Item Total Correlation | Cronbach's Alpha |
|---|---|---|---|
| Affective rumination | Q1 | .51 | .79 |
| | Q5 | .60 | |
| | Q7 | .58 | |
| | Q9 | .56 | |
| | Q15 | .59 | |
| Problem-solving pondering | Q2 | .47 | .73 |
| | Q4 | .50 | |
| | Q8 | .52 | |
| | Q11 | .50 | |
| | Q13 | .45 | |
| Detachment | Q3 | .62 | .79 |
| | Q6 | .65 | |
| | Q10 | .51 | |
| | Q12 | .63 | |
| | Q14 | .43 | |

Nunnally (1978) indicated that the acceptable reliability value is > .70. According to the results, Cronbach's Alphas for affective rumination, problem-solving pondering, and detachment were all above .70, which indicates acceptable reliability. Furthermore, item-total scale correlation of .30 or higher was considered acceptable for each item in the scale (Alpar, 2012; Sencan, 2005). It can be seen in Table 4 that all the item-total correlation coefficients were greater than .30. Hence, all items were retained in the scale.

## DISCUSSION and CONCLUSION

The aim of the study was to adapt the WRRS into Turkish. For this purpose, factor analysis and reliability analysis were utilized. When item analysis was investigated, it was found that all items in the scale had adequate discrimination. According to confirmatory factor analysis results, current study results yielded three factors; affective rumination, problem-solving pondering, and detachment, which was similar to previous research findings (Cropley et al., 2012; Querstret & Cropley, 2012). It can be interpreted that Turkish translation factor structure was consistent with the original factor structure. WRRS was translated into German and Persian. According to current study results, factor structure of the scale was similar to German translation (Syrek et al., 2017) as well as Persian translation (Firoozabadi et al., 2018). It can be said that WRRS can be utilized in different cultural contexts and present psychometrically sound results. The reliability procedure of T-WRRS was carried out by the calculation of internal consistency coefficient (Cronbach Alpha). Similar to previous study findings (Cropley et al., 2012; Firoozabadi et al., 2018; Hamesch et al., 2014; Querstret & Cropley, 2012; Syrek et al., 2017), the results demonstrated high internal consistency estimates for T-WRRS. In sum, it can be said that T-WRRS had adequate psychometric properties and can be utilized in Turkish culture. Additionally, CFA showed adequate model fit for study data providing cross-cultural evidence for the construct validity.

Although future research is required, the current study is assumed to extend the knowledge and research on work-related rumination. The T-WRRS can be utilized by experts in the field of occupational psychology, business, and administration in order to understand and assess workers' work-related rumination traits. Additionally, it is hoped that current results can aid cross-cultural studies. Previous research indicated work-related rumination has several side effects, i.e. fatigue, job strain, and it was suggested that by utilizing T-WRRS these areas, as well as other associations, can be examined in detail. Future research can further knowledge regarding possible associations, antecedents, and consequences of work-related rumination.

Despite the fact that the results of the current study are promising, there are several limitations regarding sampling and analysis. This study sample was limited to white-collar workers. Future research can focus on different samples other than white-collar workers to validate the scale. Moreover, criterion-related validity procedure was not conducted due to the lack of instruments to assess work-related rumination. Hence, further research on the psychometric properties of this scale is needed.

## REFERENCES

Alpar, R. (2012). *Uygulamalı istatistik ve gecerlik-güvenirlik*. Ankara: Detay Yayıncılık.

Brosschot, J. F., & Van Der Doef, M. (2006). Daily worrying and somatic health complaints: Testing the effectiveness of a simple worry reduction intervention. *Psychology & Health*, *21*(1) 19-31. doi: 10.1080/14768320500105346

Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 361-379). New York, NY: Guilford Press.

Byrne, B. M. (2010*). Structural equation modeling with AMOS* (2nd Ed.). New York, NY: Routledge.

Cain, M. K., Zhang, Z., & Yuan, K. H. (2016). Univariate and multivariate skewness and kurtosis for measuring nonnormality: prevalence, influence and estimation. *Behavioral Research Methods*, *49*(5), 1716-1735. doi: 10.3758/s13428-016-0814-1

Cropley, M., & Millward, L. J. (2009). How do individuals "switch-off" from work during leisure? A qualitative description of the unwinding process in high and low ruminators. *Leisure Studies*, *28*(3), 333-347. doi: 10.1080/02614360902951682

Cropley, M., & Millward-Purvis, L. (2003). Job strain and rumination about work issues during leisure time: A diary study. *European Journal of Work Organizational Psychology*, *12*(3), 195-207. doi: 10.1080/13594320344000093

Cropley, M., & Zijlstra, F. (2011). Work and rumination. In J. Langan-Fox, & C. L. Cooper (Eds.). *Handbook of stress in the occupations* (pp. 487-502). Cheltenhm, UK: Edward Elgar Publishing.

Cropley, M., Dijk, D. J., & Stanley, N. (2006). Job strain, work rumination, and sleep in school teachers. *European Journal of Work Organizational Psychology*, *15*(2), 181-196. doi: 10.1080/13594320500513913

Cropley, M., Michalianou, G., Pravettoni, G., & Millward, L. (2012). The relation of post-work ruminative thinking with eating behaviour. *Stress and Health*, *28*(1), 23-30. doi: 10.1002/smi.1397

Cropley, M., Rydstedt, L. W., Devereux, J. J., & Middleton, B. (2013). The relationship between work-related rumination and evening and morning salivary cortisol secretion. *Stress and Health*, *31*(2), 150-157. doi: 10.1002/smi.2538

Demsky, C. A., Fritz, C., Hammer, L. B., & Black, A. E. (2018). Workplace incivility and employee sleep: The role of rumination and recovery experiences. *Journal of Occupational Health Psychology*. *24*(2), 228-240. doi: 10.1037/ocp0000116

Donahue, E. G., Forest, J., Vallerand, R. J., Lemyre, P. N., Crevier-Braud, L., & Bergeron, É. (2012). Passion for work and emotional exhaustion: The mediating role of rumination and recovery. *Applied Psychology: Health and Well-Being*, *4*, 341-368. doi: 10.1111/j.1758-0854.2012.01078.x

Erdur-Baker, O., & Bugay, A. (2010). The short version of ruminative response scale: Reliability, validity and its relation to psychological symptoms. *Procedia-Social and Behavioral Sciences*, *5*, 2178-2181. doi: 10.1016/j.sbspro.2010.07.433

Erdur-Baker, O., & Bugay, A. (2012). The Turkish version of the Ruminative Response Scale: An examination of its reliability and validity. *International Journal of Education and Psychology in the Community*, *10*(2), 1-16.

Firoozabadi, A., Uitdewilligen, S., & Zijlstra, F. R. H. (2018). Should you switch off or stay engaged? The consequences of thinking about work on the trajectory of psychological well-being over time. *Journal of Occupational Health Psychology*, *23*(2), 278-288. doi: 10.1037/ocp0000068

Flaxman, P. E., Menard, J., Bond, F. W., & Kinman, G. (2012). Academics' experiences of a respite from work: Effects of self-critical perfectionism and perseverative cognition on postrespite wellbeing. *Journal of Applied Psychology*, *97*(4), 854-865. doi: 10.1037/a0028055

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2014). *Multivariate data analysis* (7th Ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Hamesch, U., Cropley, M., & Lang, J. (2014). Emotional versus cognitive rumination: Are they differentially affecting long-term psychological health? The impact of stressors and personality in dental students. *Stress and Health*, *30*(3), 222-231. doi: 10.1002/smi.2602

Hogan, B. E., & Linden, W. (2004). Anger response styles and blood pressure: At least don't ruminate about it! *Annals of Behavioral Medicine*, *27*(1), 38-49. doi: 10.1207/s15324796abm2701_6

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structural analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. doi: 10.1080/10705519909540118

Karatepe, H. T., Yavuz, F. K., & Türkcan, A. (2013). Validity and reliability of the Turkish version ruminative thoughts style questionnaire. *Bulletin of Clinical Psychopharmacology*, *23*, 231-241. doi: 10.5455/bcp.20121130122311

Kline, R. B. (2005). *Principle and practice of structural equation modelling.* New York, NY: Guilford.

Lyubomirsky, S., Caldwell, N. D., & Nolen-Hoeksema, S. (1998). Effects of ruminative and distracting responses to depressed mood on retrieval of autobiographical memories. *Journal of Personality and Social Psychology*, *75*(1), 166-177. doi: 10.1037/0022-3514.75.1.166

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519-530.

Mellings, T. M. B., & Alden, L. E. (2000). Cognitive processes in social anxiety: The effects of self-focus, rumination and anticipatory processing. *Behaviour Research and Therapy*, *38*, 243-257. doi: 10.1016/S0005-7967(99)00040-6

Meydan, C. H., & Sesen, H. (2011). *Yapısal esitlik modellemesi AMOS uygulamaları*. Ankara: Detay Yayıncılık.

Nolen-Hoeksema, S. (1987). Sex differences in unipolar depression: Evidence and theory. *Psychological Bulletin*, *101*(2), 259-282. doi: 10.1037/0033-2909.101.2.259

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. *Perspectives on Psychological Science*, *3*(5), 400-424.

Nunnally, J. C. (1978). *Psychometric theory* (2nd Ed.). New York, NY: McGrawHill.

Pravettoni, G., Cropley, M., Leotta, S. N., & Bagnara, S. (2007). The differential role of mental rumination among industrial and knowledge workers. *Ergonomics*, *50*(11), 1931-1940. doi: 10.1080/00140130701676088

Querstret, D., & Cropley, M. (2012). Exploring the relationship between work-related rumination, sleep quality, and work-related fatigue. *Journal of Occupational Health Psychology*, *17*(3), 341-353. doi: 10.1037/a0028552

Querstret, D., Cropley, M., & Fife-Schaw, C. (2016). Internet-based instructor-led mindfulness for work-related rumination, fatigue, and sleep: Assessing facets of mindfulness as mechanisms of change. A randomized waitlist-controlled trial. *Journal of Occupational Health Psychology, 22*(2), 153-169. doi: 10.1037/ocp0000028

Querstret, D., Cropley, M., Kruger, P., & Heron, R. (2015). Assessing the effect of a Cognitive Behaviour Therapy (CBT)-based workshop on work-related rumination, fatigue, and sleep. *European Journal of Work and Organizational Psychology, 25*(1), 1-18. doi: 10.1080/1359432X.2015.1015516

Roger, D., & Jamieson, J. (1988). Individual differences in delayed heart-rate recovery following stress: The role of extraversion, neuroticism and emotional control. *Personality and Individual Differences, 9*(4), 721-726. doi: 10.1016/0191-8869(88)90061-X

Rydstedt, L. W., Cropley, M., Devereux, J. J., & Michalianou, G. (2009). The effects of gender, long-term need for recovery and trait inhibition-rumination on morning and evening saliva cortisol secretion. *Anxiety Stress Coping, 22*(4), 465-474. doi: 10.1080/10615800802596378

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Erlbaum.

Sencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenilirlik ve geçerlilik*. Ankara: Seckin Yayınevi.

Sonnentag, S., & Bayer, U. V. (2005). Switching off men- tally: Predictors and consequences of psychological detachment from work during off-job time. *Journal of Occupational Health Psychology, 10*(4), 393-414.

Sonnentag, S. & Fritz, C. (2007). The recovery experience questionnaire: Development and validation of a measure for assessing recuperation and unwinding from work. *Journal of Occupational Health Psychology*, *12*(3), 204-221. doi: 10.1037/1076-8998.12.3.204

Syrek, C. J., Weigelt, O., Peifer, C., & Antoni, C. H. (2017). Zeigarnik's sleepless nights: How unfinished tasks at the end of the week impair employee sleep on the weekend through rumination. *Journal of Occupational Health Psychology, 22*(2), 225-238 doi: 10.1037/ocp0000031

Tabachnick B. G. & Fidel, L. S. (2001). *Using multivariate statistics* (4th Ed.). Boston, MA: Allyn & Bacon, Inc.

Thomsen, D. K., Mehlsen, M. Y., Christensen, S., & Zachariae, R. (2003). Rumination-relationship with negative mood and sleep quality. *Personality and Individual Differences*, *34*(7), 1293-1301. doi: 10.1016/S0191-8869(02)00120-4

Ullman, J. B. (2012). Structural equation modeling. In B. G. Tabachnick & L. S. Fidel (Eds.), *Using multivariate statistics* (6th Ed.). Boston, MA: Pearson, Inc.

Warr, P. (1990). The measurement of well-being and other aspects of mental-health. *Journal of Occupational Psychology*, *63*(3), 193-210. doi: 10.1111/j.2044-8325.1990.tb00521.x

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806-838. doi: 10.1177/0011000006288127

Zoupanou Z., Cropley M., & Rydstedt L. W. (2013). Recovery after work: The role of work beliefs in the unwinding process. *PLOS ONE, 8*(12), 1-9. doi: 10.1371/journal.pone.0081381

# İşsel Ruminasyon Ölçeğinin Türkçeye Uyarlama Çalışması

## *Giriş*

Mesai bitimindeki zaman bireylerin işleri ile ilgili görev ve sorumluluklarından ayrıştığı bir zaman dilimi olmalıdır. Fakat, birçok birey bu ayrışmayı, yaptığı işin gerekliliklerinden ötürü başaramaz. İşle ilgili düşüncelerden kopamamak *ruminasyon* olarak tanımlanmıştır (Cropley, Dijk & Stanley. 2006; Roger & Jamieson 1988). Ruminasyon alanyazında klinik psikoloji alanında sıklıkla kullanılmış ve genellikle ruminasyonun duygusal yapısından bahsedilmiştir. Nolen-Hoeksema, Wisco ve Lyubomirsky (2008) *ruminasyonu*, stress semptomları ve duygulara odaklanarak tekrar eden düşünme süreci olarak tanımlamıştır. Bireylerin işleri ile ilgili ruminatif halleri alanyazında çok yer almaması sebebiyle endüstri psikolojisi alanı bu kavram üzerine dikkat çekmiştir ve iş ile ilgili ruminasyon *işsel ruminasyon* olarak ele alınmaya başlamıştır. Bireylerin günlerinin üçte birlik kısmını işlerine ayırdıkları göz önüne alındığında (Cropley & Zijlstra, 2011), işle ilgili konularda ruminatif düşüncede olmaları beklenir. Cropley ve Zijlstra (2011) yazdıkları kitaplarında *işsel ruminasyonu* iş/işler ile ilgili tekrar eden düşünce/düşünceler olarak tanımlamışlardır. Alanyazında işsel ruminasyonun ölçülmesi için geliştirilmiş birkaç tane ölçek bulunmaktadır (Cropley ve Millward, 2003; Cropley, Michalianou, Pravettoni & Millward, 2012; Sonnentag & Fritz, 2007; Warr, 1990).

Yapılan çalışmalar işsel ruminasyon konusunun önemli olduğunu ortaya koymuştur. Türkiye'de ruminasyon kavramına ilişkin ölçek uyarlama çalışmaları yapılmıştır (Erdur-Baker & Bugay, 2010; Erdur-Baker & Bugay, 2012; Karatepe, Yavuz & Türkcan, 2013); ancak, bu ölçeklerin ruminasyonun duygusal boyutu ile ilgili olduğu görülmektedir. Cropley ve Zijlstra (2011) geleneksel ruminasyonun aksine, ruminasyonun duygusal boyutu ile ilgili, işsel ruminasyonun hem duygusal hem de bilişsel boyutu olduğunu söylemektedir. Araştırmacıların bu söylemi göz önüne alındığında alanyazında iş ile ilgili ruminatif düşüncelerin incelendiği bir araştırmaya rastlanamamıştır. Önceki araştırmalar işsel rumimasyonun farklı değişkenlerle ilişkisi olduğunu ortaya koymuştur; örneğin, uyku düzensizlikleri (Cropley ve dğerleri, 2006; Querstret, Cropley & Fife-Schaw, 2016; Querstret Cropley, Kruger & Heron 2015; Syrek Weigelt, Peifer & Antoni, 2017), yorgunluk (Querstret ve Cropley, 2012; Querstret ve diğerleri, 2015; Querstret ve diğerleri, 2016), kortizol seviyesi (Cropley Rydstedt, Devereux & Middleton, 2013), iyi oluş hali (Firoozabadi, Uitdewilligen & Zijlstra, 2016; Hamesch, Cropley & Lang, 2014; Querstret ve Cropley, 2012; Syrek ve diğerleri, 2017), iş stresi (Hamesch ve diğerleri, 2014), iş inançları (Zoupanou, Cropley & Rydstedt, 2013), işe bağlılık (Cropley ve Millward, 2009), ve iş gerginliği (Cropley, Millward-Purvis, 2003; Cropley ve diğerleri, 2006). Çalışan bireylerin ruminatif düşüncelerinin ve bu düşüncelerin sonucu olan değişkenlerin belirlenmesi ve iyileştirme çalışmalarının yapılabilmesi için Türkçe bir ölçeğe ihtiyaç duyulmaktadır. Bu araştırmanın amacı, Cropley ve diğerleri (2012) tarafından geliştirilen işsel ruminasyon ölçeğinin Türk kültürüne uyarlamaktır.

### *Yöntem*

Araştırma 582 çalışan üzerinde gerçekleştirilmiştir. Katılımcılar, Cropley ve diğerlerinin (2012) çalışmalarında bahsettiği üzere beyaz yakalı çalışanlardan oluşturulmuştur.

Veri toplama aracı olarak Cropley ve diğerleri (2012) tarafından geliştirilen işsel ruminasyon (İR) ölçeği kullanılmıştır. Toplam 15 madde ve 3 alt boyuttan oluşan ölçek, 5'li Likert tipinde geliştirilmiştir. Ölçekte yer alan birinci, beşinci, yedinci, dokuzuncu ve on beşinci maddeler "duygusal", ikinci, dördüncü, sekizinci, on birinci ve on üçüncü maddeler "problem çözme" ve üçüncü, altıncı, onuncu, on ikinci ve on dördüncü maddeler ise "kopma" alt boyutunu oluşturmuştur.

Araştırmacılar tarafından ölçek Türkçe'ye çevrilmiş ve dil geçerliği çalışmaları yapılmıştır. Ölçeğin dil geçerliğini sağladığı sonucuna varılmıştır. İşsel Ruminasyon Türkçe (İR-T) ölçeğinin yapı geçerliği için doğrulayıcı faktör analizi ve güvenirliğini belirlemek için Cronbach Alfa kullanılmıştır.

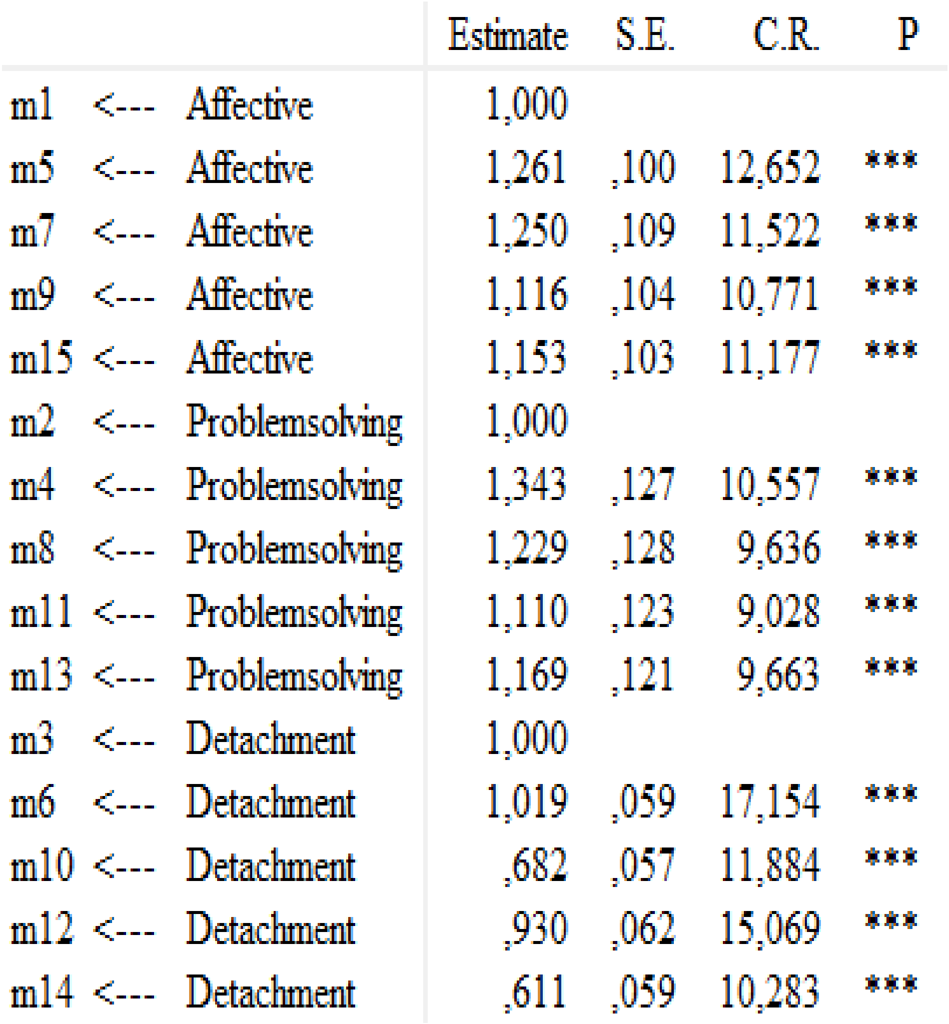
### *Sonuç ve Tartışma*

Dil geçerliği için İR ve İR-T ölçekleri İngilizce okuduğunu anlama becerisine sahip akademisyenlere üç hafta arayla uygulanmış ve her iki uygulama arasındaki Spearman Brown korelasyon katsayısı hesaplanmıştır. Analiz sonucunda duygusal alt boyutu (r = .85; *p* < .05), problem çözme alt boyutu (r = .73; *p* < .05) ve kopma alt boyutunda (r = .62, *p* < .05). ölçeğin dil geçerliğinin olduğu sonucuna varılmıştır.

Doğrulayıcı Faktör Analizinde (DFA) ilk olarak sayıltılar test edilmiştir. 607 katılımcıdan elde edilen veri setinde kayıp veri ve aykırı değer olup olmadığı araştırılmıştır ve 25 katılımcı analizden dışında tutulmuştur. Örneklem büyüklüğü > 300 olduğu için yeterli görülmüştür (Worthington ve Whittaker, 2006). Normallik sayıltısı için öncelikle AMOS'da çarpıklık, basıklık ve kritik değerler incelenmiştir. Çok değişkenli normallik için ise Mardia (1970) tarafından geliştirilen çok değişkenli basıklık değeri hesaplanmıştır ve eldeki verinin çok değişkenli normallik gösterdiği sonucuna varılmıştır (*p* > .05). Çoklu bağlantılılık sayıltısı için varyans artış faktörü (VIF) ve tolerans (T) değerleri incelenmiş ve çoklu bağlantılılık sorunu olmadığı saptanmıştır. DFA yapmak için sayıltıların sağlanmasından sonra, üç faktörün gizil değişken, bu faktörleri oluşturan ifadelerin de gösterge değişken olarak yer aldığı 1. dereceden doğrulayıcı faktör analizi modeli kurulmuştur. İkinci aşamada, model tahminlenirken yapısal eşitlik modellerinde sıklıkla kullanılan ve verilerin normal dağılmadığı durumlarda bile güvenilir sonuçlar veren en çok olabilirlik yöntemi kullanılmış, gözlemlenen değişkenlerin hatalarının, gizil değişkenlerin varyansları ve gizil değişkenlerden gözlenen değişkenlere doğru çizilen yollara ilişkin regresyon katsayılarını kapsayan parametrelerin tahmin edilebilmesi amaçlanmıştır. Son aşamada ise üç faktörlü 1. dereceden oluşturulan doğrulayıcı faktör analizi modeli için uyum indeksleri incelenmiştir. Elde edilen uyum değerlerine bakıldığında, $\chi^2$ / sd (4.04), GFI (.92), IFI (.91), CFI (.91), TLI (.91), RMSEA (.072) ve SRMR (.059) değerlerinin iyi olduğu görülmüş ve işsel ruminasyon ölçeğinin 15 ifadeden oluşan 3 faktörlü yapısının (duygusal, problem çözme, kopma) genel olarak iyi uyum sağladığı görülmektedir. İR-T için elde edilen sonuçlar önceki araştırmalarla (Cropley ve diğerleri, 2012; Querstret ve Cropley, 2012) benzerlik göstermiş ve üç boyut doğrulanmıştır.

İR-T ölçeğinin güvenirliğini belirlemek amacıyla Cronbach Alfa katsayısı hesaplanmıştır. Duygusal, problem çözme ve kopma boyutlarının güvenirlikleri sırasıyla .79, .73 ve .79 olarak hesaplanmıştır. Bu değerler daha önceki araştırmalarla benzerlik göstermektedir (Cropley ve diğerleri, 2012; Firoozabadi ve diğerleri, 2018; Hamesch ve diğerleri, 2014; Querstret ve Cropley, 2012; Syrek ve diğerleri, 2017). Uyarlanan ölçeğin güvenirliğinin olduğu sonucuna varılmıştır. Geçerlik ve güvenirlik çalışmaları sonucunda İR-T ölçeğinin Türkçe adaptasyonunun geçerli ve güvenilir olduğu sonucuna varılmıştır. Yapılan araştırmada İR-T ölçeğinin uygulandığı grup orijinal ölçektekine benzer şekilde beyaz yakalılardan oluşturulmuştur. Türkiye'deki farklı meslek grupları üzerinde de uyarlanan İR-T formunun uygulanması önerilebilir.

## Appendix A: Regression Weights of T-WRRS CFA Model



| | | | Estimate | S.E. | C.R. | P |
|---|---|---|---|---|---|---|
| m1 | <--- | Affective | 1,000 | | | |
| m5 | <--- | Affective | 1,261 | ,100 | 12,652 | *** |
| m7 | <--- | Affective | 1,250 | ,109 | 11,522 | *** |
| m9 | <--- | Affective | 1,116 | ,104 | 10,771 | *** |
| m15 | <--- | Affective | 1,153 | ,103 | 11,177 | *** |
| m2 | <--- | Problemsolving | 1,000 | | | |
| m4 | <--- | Problemsolving | 1,343 | ,127 | 10,557 | *** |
| m8 | <--- | Problemsolving | 1,229 | ,128 | 9,636 | *** |
| m11 | <--- | Problemsolving | 1,110 | ,123 | 9,028 | *** |
| m13 | <--- | Problemsolving | 1,169 | ,121 | 9,663 | *** |
| m3 | <--- | Detachment | 1,000 | | | |
| m6 | <--- | Detachment | 1,019 | ,059 | 17,154 | *** |
| m10 | <--- | Detachment | ,682 | ,057 | 11,884 | *** |
| m12 | <--- | Detachment | ,930 | ,062 | 15,069 | *** |
| m14 | <--- | Detachment | ,611 | ,059 | 10,283 | *** |

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(4); 435-450

# The Turkish Adaptation of the Statistics Anxiety Scale for Graduate Students *

Neşe GÜLER ** Gülşen TAŞDELEN TEKER *** Mustafa İLHAN ****

**Abstract**
In this study, it was aimed to adapt the Statistical Anxiety Scale (SAS) developed for graduate students by Faber, Drexler, Stappert and Eichhorn to Turkish. The research was carried out on 375 students attending graduate education in any field in Turkey. In the study, construct validity of the SAS was investigated via exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Parallel analysis method was also used in making decision about factor number of the scale. In the EFA and parallel analysis, a unidimensional structure was obtained in line with the results acquired in the factor analysis of the original form of the SAS. However; since the original form of the SAS was designed by foreseeing a three-dimensional structure of worry, avoidance and emotionality, both unidimensional and three-dimensional structures were tested in CFA. The fit indices reported in CFA were found to be within acceptable limits for both models. In the reliability analysis, Cronbach Alpha internal consistency coefficient was calculated as .91 for the whole scale, and it was found to be .91, .83, and .91 for worry, avoidance and emotionality dimensions, respectively. It was determined that item correlations exceed the lower limit of .30 for all items in the scale. Ferguson Delta statistic, which provide evidence for the discriminatory power of the entire scale, was determined as .98. These results suggest that the Turkish form of the SAS yields valid and reliable measures.

*Key Words:* Statistics anxiety, graduate students, scale adaptation, validity, reliability.

## INTRODUCTION

One of the most important stages of scientific research process is to analyse the collected data via appropriate methods (Gürbüz & Şahin, 2017). The appropriate method for data analysis differs depending on the way the data is collected and the problems sought in the research. In the most general sense, the data are analysed through descriptive analysis or content analysis if a qualitative study is conducted (Yıldırım & Şimşek, 2016); but statistical techniques are used in the quantitative studies. In this context, a researcher conducting a quantitative study needs to be knowledgeable about statistics. Of course, it does not mean that a researcher conducting qualitative study does not need knowledge of statistics. This is because knowledge of statistics is necessary not only for analysing a researcher's own data but also for following the literature and understanding the conducted studies (Tan, 2016). For this reason, statistics is considered as an instrument complement scientific research (Sutarso, 1992), and anybody doing scientific study is expected to be trained in statistical techniques beside research methods (Erkuş, 2011). Due to this, at least one statistical course is compulsory in almost all of the graduate education programmes in the social, educational, and behavioural sciences. Yet, taking a statistics course can turn into a negative experience for many students attending graduate programmes (Collins & Onwuegbuzie, 2007). Therefore, most students postpone taking statistics related courses as far as possible and prefer taking them at the last semester (Roberts & Bilderbeck, 1980). Such behaviours displayed by students against statistics is referred to as *statistics anxiety*.

* It was orally presented at the 4th International Scientific Research Congress.

** Prof., İzmir Democracy University, Faculty of Education, İzmir-Turkey, gnguler@gmail.com, ORCID: 0000-0002-2836-3132
*** Assist. Prof., Hacettepe University, Faculty of Medicine, Ankara-Turkey, gulsentasdelen@gmail.com, ORCID: 0000-0003-3434-4373
**** Assist. Prof., Dicle University, Ziya Gökalp Faculty of Education, Diyarbakır-Turkey, mustafailhan21@gmail.com, ORCID: 0000-0003-1804-002X

To cite this article:
Güler, N., Taşdelen-Teker, G., & İlhan, M. (2019). The Turkish adaptation of the statistics anxiety scale for graduate students. *Journal of Measurement and Evaluation in Education and Psychology, 10*(4), 435-450. doi: 10.21031/epod.550765

*Received: 09.04.2019*
*Accepted: 01.09.2019*

### Statistics Anxiety

*Statistics anxiety* is described as situational anxiety arising while taking a statistics course or doing the statistical operations such as collecting and analysing the data, and interpreting the outputs of the analyses (Cruise, Cash & Bolton, 1985; Onwuegbuzie, Da Ros, & Ryan, 1997). The study conducted by Onwuegbuzie (2004) reports that approximately 80% of graduate students have statistics anxiety. Statistics anxiety can influence students' ability to comprehend the articles, analyse and interpret the data (Onwuegbuzie, 1997a) and thus their achievement in statistics (Fitzgerald, Jurs & Hudson, 1996; Lalonde & Gardner, 1993; Onwuegbuzie & Seaman, 1995) and research methods courses (Onwuegbuzie, Slate, Paterson, Watson, & Schwartz, 2000), and even whether or not they will graduate from the programme they have enrolled in the long run (Onwuegbuzie, 1997b as cited in Rodarte-Luna & Sherry, 2008).

A review of relevant literature demonstrates that several studies concerning statistics anxiety have been conducted especially in the last 30 years in social sciences (Beurze, Donders, Zielhuis, Vegt & Verbeek, 2013). The remarkable results obtained from relevant studies can be summarized as followings: Students with weak mathematical background or limited education in mathematics have higher statistics anxiety (Baloğlu, 2003; Baloğlu & Zelhart, 2004; Primi & Chiesi, 2018; Roberts & Saxe, 1982; Wilson, 1997; Zeidner, 1991); there are positive correlations between statistics anxiety and tendencies to put off assignments in graduate education (Onwuegbuzie, 2004); students consider statistics as a barrier in front of academic career (Onwuegbuzie, 1997b as cited in Rodarte-Luna & Sherry, 2008); reading skills significantly affect statistics anxiety (Collins & Onwuegbuzie, 2007). The studies intending to determine the effects of such demographic variables as gender and age, on the other hand, has obtained differing findings. Sutarso (1992) found that there were no significant differences between male and female students' statistics anxiety; Baloğlu (2003), Benson (1989) and Rodarte-Luna and Sherry (2008), however, found that female students had significantly higher statistics anxiety than male students. While Beurze et al. (2013) found that statistics anxiety did not differ according to age, Baloğlu (2003) found that there was increase in statistics anxiety through age.

### Measuring Statistics Anxiety

Measurement tools created by using mathematics anxiety scales were used in earlier studies on statistics anxiety (Pan & Tang, 2005). Statistics anxiety scale developed by Pretorius and Norman (1992) and statistics anxiety inventory developed by Zeidner (1991) can be given as examples to such measurement tools (Chiesi, Primi & Carmona, 2011). In later studies, however, it was emphasised that mathematics anxiety and statistics anxiety were related but that they were distinct structures, and thus the validity of statistics anxiety scales prepared with reference to mathematics anxiety scales was questioned (Onwuegbuzie & Wilson, 2003). Thus, scales intended to measure directly statistics anxiety were developed. Of them the most frequently used one is the Statistics Anxiety Rating Scale which was developed by Cruise et al. (1985) and whose psychometrical properties were analysed more recently by Baloğlu (2002); Chew, Dillon and Svinbourne (2018); Hanna, Shevlin and Dempster (2008); Liu, Onwuegbuzie and Meng (2011); Maat and Rosli (2016); Nesbit and Bourne (2018) and Teman (2013). This five-pointed Likert type scale contains 51 items and six subscales labelled as worth of statistics, interpretation anxiety, test and class anxiety, computational self-concept, fear of asking for help, and fear of statistics teachers.

Onwuegbuzie and Wilson (2003) stated in their review study that the Statistical Anxiety Rating Scale (Cruise et al., 1985) was the most known and widely used scale on the subject. However, the fact that this scale was very long in length and also considered constructs such as attitude and self-concept in addition to anxiety (Chiesi et al., 2011) paved the way for studies aiming to develop measurement tools which were more useful and which were to measure only statistics anxiety. One of those studies was performed by Vigil-Colet, Lorenzo-Seva and Condon (2008). The researchers aimed to include in the literature a measurement tool which contained items reflecting only statistics anxiety and which was short enough to use easily. In accordance with their purpose, they developed a 24-item, three-factor (test anxiety, asking for help anxiety and interpretation anxiety) statistics anxiety scale in Spanish sample. Another contemporary measurement tool for statistics anxiety is the 17-item scale developed by Faber, Drexler, Stappert and Eichorn (2018). The scale was developed with the participation of graduate students in educational sciences and in special education. A close examination of the items in the scale makes it clear that the audience is not restricted only to students in the field of education. Hence, the scale is applicable with graduate students in diverse areas who come across statistics in the papers they read or in the research they do.

### *Statistics Anxiety Scales Available in Turkish Literature*

Four different measurement tools are found on searching for the concept of statistics anxiety (istatistik kaygısı) on Turkish pages in Google search engine. One of them is Statistics Attitudes Scale developed by Köklü (1994). The researcher concluded that the scale can be considered as both single factor and four factors as a result of the principal components analysis applied to the statistical attitude scale and called one of the factors in the four-factor scale as statistics anxiety. The second scale was developed by Köklü (1996) and the third one was developed by Yaşar (2014). The one developed by Köklü (1996) is intended directly to measure statistics anxiety. The scale developed by Yaşar (2014), on the other hand, was prepared to measure attitudes towards statistics and statistics anxiety is only one of its five factors. The property in common in the scales developed by Köklü (1994, 1996) and Yaşar (2014) is that they both are directed to undergraduate students and that they do not contain items corresponding to the basic components of graduate education such as reading scientific articles, doing scientific research and presenting it. The fourth measurement tool available on the Turkish pages of Google search engine is the statistics anxiety rating scale. Yet, on examining the studies using the scale, it was found that there was no mention of a form of adaptation into Turkish. That is to say, even though there were studies in Turkish using the statistics anxiety rating scale (Baloğlu & Zelhart, 2004; Baloğlu, Koçak & Zelhart, 2007), the studies were performed in Texas in the USA by using the original form of the scale. No studies in which the Turkish adaptation of the scale was used were available.

### *Purpose of the Study*

The objectives and contents of statistics courses taught at undergraduate and graduate levels are different. The main reason for this difference is related to the competencies that graduates should have. At the undergraduate level the topics such as basic concepts of statistics, reading and interpretation of tables and graphs, calculation of descriptive statistics, calculation and interpretation of simple correlation coefficients are covered. On the other hand, at the graduate level individuals are expected to carry out the statistical process from start to finish by planning a scientific research and so the scopes expand. In other words, the graduate student is a researcher who is accepted as an expert in the related field. For this reason, statistical anxiety scales for graduate students must contain items that correspond to the basic elements of graduate education such as reading, conducting and presenting scientific studies.

Differences in the content of statistics courses taught at undergraduate and graduate levels make it inevitable that the scales related to the anxiety, attitude or self-efficacy towards statistics as prepared for these educational levels will also differ. In this sense, it is considered that the use of statistical anxiety scales developed for undergraduate students to measure the statistical anxiety of graduate students is not correct. When the Turkish literature was analysed from this perspective, it has seen that the measurement tools developed to determine the statistical anxiety were limited to the scales for the undergraduate students. Therefore, a Turkish scale usable in determining graduate students' statistics anxiety was needed. In this context the present study aims to adapt the Statistics Anxiety Scale (SAS) developed by Faber et al. (2018) for graduate students into Turkish.

## METHOD

This research, which aims to adapt SAS into Turkish, is a descriptive study. Descriptive research aims to present and interpret the current situation as it is. These researches give a snapshot of beliefs, thoughts, emotions and behaviours at a given time and place (Stangor, 2010). Descriptive research can be quantitative or qualitative oriented. Generating numerical data, requiring selection of a sample that can represent a large population, providing inferential and explanatory information, gathering standardized information obtained by applying the same measurement tool to all participants, capturing data mostly from scales, multiple choice tests, questionnaires, etc. are typical features of quantitative-oriented descriptive research (Cohen, Manion & Morrison, 2007). When these features are taken into consideration, studies aimed at developing, adapting or revising the measurement tools can be expressed as quantitative oriented descriptive studies.

### *Study Group*

In reaching the participants of the research, three different paths were followed. First of all, the scale was applied face to face to the students who have taken the statistics course and who continue their graduate education in the faculty where the researchers work. The number of participants to whom the scale was applied face to face was 25. Then, the researchers searched as *master student* and *doctoral student* in google scholar and they limited search results to 2019. In this manner it was reached to the articles with postgraduate student(s) among its authors. Subsequently, these articles were reviewed to see if they contain statistical analyzes or whether the relevant field of the article requires statistical information. If the article contains statistical analyzes, or it is related to a field (educational sciences, field education, biostatistics etc.) where its authors are expected to have knowledge of statistics, the e-mail address of the article's author(s) who is at graduate level was recorded and the scale was sent to this author(s) via e-mail. Finally, the websites of universities were scanned and the e-mail addresses of the research assistants who indicated that they were continuing their graduate education in their resumes and that they required statistical information of the graduate program in which they were registered were recorded, and the scale was delivered electronically to these research assistants. The number of participants who answered the scale electronically was 350. Finally, a total of 375 participants who continue graduate education at any university in Turkey was reached. Of the participants 233 (62.10%) were female and 142 (37.90%) were male. The participants' ages ranged between 22 and 57 ($\bar{X}$ = 30.06, SD = 5.58), but two of them did not indicate their age. The distribution of the participants according to the institute where they are registered, the stage of graduate education they were at and whether they had taken a statistics course is shown in Table 1.

Table 1. Information on the Institute where the Participants are Registered, the Stage of Graduate Education They are at and Whether They Have Taken a Statistics Course Before

| Variable | Categories of the Variable | Frequency | Percent |
|---|---|---|---|
| The institute where the participants are registered | Educational Sciences | 276 | 73.60 |
| | Social Sciences | 62 | 16.53 |
| | Health Sciences | 25 | 6.67 |
| | Pure Science | 11 | 2.93 |
| | Uncertain | 1 | .27 |
| Stage of the participants in graduate education | Master-course | 116 | 30.90 |
| | Master-theses | 56 | 14.90 |
| | PhD course | 58 | 15.50 |
| | Preparation for PhD proficiency exam | 21 | 5.60 |
| | PhD theses | 124 | 33.10 |
| Whether to take the courses related to statistics before | Who takes courses related to statistics neither at undergraduate education nor at graduate education | 43 | 11.50 |
| | Who takes at undergraduate level only | 74 | 19.70 |
| | Who takes courses related to statistics at graduate level only | 97 | 25.90 |

| | | | |
|---|---|---|---|
| | Who takes courses related to statistics at both undergraduate and graduate level | 161 | 42.90 |

The majority (73.87%) of the participants in the study group have been registered in one of the graduate programmes of educational sciences and teacher training basic field. Yet, there were also graduate students registered in such diverse programmes as medical training, tourism and hotel management, private law, and finance. They were included in the study group due to the fact that they also needed knowledge of statistics in their graduate courses and in their scientific studies.

### *Data Collection Tool*

The research data were collected through SAS-which was developed by Faber et al. (2018) and which this study aims to adapt into Turkish. The scale is in four-pointed Likert type and it contains 17 items. There is no reverse scored item in the scale. While developing the original form of the scale a three-dimensional structure has been foresighted. Table 2 shows information on this three-dimensional structure.

Table 2. Foresighted Structure while Developing the Original Form of the SAS

| Dimension | Number of Items | Sample Item |
|---|---|---|
| Worry | 8 | If I had to comment on statistical data in a course, I would be worried that I would make a fool of myself. |
| Avoidance | 4 | When presentation topics are being assigned in the course, I would make sure that I receive a topic that doesn't involve statistics. |
| Emotionality | 5 | I would be quite nervous if I were asked to explain a chart from a research report. |

Although the scale was designed as having three factors as is shown in Table 2, the principal components analysis could not statistically separate the three anxiety components and thus the SAS had a single-factor structure. In unidimensional structure, the explained variance rate was determined as 43.59% and it was found that the factor loadings of the scale items ranged from .49 to .76. The reliability of the measures obtained with SAS was tested through Cronbach’s Alpha internal consistency coefficient and was detected as .92. The corrected total item correlations calculated for item discrimination were reported to range between .44 and .70.

Faber et al. (2018) stated that the fact that the SAS showed a statistically single-factor structure does not prevent commenting on the basis of subscales and that evaluation can be made on the subscales’ scores in addition to the total score. SAS scores range from 17 to 68. High scores from both the whole scale and the subscales indicate a high level of statistical anxiety.

### *Translating the Scale into Turkish*

Primarily the researchers who had developed the original form of the scale were contacted in adapting the scale into Turkish. Thus, Günter Faber was sent an e-mail on 10 November 2018 to get the permission for Turkish adaptation of the scale. The e-mail of Günter Faber’s approval of the adaptation was received on 11 November 2018 and the process of adaptation was thus started.

The first step in the adaptation process is to translate the scale from English to Turkish. When translating the measurement tool from the source language to the target language, there are four different methods that can be used: judgmental single-translation, judgmental back-translation, statistical single-translation and statistical back-translation (Hambleton & Bollwark, 1991). In present study, judgmental single-translation method was used. In this method, one or more translators translate the scale from the source language to the target language, then another group compare the original form with the translation form to determine whether the two forms are linguistically equivalent and they change the translation form if deemed necessary (Hambleton & Kanjee, 1993). Accordingly, the items of the SAS were translated into Turkish by five experts three of whom were experts in measurement and evaluation, one of whom was an expert in social studies education and one of whom was an expert in curriculum and instruction. Another expert in English language was not needed because the expert in curriculum and instruction was a graduate of English Language Teaching. After the five experts had translated the scale independently of each other, the translations were brought together and the Turkish equivalents which were thought to reflect the items in the best way were chosen. Then, the Turkish form was presented to the two different experts together with the original form of the scale and the experts were asked to examine whether the two forms were equivalent. Both experts stated that the two forms were generally equivalent to each other. Only one of the experts stated that the item-15 in the scale did not fully reflect the original form and proposed revision for the relevant item. The revision proposed by the expert has been adopted by the researchers and the necessary translation has been changed.

Four-pointed rating was adopted in the Turkish version of the scale as in its original version and the scale categories were labelled as *absolutely disagree* (1), *slightly agree* (2), *quite agree* (3) and *absolutely agree* (4). To test the intelligibility of the translations, the scale was applied to three research assistants who were studying for their PhD. After the feedback from the three research assistants that the scale items were clear and comprehensible, the Turkish form of the SAS (Appendix A) was ready for use. It was difficult to reach a large sample of graduate students. That’s why, the researchers thought it was unlikely to reach two different study groups, one in the pilot and the other in the actual application. Consequently, after testing the intelligibility of the scale items on a small group, the actual application of the scale was started; no pilot study was included.

### *Data Collection and Analysis*

The research data were collected online in the period between 27 November 2018 and 05 February 2019. Within the scope of psychometric properties of the measures collected by the Turkish form of the SAS; construct validity, internal consistency reliability and discrimination power have been tested. Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were done for the construct validity of the SAS, and additionally, parallel analysis method was used to determine the number of factors. The studies in the literature (Fabrigar, Wegener, MacCallum & Strahan, 1999; Macfarlane, Meach & Leroy, 2014; Raykow & Marcoulides, 2011) recommend that EFA and CFA be conducted with data obtained from different samples. The reason for this is that EFA includes some subjective decisions by the researcher. Considering that the EFA is based on a single sample, it is critical to retest the factor structure obtained in EFA on a fresh data. For this purpose, the data set is randomly splitted in half, so that the first half is used for EFA and the second half is used for CFA. Essentially, CFA tries to recreate the structure found in EFA in a different dataset. Hence, the data set was randomly divided into two according to the participant numbers prior EFA and CFA were performed. Accordingly, the data files with odd numbers were used for EFA whereas the data files with even numbers were used for CFA. Thus, there were 188 participants in the data set to which EFA was applied and there were 187 participants in the data set to which CFA was applied. The data set used in EFA was used also in parallel analysis. Because in parallel analysis, the eigenvalues obtained as a result of EFA are used when deciding the number of factors (Pallant, 2005).

Before starting the analyses, the skewness and kurtosis coefficients were examined to get an idea about the distribution of the data. Table 3 shows the skewness and kurtosis coefficients obtained for the overall and sub-scales of SAS in the data sets where AFA and CFA are conducted.

Table 3. Skewness and Kurtosis Coefficients of Data Sets in which EFA and CFA Conducted

| | **Data set used in EFA** | | **Data set used in CFA** | |
|---|---|---|---|---|
| | **Skewness** | **Kurtosis** | **Skewness** | **Kurtosis** |
| Worry | .92 | .31 | .91 | .02 |
| Avoidance | 1.47 | 1.45 | 1.46 | 1.47 |
| Emotionality | 1.06 | .45 | 1.02 | .03 |
| The whole scale | 1.08 | .50 | 1.00 | -.00 |

When the skewness and kurtosis coefficients in Table 3 are examined, it is seen that they are all within ±2 range. In perfectly symmetrical normal distribution, the coefficients of skewness and kurtosis are equal to zero. However, as a rule of thumb values for skewness and kurtosis between ±2 is interpreted as the distribution does not show a significant deviation from normal (Bachman, 2004). Accordingly, it can be said that the research data meet the assumption of normality.

Another indicator that can provide evidence for the normality of the research data is the number of participants in the study group. Indeed, Kirk (2007) points out that in large enough samples, the data approach normal distribution and that a sample of 100 people is sufficient to reach a normal distribution. Similarly, Waternaux (1976) found that when the sample size was over 100, the effect of skewness and kurtosis of the data on the results of the analysis was reduced, and that the effect was almost completely abolished in over 200 samples. Therefore, not only the calculated skewness and kurtosis coefficients; but also, the size of the study group is sufficient to say that the research data is suitable for normal distribution.

Following the examining the skewness and kurtosis coefficients, whether the data are appropriate for factor analysis was checked. For this purpose, Kaiser-Meyer-Olkin (KMO) coefficient and the results of Bartlett test were examined. The KMO was found to exceed the lower limit .60 with a value of .94 (Büyüköztürk, 2010), and Bartlett test was found significant ($\chi^2$ = 2536.07, df = 136, $p$ < .001). The results showed that the data are appropriate for factor analysis. Following this finding, EFA was conducted and principal components method was chosen in the analysis. When interpreting factor loadings in EFA, .32 value recommended by Tabachnick and Fidell (2007) was taken as a criterion.

After EFA, parallel analysis and CFA were done respectively. Two different models were tested in CFA. One of them was the three-factor structure on which the original version of the SAS was based, and the second was the single-factor structure which was reached in EFA conducted in both original and Turkish forms of the SAS. RMSEA, SRMR, CFI, IFI, RFI, NFI and NNFI (TLI) were used to find whether those tested models had been confirmed or not and to see which model fitted the data better. Considering Kline's (2016) explanation that the use of $\chi^2$ / df value as a criterion for model fit does not have a strong logical and statistical foundation, this fit index was not taken into consideration in the study. The acceptable ranges of the fit indices examined are presented in Table 4.

Table 4. The Recommended Criterion Values for the fit Indices Examined in CFA

| Fit Indices | Recommended Criteria | References |
|---|---|---|
| RMSEA | < .10 | Hoyle (2012) |
| SRMR | < .08 | Kline (2016) |
| CFI | > .90 | Wang and Wang (2012) |
| IFI | > .90 | Meyers, Gamst and Guarino (2006) |
| RFI | > .90 | Kelloway (1998) |
| NFI | > .90 | Schumacker and Lomax (2016) |
| NNFI | > .90 | Hancock and Mueller (2013) |

Factor loadings beside the model-data fit in CFA were assessed. When deciding whether the factor loading of an item was sufficient or not, the criterion of .32 was considered as in EFA. After completing the analyses for testing construct validity, reliability analysis was started. The reliability of the measures in the Turkish form of the SAS was calculated with Cronbach's Alpha internal consistency coefficient. The values of .70 and above (Tezbaşaran, 1997) were interpreted as evidence for the reliability of the measures. The discrimination of the SAS items in the Turkish sample were analysed with corrected total item correlation; and the items with correlation values above .30 (Field, 2009) were considered as discriminant enough. Ferguson Delta statistic was used to determine the discriminatory of the entire of the SAS. Calculation of Ferguson Delta, reliability and item analysis was performed on the data from all 375 participants in the study group in contrast to EFA, parallel analysis and CFA. While LISREL 8.54 package programme was used for CFA; IBM SPSS 22 package programme was employed for EFA, reliability and item analysis. Parallel analysis was done by using

Monte Carlo PCA software developed by Watkins (2000). Ferguson Delta statistics, on the other hand, was calculated on Microsoft Excel.

## RESULTS

This section includes analysis outputs for the psychometric properties of the Turkish form of the SAS. The findings obtained from the statistical analyses done for construct validity, reliability and discrimination are offered below under relevant headings.

### *Construct Validity*

First, EFA was performed for the construct validity of the SAS and the findings obtained are shown in Table 5. The results of EFA demonstrated that the Turkish version of the SAS had single-factor structure, like the original version. The variance explained for single-factor structure was found as 59%. As is clear from Table 5, the factor loadings of the scale items range between .60 and .87.

Table 5. The Findings Obtained in EFA for the Turkish Version of the SAS

| Item Number | Factor Loading | Item Number | Factor Loading | Item Number | Factor Loading |
|---|---|---|---|---|---|
| I-1 | .74 | I-7 | .85 | I-13 | .82 |
| I-2 | .77 | I-8 | .84 | I-14 | .75 |
| I-3 | .60 | I-9 | .81 | I-15 | .87 |
| I-4 | .74 | I-10 | .71 | I-16 | .64 |
| I-5 | .80 | I-11 | .77 | I-17 | .68 |
| I-6 | .76 | I-12 | .86 | | |

The single-factor structure obtained in EFA was supported by the parallel analysis results. Averages for eigenvalue are calculated from the correlation matrix which contains the number of variables and participants equal to the real data and which is formed randomly in the method of parallel analysis developed by Horn (1965), (Yavuz & Doğan, 2015). While determining the number of factors, the number of steps where the eigenvalues obtained from the actual data are greater than the eigenvalues that are estimated from random data are taken as basis (O'Connor, 2000).

Table 6. Eigenvalues Obtained from Parallel Analysis

| Number | Real Eigenvalue | Estimated Eigenvalue from Random Data |
|---|---|---|
| 1 | 10.030 | 1.563091 |
| 2 | 1.026 | 1.429725 |

According to Table 6, first eigenvalue is greater than actual data in comparison to random data. On comparing the second eigenvalues, it is found that the value estimated from the random data is higher. Thus, the single-factor structure of the scale was also confirmed through parallel analysis method. Following EFA and parallel analysis, CFA was done. The first model tested in CFA was the three-factor structure (worry, avoidance and emotionality) which was considered while developing the original version of the SAS. The fit indices reported for the three-factor structure as a result of CFA are given in Table 7.

Table 7. The Fit Indices for the Three-Factor Structure

| | Fit indices | | | | | | |
|---|---|---|---|---|---|---|---|
| | RMSEA | SRMR | CFI | IFI | RFI | NFI | NNFI |
| Value | .099 (90% confidence interval; .087; .11) | .045 | .98 | .98 | .96 | .96 | .97 |

The fit indices in Table 7, mean that the three-factor model is confirmed. The measurement model obtained for the three-factor structure of the Turkish version of SAS is shown in Figure 1.

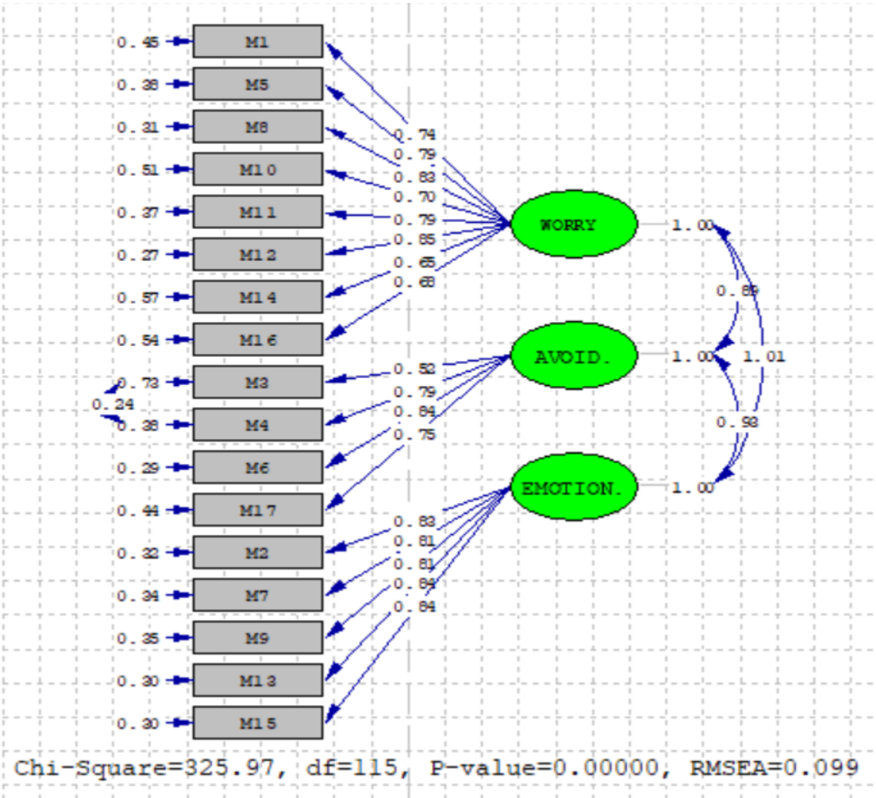


Figure 1. The Measurement Model Obtained for The Three-factor Structure in The Turkish Version of the SAS

On examining Figure 1, it is evident that the factor loadings range between .65 and .85 in the factor of worry, that they range between .52 and .84 in the factor of avoidance and that they range between .81 and .84 in the factor of emotionality. As can be seen in Figure 1, the modification was applied by correlating the error variances of item-3 and item-4 in the avoidance dimension. Item-3 contains the expression of selecting another course instead of statistics, and item-4 refers to choosing a topic that does not include statistics while sharing presentation topics. Therefore, statistical modification is supported theoretically. After the three-factor model, the single-factor model of the SAS was tested because the structure encountered in EFA was found to have single factor in its original version and in its Turkish form even though the scale items had been written on the basis of three-factor structure. The fit indices for the single-factor structure were given in Table 8.

Table 8. The Fit Indices for the Single-Factor Structure

| | Fit indices | | | | | | |
|---|---|---|---|---|---|---|---|
| | RMSEA | SRMR | CFI | IFI | RFI | NFI | NNFI |
| Value | .096 (90% confidence interval; .083; .11) | .046 | .98 | .98 | .96 | .97 | .98 |

The values in Table 8 demonstrate that the measures made with the Turkish version of SAS also fitted the single-factor model. The measurement model reached for the single-factor structure in the Turkish version of SAS is shown in Figure 2.

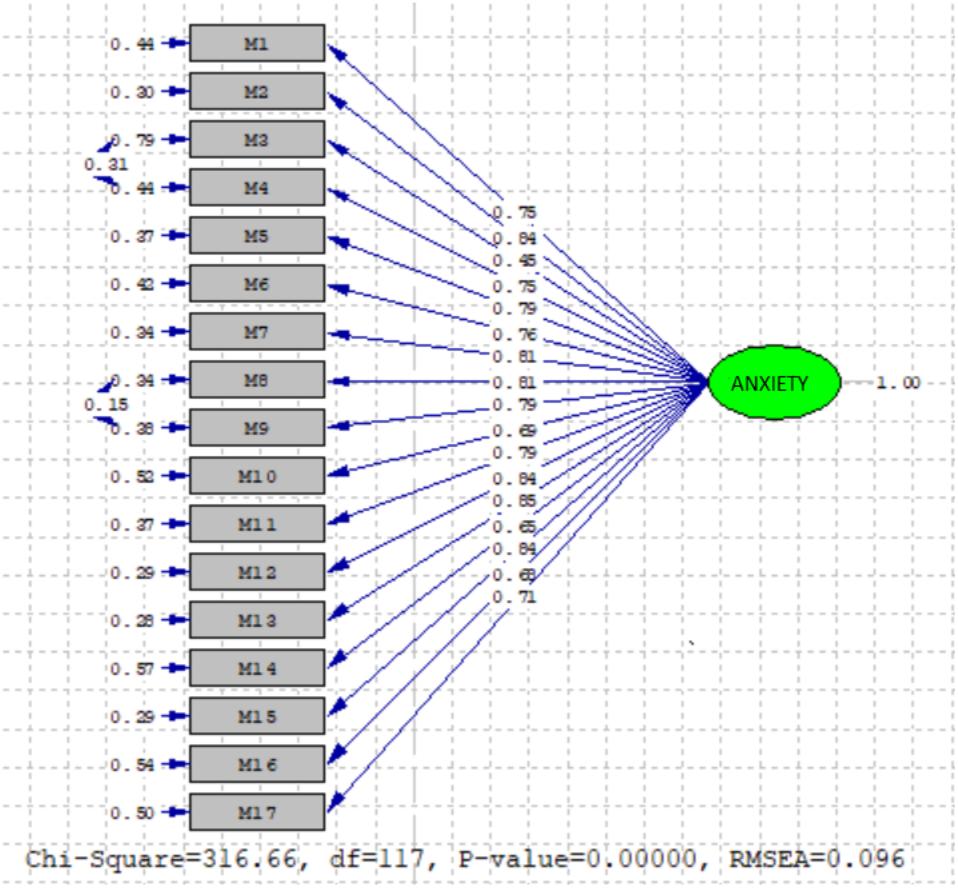


Figure 2. The Measurement Model Obtained for the Single-factor Structure in the Turkish Version of the SAS

As is clear from Figure 2, the factor loadings in the single-factor model of the Turkish version range between .45 and .85. Also, as shown in Figure 2, in addition to the modification in the three-factor model, the error variances of the eighth and ninth items of the scale were also related to each other. While the eighth item of the scale is related to the difficulties in understanding the statistical contents of the courses; ninth items is about the problems experienced in the interpretation of statistical tables. Accordingly, the modifications applied to improve model-data fit are also theoretically explainable.

### *Reliability Analysis*

Considering the fact that the Turkish version of the SAS fitted both the three-factor and the single-factor structure in CFA, internal consistency coefficient was calculated not only for the whole scale, but also reliability analyses were done for the subscales. The internal consistency coefficients calculated for the three factors of the scale and for the overall scale are shown in Table 9. Accordingly, the internal consistency coefficients range between .83 and .96.

Table 9. The Internal Consistency Coefficients for the Measures Obtained by the Turkish Version of the SAS

| Dimension | Overall Scale | Worry | Avoidance | Emotionality |
|---|---|---|---|---|
| Cronbach Alpha | .96 | .91 | .83 | .91 |

### *Item Analysis*

The corrected total item correlations ($r_{jx}$) calculated to test the item discrimination index in the Turkish version of the scale are shown in Table 10. An examination of Table 10 makes it clear that the item correlations take on values between .52 and .84.

Table 10. Discrimination Indexes for the Items in the Turkish Version of the SAS

| Item Number | $r_{jx}$ | Item Number | $r_{jx}$ | Item Number | $r_{jx}$ |
|---|---|---|---|---|---|
| I-1 | .71 | I-7 | .80 | I-13 | .81 |
| I-2 | .78 | I-8 | .80 | I-14 | .67 |
| I-3 | .52 | I9 | .77 | I-15 | .84 |
| I-4 | .73 | I-10 | .67 | I-16 | .63 |
| I-5 | .77 | I-11 | .75 | I-17 | .67 |
| I-6 | .74 | I-12 | .82 | | |

### *Ferguson Delta Statistics*

Ferguson Delta (δ) statistics in addition to item correlations were also used to demonstrate the discrimination of the SAS. According to this statistic, high variability in scores received from the scale (heterogeneity of the group) displays that the measurement tool is discriminant (Zhang & Lidbury, 2013). The variability in scores the participants receive from the scale are divided into the highest variability probable to be observed in calculating the Ferguson Delta statistics (Day & Bonn, 2011). While δ = .00 when all the participants receive the same scores from the scale, δ = 1.00 when the variability between participants' scores is equal to the highest variability probable to be observed (Hankins, 2008). Kline (2000) states that Ferguson Delta corresponds to .93 in normal distribution and suggests that the value of .90 should be taken as the criterion for the statistics. The Equation 1 is used in calculating the Ferguson Delta statistics for the measurement tools with more than two response options (Hankins, 2008).

$$\delta = \frac{[\,1+k(m-1)][n^2-\sum_i f_i{}^2\,]}{n^2k(m-1)} \qquad (1)$$

*k* = number of items in the measurement tool
*n* = sample size
*f* = frequency of each score
*m* = number of response category

As is apparent from the Equation 1, first the frequency table should be drawn for the scores received from the measurement instrument to be able to calculate the Ferguson Delta statistic (Ramsay & Reynolds, 2000). The frequencies for the scores the 375 participants received from the SAS are shown in Table 11. On placing the frequencies along with the values k = 17, m = 4 and n = 375 in the formula, the Ferguson Delta statistics was found as .98.

Table 11. Frequencies of Participants' Scores on the SAS

| Score | Frequency | Score | Frequency | Score | Frequency | Score | Frequency | Score | Frequency |
|---|---|---|---|---|---|---|---|---|---|
| 17 | 36 | 27 | 19 | 37 | 8 | 47 | 4 | 57 | 1 |
| 18 | 20 | 28 | 10 | 38 | 3 | 48 | 4 | 58 | 2 |
| 19 | 23 | 29 | 12 | 39 | 3 | 49 | 4 | 59 | 3 |
| 20 | 22 | 30 | 11 | 40 | 4 | 50 | 4 | 60 | 1 |
| 21 | 19 | 31 | 15 | 41 | 4 | 51 | 8 | 61 | 1 |
| 22 | 10 | 32 | 9 | 42 | 4 | 52 | 3 | 62 | 2 |
| 23 | 15 | 33 | 10 | 43 | 2 | 53 | 4 | 68 | 1 |
| 24 | 15 | 34 | 8 | 44 | 1 | 54 | 3 | | |
| 25 | 8 | 35 | 9 | 45 | 2 | 55 | 4 | | |
| 26 | 14 | 36 | 6 | 46 | 1 | 56 | 3 | | |

### *The Interpretation of the SAS Scores*

As all of the items in the original form of SAS had sufficient factor loadings and discriminative values also in the Turkish version of the scale, no item was removed from the scale. Thus, as in the original form, the scores that can be obtained from the overall SAS vary between 17 and 68. High scores from the scale reflect high level of statistical anxiety. Similarly, the increase in scores obtained from the subscales indicates high levels of worry, avoidance and emotionality.

## DISCUSSION and CONCLUSION

In this study, the SAS developed by Faber et al. (2018) for graduate students was adapted into Turkish. The construct validity of SAS was tested with EFA and CFA; and parallel analysis method was also used in deciding about the number of factors in the scale. A single-factor structure was found in EFA and the rate of explained variance was found to be 59%. There are various criteria set in the literature by researchers about what the rate of explained variance should be at least. While Bayram (2010) and Büyüköztürk (2010) say that the explained variance should be at least 30%; Aksu, Eser and Güzeller (2017) say that the values of 40% and above are acceptable. According to Sönmez and Alacapınar (2016), however, the rate of explained variance should be higher than the rate of unexplained variance. The rate of variance reported after EFA meets all these criteria. Besides, the factor loadings for all of the items in the SAS were found to be above the threshold level of .32 (Tabachnick & Fidell, 2007). These results indicate that the construct validity was achieved in the Turkish version of the SAS. The single-factor structure found in EFA was also supported by the results of parallel analysis.

Conclusions that there was evidence to show the construct validity of the Turkish version of the SAS in CFA as in EFA were reached. According to the fit indices reported in CFA, both the three-factor structure (labelled as worry, avoidance and emotionality) taken into consideration when developing the original form of the scale and the unidimensional structure emerging as a result of EFA were confirmed. In addition to that, it was also found that the factor loadings for both models were above .32. On considering these results about CFA along with the findings obtained in EFA and parallel analysis, it may be said that the three factors of the scale can be interpreted separately in addition to the total scores received from the scale and that it would not be very correct to make an evaluation based on the subscales only without obtaining a total score for anxiety.

It was concluded that internal consistency coefficients calculated in reliability analysis for the subscales in the SAS and for the whole scale met the criterion of .70 (Pallant, 2005; Tekindal, 2009). Accordingly, it can be stated that the Turkish version of SAS is an instrument yielding reliable measures. According to item analysis results, the corrected item correlations met the threshold value of .30 (Erkuş, 2012) for all the items in the SAS. The value found for Ferguson Delta statistics also met the criterion of .90 (Kline, 2000). Therefore, it may be said that the SAS is discriminant enough-that is to say, it is capable of discriminating between graduate students having different levels of statistics anxiety. In conclusion, the results obtained in this study indicate that the statistics anxiety of graduate students can be measured by using SAS in a valid and reliable way.

### *Recommendations for Further Studies*

This study analysed the construct validity of the Turkish version of the SAS with EFA and CFA. Convergent and divergent validity analyses can be included in further studies. Because the reliability of the SAS was analysed only on the basis of internal consistency in this study, it can be recommended that the further studies could test the test-retest reliability of the scale. Besides, since this study was conducted within the framework of classical test theory, it can be suggested that the reliability and validity of the SAS be analysed on the basis of item response theory.

By using SAS, studies can be conducted to compare the statistical anxiety levels of the researchers who continue their graduate education in any of the fields of educational, social and health sciences, field education or pure science. In this way, it can be determined whether there is a significant difference between the statistical anxieties of the individuals attending graduate education in different fields and if significant difference is detected, the rationale of the observed differences can be revealed by qualitative analysis.

## REFERENCES

Aksu, G., Eser, M. T., & Güzeller, C. (2017). *Açımlayıcı ve doğrulayıcı faktör analizi ile yapısal eşitlik modeli uygulamaları*. Ankara: Detay.

Bachman, L. F. (2004). *Statistical analyses for language assessment book*. Cambridge: Cambridge University.

Baloğlu, M. (2002). Psychometric properties of the statistics anxiety rating scale. *Psychological Reports*, *90*(1), 315-325. Retrieved from https://www.researchgate.net/publication/11464881_Psychometric_properties_of_the_statistics_anxiety_rating_scale

Baloğlu, M. (2003). Individual differences in statistics anxiety among college students. *Personality and Individual Differences, 34*(5), 855-865. doi: 10.1016/S0191-8869(02)00076-4

Baloğlu, M., & Zelhart, P. F. (2004). Üniversite öğrencileri arasında yüksek ve düşük istatistik kaygısının ayrıştırıcıları. *Eğitim ve Bilim*, *29*(133), 47-51. Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/download/5093/1176

Baloğlu, M., Koçak, R., & Zelhart, P. F. (2007). İstatistik kaygısı ve istatistiğe yönelik tutumlar arasındaki ilişki. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *40*(2), 23-39. Retrieved from http://dergipark.gov.tr/download/article-file/509051

Bayram, N. (2010). *Yapısal eşitlik modellemesine giriş AMOS uygulamaları*. Bursa: Ezgi.

Benson, J. (1989). Structural components of statistical test anxiety in adults: an exploratory model. *The Journal of Experimental Education*, *57*(3), 247-261. doi: 10.1080/00220973.1989.10806509

Beurze, S. M., Donders, A. R. T., Zielhuis, G. A., Vegt, F., & Verbeek, A. L. M. (2013). Statistics anxiety: A barrier for education in research methodology for medical students? *The Journal of the International Association of Medical Science Educators*, *23*(3), 377-384. doi: 10.1007/BF03341649

Büyüköztürk, Ş. (2010). *Sosyal bilimler için veri analizi el kitabı*. Ankara: Pegem Akademi.

Chew, P. K. H., Dillon, D. B., & Swinbourne, A. L. (2018) An examination of the internal consistency and structure of the Statistical Anxiety Rating Scale (STARS). *PLoS ONE*, *13*(3), 1-12. doi: 10.1371/journal.pone.0194195

Chiesi, F., Primi, C., & Carmona, J. (2011). Measuring statistics anxiety: Cross-Country validity of the statistical anxiety scale (SAS). *Journal of Psychoeducational Assessment*, *29*(6), 559-569. doi: 10.1177/0734282911404985

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. New York, NY: Routledge.

Collins, K. M. T., & Onwuegbuzie, A. T. (2007). I cannot read my statistics textbook: The relationship between reading ability andstatistics anxiety. *The Journal of Negro Education*, *76*(2), 118-129. Retrieved from http://www.jstor.org/stable/40034551

Cruise, R. J., Cash, R. W., & Bolton, D. L. (1985, August). *Development and validation of an instrument to measure statistical anxiety*. Paper presented at the proceedings of the American Statistical Association, Las Vegas, Nevada. Retrieved from https://www.causeweb.org/cause/research/literature/development-and-validation-instrument-measure-statistical-anxiety

Day, J., & Bonn, D. (2011). Development of the concise data processing assessment. *Physical Review Special Topics – Physics Education Research*, *7*(1), 1-14. doi: 10.1103/PhysRevSTPER.7.010114

Erkuş, A. (2011). *Davranış bilimleri için bilimsel araştırma süreci*. Ankara: Seçkin.

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme-I*. Ankara: Pegem Akademi.

Faber, G., Drexler, H., Stappert, A., & Eichhorn, J. (2018). Education science students' statistics anxiety: Developing and analyzing a scale for measuring their worry, avoidance, and emotionality cognitions. *International Journal of Educational Psychology*, *7*(3), 248-285. doi: 10.17583/ijep.2018.3340

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. doi: 10.1037/1082-989X.4.3.272

Field, A. (2009). *Discovering statics using SPSS*. London: SAGE.

Fitzgerald, S. M., Jurs, S. J., & Hudson, L. M. (1996). A model predicting statistics achievement among graduate students. *College Student Journal*, *30*(3), 361-366. Retrieved from https://psycnet.apa.org/record/1997-07633-006

Gürbüz, S., & Şahin, F. (2017). *Sosyal bilimlerde araştırma yöntemleri: Felsefe – yöntem – analiz*. Ankara: Seçkin.

Hambleton, R. K. & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. *Bulletin of the International Test Commission*, *18*, 3-32. Retrieved from https://files.eric.ed.gov/fulltext/ED337481.pdf

Hambleton, R. K., & Kanjee, A. (1993, April). *Enhancing the validity of cross-cultural studies: Improvements in instrument translation methods*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Hancock, G. R., & Mueller, R. O. (2013). *Structural equation modeling: A second course.* Charlotte, NC: Information Age.

Hankins, M. (2008). How discriminating are discriminative instruments? *Health and Quality of Life Outcomes*, *6*(36). doi: 10.1186/1477-7525-6-36

Hanna, D., Shevlin, M., & Dempster, M. (2008). The structure of the statistics anxiety rating scale: A confirmatory factor analysis using UK psychology students. *Personality and Individual Differences*, *45*(1), 65-74. doi: 10.1016/j.paid.2008.02.021

Horn, J. L. (1965). A rationale for the number of factors in factor analysis. *Psychometrica*, *30*(2), 179-185. doi: 10.1007/BF02289447

Hoyle, R. H. (2012). *Handbook of structural equation modeling*. New York, NY: Guilford.

Kelloway, E. K. (1998). *Using LISREL for structural equation modeling: A researcher's guide.* Thousand Oaks, CA: Sage.

Kirk, R. E. (2007). *Statistics: An introduction.* Belmont, CA: Wadsworth.

Kline, P. (2000). *The handbook of psychological testing*. London: Routledge

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. New York, NY: Guilford.

Köklü, N. (1994). Bir istatistik tutum ölçeğinin geçerlik ve güvenirliği. *Eğitim ve Bilim*, *18*(93), 42-47. http://egitimvebilim.ted.org.tr/index.php/EB/article/download/5906/2041 adresinden elde edilmiştir.

Köklü, N. (1996). İstatistik kaygı ölçeği: Psikometrik veriler. *Eğitim ve Bilim*, *20*(102), 45-49.

Lalonde, R. N., & Gardner, R. C. (1993). Statistics as a second language? A model for predicting performance in psychology students. *Canadian Journal of Behavioural Science*, *25*(1), 108-125. doi: 10.1037/h0078792

Liu, S., Onwuegbuzie, A. J., & Meng, L. (2011). Examination of the score reliability and validity of the statistics anxiety rating scale in a Chinese population: Comparisons of statistics anxiety between Chinese college students and their Western counterparts. *Journal of Educational Enquiry*, *11*(1), 29-42. Retrieved from https://www.ojs.unisa.edu.au/index.php/EDEQ/article/view/662/585

Maat, S. M., & Rosli, M. K. (2016). The rasch model analysis for statistical anxiety rating scale (STARS). *Creative Education*, *7*(18), 2820-2828. doi: 10.4236/ce.2016.718261

Macfarlane, I., Meach, P. M., & Leroy, B. S. (2014). *Genetic counseling research: A practical guide*. New York, NY: Oxford University.

Meyers, L. S, Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research: Design and interpretation*. London: SAGE.

Nesbit, R. J., & Bourne, V. J. (2018). Statistics anxiety rating scale (STARS) use in psychology students: A review and analysis with an undergraduate sample. *Psychology Teaching Review*, *24*(2), 101-110. Retrieved from https://psycnet.apa.org/record/2018-57771-011

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. doi: 10.3758/BF03200807

Onwuegbuzie, A. J. (2004) Academic procrastination and statistics anxiety, *Assessment & Evaluation in Higher Education*, *29*(1), 3-19. doi: 10.1080/0260293042000160384

Onwuegbuzie, A. J. (1997a) Writing a research proposal: the role of library anxiety, statistics anxiety, and composition anxiety, *Library & Information Science Research*, *19*(1), 5–33. doi: 10.1016/S0740-8188(97)90003-7

Onwuegbuzie, A. J., & Seaman, M. A. (1995). The effect of time constraints and statistics test anxiety on test performance in a statistics course. *The Journal of Experimental Education*, *63*(2), 115-124. doi: 10.1080/00220973.1995.9943816

Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics anxiety: Nature, etiology, antecedents, effects, and treatments--a comprehensive review of the literature. *Teaching in Higher Education*, 8(2), 195-209. doi: 10.1080/1356251032000052447

Onwuegbuzie, A. J., Da Ros, D., & Ryan, J. (1997). The components of statistics anxiety: A phenomenological study. *Focus on Learning Problems in Mathematics*, *19*(4), 11-35. Retrieved from https://eric.ed.gov/?id=EJ558838

Onwuegbuzie, A. J., Slate, J. R., Paterson, F. R. A., Watson, M. H., & Schwartz, R. A. (2000). Factors associated with achievement in educational research courses. *Research in the Schools*, *7*(1), 53-65. Retrieved from https://eric.ed.gov/?id=EJ644255

Pallant, J. (2005). *SPSS survival manual: A step by step guide to data analysis using SPSS for windows*. Australia: Australian Copyright.

Pan, W., & Tang, M. (2005). Students' perceptions on factors of statistics anxiety and instructional strategies. *Journal of Instructional Psychology*, *32*(3), 205-214. Retrieved from http://eds.a.ebscohost.com/eds/pdfviewer/pdfviewer?vid=0&sid=f0e9aed6-3bf9-497e-ba42-86d83215d236%40sessionmgr4008

Pretorius, T. B., & Norman, A. M. (1992). Psychometric data on the statistics anxiety scale for a sample of South African students. *Educational and Psychological Measurement*, *52*(4), 933–937. doi: 10.1177/0013164492052004015

Primi, C., & Chiesi, F. (2018, July). *The role of mathematics anxiety and statistics anxiety in learning statistics*. Paper presented at the 10th International Conference on Teaching Statistics, Kyoto, Japan. Retrieved from https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_5E2.pdf

Ramsay, M. C., & Reynolds, C. R. (2000). Development a scientific test: A practical guide. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 21-42). New York, NY: Elsevier.

Raykow, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.

Roberts, D. M., & Bilderbeck, E. W. (1980). Reliability and validity of statistics attitude survey. *Educational and Psychological Measurement*, *40*(1), 235-238. doi: 10.1177/001316448004000138

Roberts, D. M., & Saxe, J. E. (1982) Validity of a statistics attitude survey: A follow up study. *Educational and Psychological Measurement*, *42*(3), 907-912. doi: 10.1177/001316448204200326

Rodarte-Luna, B., & Sherry, A. (2008). Sex differences in the relation between statistics anxiety and cognitive/learning strategies. *Contemporary Educational Psychology*, *33*(2), 327-344. doi: 10.1016/j.cedpsych.2007.03.002

Schumacker, R. E., & Lomax, R. G. (2016). *A beginner's guide to structural equation modeling*. New York, NY: Routledge.

Sönmez, V., & Alacapınar, F. G. (2016). *Sosyal bilimlerde ölçme aracı hazırlama*. Ankara: Anı.

Stangor, C. (2010). *Research methods for the behavioral sciences*. Boston, MA: Houghton Mifflin.

Sutarso, T. (1992). *Some variables in relation to students' anxiety in learning statistics*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Knoxville, TN (ERIC document number ED-353334). Retrieved from https://files.eric.ed.gov/fulltext/ED353334.pdf

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston, MA: Pearson Education, Inc.

Tan, Ş. (2016). *SPSS ve Excel uygulamalı temel istatistik-I*. Ankara: Pegem Akademi.

Tekindal, S. (2009). *Duyuşsal özelliklerin ölçülmesi için araç oluşturma*. Ankara: Pegem Akademi.

Teman, E. D. (2013). A rasch analysis of the statistical anxiety rating scale. *Journal of Applied Measurement*, *14*(4), 414-434. Retrieved from https://www.researchgate.net/publication/257071159_A_rasch_analysis_of_the_statistical_anxiety_rating_scale

Tezbaşaran, A. (1997). *Likert tipi ölçek geliştirme kılavuzu*. Ankara: Türk Psikologlar Derneği.

Vigil-Colet, A., Lorenzo-Seva, U., & Condon, L. (2008). Development and validation of the statistical anxiety scale. *Psicothema, 20*(1), 174-180. Retrieved from http://www.psicothema.com/PDF/3444.pdf

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. Chichester, UK: Wiley.

Waternaux, C. M. (1976). Asymptotic distribution of the sample roots for a nonnormal population. *Biometrika*, *63*(3), 639-645. doi: 10.1093/biomet/63.3.639

Watkins, M. W. (2000). Monte Carlo PCA for parallel analysis [Computer software]. State College, PA: Ed & Psych Associates.

Wilson, V. (1997, November). *Factors related to anxiety in the graduate statistics classroom*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis, TN (ERIC document number ED415288). Retrieved from https://files.eric.ed.gov/fulltext/ED415288.pdf

Yaşar, M. (2014). İstatistiğe yönelik tutum ölçeği: Geçerlilik ve güvenirlik çalışması. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, *36*(2), 59-75. http://pauegitimdergi.pau.edu.tr/Makaleler/1428620766_5.pdf adresinden elde edilmiştir.

Yavuz, G. & Doğan, N. (2015). Boyut sayısı belirlemede Velicer'in map testi ve Horn'un paralel analizinin kullanılması. *Hacettepe Üniversitesi Eğitim Fakültesi, 30*(3), 176-188. Retrieved from http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/674-published.pdf

Yıldırım, A., & Şimşek, H. (2016). *Sosyal bilimlerde nitel araştırma yöntemleri*. Ankara: Seçkin.

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology*, *61*(3), 319-328. doi: 10.1111/j.2044-8279.1991.tb00989.x

Zhang, F., & Lidbury, B. A. (2013). Evaluating a genetics concepts inventory. In F. Zhang (Ed.), *Sustainable language support practices in science education: Technologies and solutions* (pp. 116-128). USA, Hershey: Medical Information Science Reference.

## Appendix A. Turkish Form of Statistics Anxiety Scale for Graduate Students *

| | Hiç Katılmıyorum | Biraz Katılıyorum | Oldukça Katılıyorum | Tamamen Katılıyorum |
|---|---|---|---|---|
| **1.** Kayıtlı olduğum lisansüstü programın istatistiksel gerekliliklerini karşılamakta zorlanırım. | 1 | 2 | 3 | 4 |
| **2.** İstatistiksel bir problem üzerinde çalışmam gerektiğinde kendimi çok rahatsız hissederim. | 1 | 2 | 3 | 4 |
| **3.** Mümkün olsa bir istatistik dersi almak yerine başka iki ders almayı tercih ederim. | 1 | 2 | 3 | 4 |
| **4.** Derslerde sunum konuları paylaşılırken istatistik içermeyen bir konu aldığımdan emin olmaya çalışırım. | 1 | 2 | 3 | 4 |
| **5.** Çalışmalarımda istatistiksel içerikleri yeterli derecede tartışmak benim için zordur. | 1 | 2 | 3 | 4 |
| **6.** Sunum hazırlarken istatistikle ilgili olan kısımları sunum dışında tutmayı tercih ederim. | 1 | 2 | 3 | 4 |
| **7.** Bir araştırma raporundaki tabloları/grafikleri açıklamam istendiğinde oldukça gerilirim. | 1 | 2 | 3 | 4 |
| **8.** Derslerdeki istatistiksel içerikleri anlamakta zorlanırım. | 1 | 2 | 3 | 4 |
| **9.** İstatistiksel değerler içeren bir tablodan gerekli bilgileri seçip ayırmada sorun yaşarım. | 1 | 2 | 3 | 4 |
| **10.** Bir derste istatistiksel verileri yorumlamam gerektiğinde komik duruma düşmekten korkarım. | 1 | 2 | 3 | 4 |
| **11.** Bir derste istatistiksel bulgular içeren sunum yapmam gerektiğinde sunumdan sonra kimsenin soru sormamasını umut ederim. | 1 | 2 | 3 | 4 |
| **12.** İstatistiksel araştırma bulgularına ilişkin tatmin edici bir rapor sunmakta güçlük çekerim. | 1 | 2 | 3 | 4 |
| **13.** İstatistiksel bir formülü uygulamak zorunda kaldığımda çok gergin hissederim. | 1 | 2 | 3 | 4 |
| **14.** Bir istatistik sınavına dikkatli bir şekilde hazırlanmış olsam da dersi geçemeyeceğim diye endişelenirim. | 1 | 2 | 3 | 4 |
| **15.** Bir derste istatistiksel bir problemi açıklamak zorunda kalma düşüncesi beni oldukça tedirgin eder. | 1 | 2 | 3 | 4 |
| **16.** Bir istatistik dersi aldığımda öğrendiğim her şeyi hemen unutacağım endişesi yaşarım. | 1 | 2 | 3 | 4 |
| **17.** Eğer mümkünse bilimsel metinlerdeki istatistiksel tabloları ve grafikleri atlarım. | 1 | 2 | 3 | 4 |

* It is sufficient to reference the article for the use of the scale. Furthermore, there is no need for permission from the authors.

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(4); 451-466

# Adult Daughter-Mother Attachment: Psychometric Properties of Turkish Version of Adult Attachment Scale *

Meltem ANAFARTA ŞENDAĞ ** Funda KUTLU ***

**Abstract**
The present study aims to assess the psychometric qualities of the Turkish version of the Adult Attachment Scale (AAS) assessing adult daughter's current attachment to their elderly mother. In total, 560 women with the mean age of 39.6 have participated. Parallel to the original study, exploratory factor analysis was conducted with adult daughters (*N* = 304) who were providing instrumental help to their mothers regularly. Results yielded 2 correlated factors (secure base and safe haven). Confirmatory factor analysis revealed that the factor structure is applicable to the adult daughters who were not providing regular help to their mothers (*N* = 256). Measurement invariance was established across two groups constructed in terms of the presence of instrumental help provided to the elderly mothers by their adult daughters. Internal consistency and 6-month stability for the scale are satisfactory. Further evidence for convergent and concurrent validity has been supported by presenting a positive correlation of AAS with the level of significance of the mother in the adult daughter's attachment hierarchy as compared to other attachment figures, levels of quality of the current relationship and the frequency of contact with the mother. Results are discussed in terms of AAS's appropriateness for Turkish culture and possible contribution in an understanding attachment to a parent in late adulthood, a critical emerging need for the aging world.

*Key Words*: Attachment, adult daughter, elderly mother.

## INTRODUCTION

Given that the global population has aged at an unprecedented rate and that 28% of the European population will be over 65 in 30 years (He, Goodkind & Kowall, 2016), it becomes a critical and urgent task to question and to improve our scientific understanding of aging and old age. The major reasons for such global demographic change are reported as being the aging of "baby boom" generation and decreased fertility rates (Bloom, Canning & Lubet, 2015; Lowenstein, 2005; Trommsdorff & Nauck, 2006). In addition, increasing life expectancy due to development in medicine and preventive health care practices is also mentioned as an important reason for the increase in the elderly population (Kontis et al., 2017). The number of people reaching the age of 100 is increasing every year in the world (Martin & Baek, 2018; Rochon et. al., 2014). In previous times, it was not normal for a person to live enough to see his/her grandson's child, but nowadays it is considered as normal.

There are two major consequences of such demographic change for the future. First, health and insurance sectors are under great pressure. Especially for women who have a longer life expectancy than men, but generally have lower education and income levels, health care costs are a critical problem that needs to be addressed and the renovation of an administrative structure is unavoidable. Secondly, and similar to administrative structures, significant changes are also expected in family structures, and even today these changes are remarkable. For example, in a family, the years that three or even four generations can live together are increasing. Improved health quality of life allows grandparents spending more and quality time with their children and grandchildren. On the other hand, the necessity to provide both instrumental and emotional care to the aging member of the family increases as well.

* The current study is part of a larger project funded by Turkish Academy of Sciences (TUBITAK, Support Code: 113K537)

** Asst. Prof., Ufuk University, Faculty of Arts and Sciences, Ankara-Turkey, meltem.anafarta@ufuk.edu.tr, ORCID ID: 0000-0002-4810-893X
*** Asst. Prof., Ufuk University, Faculty of Arts and Sciences, Ankara-Turkey, funda.kutlu@ufuk.edu.tr, ORCID ID: 0000-0002-6182-843X

To cite this article:
Anafarta-Şendağ, M. & Kutlu, F. (2019). Adult daughter-mother attachment: Psychometric properties of Turkish version of adult attachment scale. *Journal of Measurement and Evaluation in Education and Psychology, 10*(4), 451-466. doi: 10.21031/epod.553763

*Received: 15.04.2019*
*Accepted: 9.11.2019*

Previous research shows that the responsibility for providing care for the elderly parent, regardless of Eastern or Western cultures, is predominantly on daughters or daughter in laws' shoulders (Ataca, Kağıtçıbaşı & Diri, 2005; Finley, 1989; Ingersoll-Dayton, Starrels & Dowler, 1996; İmamoğlu, 1987; Kagıtçıbaşı, 1985). In the last decade more daughters have left work to take care of their elderly parents (Manuela, Emmanuele & Cristina, 2016). These findings point to the importance of dwelling on the relationship between an adult daughter and aging mother in various dimensions.

Considering the aging literature up to date, it could be seen that studies mostly focus on the deterioration of physical and psychological health, health care practices, insurance policies, and stress experienced by the caregivers (e.g., Anderson & Hussey, 2000; Feng, Liu, Guan & Mor, 2012; Rowland & Bellizzi, 2014; Shulz & Sherwood, 2008). In addition to that, sociological studies focus mostly on concepts as intergenerational solidarity, filial obligation and/or piety to investigate the role of culture in support of the elderly members (e.g., Bengtson, Rosenthal & Burton, 1990; Bengston & Oyama, 2007; Rossi & Rossi, 1990). Although the critical importance of all these factors cannot be denied, the researchers (e.g., Bengtson, Giarrusso, Mabry & Silverstein, 2002; Lowenstein, Katz & Gur-Yaish, 2007; Schwarz, Trommsdorff, Kim & Park, 2006) themselves stated that these factors were insufficient to understand the whole picture and implied that quality of dyadic emotional bond had a central role in completing this picture. Yet the literature has provided little about this issue**.**

Multiple factors, such as a sense of responsibility, filial obligation, necessity, and respect are determinant in the behaviors of an adult daughter in caring for her mother. However, these factors do not strongly relate to the quality and effectiveness of care provided by them and the emotional burden experienced by both parties, those of which are major determinants in both psychological and physical quality of life both for care givers and takers. At this point, the quality of the emotional bond between adult daughter and mother becomes critical and attachment theory has much to offer about this.

Attachment theory is considered to be unique and informative in terms of highlighting the survival value of the attachment bond, explaining the difference between attachment and dependence, and emphasizing the normalcy of lifelong need for attachment. As stated, an independent/autonomous self is established through a functional bond early in life (Sroufe, Fox & Pancake, 1983) that in turn facilitates the beliefs about dependability of others and so-called secure attachment (Bowlby, 1969).

It is proposed that early attachment relationships shape the capacity to love someone, to care for someone, and ask someone's care when needed (Waters, Kondo-Ikemura, Posada & Richters, 1991), thus it organizes emotions and behaviors in close relationships throughout life (Ainsworth, 1989; Bowlby, 1979; Waters, Merrick, Treboux, Crowell & Albersheim, 2000).

Considering its survival value, an attachment bond is stated to be established not only with one figure but is constructed in a hierarchy composed of a finite number of significant others (Bowlby, 1969/1982). Within this dynamic hierarchy, the primary attachment figure changes from parents to friends and partners, from childhood to adulthood (Rosenthal & Kobak, 2010). However, attachment to the mother (primary caregiver) was proposed to be unique and non-replaceable (Ainsworth, 1989). Unlike fathers, mothers were shown to preserve a place in the attachment hierarchy throughout their children's lives although their primary status might be replaced by the romantic partners (Doherty & Feeney, 2004; Rosenthal & Kobak, 2010). Moreover, it was stated that mothers might regain their primary status in the attachment hierarchy during certain developmental milestones of their adult children such as becoming a parent for the first time (Doherty & Feeney, 2004).

Despite the empirical evidence for the continuous role of the mother as an attachment figure, the literature has little to provide in understanding the dynamics of this continuous emotional bond between adult children and their elderly parents. In that sense, Cicirelli's early works (1983, 1991, 1993, 1995, 2010) provided some valuable insights about the motivation of adult children to provide both instrumental and emotional care for their elderly parents and whether this role reversal could be interpreted as reciprocity of attachment bond for the sake of others or as an attempt for the adult child to protect the attached figure for the sake of himself or herself. Especially when an elderly parent needs health care due to old age, losing an attachment figure becomes a salient and realistic threat for an adult child. In that case, the dynamics of both instrumental and emotional caregiving provided to an elderly parent by an adult child might be more complex than it has been known to be. To answer such questions Cicirelli (1991) pointed out the absence of appropriate measurement tool, and thus constructed the Adult Attachment Scale (AAS) that aims to assess the level of an adult child's current attachment to the mother.

The scale has been developed and validated in various studies conducted with adult daughters who were providing instrumental care for their parents (Cicirelli, 1991, 1993, 1995). By utilizing AAS, Cicirelli (1993) has examined the adult daughter's helping behaviors and the subjective burden associated with it. Results have revealed that as a parent's need for help, the daughter's feeling of obligation and feeling of attachment increases, the frequency of helping behavior increases as well. On the other hand, independent of the frequency of helping behavior, it was found that the subjective burden was positively correlated with the feeling of obligation but negatively correlated with feeling of attachment. These findings not only emphasize the critical role of the emotional bond between an adult child and elderly parent but also support that AAS can be a valuable tool in future studies as well.

### *Purpose of the Study*

As stated above, independent of cultures, the responsibility for providing care for elderly parents was shown to be predominantly on daughters (Ataca et al., 2005; Finley, 1989; Ingersoll-Dayton, et al., 1996; İmamoğlu, 1987; Kağıtçıbaşı, 1985; Zhan & Montgomery, 2003) yet little is known about the emotional dynamics and process of attachment between them. However, when the aging world population and foreseeable societal and familial changes in the future are taken into consideration, understanding of the dynamics of attachment between daughters and aging parents is becoming more and more critical. At this point, AAS that aims to assess the attachment between adult daughters and aging mothers is providing a valuable starting point and opportunities for future studies. Therefore, the purpose of the present study is to adapt AAS into Turkish and test its psychometric properties and factorial structure with adult daughters.

In addition, Cicirelli's (1995) comments about the sample feature in the original study and his suggestions for future studies in this regard were also taken into consideration in this study. Even though AAS was developed with adult women who regularly provide help for their parents, Cicirelli (1995) emphasized that it is important to test this scale with adult women who do not provide help to their parents for many different reasons (such as living away, workload of daughter, absence of parents need, presence of other children helping mother). Therefore, in the present study, the structure invariance of AAS was examined with two groups of adult daughters who did and did not provide instrumental help to their mothers on a regular base.

## METHOD

The research was conducted with a cross-sectional survey method that aims to examine the existing aspects of participants. The sampling procedure has completed in two stages. At first, the Turkish Statistical Institute enlisted 650 street numbers for each income level (low, average, and high-income levels) which totals to 1950 street numbers among 456 neighborhoods in 7 major municipalities in Ankara by Stratified Random Sampling. Although the neighborhoods were determined by stratified random sampling, the participants in those neighborhoods were selected according to the purpose of the study and their approval. Therefore, the second stage of the sampling procedure was purposive. Accordingly, adult women who were being married and having mothers alive at the time of the study were invited to participate.

### *Participants*

In total, 560 married women ($\bar{X}_{\text{age}}$ = 38.6, $SD_{\text{age}}$ = 8.68, Range: 25-65 years) whose mothers were alive and did not need care at the time of the study were participants of the study.

The participants were divided into two groups according to whether they regularly helped their mothers in daily tasks (e.g., cleaning, shopping, arranging doctor appointments, handling bank accounts and payments etc.,) in the last 3 months. The first group included 304 women ($\bar{X}_{\text{age}}$ = 38.8, $SD_{\text{age}}$ = 8.57, Range: 23-65 years) labeled as "Daily Help (DH)". The second group included 256 women ($\bar{X}_{\text{age}}$ = 38.5, $SD_{\text{age}}$ = 8.78, Range: 25-65 years) and labeled as "No Daily Help (NDH)". In the DH group 60% (*n* = 184) of the women had a university degree or more and 65% (*n* = 195) were employed at the time of the study. Similarly, in the NDH group, 53% (*n* = 137) of the women had a university degree or more and 56% (*n* = 140) were employed at the time of the study. Considering the residency status of women in DH Group, 53% (*n* = 160) were living very close by with their mothers (in the same house, building or neighborhood), 30% (*n* = 92) were living in the same city but in distant neighborhoods, and 17% (*n* = 52) were living in a different city than their mothers. Considering the residency status of women in NDH Group, 29% (*n* = 73) were living very close by with their mothers, 35% (*n* = 90) were living in the same city but in distant neighborhoods, and 36% (*n* = 92) were living in a different city than their mothers.

## *Data Collection Instruments*

### *Adult attachment scale (AAS)*

The original scale (Cicirelli, 1991, 1995) consists of 16 items representing the basic aspects of secure attachment as seeking security or comfort (e.g. "At times when I have some trouble or difficulty, my mother's image seems to come to my mind"), distress upon separation (e.g. "If I am unable to see my mother for a long time, it bothers me a lot"), joy upon reunion (e.g. "When I have not seen my mother for a while, I feel happy when I see her again"), and feelings of love and closeness (e.g. "Being with my mother makes me feel very happy"). The factorial structure and psychometric properties of the scale were tested with a sample of adult daughters who were providing care for their elderly mothers at the time of the study and the exploratory factor analysis indicated 2 factors. After the elimination of one item that loaded heavily on both factors (Item 12: "When I have been away from my mother for a long time, I feel a sense of security to be with her again*"*) considerable overlap between the two factors was observed so the scale was regarded as unidimensional. The significant correlation of AAS with love (*r* = .73), trust (*r* = .60), and interpersonal antagonism (*r* = -.28) were stated in support of validity. Furthermore, adult daughters' attachment to their mothers assessed by the AAS was stated to be a better predictor of daughters' helping behavior than love, trust, and interpersonal antagonism. Lastly, AAS was reported to have considerable stability assessed by internal consistency reliability ($\alpha$ = .95) and one year test–retest reliability (*r* = .73).

The final version of the scale consists of 15 items and is rated on a 7-point Likert Scale (1 = *Totally Disagree* and 7 = *Totally Agree*). Higher scores are pointing to a stronger level of attachment to the mother.

### *Mother–adult daughter questionnaire (MAD)*

Originally developed by Rastogi (2002), MAD aims to assess the various aspects of adult daughters' current relationship with their mothers across different cultures. The questionnaire has 18 items rated on a 5-point Likert scale (1 = *Very False* and 5 = *Very True*) and 7 single items that are not included in the scale score but providing extra information about the mother-daughter relationship and recommended to be selected following the purpose of the research.

The Turkish version of MAD has shown to have good psychometric qualities (Onaylı, Erdur-Baker & Aksöz, 2010) and composed of 2 factors. The first factor is "Connectedness" (10 items) and represents the mutual ability to share feelings and opinions, as well as to make sacrifices within the context of the adult daughter-mother relationship (e.g. "I can share my intimate secrets with my mother; My mother can share her intimate secrets with me"). The second factor is "Trust in Hierarchy" (8 items) and represents the respect for the mothers' wisdom and her higher status in the family hierarchy; a reported positive aspect of an adult-parent relationship in collectivistic cultures (e.g. "I feel I can use my mother's wisdom as a resource when making decisions"). The strong correlation between MAD and parental bonding was established ($r$ = .69) to support validity. Test-retest reliability was satisfactory for 3 weeks interval ($r$ = .90) and internal consistency for the whole scale and two factors were between .88 and .91. For the present study, internal consistency was found .88 for "Connectedness", .87 for "Trust in Hierarchy", and .91 for the whole scale.

Regarding the purpose of the study, 2 items (among 7 single items of MAD) questioning the feeling of closeness and overall relationship satisfaction with the mother were selected.

*WHOTO*

This instrument (Hazan & Zeifman, 1994) was constructed to investigate individuals' attachment network and the relative primacy of significant others in the attachment hierarchy. In the present study, the revised version of WHOTO (Fraley & Davis, 1997; Trinke & Bartholomew, 1997) including 6 items for three attachment functions (proximity seeking, secure base, safe haven) was used. Example items can be listed as; "People you make sure to see or talk to frequently" for physical proximity seeking (PP), "People you immediately think of contacting when something bad happens" for safe haven (SH), and "People you know always wants the best for you" for secure base (SB). For each item, the participants are required to give 4 names in order of significance. The scores are from 4 (the first person listed) to 1 (the last person listed) with higher scores indicating the primacy of the figure. The primacy score for each attachment figure could be obtained both for each function separately and totally by averaging scores across each item.

WHOTO was tested with Turkish married women (Gündoğdu-Aktürk, 2010) and internal consistency for overall attachment primacy was established between .85 and .90 for primary attachment figures (spouse, mother, father, and children). Furthermore, satisfactory correlation between WHOTO and attachment avoidance ($r$ = -.43), marital satisfaction ($r$ = .40), and emotional caregiving style ($r$ = .40) was established to support the validity. For the study, only attachment primacy of mother was calculated for all attachment functions separately.

*Personal information form*

In addition to the scale items, the participants' age, level of education, employment status, residency status, frequency of contact with the mother in one week (face to face, phone, email etc.), whether the mother has any age-related health condition that needs care, and whether they provided regular help to their mothers in the daily tasks during the last 3 months were asked.

***Procedure***

The present study was approved by the ethical committee of the university where the research was being held. The study was completed as part of a larger project titled "Adult Daughter-Mother Attachment: The Relationship between Caregiving Style of Adult Daughter, Mothers' Psychological Well-Being and Future Care Seeking", granted by Turkish Academy of Sciences between 2014 and 2017.

Before the data collection, AAS (Cicirelli, 1995) was translated into Turkish utilizing translation and back-translation procedure by the researchers who had the command of both languages. Considering the range of SES and education level, the scale was constructed as 5 point Likert Scale (1 = *Totally Disagree;* 5 = *Totally Agree*) rather than 7 point as in the original form, in order to control the extreme responses and the high level of skewness and kurtosis (Hui & Triandis, 1989, Lozano, Garcia-Cueto & Muniz, 2008).

The data was collected from the psychology undergraduate students. Health centers, mukhtars, pharmacies, shopping malls, parks and other similar public areas in the neighborhoods listed by Turkish

Statistical Institute were visited and the research was completed with women who met the inclusion criteria of the research and volunteered to participate in the study. After signing the informed consent form, the scales were handed to the participants in an open envelope and asked to fill in that instant. The completed scales were received in a sealed envelope. Applications lasted approximately 15 minutes. In the end, participants were given a gift voucher of 10 Turkish Liras as a token of appreciation and were asked if they are willing to participate for the second time. Women who agreed to participate in the retest were asked for their contact number or email address.

Data collection process had been completed with 600 participants however, 40 of them were discarded due to missing data. The analyses were completed with 560 participants. Six months after the first test, the retest was completed with women (21%, $n$ = 120) who agreed to participate for the second time.

### *Data Analysis*

The factor structure of AAS was first tested for DH Group using EFA. Secondly, the factor structure obtained for the DH Group was confirmed for the NDH Group by CFA and the measurement invariance was tested.

The internal consistency of the scale was computed by Cronbach's Alpha coefficient. Also, test-retest reliability was assessed by Pearson correlation coefficients for the 6-month interval. Finally, Pearson's correlation coefficients of AAS with MAD and WHOTO as theoretically and empirically related and similar constructs were tested in support of convergent and concurrent validity.

## RESULTS

Prior to analysis data were screened for missing data and was found to be no more than 5% of the total number of items. Mean replacement was preferred for interval variables. Univariate and multivariate outliers, normality, and linearity were examined and assured for the data.

### *Exploratory Factor Analysis for DH Group (EFA)*

Before EFA, inter-item correlations for singularity, VIF (variance inflation factors), CI (condition indices), and TI (Tolerance Indices) for multicollinearity problems were examined for 15 items in the original AAS. For multicollinearity, VIF > 5, CI > 30, and TI < .20 were accepted as critical levels (James, Witten, Hastie & Tibshirani, 2014). Accordingly, the item 14 ("When I am with my mom I feel I am with someone whom I can totally depend on") was found to be critical in terms of multicollinearity (VIF = 4.75, CI = 36.46, TI = .21). This item also was found to have a high level of inter-item correlation ($r$ = .85) with item 9 ("When I am with my mom I feel I am with someone I can lean on"). When examined closely, it was noted that there was a subtle difference between the expressions of these two items in English, which might be lost in translation. Since multicollinearity statistics for item 9 were satisfactory, item 14 was decided to be discarded from further analyses.

After the elimination of one item from the original scale, the factorability of 14 AAS items for DH Group (N = 304) was examined. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = .95 (Field, 2013; Tabachnick & Fidell, 2013). The result of the Barlett test of sphericity ($\chi^2$ = 3706.1 $p$ < .001) were ensured that the data was appropriate for factor analysis. EFA was conducted initially regarding the 4-factor structure implied in the original study. However, it was found that the last two factors had eigenvalues less than Kaiser's criterion of 1 and the distribution of items was not conceptually and theoretically meaningful. Therefore, the research assumption about two-factor structure of the scale was tested.

A principal component analysis (PCA) was conducted to determine the factor structure. Based on the eigenvalues and the scree plot, a two-factor solution was indicated. Item 6 has been eliminated due to cross-loading, considering the loadings were above .30 for both factors (Tabachnick & Fidell, 2013). After the elimination of Item 6, interpretation of the two factors was conducted by Direct Oblimin rotation since the factors were conceptually related. Results showed that the variance explained by the first and second factor were 64.32% and 7.53% respectively, making 71.86% of total variance explained by 13 items of AAS. The factor loadings of the items are presented in Table 1.

Table 1. Factor Loadings, Mean and Standard Deviation for AAS

| | **Factor I** | **Factor II** | | |
|---|---|---|---|---|
| | **Secure Base** | **Safe Haven** | ***M*** | ***SD*** |
| Item 4 | .99 | -.16 | 4.07 | .83 |
| Item 9 | .87 | .00 | 3.94 | 1.00 |
| Item 7 | .87 | .04 | 3.93 | .94 |
| Item 15 | .82 | .10 | 3.92 | .93 |
| Item 11 | .79 | -.01 | 3.94 | .88 |
| Item 1 | .77 | .08 | 3.88 | .99 |
| Item 3 | .75 | .10 | 4.07 | .89 |
| Item 6 | .48 | .37 | 3.62 | 1.09 |
| Item 10 | -.08 | .95 | 3.33 | 1.23 |
| Item 5 | -.02 | .89 | 3.00 | 1.34 |
| Item 8 | -.05 | .87 | 3.07 | 1.27 |
| Item 13 | .18 | .74 | 3.31 | 1.23 |
| Item 2 | .13 | .66 | 3.26 | 1.15 |
| Item 16 | .06 | .66 | 3.29 | 1.34 |
| Eigenvalues | 9.00 | 1.10 | | |
| Exp. Variance | 64.32 | 7.53 | | |

When the distribution of items was examined, Factor 1 was seen to include items that represent the internalized aspect of attachment of which the mother was perceived as a sense of security even without the presence of active threat; therefore labelled as “Secure Base” (SB) (e.g. “When I am with my mother, I feel that I am with someone I can depend on”). Furthermore, Factor 2 was labeled as “Safe Haven” (SH) considering the items loaded on this factor were about the actual support seeking in the presence of threat (e.g. “If I am in trouble, the first person I want to talk to is my mother”).

### *Confirmatory Factor Analysis for NDH Group (CFA)*

Similar to the original study (Cicirelli, 1995), the EFA was conducted with adult women who provided instrumental help to their mothers in the last 3 months (DH Group). Additionally, and as suggested, the factor structure obtained by EFA was confirmed in the second group of adult women that was similar to the first group in terms of age, education level, and occupational status, but different in terms of daily help provided to the mothers. The second group (NDH) was composed of adult women who did not provide instrumental help to their mothers in the last 3 months for several different reasons (such as living away, the workload of a daughter, absence of parents need, presence of other children helping mother). The purpose of conducting CFA for the NDH group is to confirm the similar factor structure established by EFA in the DH Group and to establish measurement invariance.

Before CFA, inter-item correlations for singularity, VIF, CI, and TI for multicollinearity problems were examined and assured for the NDH Group.

CFA for NDH Group ($N = 256$) was conducted as a higher order construct of adult attachment composed of two factors as SB (7 items) and SH (6 items) established by EFA for DH Group. The model indices of 13 items were as follows: $\chi^2(64) = 215.52$ $p < .01$, $\chi^2$/sd = 3.37, GFI = .88, AGFI = .84, CFI = .94, TLI = .93, RMSEA = .09, suggesting a poor fit. Hence, the regression weights and modification indices pointed out that the fit of the model could be improved. Considering the modification indices, the error term of the item 4 was correlated with items 7 and 3 within the AAS-SB factor. Also, the error terms of items 10 and 13, 8 and 16 in the AAS-SH factor were correlated. As a result, the second model presented an adequate fit for 13 items with two factors ($\chi^2(59) = 232.92$, $p < .01$, $\chi^2$/sd = 3.9, GFI = .94, AGFI = .90, CFI = .96, TLI = .95, RMSEA = .07). As shown in Figure 1, weights for the regressions of item scores on their respective factors were between .56 and .88.

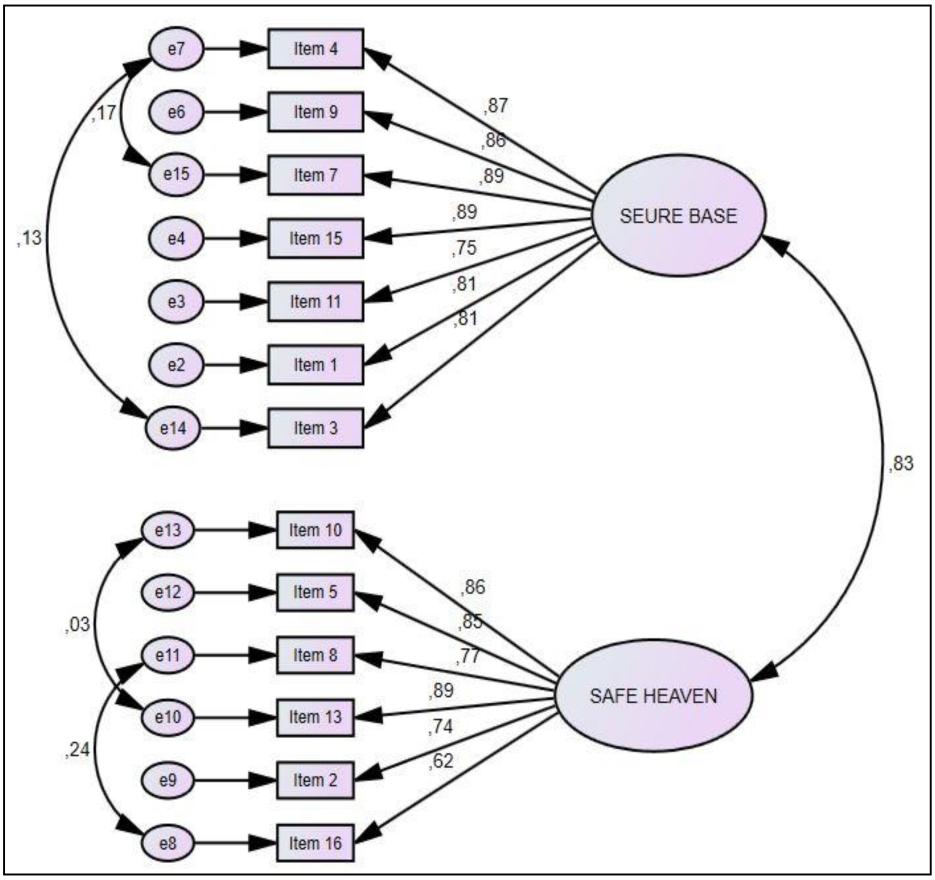


Figure 1. General Measurement Model for NDH Group

### *Measurement Invariance for DH and NDH Group*

Based on the acceptable results of CFA for NDH Group ($\chi^2(59) = 232.92$, $p < .01$, $\chi^2$/sd = 3.9, GFI = .94, AGFI = .90, CFI = .96, TLI = .95, RMSEA = .07) multi-group analysis was conducted for measurement invariance. At first, the CFA for DH Group was established as well for comparison ($\chi^2(60) = 172.95$, $p < .01$, $\chi^2$/sd = 2.88, GFI = .92, AGFI = .86, CFI = .97, TLI = .96, RMSEA = .07). The configural model yielded an adequate fit to the data as seen in Table 2.

Table 2. The Goodness of Fit Indices for Invariance Test and Results of $\chi^2$ Difference Tests

| | $\chi^2$ | *df* | $\triangle\chi^2$ | ($\triangle$*df*) | RMSEA | SRMR | CFI | $\triangle$CFI |
|---|---|---|---|---|---|---|---|---|
| CI | 332.14 | 120 | - | - | .05 | .04 | .965 | - |
| MI | 339.76 | 131 | 7.62 | 11 | .05 | .05 | .965 | .000 |
| FVI | 344.01 | 134 | 11.89 | 14 | .05 | .06 | .965 | .000 |
| FCI | 395.97 | 151 | 63.83 | 31 | .05 | .06 | .959 | .006 |

*Note 1*. DH Group *N* = 304; NDH Group *N* = 256.

*Note 2*. CI = configural invariance; MI = measurement invariance; FVI = factor variance invariance; FCI = factor covariance invariance.

Comparing the MI, FVI, and FCI models with the CI model, the changes in $\chi^2$ were nonsignificant. Further, the changes in CFI between CI and MI, between MI and FVI, and between FVI and FCI were either smaller than or equal to .01. These findings of invariance testing provided support for factorial invariance of the AAS scale across two groups.

### *Reliability: Internal Consistency and Test-Retest Reliability*

The final version AAS is composed of 13 items with Cronbach's Alfa coefficient of .95 for the total scale, .94 for the AAS-SB, and .89 for the AAS-SH. Item-total correlations ranged between .63 and .82 for AAS-SB and .42 and .73 for AAS-SH. Also, the correlation coefficients of AAS-SB and AAS-SH with total scale were .93 and .96 respectively. The correlation coefficient between the two factors was .78.

Stability of AAS was tested over the 6-month interval for the 21.4% of the participants (*n* = 120) and the significant correlation coefficient between time 1 and time 2 was established for AAS-SB (*r* = .75), AAS-SH (*r* = .69), and for the total scale score (*r* = .78).

### *Further Support for AAS: Convergent and Concurrent Validity*

To provide further support for the validity of AAS, correlation coefficients with theoretically and empirically related variables were computed. As presented in Table 3, the AAS scores were found to be weakly and negatively correlated with age (AAS total *r* = -.14, AAS-SB *r* = -.10, ASS-SH *r* = -.17), moderately and positively correlated with the frequency of contact with the mother (AAS total *r* = .40, AAS-SB *r* = .37, ASS-SH *r* = .38). Furthermore, AAS scores were found to be significantly and strongly correlated with MAD assessing the quality of the current relationship between mother and adult daughters in 4 subdomains, the correlation coefficients were ranging from .54 to .74.

Lastly, in support of the concurrent validity, AAS was found to be significantly correlated with WHOTO which assesses the primacy of the mother in the attachment hierarchy both in general and separately for each basic function of attachment (Physical Proximity, Secure Base, and Safe Haven). Although generally and consistently significant, the correlation coefficient of AAS-SB with WHOTO-SB (*r* = .45) were relatively stronger than WHOTO-SH (*r* = .35) and WHOTO-PP (*r* = .33). Also, the correlation coefficient of AAS-SH with WHOTO-SH (*r* = .54) and WHOTO-PP (*r* = .56) were relatively stronger than WHOTO-SB (*r* = .35) implicating the differential pattern of relationship for the 2 factors of AAS.

Table 3. Pearson Correlation Coefficients of AAS with Related and Similar Constructs

| | **AAS-SB** | **AAS-SH** | **AAS-TOT** |
|---|---|---|---|
| Age | $-.10^{*}$ | $-.17^{**}$ | $-.14^{**}$ |
| Frequency of Contact | $.37^{**}$ | $.38^{**}$ | $.40^{**}$ |
| MAD-Connectedness | $.70^{**}$ | $.69^{**}$ | $.74^{**}$ |
| MAD-Hierarchy Trust | $.48^{**}$ | $.66^{**}$ | $.62^{**}$ |
| MAD-Feeling of Closeness | $.55^{**}$ | $.58^{**}$ | $.60^{**}$ |
| MAD-Rlt. Satisfaction | $.54^{**}$ | $.55^{**}$ | $.58^{**}$ |
| WHOTO-SB | $.45^{**}$ | $.35^{**}$ | $.42^{**}$ |
| WHOTO-SH | $.35^{**}$ | $.54^{**}$ | $.50^{**}$ |
| WHOTO-PP | $.33^{**}$ | $.56^{**}$ | $.51^{**}$ |

*Note 1*. MAD=Mother-Adult Daughter Questionnaire, WHOTO-PP = Physical Proximity, WHOTO-SB = Secure Base, WHOTO-SH= Safe Haven, WHOTO-TOT= Total.
$^{*}p < .05$, $^{**}p < .01$

## DISCUSSION and CONCLUSION

In this study, the Turkish validity and reliability of AAS which was specifically developed for the purpose of assessing adult daughters' attachment to their aging mothers, was tested.

When the structural analyses were considered, results were consistent with the original study, and beyond that, a more coherent picture was provided in support of the construct validity. Although AAS was originally developed depending on the four basic aspects of secure attachment (seeking security or comfort, distress upon separation, joy upon reunion, and feelings of love and closeness), the result of the factor analysis was reported to be unexpected and inconclusive. Thus, AAS has been tentatively regarded as a unidimensional construct. By pointing out the small size and homogeneous structure of the sample in the original study, Cicirelli (1995) had stated that the construct validity of AAS should be tested in a larger and heterogeneous sample. More specifically, Cicirelli (1995) emphasized the importance of replication with a group of adults who could not help their parents for any reason, recalling that AAS was only developed with adult daughters who regularly helped their mothers on daily tasks. Regarding this, the present study was designed with a relatively larger sample size that was separated into two groups as adult daughters who did and did not provide regular help to their mothers on daily tasks.

First, EFA was conducted with a sample of adult women, who had similar characteristics to the sample of the original study in terms of the instrumental help provided regularly to mothers (DH Group). Depending on the preliminary results, two items were discarded from the Turkish version of the scale either due to multicollinearity or high cross-loading. After discarding the 2 items, the two-factor solution and the distribution of items were found to be similar to the original study only with some minor differences that made the interpretation conceptually more meaningful. When the content of the factors examined closely, rather than four-factor structure proposed by Cicirelli (1995) an alternative and theoretically more meaningful perspective for labeling the factors popped out. Accordingly, factors were labeled as “Secure Base” and “Safe Haven”; two basic functions of attachment one of which represents the source of security without the presence of active support seeking and the other one of which represents the actual seek of support in the presence of a threat. This factor structure established by EFA was further validated by CFA with the second group of adult daughters who were not providing instrumental help to their mothers regularly (NDH Group). The results revealed that the factor structure obtained was valid for both samples. Furthermore; measurement, factor variance, and factor covariance invariance were demonstrated for two samples by multi-group analysis and the results were considered as strong support for the structural validity of AAS.

Concurrent validity of AAS was tested by examining its correlation with WHOTO a scale of which, assesses attachment network and the relative primacy of significant others in the attachment hierarchy. AAS and WHOTO are similar not only conceptually but also structurally. However, the major difference between these two scales is that WHOTO requires an evaluation of the relative importance of the many significant others in one’s life. Because of that, WHOTO is sensitive to the width of attachment network and the scale score might vary according to marital status, death of a family member, number of siblings, children, friends, and relatives, etc., which might complicate the interpretation of scale score. In contrast, AAS requires a relationship-specific evaluation independent of the presence of other attachment figures. In favor of concurrent validity, results presented that AAS, in general, is positively correlated with WHOTO. This means that as adult daughters’ level of attachment to mother increases, the mothers’ priority in the attachment hierarchy increases compared to other attachment figures. Besides, it was noted that similar dimensions (e.g. AAS-SB & WHITE-SB) on both scales were more strongly related to each other than non-similar dimensions (e.g. AAS-SB & WHOTO-SH). Although the statistical significance of these differences in correlation coefficients has not been tested, such a pattern could be considered as remarkable.

Convergent validity of AAS was tested by examining its correlation with theoretically related concepts such as general relationship quality assessed by MAD and frequency of contact. As expected, AAS was found to be strongly correlated with all MAD subscales and moderately correlated with the frequency of contact. Although weak, significant negative correlation between AAS and the age of adult daughter also has been found. This finding is consistent with the attachment theory which states that the importance of the mother as attachment figure decreases in time and that attachment is transferred to friends, romantic partners, and spouses over time (Rosenthal & Kobak, 2010). This result should also be considered as critical in pointing out the appropriate way of interpretation of the scale score. Accordingly, it should be noted that attachment level which is sensitive to developmental changes in the attachment network, is not necessarily accounted as the attachment security that is resistant to change. Thus, for future research, it is strongly advised that AAS is used and interpreted in its conceptual limits.

To sum up, the Turkish version of AAS could be accepted as a reliable and valid measure of the level of adult daughters’ current attachment to their mothers. AAS-Turkish is composed of two conceptually related factors that can be utilized both as separate scores to point out the significance of mother either as a secure base or haven and as a total score pointing out the significance of mother as an attachment figure. The scale should be fruitful for researchers in testing certain predictions and understanding the dynamics of attachment bond between adult children and their elderly parents. Given that the global population is rapidly aging (He et al., 2016), that the cultural expectations for adult daughters as being primary caregivers for the elderly are increasing (Ataca et al., 2005; Finley, 1989; Ingersoll-Dayton et.al., 1996; İmamoğlu, 1987; Kağıtçıbaşı, 1985; Zahn & Montgomery, 2003), and that attachment theory still offers little to understand the dynamics of attachment relationships in late life AAS could be a valuable tool in providing preliminary answers.

## REFERENCES

Ainsworth, M. D. S. (1989). Attachment beyond infancy. *American Psychologist, 44*(4), 709-716. Retrieved from https://pdfs.semanticscholar.org/ 5431/41e657bda74736ff87ac10d70643cd639892.pdf

Anderson, G. F., & Hussey, P.S. (2000). Population aging: A comparison among industrialized countries populations around the world are growing older, but the trends are not cause for despair. *Health Affairs, (19)*3, 191-203. Retrieved from http://hcim.di.fc.ul.pt/hcimwiki/images/3/3a/Anderson2000-PopulationAging.pdf.

Ataca, B., Kağıtçıbaşı, Ç., & Diri, A. (2005). The Turkish family and the value of children: Trends over time. In G. Trommsdorff, & B. Nauck (Eds.), T*he value of children in cross-cultural perspective: Case studies from eight societies* (pp. 91-119). Lengerich, Germany: Pabst Science.

Bengston, V. L., & Oyama, P. S. (2007). Intergenerational solidarity and conflict: Strengthening economic and social ties. *Department of Economic and Social Affairs Division for Social Policy and Development.* Retrieved from https://www.un.org/esa/socdev/unyin/documents/egm_unhq_oct07_bengtson.pdf.

Bengtson, V. L., Rosenthal, C. J., & Burton, L. M. (1990). Families and aging: Diversity and heterogeneity. In R. Binstock, & L. George (Eds.), *Handbook of Aging and the Social Sciences* (pp. 263-287). New York, NY: Academic.

Bengtson, V., Giarrusso, R., Mabry, J. B., & Silverstein, M. (2002). Solidarity, conflict, and ambivalence: complementary or competing perspectives on intergenerational relationships? *Journal of Marriage and Family, 64*(3), 568-576 doi: 10.1111/j.1741-3737.2002.00568.x

Bloom, D. E., Canning, D., & Lubet, A. (2015). Global population aging: Facts, challenges, solutions & perspectives. *Daedalus, 144*(2), 80–92. doi:10.1162/daed_a_00332

Bowlby, J. (1969/1982). *Attachment and loss: Vol 1: Attachment*. New York, NY: Basic Books.

Bowlby, J. (1979). *The making and breaking of affectional bonds.* London: Tavistock.

Cicirelli, V. G. (1983). Adult children’s attachment and helping behavior to elderly parents: A path model. *Journal of Marriage and the Family, 45*(4), 815-825. Retrieved from https://www.jstor.org/stable/351794

Cicirelli, V. G. (1993). Attachment and obligation as daughters’ motives for caregiving behavior and subsequent effect on subjective burden. *Psychology and Aging, 8*(2), 144-155. doi: 10.1037/0882-7974.8.2.144

Cicirelli, V. G. (1995). A measure of caregiving daughters’ attachment to elderly parents. *Journal of Family Psychology, 9*(1), 89-94. doi: 10.1037/0893-3200.9.1.89

Cicirelli, V. G. (2010). Attachment relationships in old age. *Journal of Social and Personal Relationships, 27*(2), 191-199. doi: 10.1177/0265407509360984

Cicirelli, VG. (1991). *Family caregiving: Autonomous and paternalistic decision making.* Newbury Park, CA: Sage.

Doherty, N. A., & Feeney, J. A. (2004). The composition of attachment networks throughout the adult years. *Personal Relationships, 11*(4), 469-488. doi: 10.1111/j.1475-6811.2004.00093.x

Field, A. (2013). *Discovering statistics using IBM SPSS statistics: And sex and drugs and rock “n” roll*, (4th ed.). London: Sage.

Finley, N. J. (1989). Theories of family labor as applied to gender differences in caregiving for elderly. *Parents Journal of Marriage and Family, 51*(1), 79-86. Retrieved from https://www.jstor.org/stable/352370?seq=1&cid=pdf-reference#references_tab_contents

Fraley, R. C., & Davis, K. E. (1997). Attachment formation and transfer in young adults’ close friendships and romantic relationships. *Personal Relationships, 4*(2), 131-144. doi: 10.1111/j.1475-6811.1997.tb00135.x

Gündoğdu-Aktürk, E. (2010). *Attachment figure transference, caregiving styles and marital satisfaction in arranged and love marriages* (Unpublished master’s thesis). Middle East Technical University, Ankara Retrieved from https://toad.halileksi.net/sites/default/files/pdf/yakin-iliskilerde-bakim-verme-olcegi-toad.pdf

Hazan, C., & Zeifman, D. (1994). Sex and the psychological tether. In K. Bartholomew & D. Perlman (Eds.), *Advances in personal relationships, Vol. 5: Attachment processes in adulthood* (pp. 151-177). London: Jessica Kingsley.

He, W., Goodkind, D., & Kowal, P. R. (2016). *An aging world: 2015*. Washington, DC: United States Census Bureau.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventionalcriteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*(3), 296-309. doi: 10.1177/0022022189203004

İmamoğlu, E. O. (1987). An interdependence model of human development. In Ç. Kağıtçıbaşı (Ed), *Growth and progress in cross-cultural psychology* (pp.138-145). Lisse, Netherlands: Swets & Zeitlinger.

Ingersoll-Dayton. B, Starrels, M. E., & Dowler, D. (1996). Caregiving for parents and parents-in-law: Is gender important. *The Gerontologist*, *36*(4), 483-491. doi: 10.1093/geront/36.4.483

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning: With applications in R* (pp. 59-120). New York, NY: Springer.

Kağıtçıbaşı, Ç. (1985). Intra-family interaction and a model of change. In T. Erder (Ed.), *Family in Turkish society* (pp. 149-166). Ankara: Turkish Social Science Association.

Kontis, V., Bennett, J. E., Mathers, C. D., Li, G., Foreman, K., & Ezzati, M. (2017). Future life expectancy in 35 industrialised countries: Projections with a Bayesian model ensemble. *The Lancet*, *389*(10076), 1323-1335. doi: 10.1016/S0140-6736(16)32381-9

Lowenstein, A. (2005). Global ageing and challenges to families. In M. Johnson (Ed.), *The Cambridge handbook of age and ageing* (pp. 403-412). Cambridge: Cambridge University. doi:10.1017/CBO9780511610714.042

Lowenstein, A., Katz, R., & Gur-Yaish, N. (2007). Reciprocity in parent-child exchange and life satisfaction among the elderly: A cross-national perspective. *Journal of Social Issues*, *63*(4), 865-883. doi:10.1111/j.1540-4560.2007.00541.x

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, *4*(2), 73-79. doi: 10.1027/1614-2241.4.2.73

Manuela, N., Emmanuele, P., & Cristina, S. (2016). Female employment and elderly care: the role of care policies and culture in 21 European countries. *Work Employment and Society*, *30*, 607-630. Retrieved from https://iris.unito.it/retrieve/handle/2318/1559499/132926/Naldini%20et%20al_Elderly%20care_WES_2016.pdf

Martin, P., & Baek, Y. (2018). *Centenarian in the SAGE encyclopedia of lifespan human development*. Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/ 9781506307633.n126.

Onaylı, S., Erdur-Baker, Ö., & Aksöz, A. (2010). The Turkish adaptation of the mother-adult daughter questionnaire. *Procedia-Social and Behavioral Sciences, 5,* 1516-1520. doi: 10.1016/j.sbspro.2010.07.318

Rastogi, M. (2002). The mother-adult daughter questionnaire (MAD): Developing a culturally sensitive instrument. *The Family Journals, 10*(2), 145-155. doi: 10.1177/1066480702102004

Rochon, P. A., Gruneir, A., Wu, W., Gill, S. S., Bronskill, S. E., Seitz, D. P., …& Anderson, G. M. (2014). Demographic characteristics and healthcare use of centenarians: A population-based cohort study. *Journal of the American Geriatrics Society, 62*(1), 86-93. doi:10.1111/jgs.12613

Rosenthal, N. L., & Kobak, R. (2010). Assessing adolescents’ attachment hierarchies: Differences across developmental periods and associations with individual adaptation. *Journal of Research on Adolescence, 20*(3), 678-706. doi: 10.1111/j.1532-7795.2010.00655.x

Rowland, J. H., & Bellizzi, K. M. (2014). Cancer survivorship issues: Life after treatment and implications for an aging population. *Journal of Clinical Oncology, 32*(24), 2662-2668. doi:10.1200/jco.2014.55.8361

Sroufe, L. A., Fox, N. E., & Pancake, V. R. (1983). Attachment and dependency in developmental perspective. *Child Development, 54*(6), 1615-1627. Retrieved from https://www.jstor.org/stable/1129825

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th. Ed.). Boston, MA: Pearson.

Trinke, S. J., & Bartholomew, K. (1997). Hierarchies of attachment relationships in young adulthood. *Journal of Social and Personal Relationships, 14*(5), 603-625. doi: 10.1177/0265407597145002

Trommsdorff, G., & Nauck, B. (2006) Demographic changes and parent–child relationships. *Parenting, 6*(4), 343-360. doi: 10.1207/s15327922par0604_4

Waters, E., Kondo-Ikemura, K., Posada, G., & Richters, J. E. (1991). Learning to love: Mechanisms and milestones. In M. Gunner & L.A. Sroufe (Eds.), *Self process in development*. The Minnesota Symposia on Child Psychology, Vol. 23. Hillsdale, NJ: Erlbaum.

Waters, E., Merrick, S., Treboux, D., Crowell, J., & Albersheim, L. (2000). Attachment security in infancy and early adulthood: A twenty- year longitudinal study. *Child Development*, *71*(3), 684-689. doi:10.1111/1467-8624.00176

# Yetişkin Kız - Anne Bağlanması: Yetişkin Bağlanma Skalası Türkçe Versiyonunun Psikometrik Özellikleri

## *Giriş*

Dünya nüfusu gittikçe artan bir hızda yaşlanmaktadır (Bloom, Canning & Lubet, 2015; Lowenstein, 2005; Trommsdorff & Nauck, 2006). Gelecek 30 yıl içerisinde Avrupa nüfusunun %28'nin 65 yaş üstü olacağına işaret edilmektedir (He, Goodkind & Kowall, 2016). Gün geçtikçe hızlanan bu değişimin sağlık sektörü için taşıdığı risklerle birlikte, sosyal yaşantı ve aile yapısında, bugünden gözlemlenmeye başlayan, olumlu ve olumsuz sonuçları da tahmin edilmektedir. Buna göre; bir ailede üç ve hatta dört neslin bir arada geçireceği yıllar artmakta, artan sağlık yaşam kalitesi büyük ebeveynlerin aile dinamiklerindeki rol ve sorumluluklarını farklılaştırmaktadır. Öte yandan, yaşlanan ve bakıma ihtiyaç duyan aile üyelerine bakım sorumluluğu ve süresi de artmaktadır. Birçok farklı kültürde, yaşlanan aile üyesine bakım verme sorumluluğunun kız çocuğunda olduğu (Ataca, Kağıtçıbaşı, & Diri, 2005; İmamoğlu, 1987; Ingersoll-Dayton, Starrels, & Dowler, 1996; Zhan & Montgomery, 2003) ve son yıllarda daha fazla kadının yaşlı ebeveynine bakım vermek için işten ayrıldığı rapor edilmektedir (Manuela, Emmanuele, & Cristina, 2016).

Bu değişimlerin işaret ettiği riskler ve çözüm ihtiyacı ile paralel olarak alan yazındaki araştırmaların arttığı, psikobiyolojik boyutta yaşlı sağlığı, yaşlı bakım uygulamaları, bakım veren yükü ve stresi (örn., Feng, Liu, Guan & Mor, 2012; Rowland & Bellizzi, 2014), sosyolojik boyutta ise nesiller arası dayanışma ve evlat yükümlülüğü (örn., Bengtson, Rosenthal & Burton, 1990; Bengston & Oyama, 2007) gibi konuların ağırlıklı olarak vurgulandığı dikkat çekmektedir. Bu konuların önemi yadsınamaz olmakla birlikte, bazı araştırmacılar (örn., Bengtson, Giarrusso, Mabry & Silverstein, 2002; Lowenstein, Katz & Gur-Yaish, 2007; Schwarz, Trommsdorff, Kim & Park, 2006) büyük resimdeki önemli bir boşluğa işaret etmekte; bakım alan ve veren arasındaki duygusal bağın göz ardı edildiğine vurgu yapmaktadır. Yaşlanan aile bireyine bakım vermede her ne kadar sorumluluk duygusu, evlat yükümlülüğü, zorunluluk, gereklilik ve saygı gibi faktörler belirleyici olsa da, bu faktörlerin sunulan bakımın kalitesi ve etkinliği ile ilişkili olmadığı, her iki tarafın deneyimlediği duygusal stres, fiziksel ve psikolojik yaşam kalitesinde de temel etken olmadığı görülmektedir. Bu noktada, yetişkin çocuğun yaşlı ebeveyni ile kurduğu duygusal bağ kalitesinin kritik bir öneme sahip olduğu ifade edilmekte ve bağlanma kuramına işaret edilmektedir (örn., Bengtson, Giarrusso, Mabry & Silverstein, 2002; Lowenstein, Katz & Gur-Yaish, 2007).

Bağlanma kuramı, bağlanmanın yaşamsal önemini vurgulaması ve ömür boyu bağ kurma ihtiyacının normalliğine işaret etmesi açısından özgün ve etkili bir yaklaşım ortaya koymaktadır. Yaşamın en erken döneminde, ilk olarak çoğunlukla anneyle kurulan güvenli bağın sevme, önemseme, umursama, yardım etme ve yardım isteme potansiyelini şekillendirdiği (Waters, Kondo-Ikemura, Posada, & Richters, 1991) dolayısıyla, yaşam boyu tüm yakın ilişkilerdeki duygu ve davranışların düzenlenmesinde kritik bir faktör olduğu belirtilmektedir (Ainsworth, 1989; Bowlby, 1979; Waters, Merrick, Treboux, Crowell, & Albersheim, 2000).

Yalnızca tek bir kişiyle kurulmayan bağlanmanın, sınırlı sayıda önemli diğerlerinden oluşan hiyerarşik bir ağ olduğu (Bowlby,1969/1982), bu ağın gelişimsel süreçte değiştiği ve birincil bağlanma figürünün zamanla ebeveynden arkadaşa, arkadaştan eşe transfer edildiği ifade edilmektedir (Rosenthal & Kobak, 2010). Ancak babanın aksine, annenin statüsü değişse de her zaman bağlanma ağında yeri olduğu ve hatta yaşamdaki gelişimsel dönüm noktalarında (ilk kez anne/baba olmak gibi) annenin hiyerarşideki birincil pozisyona tekrar yükselebildiği belirtilmektedir (Doherty & Feeney, 2004; Rosenthal & Kobak, 2010). Bir bağlanma figürü olarak annenin yaşam boyu devam eden rolü ampirik olarak desteklenmiş olsa da yetişkin çocuk ve ebeveyni arasındaki duygusal bağın dinamikleri üzerine alan yazında oldukça sınırlı sayıda çalışma bulunmaktadır. Bu noktada Cicirelli'nin araştırmaları (1983, 1991, 1993, 1995, 2010), yetişkin çocukların yaşlı ebeveynlerine hem duygusal anlamda ilgi göstermeleri hem de gündelik işler açısından yardımcı olmalarındaki motivasyonun anlaşılmasında bir başlangıç olmuştur. Alan yazındaki boşluğa ve yetişkin çocuğun anneyle devam eden bağlanmasının değerlendirilmesinde uygun bir aracın yokluğuna vurgu yapan Cicirelli (1991), bu amaçla Yetişkin Bağlanma Skalası'nı (YBS) geliştirmiştir.

YBS, gündelik işlerde annelerine düzenli olarak yardım eden kadınlarla geliştirilmiş ve farklı araştırmalarda amaca yönelik test edilmiştir (Cicirelli, 1991, 1993, 1995). Buna göre; yetişkin kızların yardım sıklığı ile algıladıkları bakım verme yükü arasındaki ilişkiyi inceleyen Cicirelli (1993), öncelikle yardım sıklığının ebeveynin yardıma ihtiyaç duyması, yetişkin kızın yükümlülük hissi ve bağlanma düzeyi ile ilişkili olduğunu belirtmiştir. Ancak, yardım sıklığından bağımsız olarak algılanan bakım verme yükünün hissedilen yükümlülük ile pozitif, bağlanma düzeyiyle ise negatif yönde ilişkili olduğunu ortaya koymuştur. Alan yazındaki bu ilk çalışmalar, bir yandan yetişkin çocuk - anne arasındaki duygusal bağın anlaşılmasının kritik önemine bir yandan da YBS' nin ilerideki araştırmalarda değerli bir araç olarak kullanılabileceğine işaret etmektedir.

Türk kültüründe, yaşlanan ebeveyne bakım verme yükümlülüğünün ağırlıklı olarak kız çocuğunda olması (Ataca, Kağıtçıbaşı, & Diri, 2005; İmamoğlu, 1987), bakım verme sürecinde karşılıklı kurulan duygusal bağın önemi yadsınamaz olmakla birlikte araştırmacılar tarafından konunun göz ardı edilmiş olması, yetişkin çocuk - anne bağlanması üzerine ulusal ve uluslararası alan yazında sınırlı sayıda bilgiye ulaşılması nedeniyle ve konuyla ilgili araştırmalara bir başlangıç noktası olması amacıyla, bu çalışmada YBS' nin Türkçe 'ye uyarlaması hedeflenmiştir.

## *Yöntem*

Araştırmaya, yaşları 25-65 arasında değişen, evli ve anneleri halen hayatta olan yetişkin kadınlar katılmıştır. Katılımcılar, son 3 ayda gündelik işlerde (örn., temizlik, alışveriş, banka işleri, sağlık işlerinin takibi vb.) annelerine düzenli yardım etme durumuna göre iki gruba ayrılmıştır. Düzenli yardım eden (DY) grup 304 ($\bar{X}_{\text{yaş}}$ = 38.8), düzenli yardım etmeyen (DYY) grup ise 256 ($\bar{X}_{\text{yaş}}$ = 38.5) kişiden oluşmuştur. Hiçbir annenin çalışmanın yapıldığı dönemde ileri yaşa bağlı bakım ihtiyacı olmaması araştırmaya katılma kriteri olarak dikkate alınmıştır.

### *Veri toplama araçları*

Yetişkin Bağlanma Skalası (YBS) (Cicirelli, 1991, 1995), yetişkin bireyin ebeveynine bağlanma düzeyini değerlendirmeyi amaçlayan; güvenli bağlanmanın 4 tanımlayıcı özelliği olan güvenlik arayışı, ayrılık stresi, fiziksel yakınlıktan keyif alma, sevgi ve yakınlık hissi boyutları temel alınarak oluşturulmuş 16 maddelik, 7 noktalı Likert tipi bir ölçektir. Orijinal çalışmada, açımlayıcı faktör analizi 2 boyut ortaya koymuş ancak, bu boyutların önemli düzeyde örtüşmesi nedeniyle ölçek tek boyut olarak kabul edilmiştir. Güvenirliği düşük olan bir maddenin çıkarılmasıyla YBS, 15 maddelik tek boyutlu bir ölçek olarak sunulmuştur. Ölçeğin geçerliği kapsamında sevgi ($r$ = .73), güven ($r$ = .60) ve kişilerarası çatışma ($r$ = -.28) kavramlarıyla ilişkisi desteklenmiştir. Ayrıca, YBS ile değerlendirilen bağlanma düzeyinin, sevgi, güven ve kişilerarası çatışma düzeyinden daha güçlü olarak anneye yardım etme davranışlarını yordadığı da gösterilmiştir. Ölçeğin iç tutarlık katsayısı .95, bir yıllık test-tekrar test güvenirliği ise .73 olarak tespit edilmiştir.

Anne-Yetişkin Kız Ölçeği (AYKÖ) (Rastogi, 2002), yetişkin kızların anneleriyle bugünkü ilişkilerini kültürel farklılıkları da dikkate alarak değerlendirmeyi hedefleyen, 18 maddelik, 5 noktalı Likert tipi bir ölçektir. Türkçe uyarlaması yapılan AYKÖ (Onaylı, Erdur-Baker, & Aksöz, 2010), "Bağlılık" ve "Hiyerarşiye Güven" olarak 2 alt ölçekten oluşmaktadır. Test-tekrar test güvenirlik katsayısı .90, iç tutarlık katsayısı ise "Bağlılık" alt ölçeği için .88, "Hiyerarşiye Güven" alt ölçeği içinse .87 olarak belirtilmiştir.

Bağlanma ağı ve bu ağa dahil olan kişilerin hiyerarşik pozisyonunu değerlendirmeyi hedefleyen KİME ölçeği (Fraley & Davis, 1997; Hazan & Zeifman, 1994; Trinke & Bartholomew, 1997), bağlanmanın fiziksel yakınlık arayışı (FY), güvenli sığınak (GS) ve güvenli üs (GÜ) işlevini temel alan 6 maddeden oluşmaktadır. Her madde için önem sırasına göre dört kişinin sıralanması beklenmektedir. Puanlar 1 (listelenen son kişi) ila 4 (listelenen ilk kişi) arasında değişmekte ve yüksek puan, listelenen kişinin bağlanma hiyerarşisindeki önceliğine işaret etmektedir. Puanlar hem bağlanma işlevleri için ayrı olarak hem de toplam puan olarak hesaplanabilmektedir. Ölçeğin Türkçe uyarlaması yapılmış (Gündoğdu-Aktürk, 2010) ve iç tutarlık katsayısı .85 - .90 aralığında rapor edilmiştir.

### *İşlem*

Bu çalışma, 2014-2017 yılları arasında TÜBİTAK tarafından desteklenen bir proje kapsamında yapılmıştır. Çalışma öncesi gerekli etik kurul onayı alınmıştır.

YBS' nin (Cicirelli, 1995) Türkçe çevirisi, her iki dile hakim uzmanlarca, çeviri ve geri-çeviri yöntemiyle yapılmıştır. Ölçek, aşırı uç değerlere yığılmayı önlemek amacıyla 5 noktalı Likert tipi olarak düzenlenmiştir.

## *Sonuç ve Tartışma*

Çalışmada ilk olarak YBS'nin faktör yapısı, Temel Bileşenler Analizi ile DY grubunda (N = 304) test edilmiştir. Özdeğeri 1'den büyük olan 2 faktör yapısı gözlemlenmiştir. Özdeğerlerin çizgi grafik dağılımı ve madde dağılımının kuramsal tutarlılığı dikkate alınarak 2 faktörlü çözümlemenin uygunluğuna karar verilmiştir. İki maddenin çapraz yükleme ve çoklu bağlantı nedeniyle çıkarılması sonrasında, toplam 13 madde için 2 faktörlü yapı, eğik rotasyon ile tekrar test edilmiştir. Açıklanan varyans tüm ölçek için %71.86, birinci faktör (7 madde) için %64.3, ikinci faktör (6 madde) için ise %7.53 olarak tespit edilmiştir (Tablo1). Madde dağılımları incelendiğinde birinci faktörün altında, tehdit/tehlike olmadığı anlarda hissedilen içselleştirilmiş güvenlik hissiyle ilgili maddelerin (örn., *Annemle birlikte olduğum zaman güvenebileceğim biri ile birlikte olduğumu hissederim)* toplandığı görülmüş ve bu faktör 'Güvenli Üs' (GÜ) olarak isimlendirilmiştir. İkinci faktörde ise tehlike anında aktif destek arayışıyla ilişkili maddelerin (örn., *Bir zorluk yaşadığımda konuşmak istediğim ilk kişi annemdir*) toplandığı görülmüş ve bu faktör de "Güvenli Sığınak" (GS) olarak isimlendirilmiştir.

DY grubunda elde edilen ölçek yapısının, DYY grubunda doğrulanması ve ölçüm değişmezliğinin test edilmesi amacıyla Doğrulayıcı Faktör Analizi yapılmıştır. Global uyum iyiliği gösterge değerlerine göre, verinin ilk modele iyi uyum göstermediği bulunmuştur ($\chi^2(64)$ = 215.52 p<.01, $\chi^2$/sd = 3.37, GFI=.88, AGFI = .84, CFI = .94, TLI = .93, RMSEA = .09). Modifikasyon göstergeleri doğrultusunda yapılan düzenleme sonucunda verinin modele kabul edilebilir düzeyde uyum sağladığı bulunmuştur ($\chi^2(59)$ = 232.92, p<.01, $\chi^2$/sd = 3.9, GFI = .94, AGFI = .90, CFI =.96, TLI =.95, RMSEA = .07). Doğrulayıcı faktör analizi sonucunda elde edilen bu modelin DYY ve DY grupları için değişmezliği, çoklu grup analizi ile test edilmiştir. Metrik, ölçek ve katı değişmezlik modellerinin biçimsel değişmezlik modelinden anlamlı düzeyde farklı olmadığı tespit edilmiştir. Elde edilen uyum istatistikleri ve uyum katsayılarına ait fark değerleri Tablo 2'de sunulmuştur.

Açımlayıcı ve doğrulayıcı faktör analizler sonucunda, 2 boyutlu 13 maddelik bir ölçek olarak yapılandırılan YBS'nin tüm ölçek için iç tutarlık katsayısı .95, YBS-GÜ için .94 ve YBS-GS için .89 olarak tespit edilmiştir. YBS'nin test tekrar-test güvenirliği 6 ay arayla yapılmış ve YBS-GÜ için .75, YBS-GS için . 69 ve toplam puan için .78 olarak tespit edilmiştir.

Yakınsak geçerliği destekler nitelikte, YBS'nin anneyle iletişime geçme sıklığı (*r* = .37 - .40), ilişki tatmini (*r* = .54 - .58), yakınlık hissi (*r* = .55 - .60), AYKÖ-Bağlılık (*r* = .69 - .74) ve AYKÖ-Hiyerarşiye Güven (*r* = .48 - .66) ile pozitif yönde anlamlı ilişkisi gösterilmiştir. Ayrıca, eş zaman geçerliğini destekler nitelikte YBS'nin, bağlanmada hiyerarşi önceliğini değerlendiren KİME'nin tüm alt boyutlarıyla anlamlı, pozitif ve tutarlı ilişki tespit edilmiştir. Buna ek olarak, YBS-GÜ'nün KİME-GÜ (*r* = .45) ile korelasyon katsayısının KİME-GS (*r* = .35) ve KİME-FY'ye (*r* = .33) kıyasla görece daha yüksek olduğu, benzer şekilde YBS-GS'nin KİME-GS (*r* = .54) ve KİME-FY (*r* = .56) ile korelasyon katsayısının KİME-GÜ'ye (*r* = .35) kıyasla görece daha yüksek olduğu dikkat çekmiştir.

Sonuç olarak, yetişkin kızların yaşlanmakta olan annelerine bağlanma düzeylerini değerlendiren YBS Türkçe versiyonunun geçerliği ve güvenirliği ampirik olarak desteklenmiştir. Birbiriyle ilişkili 2 alt ölçekten oluşan YBS-Türkçe, yetişkin kız için annenin güvenli üs ve sığınak olarak önemini vurgulamak amacıyla ayrı puanlanabildiği gibi, annenin güvenli bağlanma figürü olarak önemini vurgulamak amacıyla toplam puan olarak da değerlendirilebilmektedir. Yaşlanan dünya nüfusu, hızla değişmekte olan aile yapısı ve dinamikleri, yaşanan sosyodemografik değişimlere yönelik öngörülen riskler, yaşlanan ebeveyn ve yetişkin çocuğun yaşam boyu devam eden ilişkilerinin duygusal niteliği ve bağlanma dinamiklerine yönelik alan yazındaki sınırlı bilgi dikkate alındığında, YBS'nin bu alandaki çalışmaların artmasına öncü olacağı ve önemli bilgilerin elde edilmesi için değerli bir araç olacağı düşünülmektedir.

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(4); 467-479

# Academic Jealousy Scale: Validity and Reliability Study

Duygu KOÇAK *

**Abstract**
This study aims to develop a measurement tool to determine the level of academic jealousy. For this purpose, firstly, the literature was examined, and a theoretical framework was formed, and then an item set of 47 items was created. The items that were submitted to expert opinion were eliminated and corrected, and 41 items were decided to use, and a trial form was obtained. In this study, 478 university students were reached. One-on-one interviews were conducted with ten students who are studying at Alanya Alaaddin Keykubat University before the trial application. Then, the trial form was applied to 254 university students, and the data obtained were analyzed with Exploratory Factor Analysis, and as a result of this analysis, a structure with three factors (maturity, self-denigration, and envy) was revealed. In order to test the defined structure, the final form of the scale was applied to another group of 154 people, and the data obtained were subjected to Confirmatory Factor Analysis, and goodness of fit indices of the scale was found to be between good fit or acceptable fit. Accordingly, the structure with 19-item and three factors was confirmed Cronbach Alpha internal consistency coefficient of the whole scale was found .779, Envy subfactor was found .840; self-denigration subfactor was found .840 and Maturity subfactor was found .817. In order to determine the reliability of the scale in terms of stability, the scale was applied to a different group of 57 people at two-week intervals, and the correlation between the two applications was recorded as .89. It was concluded that the Academic Jealousy Scale developed according to these findings, is a valid measurement tool and will give reliable scores in measuring academic jealousy.

*Key Words:* Jealousy, envy, academic jealousy, academic achievement, scale development.

## INTRODUCTION

Jealousy is a concept that is dealt with in many areas such as psychology, sociology, anthropology, especially in emotional relations between people (Pines & Aronson, 1983). It is known that the first theoretical study on jealousy was made by Lewin (1948), and it was an emotion or behavior that came up, especially in the relationships between married couples. Pines (1998) described *jealousy* as a response to a hazard element that could lead to the breakdown or end of a valued relationship. In a relationship, the emotional state resulting from the relationship of the person's partner with another person (Buunk & Bringle, 1987; White, 1981), feelings of anger, unhappiness and fear caused by the deterioration or end of the relationship (DeSteno & Salovey, 1996) definitions were made. It can be said that *jealousy* is often defined as the reaction to the possibility of an end to a relationship as a result of the presence of a competitor in an emotional relationship or a marriage (Buunk, Angleitner, Oubaid & Buss, 1996; Mathes & Severa, 1981).

All these definitions explain *jealousy* as a reaction to the possibility of ending or ending the relationship based on an emotional relationship situation. Pines and Bowes (1992) state that *jealousy* is a complex set of emotions and is extremely painful for most people. There are also various approaches to cause jealousy. For example, Mead (1977) argues that jealousy results from feelings of insecurity and inadequacy of cultural or individual origin. Greenberg and Pyszczynski (1985) state that love, low self-esteem, fear of losing, and insecurity are at the basis of jealousy. Freud, on the other hand, made four different explanations of the basis of jealousy (Pines, 1998): the sadness of the fear of the loss of the loved one, the realization that we could not have everything we wanted (painful awareness), the feelings of envy for successful opponents and the self-indulgence of feeling responsible for losing ourselves. It is seen that Freud pointed to a different point in his statement about jealousy. It shows that envy against successful opponents. Although the concept of envy is similar to

* Assist. Prof. PhD., Alanya Alaaddin Keykubat University, Faculty of Education, Antalya-Turkey, duygu.kocak@alanya.edu.tr, ORCID ID: 0000-0003-3211-0426

To cite this article:
Koçak, D. (2019). Academic jealousy scale: Validity and reliability study. *Journal of Measurement and Evaluation in Education and Psychology, 10*(4), 467-479. doi: 10.21031/epod.566752

*Received: 17.05.2019*
*Accepted: 31.10.2019*

jealousy in daily life, it has significant similarities to jealousy in emotional relationships. However, it is different because it is aimed at something that is not owned and a perceived opponent.

*Jealousy* refers to an individual wants to have what other people have, the individual compares the material opportunities, success, physical characteristics of others with his own, and ultimately the superiority or quantity of someone else, and ultimately, describes a situation where the individual cannot accept the quality or quantity of someone else superior to themself (Anderson, 2002; Kim & Hupka, 2002; Parrott & Smith, 1993; Pines, 1998). While in jealousy, the individual reacts to maintain a relationship that he or she has, in case of envy, the individual wants a situation that he or she does not have (Pines & Aronson, 1983). One of the main differences between jealousy and envy is that jealousy involves three people, not two people as envy. An individual can envy others and aim at what other people have; property, beautiful eyes, personality traits, success, and so on. The focus of envy is an object or property. The focus of jealousy is a third person who is perceived as a threat to the existing relationship (Brehm, 1992; Friday, 1985; Pines, 1998; Salovey & Rodin, 1986). According to Spielman (1971), *envy* is the desire of the individual to have what another person has, and the unhappiness and feeling of badness are given by something that someone else has what she or he wants to have. According to this, envy shows itself with the anger and sadness of not having.

Salovey and Rodin (1986) made the difference between jealousy and envy by defining social relationship jealousy and social comparison jealousy. Accordingly, the reaction of an individual's relationship with another person (this can also be an object) is threatened by another person is *social relationship jealousy*. This can be considered as a reaction to the risk of something that an individual has; it is taken away. What is owned can be a relationship, home, car, success, professional position, but it is often defined as a reaction to the risk of loss or break down of an emotional relationship. In *jealousy of social comparison*, there is the relationship, professional position, success, home, car, personality trait, or physical trait that the individual wishes to possess, and is the effort to be nurtured and replaced by another person who has this condition. Although both definitions are called *jealousy in daily life*, it can be stated that social relationship jealousy corresponds to jealousy and social comparison jealousy corresponds to envy.

When definitions of jealousy and envy are examined, it is seen that both concepts are directly related to each other. One concept is the tendency to protect something that is owned, and the other is the tendency to obtain something that does not. With X and Y persons and object A, these two concepts can be summarized as follows: Person X has A and knows that Y wants to have A. In this case, X's sense of protecting A from Y is jealousy. Person Y wants to have A, but X owns it. In this case, Y's aim and reaction to obtain A from X is envy. Jealousy and envy involve complex emotions experienced during these desires to obtain or not to lose something. Various studies and scales are found in the literature in order to reveal these complex emotions and the variables they are associated with, especially in order to measure jealousy in emotional relationships. With the scales such as Cognitive Distortions Related to Relationships developed by Hamamcı (2002), Multidimensional Jealousy Scale was developed by Pfeiffer and Wong (1989) and was adapted into Turkish by Karakurt (2001), Emotional Jealousy Scale was developed by Kızıldağ (2017), Partner Emotional Jealousy Scale was developed Kızıldağ and Yıldırım (2017), jealousy, especially in emotional relations, was tried to be defined and measured.

Although jealousy comes to mind when jealousy is mentioned, one of the critical points in this regard is the concepts of jealousy and envy among individuals encountered in education. The envy for successful opponents and the self-criticism that led us to hold ourselves responsible for being lost constitute which was mentioned in Freud's statement of jealousy are an essential and frequently encountered dimension of jealousy (Pines, 1998). Massé and Gagné (2002) found that the students who successfully stand out from their peers were jealoused by their peers (they describe it as a jealousy corresponding to the concept of envy) and they showed that students were jealous of their peers' social and academic achievements depending on their academic achievement or intelligence. Rentzsch, Schröder-Abé and Schütz (2015) showed that students develop a sense of hostility towards others with academic self-esteem, especially in competitive environments, and the envy mediates this. González-Navarro, Zurriaga-Llorens, Tosin-Olateju, and Llinares-Insa (2018) have demonstrated that envy governs interpersonal relationships in working and competitive environments. In these limited numbers of studies, it was seen that qualitative approaches were used to measure envy or the sense of jealousy, and there was no quantitative measurement tool to measure the state of jealousy encountered in academic settings in the literature.

Today, within the number of educational institutions and university graduates increase each year, it is thought that there is competition in both educational institutions and professional institutions, and this will bring the concept of envy. The research results support this. When the studies which were done and the measurement tools which are used are examined, it is seen that there is not a measurement tool to directly reveal academic jealousy. The indirect consequence of this is the theoretical framework for the concept of academic jealousy could not be developed. Many measurement tools to measure emotional jealousy provide to definition of this feature and investigate of the relationship between variables that may be related. Therefore, theories about jealousy in emotional relations have been developed. The lack of a measurement tool to measure academic jealousy in the literature, it was cause that this feature has not been investigated. The use of the scale by researchers is an important starting point in terms of defining the concepts and structures to which the structure is related. In other words, being able to measure the concept of academic jealousy with a measurement tool will also provide to determine the other structures in which it is associated and characteristics of the structure. For this reason, it is thought that the scale plays an important role in the development of the theoretical structure of academic jealousy. Determining an individual's level of academic jealousy will make it easier to determine how this trait will affect one’s academic achievement, course of education, peer relationships, and other academic situations. In this sense, the use of the academic jealousy scale by the guidance units in schools is important in terms of recognizing the students and being able to consult them accordingly.

### *Purpose of the Study*

Although the concept of jealousy is frequently examined in the literature, there is no theoretical study on the concept of academic jealousy. At the same time, the existence of many measurement tools to measure the concept of jealousy, especially in emotional relationships, is the basis for the development of the relevant theoretical structure. As a reason why the theoretical structure of the concept of academic jealousy is not defined, it can be considered that there is no measurement tool for measuring the related structure. In this respect, the primary purpose of this research is to develop the scale of academic jealousy. With this primary purpose, it is aimed to form the basis of the theoretical infrastructure related to the concept of academic jealousy. Accordingly, in this study, determining the indicators related to the concept of academic jealousy and developing the measurement tool are forming the basis of the research.

## METHOD

This study is a descriptive study in which the validity and reliability analyses of Academic Jealousy Scale were conducted and the psychometric properties of academic jealousy were determined.

### *Working Groups*

The study group of this study consists of 478 students who studied at Alanya Alaaddin Keykubat University in the Fall Semester of 2018-2019 Academic Year selected by random sampling method. Four different study groups were formed at the data collection stage. The first study group was determined to reveal how the items were understood by the students during the writing phase of the scale items. One-on-one interviews were conducted with ten students. Before the item pool was generated, the students in this study group wrote an essay about their feelings and behaviors in case of academic jealousy. The second study group consisted of 254 students who participated in the application of pre-testing after writing the scale items. The data obtained from this application were used for Exploratory Factor Analysis (EFA) and reliability calculations in terms of internal consistency. The information about the participants in the second study group is presented in Table 1.

Table 1. Information About the Participants in the Second Study Group

| | Grade | Faculty of Education | Vocational School | Faculty of Engineering | Faculty of Management | School of Medicine |
|---|---|---|---|---|---|---|
| Female | 1. | 24 | 6 | 1 | - | 5 |
| | 2. | 50 | 8 | 7 | 9 | - |
| | 3. | 21 | - | 1 | 17 | - |
| | 4. | 5 | - | - | - | - |
| Male | 1. | 5 | 2 | 1 | 1 | 1 |
| | 2. | 30 | 14 | 9 | 8 | 5 |
| | 3. | 16 | - | 2 | 6 | - |
| | 4. | 2 | - | - | 2 | - |

A total of 254 students who are studying at different faculties and vocational schools and at different grade levels consist of the second study group of the present study. EFA was applied to the data obtained from the second study group in order to determine the construct validity of the scale. In addition, the data obtained from the second study group were used to determine the reliability of the scale and its sub-factors in terms of internal consistency and to calculate item discrimination values.

The third study group was formed to determine the reliability of the scale in terms of stability and consisted of 57 people. Information about the participants in the third study group is presented in Table 2.

Table 2. Information About the Participants in the Third Study Group

| | Grade 2 | Grade 4 |
|---|---|---|
| Female | 14 | 18 |
| Male | 9 | 16 |

Table 2 provides information about the students in the third study group. In order to determine the reliability of the scale in terms of stability, a total of 57 students who are studying at the Faculty of Education and not in the first group were determined as the third study group.

The fourth study group consisted of 157 students who were reapplication done to confirm the structure, which was determined by EFA. The data obtained from this group were analyzed by Confirmatory Factor Analysis (CFA). Information about the participants who are in the fourth study group is given in Table 3.

Table 3. Information About the Participants in the Fourth Study Group

| | Grade | Faculty of Education | Vocational School | Faculty of Engineering | Faculty od Management | School of Medicine |
|---|---|---|---|---|---|---|
| Female | 1. | 10 | 2 | 5 | 10 | 8 |
| | 2. | 4 | 13 | 5 | 5 | - |
| | 3. | - | - | 3 | 7 | - |
| | 4. | 5 | - | 4 | 2 | - |
| Male | 1. | 6 | 3 | 9 | 4 | 7 |
| | 2. | 5 | 6 | 7 | 4 | 2 |
| | 3. | - | - | 10 | 6 | - |
| | 4. | - | - | 2 | 3 | - |

Table 3 provides information about the students who are in the fourth working group. In order to test the structure obtained with the scale, a total of 157 students were identified as the fourth study group.

The responses of 4 participants were excluded from the analysis due to they create extreme value in the data set. For this reason, CFA was applied with data of 153 people.

### *Process*

In the development phase of the scale, firstly the theoretical framework was formed by examining the literature and provided 10 students who are in the first study group of the study write a composition about how do you feel if your friends get higher scores, perform better or be more successful than you, on issues like exam results, participate to lesson and academic achievement. The written statements of the students were taken into consideration in the creation of the scale items. There is no scale related to academic jealousy in the literature. For this reason, the theoretical structures about envy and jealousy were examined, and items that may be indicative of academic jealousy are written. According to this, a total of 47 items were written. The opinions of one expert in the field of psychology, two experts in the field of guidance and psychological counseling were consulted, and three items were excluded from the scope on the grounds that they could not measure academic jealousy. Then, the opinions of three different measurement and assessment experts were consulted, and two more items were excluded from the scale because they measured both emotional and behavioral dimensions. Necessary arrangements were made in line with the recommendations, and a total of 41 items were decided. Finally, based on the opinion of one Turkish language expert, the items were checked for grammatical rules. The items were asked by one-on-one interviews with 10 participants in the first study group and how the items were understood was determined, and the requisite modifications were done. This process provided significant findings to determine the structural validity of the items in the scale.

The experimental form was applied to 254 students who were in the second study group, and the data obtained from the second study group were analyzed with CFA in order to define the structure statistically. In addition, item statistics were determined, the final version of the scale was decided considering the theoretical structure, and the internal consistency of the scale was estimated. In order to strengthen the evidence about the reliability of the scale, the reliability in the test-retest of the scale was determined by applying the scale to 57 students who are in the third study group twice in two weeks interval. One more evidence of the structural validity of the scale was obtained by confirming the structure. A total of 153 students who were in the fourth study group were applied the final form of the scale and CFA was applied to the data. The accuracy of the structure created in this way has been tested.

In the development phase of the scale, the literature review was conducted in order to determine criterion validity and whether a scale which proven reliability and validity and developed or adapted to Turkish to measure similar or opposite structures was investigated. However, although the concept of jealousy in emotional relations is a frequently discussed issue, academic jealousy has not been the subject of research. For this reason, criterion-based validity could not be determined due to the insufficiency of the literature, and only construct validity and content validity were investigated.

### *Data Analysis*

Expert opinion was consulted to determine the content validity of the scale. In order to determine the construct validity, how the items were understood by students was examined. For this purpose, ten students were interviewed about the intelligibility of the items. As statistically evidence was presented about structural validity with EFA and CFA. In the EFA process, Horn's Parallel Analysis method was used in addition to the K1 rule, which is known to as eigenvalue above 1, in determining the factor number of the scale. Cronbach Alpha was used to determine the reliability of the scale in terms of internal consistency, and in order to determine the reliability in terms of stability, the relationship between the data obtained by the test-retest method was determined by Pearson Product Moment Correlation Coefficient.

Before performing the data analysis, the missing value was examined in the obtained data. The total missing value rate was approximately 2% in the second study group, 0% in the third study group, and approximately 1.5% in the fourth study group. Massing values were completed by Expectation-Maximization Algorithm method due to the method performs well in case of low percentage of missing data in every missing data mechanism (Koçak & Çokluk-Bökeoğlu, 2017). Then, the extreme value analysis was performed, and four participants in the fourth study group were excluded from the analysis due to their extreme values. The normality of the data sets was tested, and the analysis process was initiated. Lisrel 8.51 program was used for DFA. Other analyses were performed using "psych" package in R program.

## RESULTS

In this section, firstly, findings related to EFA, then findings related to CFA, and finally, findings related to reliability of the scale are given. In the process of obtaining the findings, the first EFA was performed. Before evaluating the results of EFA, it is necessary to examine whether the data are suitable for factor analysis. Whether the data are suitable for EFA can be explained by Kaiser-Meyer-Olkin (KMO) and Bartlett Sphericity Test (Çokluk, Şekercioğlu & Büyüköztürk, 2012). Kaiser-Meyer-Olkin (KMO) coefficient and Bartlett Sphericity test results obtained in accordance with this requirement are presented in Table 4.

Table 4. KMO and Bartlett Sphericity Test Results

| **Sample Value of Kaiser-Meyer-Olkin** | | .837 |
|---|---|---|
| **Bartlett Sphericity Test** | Approximate Chi-Square | 1444 |
| | Degree of freedom | 171 |
| | Significance level | .000 |

In accordance with the values presented in Table 4, it was decided that the data were suitable for factor analysis. Field (2000) states that the Kaiser-Meyer-Olkin value should be above .50. A three-factor structure was obtained as a result of EFA. Information about the factors is given in Table 5.

Table 5. Results Related about Number of Factors and Total Variance Explained

| **Factors** | **Eigenvalue** | **(%) of Variance** | **Total Variance Explained** |
|---|---|---|---|
| Factor 1 | 4.989 | 26.259 | 48.007 |
| Factor 2 | 2.603 | 13.698 | |
| Factor 3 | 1.529 | 8.050 | |
| Factor 4 | 0.998 | 5.250 | |
| Factor 5 | 0.980 | 5.156 | |

In Table 5, the factors obtained as a result of EFA, the variance explained by the factors, and the total variance ratio explained by three factors with an eigenvalue above 1 are presented. To accept factors with eigenvalues above 1 as determinative factors are called the K1 rule (Çokluk & Koçak, 2016). According to this method, factors with eigenvalues above 1 were accepted as valid factors. A total variance ratio between 40 and 60 percent is accepted as ideal (Scherer, Luther, Wiebe & Adams, 1988), as a result of the analysis, the explained total variance ratio by three factors is approximately 48 percent. When the eigenvalues of the factors are examined, it is seen that the eigenvalue of Factor 4 is very close to 1. According to this method, the number of factors must be determined as three due to the rule that the eigenvalue is above 1. In order to find additional proof for number of factors, Horn's Parallel Analysis method was used. Horn's parallel analysis method can be used to determine how many factors the structure has, especially when a structure is defining for the first time. In this method, EFA is performed in parallel in both data by producing artificial data reflecting the characteristics of the real data, and eigenvalues of factors are compared. It is one of the most powerful methods used to determine the number of factors. The structure of the academic jealousy scale is dimensioned for the first time in this study; hence Horn's parallel analysis method is used to determine the number of factors. The results obtained are presented in Table 6.

Table 6. Results about Parallel Analysis Method

| | Eigenvalue | | | | |
|---|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 |
| Real data | 4.989 | 2.603 | 1.529 | 0.998 | 0.980 |
| Parallel data | 1.276 | 1.158 | 1.073 | 1.055 | 1.021 |

The eigenvalues were obtained from Horn's Parallel Analysis Method are presented in Table 6. In this method, simulated data are generated parallel to the real data, which were analyzed with EFA, then the eigenvalues of factors are compared by performing CFA on both data. The stage in which the eigenvalue of the simulated data begins to be higher than the eigenvalue of real data is determined as factor number (Çokluk & Koçak, 2016; Koçak, Çokluk & Kayri, 2016). In Table 6, the eigenvalue of the real data is higher than the eigenvalue of the simulated data in the first three factors. As for the fourth factor, the eigenvalue of the simulated data started to take higher value than the eigenvalue of the real data. Accordingly, this method indicates that the number of factors is three.

After the number of factors was determined, factor loadings, cross-loading, serving the same purpose, and item discrimination of 41 items were examined. Items with an item discrimination index above .30 are proper discriminating items (Turgut & Baykul, 2010). Accordingly, items with item discrimination of less than .30 were excluded from the test. Item with high and close factor loadings in more than one factor is cross-loaded items. Therefore, items with a factor load high in more than one factor were excluded from the test. Finally, the other items were removed from the test by holding one of the items of the same purpose and parallel in the test. Following these procedures, a three-factorial structure was obtained with 19 items. The theoretical compatibility of the items which are in the factors has been the main criterion. When deciding to keep the items on the scale, it was decided by considering the compatibility of the item with the relevant factor, both theoretically and statistically.

After the 3-factor structure of the scale was determined, the stage of the labeling of the factors was started. When the factors were labeled, the studies in the literature were taken into consideration. According to this, when a student's friend scores higher than the student's, is more successful than the student or achieves success, the factor which includes the items which measure the student's angry with himself, the student's self-blame, this situation causes him discomfort was labeled as "Self-denigration". This dimension corresponds to the self-criticism of Freud's definition of jealousy in Pines (1988), in which the individual blames himself. For example, an item which is "I get angry with myself when I get a lower score than my friend in an exam" is included in this factor. Guerrero and Afifi (1998) state that self-blame and emotional destruction are negative and destructive emotions that may arise during the situation of jealousy. Similarly, Brehm (1992) states too that in the case of jealousy, self-blame, and situation of emotional depression occur. These items point to situations in which the individuals blame themself and create negative feelings towards themself rather than their friend whom they consider to be a competitor. Theoretically, in cases of jealousy, it is possible for individuals to seek the fault in themselves, to get angry, or to blame themselves (Brehm, 1992; Demirtaş, 2004; Guerrero & Afifi, 1998).

The items in the factor which is named as "Envy" use reflect anger towards the owner of the success that the individual cannot have. Envy refers to individual want to have something that others have, and compare an individual's own quality and quantity with other's, as a result of this, the feeling of individuals reach the point of envy (Anderson, 2002; Demir, 2004; Kim & Hupka, 2002; Parrott & Smith, 1993; Pines, 1998). This factor entirely coincides with Freud's concept of envy in Pines (1988). For example, items that "I am grudging my friends who score higher than me in exams" and "I want to prevent my friends from studying" are included in this factor. The items in this factor reveal a sense of envy. Another factor, called "Maturity", is the reaction in which individuals want to have what another person has but turns it into behavior in an insightful or level-headed way, without anger at themselves or their friends. These reactions are named as *mature behaviors*. Some studies which revealed the relationship between jealousy and age and indirectly maturity (as cited in Bringle & Williams 1979; Bringle, Roach, Andler & Evenbeck, 1979; Demirtaş, 2004; Mathes, Phillips, Skowran & Dick, 1982; Sullivan, 1953) have revealed too that individuals show more mature attitudes in case of possible jealousy as the age increases. Items that "I ask my successful classmates about their working techniques ask my successful classmates about their working techniques" and "When my friend is more successful than me, I will be happy for her" are included in this factor and indicate mature reactions.

Item and factor loading distributions of the 19-item three-factor structure, which was determined by considering factor loadings, item discrimination, theoretical structure, are presented in Table 7.

Table 7. Results Related to Distribution of Items to Factors, And Factor Loadings

| | **Factor Loadings** | | |
|---|---|---|---|
| **Items** | **Envy** | **Self-denigration** | **Maturity** |
| M19. If my friends are more successful than me, it will spoil our friendship. | .852 | | |
| M15. I am grudging my friends who score higher than me in exams. | .819 | | |
| M5. I want to prevent my friends from studying. | .645 | | |
| M18. It makes me angry that someone else achieves the success I cannot. | .630 | | |
| M17. I feel uncomfortable of people who perform better than me at lesson. | .597 | | |
| M10. I do not want my friends to get a master's degree. | .577 | | |
| M16. It makes me angry that someone else answers as true the question I am wrong. | .495 | | |
| M1. I feel uncomfortable being deficient at lessons. | | .772 | |
| M9. I blame myself when I score lower than my friends. | | .708 | |
| M6. It makes me sad that my average is lower than the average of my friends. | | .656 | |
| M7. It makes me ambitious that my friends score higher than me. | | .632 | |
| M3. I like to compete with my friends on lesson topics. | | .606 | |
| M4. When my friends study more than me, I feel irresponsible. | | .559 | |
| M13. I congratulate the people who scored higher than me in the exams. | | | .808 |
| M14. When my friends are more successful than me, I will be happy for them. | | | .787 |
| M12. I ask my friends who are more successful than me to study techniques. | | | .711 |
| M8. I want to be friends with people who are more successful than me. | | | .576 |
| M2. I motivate my friends about to be successful. | | | .552 |
| M11. I want help from my successful friends about the lessons. | | | .464 |

Varimax method is the most appropriate rotation method when factor loadings of the items are high in a single factor during the EFA stage and when especially some items have a very high factor loading (Kaiser, 1958). The factor loadings are fixed by rotating with the Varimax method by considering this situation. The correlations between the three factors which are presented in Table 7 were calculated as .428 between Envy and Self-denigration, -.376 between Envy and Maturity and -.133 between Maturity and Self-denigration. Item – total test score correlation and Cronbach's Alpha coefficients are presented in Table 8.

Table 8. Results about Item - Total Test Score Correlations

| | **Item - Total Test Score Correlation** | | | **Cronbach's Alpha** |
|---|---|---|---|---|
| **Items** | **Envy** | **Self-denigration** | **Maturity** | **Coefficient** |
| M19 | .537 | | | .769 |
| M15 | .372 | | | .772 |
| M5 | .392 | | | .773 |
| M18 | .549 | | | .771 |
| M17 | .535 | | | .772 |
| M10 | .420 | | | .774 |
| M16 | .461 | | | .774 |
| M1 | | .425 | | .773 |
| M9 | | .430 | | .774 |
| M6 | | .420 | | .773 |
| M7 | | .348 | | .776 |
| M3 | | .391 | | .775 |
| M4 | | .377 | | .775 |
| M13 | | | .529 | .771 |
| M14 | | | .547 | .771 |
| M12 | | | .395 | .773 |
| M8 | | | .397 | .773 |
| M2 | | | .381 | .773 |
| M11 | | | .349 | .776 |
| **Total Cronbach's Alpha Coefficient** | **.840** | **.840** | **.817** | **.779** |

When Table 8 is examined, it is seen that Cronbach's Alpha internal consistency coefficient was calculated as .779 for the whole scale, .840 for Envy factor, .840 for Self-denigration factor, and .817 for Maturity factor. The correlation coefficient between the data obtained by applying the scale to the third study group with two-week interval was calculated as .89. This coefficient provides evidence for the reliability of the scale in terms of stability. The item - total test score correlation coefficient provides evidence for the discrimination of items (Baykul, 2000). When the item total test score coefficients, which are presented in Table 8, are examined, it is seen that the item discrimination values of items in scale vary between .348 and .549.

In order to confirm the 19-item 3-factor structure, 153 participants in the fourth study group, were also applied the scales, and CFA was performed to the data obtained from this application.

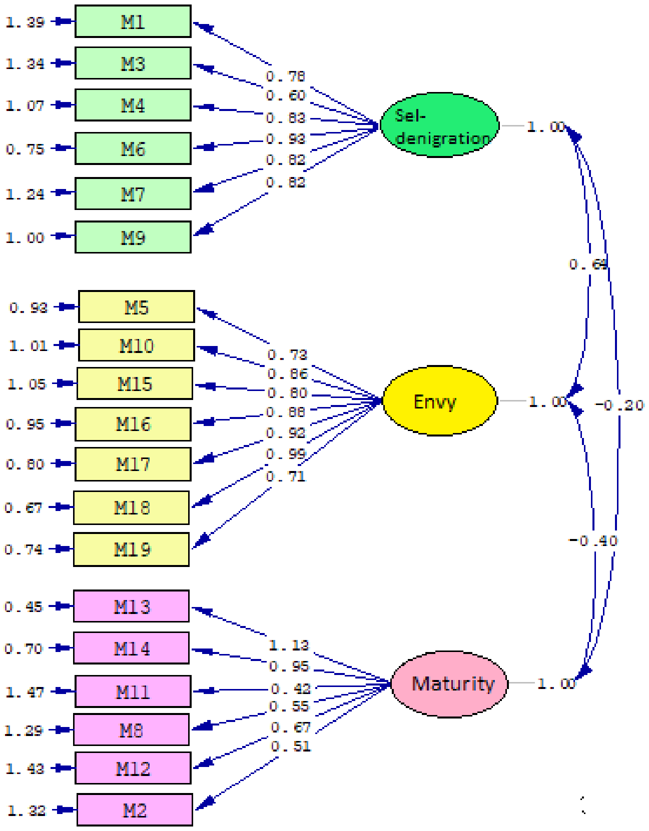


Figure 1. Path Diagram of the Academic Jealousy Scale

The path diagram in Figure 1 was obtained as a result of CFA. Fit indices related to the structure are presented in Table 9.

Table 9. Goodness of Fit Values of the Structure

| Fit indices | Value | Fit |
|---|---:|---:|
| χ2 | 264.86 | Good |
| χ2/df | 1.77 | Good |
| RMSEA | .07 | Acceptable |
| SRMR | .09 | Acceptable |
| NFI | .96 | Good |
| NNFI | .66 | Acceptable |
| CFI | .88 | Bad |
| GFI | .85 | Bad |
| AGFI | .85 | Acceptable |

When Table 9 is examined, it is seen that the fit indices except the CFI and GFI indices are within the acceptable or good fit ranges (Bentler & Bonett, 1980; Çokluk et al., 2012; Jöreskog & Sörbom, 1982; Stevens, 2002). It is seen that the ratio of the chi-square value to the degree of freedom is 1.77, and the RMSEA value is .072. The ratio of the chi-square to the degree of freedom is less than 2 indicates a good fit, and an RMSEA of less than .08 indicates an acceptable fit (Tabachnick & Fidell, 2012). Based on this, the Academic Jealousy Scale which was defined as a three-factor structure with 19 items was accepted as confirmation. It is validity because the majority of the fit indexes were between good and acceptable values.

## DISCUSSION and CONCLUSION

In this research, it was aimed to improve an Academic Scale. For this purpose, 478 university students were reached, and the related structure was defined, reliability and validity indices were determined.

In the development of the scale, a three-factor structure with 19 items was obtained by considering the theoretical structure. The scale explained approximately 48% of the total variance, and the discrimination indices of items in the scale range from .48 to .549. Cronbach's Alpha method was used to determine the reliability of the scale in terms of internal consistency, and test-retest method was used to determine the reliability of the scale in terms of stability. Cronbach Alpha internal consistency coefficient was found to be .779 for the whole scale, .840 for Envy subfactor, .840 for Self-denigration subfactor, and .817 for Maturity sub-factor. The correlation coefficient between the data, which were obtained by applying the scale to the third study group with two-week intervals is .89. This coefficient provides evidence for the reliability of the scale in terms of stability. Accordingly, it was concluded that the scale is reliable in terms of stability and internal consistency.

As a result of the CFA to test the three-factor structure which was obtained from EFA, it was found that RMSEA value is .072, ratio of chi-square to freedom degree is 1.77, SRMR value is .09, NFI value is .96, NNFI value is .66, CFI value is .88, GFI value is .85, and AGFI value is .85. The majority of the indices of model fit indicate that fit is good or acceptable, so it is concluded that the structure is confirmed.

There are statements in the literature that the reactions to be shown differ depending on the level of jealousy. Although these statements are theoretical explanations about jealousy in emotional relations, it is thought that similar reactions will be exhibited in academic jealousy. In the literature, it is stated that the reactions to jealousy are emotional, cognitive and physical (Aune & Comstock, 1991; DeWeerth & Kalma, 1993; Guerrero, 1998; Mathes & Verstraete, 1993; Pines & Aronson, 1983; Shettel-Neuber, Bryson, & Young, 1978). It is concluded that the individual's positive or negative cognitive, emotional, and physical responses constitute the concept of academic jealousy with the factors and items in the scale defined in the Academic Jealousy Scale developed. For example, "I want to prevent my friends from studying" in the scale corresponds to a cognitive and negative response, while "When my friends are more successful than me, I will be happy for them" is an emotional and positive response.

When the factors of Envy, Self-denigration, and Maturity that constitute academic jealousy are examined, it can be stated that the responses to be given vary depending on the level of jealousy. The envy factor reflects the individual's negative behaviors. The items in this factor reflect the anger towards the person who has achieved the success that the individual cannot. Envy refers to individual want to have something that others have, and compare an individual's own quality and quantity with other's, as a result of this, the feeling of individuals reach the point of envy (Anderson, 2002; Demir, 2004; Kim & Hupka, 2002; Parrott & Smith, 1993; Pines, 1998). For example, an individuals' desire to prevent their successful friends from studying, or getting angry at them reflects the envy subfactor. This factor includes negative emotions as well as behaviors that will adversely affect other people. This situation can be thought of as a physical reaction (Afifi & Reichert, 1996).

In the sub-factor of Self-denigration, it was concluded that individuals are uncomfortable when someone is more successful than themselves, but as a result of this situation, they blame themselves.

In the case of jealousy, individuals may get angry with themselves (Brehm, 1992; Demirtaş, 2004; Guerrero & Afifi, 1998). Guerrero and Afifi (1998) and Brehm (1992) state that self-blame is destructive emotions that may arise in jealousy. This factor fully reflects the individual's self-blame. Pines (1988) names this as self-criticism, but when defining it, one also states that individuals blame themselves. This dimension reveals that individuals blame themselves and not their friend whom they considered to be their rival, looking for a mistake in themselves and get mad at themselves (Brehm, 1992; Demirtaş, 2004; Guerrero & Afifi, 1998).

The Maturity factor presented that the reaction in which individuals want to have what another person has but turns it into behavior in an insightful or level-headed way, without anger at themselves or their friends. These reactions are called as mature behaviors. The studies which revealed the relationship among jealousy, age and indirectly maturity (by Bringle et al., 1979; Bringle & Williams 1979; Demirtaş, 2004, Sullivan, 1953; Mathes et al., 1982) reveal that individuals' behaviors in jealousy are related to maturity. In this study, it was also seen that individuals in mature academic jealousy exhibited a mature attitude, preferred to congratulate or sharing experiences, instead of calling guilty.

As a result, it was concluded that the Academic Jealousy Scale, which consists of three factors, Envy, Self-denigration, and Maturity, has 7 items in the Envy factor, 6 items in the Self-denigration factor, and 6 items in the Maturity factor, and the scale and the whole of the factors are reliable and the structure which was defined was confirmed. The concept of jealousy in emotional relations has been frequently investigated in the literature. The development of different scales to measure emotional jealousy formed the basis for the studies on this subject. The most critical obstacle to the study of the concept of academic jealousy is the lack of a scale to measure this feature. It was demonstrated that the developed scale was a valid and reliable scale that could measure academic jealousy in this study. By using the scale in different studies, other concepts related to academic jealousy can be searched, and related theoretical developments can be recorded. This aspect of the study is thought to contribute to the literature.

## REFERENCES

Afifi, W. A., & Reichert, T. (1996). Understanding the role of uncertainty in jealousy experience and expression. *Communication Reports, 9*(1), 93-103.

Anderson, R. E. (2002). Envy and jealousy. *American Journal of Psychotherapy, 56*(4), 455-480. doi: 10.1176/appi.psychotherapy.2002.56.4.455

Aune, K. S., & Comstock, J. (1991). Experience and expression of jealousy: Comparison between friends and romantics. *Psychological Reports, 69*(1), 315-319. doi: 10.2466/pr0.1991.69.1.315

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. Ankara: ÖSYM Yayınları.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606. doi: 10.1037/0033-2909.88.3.588

Brehm, S.S. (1992). *Intimate relationships.* New York, NY: McGraw Hill.

Bringle, R. G., & Williams, L. J. (1979). Parental off-spring similarity on jealousy and related personality dimensions. *Motivation and Emotion, 3*(3), 265-286. doi: 10.1007/BF01904230

Bringle, R. G., Roach, S., Andler, C., & Evenbeck, S. (1979). Measuring the intensity of jealousy reactions. *Catalog of Selected Documents in Psychology, 9*, 23-24.

Buunk, B. P., Angleitner, A., Oubaid, V., & Buss, D. M. (1996). Sex differences in jealousy in evolutionary and cultural perspective: Tests from the Netherlands, Germany, and the United States. *Psychological Science, 7*(6), 359-379.

Buunk, B., & Bringle, R. G. (1987). *Jealousy in love relationships.* In D. Perlman & S. Duck (Eds.), *Intimate relationships: Development, dynamics, and deterioration* (pp. 123-147). Beverly Hills, CA: Sage.

Çokluk, Ö., & Koçak, D. (2016): Using Horn's parallel analysis method in exploratory factor analysis for determining the number of factors. *Kuram ve Uygulamada Eğitim Bilimleri*, *16*(2), 537-551.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik SPSS ve Lisrel uygulamaları* (3rd Ed.). Ankara: Pegem Akademi.

Demir B (2004). Ankara atatürk eğitim ve araştırma hastanesi örneğinde hastane organizasyonu içerisinde hekim-hemşire ilişkisinin çatışma ve güç ilişkileri açısından analizi: Sosyo ekonomik düzeyin, eğitimin, cinsiyet ve çalışma süresinin etkileri (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Demirtaş, H. A. (2004). *Yakın ilişkilerde kıskançlık* (bireysel, ilişkisel ve durumsal değişkenler) (Yayımlanmamış doktora tezi). Ankara Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

DeSteno, D. A., & Salovey, P. (1996). Jealousy and envy. In A. S. R. Manstead & M. Hewstone (Eds.) *The blackwell encyclopedia of social psychology* (pp. 342-343). Oxford: Blackwell.

DeWeerth, C., & Kalma, A. P. (1993). Female aggression as a response to sexual jealousy: A sex role reversal? *Aggressive Behavior*, *19*(4), 265-279. doi: 10.1002/1098-2337(1993)19:4<265::aid-ab2480190403>3.0.co;2-p

Field, A. (2000). *Discovering statistics using SPSS for windows*. London: Sage Publications.

Friday, N. (1985). *Jealousy*. New York, NY: William Morrow.

González-Navarro, P., Zurriaga-Llorens, R., Tosin-Olateju, A., & Llinares-Insa, L. I. (2018). Envy and counterproductive work behavior: The moderation role of leadership in public and private organizations. *Int. J. Environ. Res. Public Health 15*(7). doi: 10.3390/ijerph15071455

Greenberg, J., & Pyszczynski, T. (1985). Proneness to romantic jealousy and responses to jealousy in others. *Journal of Personality, 53*(3), 468-479. doi: 10.1111/j.1467-6494.1985.tb00377.x

Guerrero, L. K. (1998). Attachement style differences in the experience and expression of romantic jealousy. *Personal Relationships*, *5*(3), 273-291. doi: 10.1111/j.1475-6811.1998.tb00172.x

Guerrero, L. K., & Afifi, W. A. (1998). Communicative responses to jealousy as a function of self-esteem and relationship maintenance goals: A test of Bryson's dual motivation model. *Communication Reports*, *11*(2), 111-122. doi: 10.1080/08934219809367693

Hamamcı, Z. (2002). *Ergenlerin yalnızlık düzeyleri ve kişiler arası ilişkilerle ilgili bilişsel çarpıtmaları arasındaki ilişkinin incelenmesi* (Yayımlanmamış yüksek lisans tezi), Ankara Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Jöreskog, K. G. & Sörbom, D. (1982). Recent developments in structural equation modeling. *Journal of Marketing Research, 19*(4), 404-416. doi: 10.2307/3151714

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*. *23*(3). 187-200. doi: 10.1007/BF02289233

Karakurt, G. (2001). *The impact of adult attachment styles on romantic jealousy* (Yayımlanmamış yüksek lisans tezi). Orta Doğu Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Kim, H. J., & Hupka, R. B. (2002). Comparison of associative meaning of the concepts of anger, envy, fear, romantic jealousy, and sadness between English and Korean. *Cross- Cultural Research, 36*(3), 229-255. doi: 10.1177/10697102036003003

Kızıldağ, S. (2017). Duygusal kıskançlık ölçeği üniversite öğrencileri formu: Geçerlik ve güvenirlik çalışmaları. *Journal of Measurement and Evaluation in Education and Psychology, 8*(1), 146-157. doi: 10.21031/epod.302673

Kızıldağ, S., & Yıldırım, İ. (2017). Eş duygusal kıskançlık ölçeği'nin geliştirilmesi. *Kuram ve Uygulamada Eğitim Bilimleri., 17*(1), 175-190.

Koçak, D., & Çokluk-Bökeoğlu, Ö. (2017). Kayıp veriyle baş etme yöntemlerinin model veri uyumu ve madde model uyumuna etkisi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8*(2), 200-223. doi: 10.21031/epod.303753

Koçak, D., Çokluk, Ö., & Kayri, M. (2016): Faktör sayısının belirlenmesinde MAP testi, paralel analiz, K1 ve yamaç birikinti grafiği yöntemlerinin karşılaştırılması. *YYÜ Eğitim Fakültesi Dergisi, 11*(1), 330-359. https://dergipark.org.tr/tr/download/article-file/253575 adresinden elde edilmiştir.

Lewin, K. (1948). *Resolving social conflicts.* New York, NY: Harper.

Massé, L., & Gagné, F. (2002). Gifts and talents as sources of envy in high school settings. *Gifted Child Quarterly, 46*(1), 15-29. doi: 10.1177/001698620204600103

Mathes, E. W., & Severa, N. (1981). Jealousy, romantic love, and liking: Theoretical considerations and preliminary scale development. *Psychological Reports, 49*(1), 23-31. doi: 10.2466/pr0.1981.49.1.23

Mathes, E. W., & Verstraete, C. (1993). Jealous aggression: Who is the target, the beloved or the rival. *Psychological Reports, 72*(3), 1071-1074. doi: 10.2466/pr0.1993.72.3c.1071

Mathes, E. W., Phillips, J. T., Skowran, J., & Dick, W. E. (1982). Behavioral correlates of the interpersonal jealousy scale. *Educational and Psychological Measurement, 42*(4), 1227-1231. doi: 10.1177/001316448204200432

Mead, M. (1977). Jealousy: Primitive and civilized. In G. Clanton & L. G. Smith (Eds.), *Jealousy* (pp. 115-127). Englewood Cliffs, NJ: Prentice Hall.

Parrott, W. G., & Smith, R. H. (1993). Distinguishing the experience of jealousy and envy. *Journal of Personality and Social Psychology*, *64*(6), 906-920. doi: 10.1037//0022-3514.64.6.906

Pfeiffer, S. M., & Wong, P. T. (1989). Multidimensional jealousy. *Journal of Social and Personal Relationships, 6*(2), 181-196. doi: 10.1177/026540758900600203

Pines, A.M. (1988). Keeping the spark alive: Preventing burnout in love and marriage. New York, NY: Free Press.

Pines, A. M. (1998). *Romantic Jealousy: Causes, symptoms, cures*. New York, NY: Routledge.

Pines, A. M., & Aronson, E. (1983). The jealousy question scale. *Psychological Reports, 50,* 1143-1147.

Pines, A. M., & Bowes, C. F. (1992). Romantic jealousy. *Psychology Today, 25*(2), 48-56. Retrieved from https://www.psychologytoday.com/us/articles/199203/romantic-jealousy

Rentzsch, K., Schröder-Abé, M., & Schütz A. (2015). Envy mediates the relation between low academic self-esteem and hostile tendencies. *Journal of Research in Personality, 58,* 143-153. doi: 10.1016/j.jrp.2015.08.001

Salovey, P., & Rodin, J. (1986). Differentiation of social-comparison jealousy and romantic jealousy. *Journal of Personality and Social Psychology, 50*(6), 1100-1112. doi: 10.1037/0022-3514.50.6.1100

Scherer, R. F., Luther, D. C., Wiebe, F. A., & Adams, J. S. (1988). Dimensionality of coping: Factor stability using the ways of coping questionnaire. *Psychological Report*, *62*(3), 76-770. doi: 10.2466/pr0.1988.62.3.763

Shettel-Neuber, J., Bryson, J. B., & Young, C. E. (1978). Physical attractiveness of the "other person" and jealousy. *Personality and Social Psychology Bulletin*, *4*(4), 612-615. doi: 10.1177/014616727800400424

Spielman, P. M. (1971). Envy and jealousy: An attempt at clarification. *Psychoanalytic Quarterly, 40*(1), 59-82. doi: 10.1080/21674086.1971.11926551

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences.* New Jersey, NJ: Lawrance Erlbaum Associates, Inc.

Sullivan, H. S. (1953). *The interpersonal theory psychiatry*. New York, NY: Norton.

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics*. Boston, MA: Allyn & Bacon/Pearson Education.

Turgut, M. F., & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme*. Ankara: PegemA.

White, G. L. (1981). Relative involvement, inadequacy, and jealousy: A test of a causal model. *Alternative Lifestyles, 4*(3), 291-309. doi: 10.1007/BF01257942