Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Kış* 2018 *Winter* 2018 *Cilt: 9- Sayı: 4* *Volume: 9- 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 sorumluğu ilgili yazarlara aittir.

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.

İletişim
e-posta: epod@epod-online.org

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

Dizinleme / Abstracting & Indexing
Emerging Sources Citation Index (ESCI), DOAJ (Directory of Open Access Journals), TÜBİTAK TR DIZIN

Hakem Kurulu / Referee Board

Ahmet Salih ŞİMŞEK (Cumhuriyet Üni.)
Ahmet TURHAN (American Institute Research)
Akif AVCU (Marmara Üni.)
Asiye Şengül Avşar (Recep Tayyip Erdoğan Üni.)
Ayfer SAYIN (Gazi Üni.)
Ayşegül ALTUN (Ondokuz Mayıs Üni.)
Arif ÖZER (Hacettepe Üni.)
Aylin ALBAYRAK SARI (Hacettepe Üni.)
Bahar Şahin Sarkın (İstanbul Okan Üni.)
Belgin DEMİRUS (MEB)
Bengu BORKAN (Boğaziçi Üni.)
Betül ALATLI (Gaziosmanpaşa Üni.)
Beyza AKSU DÜNYA (Bartın Üni.)
Bilge GÖK (Hacettepe Üni.)
Bilge BAŞUSTA UZUN (Mersin Üni.)
Burak AYDIN (Recep Tayyip Erdoğan Üni.)
Burcu ATAR (Hacettepe Üni.)
Burhanettin ÖZDEMİR (Siirt Üni.)
Cem Oktay GÜZELLER (Akdeniz Üni.)
Cenk AKAY (Mersin Üni.)
Ceylan GÜNDEĞER (Hacettepe Üni.)
Çiğdem Reyhanlioğlu Keçeoğlu
Cindy M. WALKER (Duquesne University)
Çiğdem AKIN ARIKAN (Hacettepe Üni.)
David KAPLAN (University of Wisconsin)
Deniz GÜLLEROĞLU (Ankara Üni.)
Derya ÇAKICI ESER (Kırıkkale Üni)
Derya ÇOBANOĞLU AKTAN (Hacettepe Üni.)
Didem ÖZDOĞAN (İstanbul Kültür Üni.)
Dilara BAKAN KALAYCIOĞLU (ÖSYM)
Dilek GENÇTANRIM (Kırşehir Ahi Evran Üni.)
Durmuş ÖZBAŞI (Çanakkele Onsekiz Mart Üni)
Duygu Gizem ERTOPRAK (Amasya Üni)
Duygu KOÇAK (Alanya Alaaddin Keykubat Üni.)
Ebru DOĞRUÖZ (Çankırı Karatekin Üni.)
Elif Bengi ÜNSAL ÖZBERK (Trakya Üni.)
Emine ÖNEN (Gazi Üni.)
Emrah GÜL (Hakkari Üni.)
Emre ÇETİN (Doğu Akdeniz Üni.)
Emre TOPRAK (Erciyes Üni.)
Eren Halil Özberk (Trakya Üni.)
Ergül DEMİR (Ankara Üni.)
Erkan ATALMIS (Kahramanmaras Sutcu Imam Uni.)
Esin TEZBAŞARAN (İstanbul Üni)
Esin YILMAZ KOĞAR (Niğde Ömer Halisdemir Üni.)
Esra Eminoğlu ÖZMERCAN (MEB)
Fatih KEZER (Kocaeli Üni.)
Fatih ORCAN (Karadeniz Teknik Üni)
Fatma BAYRAK (Hacettepe Üni.)
Fazilet TAŞDEMİR (Recep Tayyip Erdoğan Üni.)
Funda NALBANTOĞLU YILMAZ (Nevşehir Üni.)
Gizem UYUMAZ (Giresun Üni.)
Gonca Usta (Cumhuriyet Üni.)
Gül GÜLER (İstanbul Aydın Üni.)
Gülden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Sakarya Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan Sarıçam (Dumlupınar Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hülya YÜREKLI (Yıldız Teknik Üni.)
İbrahim Alper KÖSE (Abant İzzet Baysal Üni.)
İlhan KOYUNCU (Adıyaman Üni.)
İlkay AŞKIN TEKKOL (Kastamonu Üni.)
İlker KALENDER (Bilkent Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent YAKAR (Hacettepe. Üni.)
Mehmet KAPLAN (MEB)
Melek Gülşah ŞAHİN (Gazi Üni.)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Meltem YURTÇU (Hcettepe Üni.)
Metin BULUŞ (Adıyaman Üni)
Murat Doğan ŞAHİN (Anadolu Üni.)
Mustafa ASİL (University of Otago)
Mustafa İLHAN (Dicle Üni.)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe Üni.)
Neşe GÜLER (İzmir Demokrasi Üni.)
Neşe ÖZTÜRK GÜBEŞ (Mehmet Akif Ersoy Üni.)
Nuri DOĞAN (Hacettepe Üni.)
Nükhet DEMİRTAŞLI (Emekli Öğretim Üyesi)
Okan BULUT (University of Alberta)
Onur ÖZMEN (TED Üniversitesi)
Ömer KUTLU (Ankara Üni.)
Ömür Kaya KALKAN (Pamukkale Üni.)
Önder SÜNBÜL (Mersin Üni.)
Özge ALTINTAS (Ankara Üni.)
Özge BIKMAZ BİLGEN (Adnan Menderes Üni.)
Özlem ULAŞ (Giresun Üni.)
Ragıp Terzi (Harran Üni.)
Recep Serkan ARIK (Dumlupınar Üni.)
Sakine GÖÇER ŞAHİN (University of Wisconsin Madison)
Seçil ÖMÜR SÜNBÜL (Mersin Üni.)
Sedat ŞEN (Harran Üni.)
Seher YALÇIN (Ankara Üni.)
Selahattin GELBAL (Hacettepe Üni.)

Hakem Kurulu / Referee Board

Selen Demirtaş ZORBAZ (Ordu Üni)
Selma Şenel(Balıkesir Üni.)
Sema SULAK (Bartın Üni.)
Semirhan GÖKÇE (Niğde Ömer Halisdemir Üni.)
Seval KIZILDAĞ (Adıyaman Üni.)
Sevda ÇETİN (Hacettepe Üni.)
Sevilay KİLMEN (Abant İzzet Baysal Üni.)
Sinem Evin AKBAY (Mersin Üni.)
Sümeyra SOYSAL (HAcettepe Üni.)
Şeref TAN (Gazi Üni.)
Şeyma UYAR (Mehmet Akif Ersoy Üni.)
Tahsin Oğuz BAŞOKÇU (Ege Üni.)
Terry A. ACKERMAN (University of Iowa)
Tuğba KARADAVUT AVCI (Kilis 7 Aralık Üni.)
Tuncay ÖĞRETMEN (Ege Üni.)
Tülin ACAR (Parantez Eğitim)
Türkan DOĞAN (Hacettepe Üni.)
Yavuz AKPINAR (Boğaziçi Üni.)
Yeşim ÖZER ÖZKAN (Gaziantep Üni.)
Zekeriya NARTGÜN (Abant İzzet Baysal Üni.)

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

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Aralık 2018, Sayı: 9-4)

Journal of Measurement and Evaluation in Education and Psychology (December 2018, Issue: 9-4)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(4);312-325

Examining Invariant Item Ordering Using Mokken Scale Analysis for Polytomously Scored Items*

Hakan KOĞAR**

Abstract
The aim of the present study is to identify and compare the number of items violating the item ordering, the total number of item pairs causing violation, the test statistics averages and the H^T values of the overall test obtained from three separate Mokken IIO models in the simulative datasets generated by the graded response model. All the simulation conditions were comprised of 108 cells: 3 (minimum coefficient of a violation) x 2 (item discrimination levels) x 3 (sample sizes) x 2 (number of items) x 3 (response categories). MIIO, MSCPM and IT methods were used for data analysis. When the findings were considered in general, it was found that the MIIO method yielded the most stable values due to the fact that it was not affected by the lowest violation coefficient and was affected only slightly by simulation conditions. Especially in conditions where the violation coefficient was 0.03 (the default value in the Mokken package), it was recommended to use the MIIO method in identifying item ordering. Even though the MSCPM method yielded similar findings to those of the IT method, it generated more stable findings in particularly high sample sizes. In conditions where sample size, number of items and item discrimination were high, the MSCPM was recommended to be used.

Key Words: Invariant item ordering, mokken scale analysis, polytomous items, polytomous item response theory.

INTRODUCTION

A high score from psychological tests measuring personality or interests generally indicates positive responses regarding the related trait, while a high score from a cognitive test measuring ability indicates a better solution as regards the related cognitive trait. For example, an arithmetic question such as $\frac{3}{8} - \frac{1}{4} = ?$ on a cognitive test may seem like a simple question, but it measures two separate skills. First, the common divisors should be found, and then the numerators should be subtracted from each other (Ligtvoet, Van der Ark, Marvelde, and Sijstma, 2010). When we identify this question as an easy one in terms of item difficulty and place it among the first questions of a test, we should ask ourselves, "According to which skill level is this question easy?"

Traditionally, items in a test are ordered in terms of item difficulty. However, one item being more difficult than another item does not mean that this item is at the same difficulty level in all the subtests of the test. For instance, while a test item may be difficult for a subtest requiring a low-level skill, an exact opposite order can emerge in a subtest requiring a high-level skill (Ligtvoet, 2010). However, in measurement practices the order of items, based on item difficulty or attractiveness, should be the same for all participants. To illustrate, in intelligent tests developed for children, items are ordered according to item difficulty (Wechsler, 1999). The primary aim underlying this kind of sequencing is to prevent students from panicking when they encounter difficult questions and to enable students to reflect their performance onto the test. Another aim is to increase the difficulty level of the subtests to address the increasing age in different age groups. It is possible, in this way, to define the starting and ending points of the subtests according to age groups, which, it is claimed, an order of items that does not vary according to different age groups and individuals is possible. However, this is considered to

* A part of this study was presented as an oral presentation at the 5. International Eurasian Educational Research Congress (EJER)

** Assistant professor, Akdeniz University, Faculty of Education, Department of Educational Sciences, Antalya, Turkey, e-mail: hkogar@gmail.com, ORCID ID: 0000-0001-5749-9824

To cite this article:
Koğar, H. (2018). Examining invariant item ordering using mokken scale analysis for polytomously scored items. *Journal of Measurement and Evaluation in Education and Psychology, 9*(4), 312-325. DOI: 10.21031/epod.412689
Received: 04.04.2018
Accepted:10.07.2018

be an assumption as it is not based on experimental evidence (Ligtvoet et al., 2010). Another consideration is that in tests measuring attitude and personality, generally a structure in which psychological traits are ordered is used (Watson, Deary, and Shipley, 2008). For instance, in a measurement tool measuring introvertedness, when such items as, "I rarely talk to other people in the company" and "I prefer to do my work on my own and do not prefer to see other people" are compared.

It is possible to think that the latter indicates introvertedness more than the former does. However, in practice, many people prefer to do their work on their own, although they are not introverts. Such conditions show us that it is wrong to establish the order by considering item means. However, it is possible for a group of items to have an invariant item ordering (IIO) and to have a structure by identifying a level of grouping (Ligtvoet et al., 2010, p. 2).

IIO was developed with the aim of overcoming the problems that can stem from ordering test items based solely on item difficulty (Sijstma and Junker, 1996). IIO is the situation where the order of items is the same for all the participants. The benefits of IIO have been proven from various aspects. IIO is defined within the scope of item response theory (IRT). To determine the IIO of test items, they should have the assumptions of IRT models. Sijtsma and Junker (1996) showed that IIO could only be used in IRT models in which item response function (IRF) does not intersect. IIO can only be applied to Rasch (1960) and the double monotonicity model (DMM) in dichotomously scored datasets (Mokken and Lewis, 1982). In polytomously scored datasets, on the other hand, IIO can only be applied to the rating scale model (Andrich, 1978) and the restricted graded response model (Muraki, 1990) (Ligtvoet et al., 2010).

The IIO methods are *manifest invariant item ordering* (MIIO) model, t*he manifest scale of the cumulative probability model* (MSCPM) and i*ncreasingness in transposition* (IT) model, which is addressed within the scope of Mokken Scaling Analyses (MSA) (Van der Ark, 2012). These are nonparametric methods that require very few assumptions (unidimensionality, latent monotonicity, non-intersection). Each method can generate a fixed item order and items that violate this order (Ligtvoet et al., 2010; Ligtvoet, Van der Ark, Bergsma, and Sijtsma, 2011). The average ratios of the MIIO polytomously scored items were developed with the aim of identifying whether or not polytomously scored items intersected with the item response function. MSCPM examines the manifest item step response function for each item pair. However, this high method of IIO has some disadvantages in practice. Because it compares each item pair individually, it yields an excessive number of comparative findings. For this reason, it has the tendency to propose the fact that all the items lead to violation. The MSCPM method, when compared to the other models, has the potential to yield a higher number of violating items (McGrory, 2015). In the related literature, there is very limited information regarding the details of these methods.

The IIO violating items are initially identified and then they are sequentially removed from the test. This process is continued until there are no IIO violating items remaining in the test. Subsequently, the person scalability coefficient (H^T), which is a measure for individuals' adaptation, is calculated. This coefficient resembles the H coefficient, but it is obtained from the converted data matrix. The H^T coefficient, which has a value between $0 \leq H^T \leq 1$ was developed by Sijstma and Meijer (1992) to determine the model-data fit of DMM. The obtained high values in DMM indicate that the person ordering is invariant. In other words, the order of the items is independent of a group of individuals; it is invariant. Negative H^T values indicate the violation of the non-intersection assumption (Ligtvoet et al., 2010, 2011). According to Sijstma, Meijer and Van der Ark (2011), the H^T coefficient is as important as the other scalability coefficients (H, H_i, H_{ij}) because it shows to what extent the person ordering is independent of the Guttman error. However, it is more sensitive than the other scalability coefficients in many respects. IIO values are obtained in situations where IRFs are not close to each other. This situation shows that the H^T coefficient should not be used for the purpose of evaluating the quality of a measurement.

MIIO is the default IIO method in the Mokken package in R software. There are numerous studies in which MIIO is applied to various scales to determine the invariant item ordering (Ahmadi, Reidpath, Allotey, and Hassali, 2016; Gibbons, Small, Rick, Burt, Hann, and Bower, 2017; Lee, Chen, Jiang,

Chu, Chiu, Chen, and Chen, 2016; Ligtvoet, van der Ark, and Sijtsma, 2008; Saiepour, Najman, Clavarino, Baker, Ware, and Williams, 2014; Stewart, Allison, Baron-Cohen, and Watson, 2015; Stochl, Jones, and Croudace, 2012; Van der Graaf, Segers, and Verhoeven, 2015; Yoon, Shaffer, and Bakken, 2015). However, there are no studies in literature regarding the use of the other two methods for IIO. Sijstma and Meijer (1992) supported their research in which they developed the H^T coefficient with a simulation study. In this research conducted on dichotomously scored datasets, the higher the item difficulty and item discrimination coefficients were, the higher the H^T coefficient turned out to be. It was observed that sample size and length of test had a limited effect. The other qualities of the item response function and the ability parameter distributions remained constant.

The only study which compared and discussed these three methods based on a single real dataset belongs to Ligtvoet et al. (2011). In this study, two small datasets were used to compare the methods of MIIO, MSCPM and IT. In the eight items of the first dataset, MIIO yielded a violation in two of the total 28 item pairs. Since the common point of these two item pairs was the fifth item, it was recommended that this item be removed from the test. The MSCPM model found violation in seven of the 63 item pairs. It was recommended that the third and sixth items be removed. The IT method was applied for the remaining five items. Violation was observed in two of the 60 item pairs. It was recommended that the first item be removed. In the second dataset, the IRFs of six item pairs were examined. While the MIIO method did not yield any violations, the IT method yielded one and the MSCPM method yielded two violations. Furthermore, in this study, Ligtvoet et al. (2011) conducted a simulation study on the determination of MIIO sensitivity and specificity and the H^T coefficient. The findings of this simulation constitutes the foundation of this research study.

In a pilot study (Ligtvoet et al. (2011) on MIIO, MSCPM and IT, it was found that each of these models indicated different items to be removed. When a situation contradictory to IIO emerged, it was observed that MSCPM was more sensitive and generally proposed more items to be removed than MIIO and IT did. The item ordering obtained from IT is expected to be stricter when compared to the other models; thus, findings indicating more items to be removed is expected. For this reason, these preliminary findings are found to be surprising. Another point is that these methods are not hierarchically related; that is, they examine different features of the dataset. Hence, it is normal that they yield different items for remove (Van der Ark, 2012). This finding reported by Van der Ark (2012) seems to be the result of a single study comparing these methods. Hence, it is clear that further studies need to be conducted to compare these methods.

Purpose of the Study

The aim of the present study is to identify and compare the number of items violating the item ordering, the total number of item pairs causing violation, the test statistics averages (t, z and χ^2 values) and the H^T values of the overall test obtained from three separate Mokken IIO models in the simulative datasets generated by the graded response model.

METHOD

Data Simulation Procedures

In polytomously scored datasets, only the rating scale model (Andrich, 1978) and the restricted graded response model (Muraki, 1990) can show IIO. Ligtvoet et al., (2010) study showed that IRFs almost always intersected in dense regions of the latent variable y, so that it seemed safe to use the graded response model. So, graded response model was used to generate data in the present study. The simulation conditions were defined and the model was used to produce datasets. The simulation conditions were as follows:

1. Minimum coefficient of a violation: This value, which was 0.03 by default, was simulated as 0.03, 0.27 and 0.45. A value of 0.00 indicated that the slightest violation would be significant, whereas a

value of 0.45 indicated that only where there was a highly significant violation could a violation to be considered significant (Ligtvoet et al., 2011). In other words, this value is a criterion value. A value of or near 0.00 would lead to an increase in the number of items to be proposed for remove and a value of or near 0.45 would lead to a decrease in the number of items to be proposed for remove.

2. *Item discrimination levels*: Two item discrimination levels, namely low and high, have been defined. A low discrimination level was obtained from a normal distribution with mean of 0.5 and variance of 1; a high discrimination was obtained from a normal distribution with a mean of 1.5 and variance of 1. These coefficients were identified based on the studies by Desa, (2012) and Dodeen (2004). The item difficulty coefficients were obtained from a normal distribution with a mean of 0 and variance of 1.

3. *Sample size*: In the present study, sample sizes were identified as 100, 250 and 500. In simulation studies based on the nonparametric item response theory, sample size was defined to be approximately 200 (Van Abswoude, Van der Ark and Sijstma, 2004; Van Abswoude, Vermunt, Hemker, and Van der Ark, 2004). In the present study, sample sizes bigger and smaller than this value have also been defined. The ability distributions were obtained from the normal distributions.

4. *Number of items*: Two tests – one short (k=5) and one long (k=15) – were used (Ligtvoet et al., 2011).

5. *Response categories*: Response categories were identified as 3, 5 and 7. The response category values were adapted from the studies by Lozano, García-Cueto, and Muñiz (2008) and Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol and Coffman (2009).

20 replications (Drasgow, 1989) were applied to each dataset. 720 datasets were obtained as a result of 36 datasets * 20 replications: 2 (item discrimination levels) x 3 (sample sizes) x 2 (number of items) x 3 (response categories).

The dependent variables of the present study were the number of items violating the order, the number of item pairs leading to the total violation, the test statistics averages, and the H^T values of the overall test. Data generation was performed via the WINGEN 2.0 software program.

Data Analysis

All the simulation conditions are comprised of 108 test conditions: 3 (minimum coefficient of a violation) x 2 (item discrimination levels) x 3 (sample sizes) x 2 (number of items) x 3 (response categories). By applying the MIIO, MSCPM and IT methods, which were addressed within the scope of MSA, the number of items violating the order, the number of item pairs leading to the total violation, the test statistics averages, and the H^T values of the overall test were identified for each cell. The analyses were performed via the Mokken 2.8.10 (Van der ark, 2007) package in R software.

The H^T coefficient in dichotomously scored datasets was developed by Sijstma and Meijer (1992). In polytomously scored items, Ligtvoet et al., (2011) developed the H^T coefficient, which is the primary dependent variable of the present study, by generalizing the interpretation of the H scalability coefficient. When IIO is applied to a dataset that can show IIO, it shows that an H^T coefficient of 0.3 or below is an indication of a wrong item ordering. A coefficient between 0.3 and 0.4 shows a low degree of accuracy in item ordering, a coefficient between 0.4 and 0.5 indicates a moderate degree of accuracy in item ordering, and one above 0.5 indicates a high degree of accuracy in item ordering (Ligtvoet et al., 2011).

For IIO to be identified, first the number of items leading to significant violations according to the specified lowest violation coefficient needs to be identified. If no item causes violation, then the presence of IIO for all the k number of items is proved; otherwise, the item causing the most violation is removed from the test. Subsequently, the same method is replicated for the remaining (k-1)(k-2)/2 item pair. If this item also needs to be removed, then the method is replicated for the (k-2)(k-3)/2 item pair. This process is repeated until there are no items causing violation. If there are two or more items

with the same number of violations, which items are to be removed are identified by means of two different techniques. The first item to be removed is the one that has the lowest item scalability coefficient (H_i). The second is identified by considering the content of the item (Ligtvoet et al., 2011; Sijtsma and Molenaar, 2002).

In studies where the methods of MIIO, MSCPM and IT are used simultaneously, the items to be removed are those that violate the common order. The level of this violation is identified by means of the lowest violation coefficient and this value, by default, is considered to be 0.03. A decrease in this value indicates that even the slightest violation is accepted. The degree of the violation is determined via the t test technique (t values) in the MIIO method, the z test technique (z values) in the MSCPM method and the chi-squares technique (χ^2 values) in the IT method. The violation causing items that are statistically significant should be removed from the test sequentially; if there are more than one item that cause a high degree of violation, the item with the lowest scalability coefficient is removed from the test (Ligtvoet, 2010).

RESULTS

The findings regarding the number of items violating the order are presented in Table 1. The IT method could not yield findings in conditions with a sample size of 100. In almost all conditions of simulation, the number of items violating the order that the MSCPM and IT methods yielded was higher than that yielded by the MIIO method. Furthermore, while the MSCPM and IT methods were significantly affected by a change in the lowest violation coefficient, of these two methods, IT was mostly affected by this coefficient. In a condition where violation coefficient value was 0.45, IT hardly yielded any item for remove. For example, in one simulation condition with the lowest violation coefficient was 0.03 in the IT method, an average of 12.40 items of 15 items were yielded for remove, while in another condition with the lowest violation coefficient of 0.27, an average of 1.60 items were yielded for remove. Similar examples were present in the MSCPM method as well. However, in the MIIO method, the number of items yielded for remove was quite close for the lowest and highest violation coefficients.

The number of items causing violation in the order was high for all methods across all sample sizes and in conditions where the number of items was 15 and the response categories were 5 and 7. However, in conditions where the number of items was 15, the response category was 7, and the item discrimination level was low, the methods, particularly MIIO, yielded very few number of items to be removed. The MIIO method yielded an average of 0.05, 1.00 and 1.45 items to be removed in samples sizes of 100, 250 and 500, respectively in the specified simulation conditions. These findings are quite surprising. While an increase in the number of items yielded for remove was observed as the sample size increased, no effect of number of items, response categories, and item discrimination on the number of items to be removed for violating the item ordering was observed.

The findings regarding the number of item pairs causing violation are presented in Table 2. In all simulation conditions, the number of item pairs causing violation identified by the IT method was higher than that yielded by the other methods. Especially in conditions where the number of items is 15, and the response categories are 5 and 7, more than 1000 item pairs causing violation were detected. However, in conditions where the lowest violation coefficient was 0.03, these values that were produced in high numbers yielded rather low values (0.00 – 74.10) in conditions where the lowest violation coefficients were 0.27 and 0.45. Thus, it was revealed that IT was significantly affected by the lowest violation coefficient in these conditions as well. The MSCPM and IT methods identified a higher number of item pairs to be causing violation than the MIIO method. As the number of these item pairs has an impact on the number of items yielded for remove, it is normal that this finding shows similarity to those presented in Table 1.

As the sample size increased, the number of item pairs causing violation identified by all the methods also increased. In the MSCPM and IT methods, it is observed that as the number of response categories increased, the number of item pairs causing violation also increased. However, the same situation was

not valid for MIIO. It can be claimed that in all the methods, in all the conditions where item discrimination is high, a higher number of item pairs causing violation were identified.

Table 1. Findings from the Number of Items Violating the Order

S	NI	RC	ID	MIIO			MSCPM			IT		
				0.03	0.27	0.45	0.03	0.27	0.45	0.03	0.27	0.45
100	5	3	L	0.00	0.00	0.00	0.75	0.00	0.00	-	-	-
			H	0.05	0.05	0.00	1.00	0.10	0.00	-	-	-
		5	L	0.00	0.00	0.00	0.90	0.00	0.00	-	-	-
			H	0.05	0.05	0.05	1.95	0.05	0.00	-	-	-
		7	L	0.00	0.00	0.00	1.70	0.00	0.00	-	-	-
			H	0.05	0.05	0.00	2.65	0.05	0.00	-	-	-
	15	3	L	0.15	0.15	0.05	4.40	0.25	0.00	-	-	-
			H	1.00	1.00	1.00	7.00	2.25	0.30	-	-	-
		5	L	2.60	2.60	2.60	10.45	3.25	0.70	-	-	-
			H	1.95	1.95	1.60	9.85	2.45	0.30	-	-	-
		7	L	0.05	0.05	0.05	6.55	2.10	0.55	-	-	-
			H	2.00	2.00	2.00	11.70	3.70	0.40	-	-	-
250	5	3	L	1.40	0.90	0.10	3.00	1.80	0.10	3.00	1.15	0.00
			H	0.50	0.25	0.00	1.45	0.00	0.00	1.60	0.00	0.00
		5	L	0.00	0.00	0.00	2.40	1.00	0.00	2.00	0.00	0.00
			H	0.05	0.05	0.05	2.80	0.35	0.00	3.00	0.35	0.00
		7	L	0.30	0.30	0.30	2.95	1.05	0.05	2.95	0.00	0.00
			H	0.50	0.50	0.50	2.45	1.00	0.00	2.65	0.00	0.00
	15	3	L	2.75	1.80	0.20	7.95	1.20	0.80	9.20	1.00	0.60
			H	1.20	0.40	0.05	9.20	1.00	0.00	8.60	0.20	0.00
		5	L	5.20	5.20	4.40	11.40	4.20	1.00	9.40	0.60	0.00
			H	5.00	4.00	3.40	12.20	4.40	1.00	12.00	2.00	0.05
		7	L	1.00	1.00	1.00	10.40	4.60	0.15	10.00	0.05	0.00
			H	3.00	3.00	3.00	12.00	5.20	0.60	12.40	1.60	0.20
500	5	3	L	2.00	1.60	0.90	2.60	1.00	0.00	2.40	0.95	0.00
			H	1.20	0.10	0.00	3.00	0.00	0.00	3.00	0.00	0.00
		5	L	0.60	0.60	0.30	3.00	1.20	0.10	3.00	0.00	0.00
			H	1.20	1.00	0.45	2.60	0.25	0.00	2.60	0.00	0.00
		7	L	0.20	0.70	0.25	2.20	1.60	0.20	2.00	0.00	0.00
			H	1.40	1.40	1.40	1.90	1.20	0.35	2.00	0.00	0.00
	15	3	L	7.20	5.00	3.00	11.60	3.60	1.20	11.00	2.60	0.00
			H	5.20	4.00	1.20	10.00	2.60	0.05	9.20	1.80	0.00
		5	L	3.60	3.40	2.60	11.40	3.60	0.30	10.80	0.45	0.00
			H	5.40	5.40	3.40	12.20	8.60	2.80	12.20	4.80	0.95
		7	L	1.40	1.40	1.40	9.40	3.40	1.20	6.20	0.00	0.00
			H	6.60	6.20	5.20	12.80	9.40	5.80	12.00	6.80	1.15

S: sample size, NI: number of items, RC: response category, ID: item discrimination, L: low, H: high

Table 2. Findings from the Total Number of Item Pairs Causing Violation

S	NI	RC	ID	MIIO			MSCPM			IT		
				0.03	0.27	0.45	0.03	0.27	0.45	0.03	0.27	0.45
100	5	3	L	0.30	0.00	0.00	1.80	0.00	0.00	-	-	-
			H	0.40	0.30	0.00	3.50	0.40	0.00	-	-	-
		5	L	0.00	0.00	0.00	4.20	0.00	0.00	-	-	-
			H	0.40	0.10	0.10	12.50	0.30	0.00	-	-	-
		7	L	0.30	0.20	0.00	10.70	0.00	0.00	-	-	-
			H	0.60	0.60	0.10	20.30	0.10	0.00	-	-	-
	15	3	L	2.85	0.30	0.10	49.40	0.50	0.00	-	-	-
			H	7.10	3.50	6.95	67.90	9.50	0.80	-	-	-
		5	L	23.70	16.90	10.70	210.90	23.80	1.90	-	-	-
			H	19.40	11.60	6.30	204.80	0.70	0.00	-	-	-
		7	L	5.50	3.50	1.40	186.60	31.00	1.50	-	-	-
			H	32.50	27.10	20.30	377.00	39.70	3.50	-	-	-
250	5	3	L	9.90	2.40	0.20	25.70	6.50	0.20	39.20	4.10	0.00
			H	2.40	0.50	0.00	7.20	0.00	0.00	25.90	0.00	0.00
		5	L	0.60	0.00	0.00	31.90	8.80	0.00	70.20	0.00	0.00
			H	1.50	0.30	0.10	21.70	1.10	0.00	35.20	1.00	0.00
		7	L	5.70	3.60	1.50	50.20	5.40	0.10	76.40	0.00	0.00
			H	1.80	1.50	1.00	26.10	5.00	0.00	27.40	0.00	0.00
	15	3	L	43.70	6.00	0.90	128.80	10.90	2.50	274.10	13.70	1.30
			H	17.20	0.50	0.10	105.90	3.50	0.00	209.50	0.40	0.00
		5	L	78.20	50.20	27.40	381.10	39.20	6.30	617.70	1.00	0.00
			H	57.60	35.50	20.00	379.60	50.20	7.90	628.40	14.40	0.10
		7	L	27.90	18.90	8.00	451.80	33.80	0.40	790.20	0.10	0.00
			H	40.50	32.10	19.20	546.50	85.90	4.20	824.80	11.80	0.60
500	5	3	L	14.20	5.50	1.90	27.00	2.90	0.00	29.60	1.90	0.00
			H	11.10	0.20	0.00	33.50	0.00	0.00	46.60	0.00	0.00
		5	L	6.80	3.10	0.80	49.30	3.80	0.20	74.10	0.10	0.10
			H	8.40	3.90	1.30	38.50	0.50	0.00	75.70	0.00	0.00
		7	L	3.40	2.10	0.60	78.10	11.70	0.60	124.90	0.00	0.00
			H	9.70	8.00	5.50	53.80	14.80	1.00	90.30	0.00	0.00
	15	3	L	200.60	72.00	26.80	421.70	37.70	2.90	525.70	12.10	0.00
			H	75.60	19.00	4.20	211.60	12.90	0.10	357.80	10.90	0.00
		5	L	78.40	38.00	16.50	539.90	27.00	0.90	1004.00	0.90	0.00
			H	100.00	61.60	36.10	841.50	208.70	23.70	1315.30	42.90	4.50
		7	L	40.20	24.20	8.00	636.70	69.90	11.70	1027.70	0.00	0.00
			H	113.20	99.90	78.70	1075.00	377.60	98.50	1385.50	74.10	5.20

S: sample size, NI: number of items, RC: response category, ID: item discrimination, L: low, H: high

Table 3. Findings from the Test Statistics Averages (t, z and χ^2 Values)

S	NI	RC	ID	MIIO			MSCPM			IT		
				0.03	0.27	0.45	0.03	0.27	0.45	0.03	0.27	0.45
100	5	3	L	0.07	0.00	0.00	3.18	0.00	0.00	-	-	-
			H	0.01	0.01	0.00	7.15	0.06	0.00	-	-	-
		5	L	0.00	0.00	0.00	9.04	0.00	0.00	-	-	-
			H	0.01	0.00	0.00	3.74	0.01	0.00	-	-	-
		7	L	0.08	0.06	0.00	4.69	0.00	0.00	-	-	-
			H	0.01	0.01	0.09	3.01	0.00	0.00	-	-	-
	15	3	L	0.09	0.05	0.01	1.76	0.28	0.00	-	-	-
			H	1.55	1.84	0.59	1.12	3.00	0.05	-	-	-
		5	L	0.48	0.45	0.38	0.86	1.17	0.29	-	-	-
			H	0.56	0.48	0.44	0.90	1.20	0.05	-	-	-
		7	L	0.01	0.01	0.00	1.17	2.15	0.23	-	-	-
			H	0.73	0.72	0.68	1.61	0.96	0.14	-	-	-
250	5	3	L	2.08	0.04	0.18	0.11	0.06	0.32	0.10	0.02	0.00
			H	0.73	0.25	0.00	0.74	0.00	0.00	12.97	0.00	0.00
		5	L	0.10	0.00	0.00	3.01	1.92	0.00	29.21	0.00	0.00
			H	0.32	0.10	0.05	1.72	1.34	0.00	21.56	5.88	0.00
		7	L	0.91	0.96	0.48	2.66	3.77	0.05	7.44	0.00	0.00
			H	0.63	0.59	0.48	1.70	0.85	0.00	15.62	0.00	0.00
	15	3	L	1.69	0.88	0.23	2.95	2.48	0.61	4.96	16.75	3.54
			H	0.84	0.09	0.02	3.09	0.59	0.00	4.70	0.63	0.00
		5	L	2.72	2.63	2.35	1.75	4.55	0.94	5.17	0.84	0.00
			H	3.07	2.97	2.52	2.69	2.41	0.18	4.71	5.14	0.20
		7	L	1.08	0.98	0.70	2.14	1.25	0.14	10.19	0.17	0.00
			H	1.89	1.84	1.63	3.67	3.44	0.76	5.78	2.74	1.31
500	5	3	L	4.80	4.01	2.21	2.34	2.90	0.00	4.93	14.84	0.00
			H	1.83	0.12	0.00	4.44	0.00	0.00	10.08	0.00	0.00
		5	L	1.09	0.81	0.38	2.41	2.47	0.14	4.77	0.15	0.15
			H	1.86	1.50	0.65	3.47	0.99	0.00	4.23	0.00	0.00
		7	L	0.60	0.51	0.21	2.81	0.77	0.35	24.49	0.00	0.00
			H	2.71	2.69	2.49	6.14	2.59	1.57	12.51	0.00	0.00
	15	3	L	4.67	4.63	3.73	2.93	5.71	0.94	4.62	2.51	0.00
			H	3.48	2.77	1.29	2.73	3.56	0.04	4.49	1.91	0.00
		5	L	2.56	2.50	2.07	2.34	2.04	0.27	4.86	2.54	0.00
			H	4.74	4.71	4.36	3.67	1.08	1.72	6.26	3.09	11.29
		7	L	1.14	1.06	0.67	2.89	2.51	0.78	21.46	0.00	0.00
			H	5.87	5.84	5.77	3.39	1.76	4.77	4.66	8.30	18.07

S: sample size, NI: number of items, RC: response category, ID: item discrimination, L: low, H: high

Table 4. Findings from H^T Values of the Overall Test

S	NI	RC	ID	MIIO			MSCPM			IT		
				0.03	0.27	0.45	0.03	0.27	0.45	0.03	0.27	0.45
100	5	3	L	0.13	0.13	0.13	0.15	0.13	0.13	-	-	-
			H	0.14	0.14	0.13	0.30	0.15	0.13	-	-	-
		5	L	0.15	0.15	0.15	0.21	0.15	0.15	-	-	-
			H	0.18	0.18	0.18	0.38	0.18	0.17	-	-	-
		7	L	0.20	0.20	0.20	0.35	0.20	0.20	-	-	-
			H	0.15	0.15	0.15	0.29	0.15	0.15	-	-	-
	15	3	L	0.09	0.09	0.09	0.11	0.10	0.09	-	-	-
			H	0.27	0.27	0.27	0.40	0.29	0.24	-	-	-
		5	L	0.08	0.08	0.08	0.15	0.07	0.06	-	-	-
			H	0.23	0.23	0.22	0.52	0.23	0.20	-	-	-
		7	L	0.19	0.19	0.19	0.25	0.20	0.19	-	-	-
			H	0.16	0.16	0.16	0.40	0.18	0.13	-	-	-
250	5	3	L	0.04	0.03	0.02	0.20	0.05	0.02	0.20	0.03	0.02
			H	0.18	0.16	0.16	0.25	0.16	0.16	0.27	0.16	0.16
		5	L	0.05	0.05	0.05	0.04	0.05	0.05	0.04	0.05	0.05
			H	0.39	0.39	0.39	0.83	0.42	0.38	0.89	0.42	0.38
		7	L	0.06	0.06	0.06	0.23	0.08	0.05	0.23	0.05	0.05
			H	0.18	0.18	0.18	0.36	0.18	0.16	0.28	0.16	0.16
	15	3	L	0.05	0.04	0.04	0.06	0.04	0.04	0.06	0.04	0.04
			H	0.32	0.31	0.30	0.58	0.32	0.30	0.51	0.30	0.30
		5	L	0.06	0.06	0.06	0.16	0.04	0.03	0.01	0.03	0.03
			H	0.17	0.17	0.16	0.24	0.15	0.14	0.20	0.13	0.12
		7	L	0.07	0.07	0.07	0.09	0.07	0.07	0.06	0.07	0.07
			H	0.20	0.20	0.20	0.47	0.23	0.14	0.38	0.14	0.13
500	5	3	L	0.04	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.05
			H	0.05	0.04	0.04	0.11	0.04	0.04	0.08	0.04	0.04
		5	L	0.06	0.06	0.06	0.22	0.07	0.05	0.13	0.05	0.05
			H	0.07	0.07	0.08	0.15	0.06	0.07	0.05	0.07	0.07
		7	L	0.08	0.08	0.08	0.15	0.11	0.08	0.14	0.08	0.08
			H	0.03	0.03	0.03	0.03	0.03	0.04	0.03	0.04	0.05
	15	3	L	0.03	0.02	0.02	0.08	0.02	0.01	0.02	0.02	0.01
			H	0.31	0.30	0.24	0.51	0.28	0.21	0.49	0.27	0.21
		5	L	0.19	0.19	0.18	0.51	0.18	0.14	0.33	0.14	0.13
			H	0.14	0.14	0.12	0.41	0.21	0.11	0.39	0.11	0.11
		7	L	0.13	0.13	0.13	0.16	0.12	0.12	0.13	0.12	0.12
			H	0.30	0.29	0.29	0.52	0.36	0.23	0.25	0.16	0.13

S: sample size, NI: number of items, RC: response category, ID: item discrimination, L: low, H: high

The findings regarding the average test statistics are presented in Table 3. Because each method utilizes different hypotheses to identify the items to be removed for violating the item ordering, each method yielded different test statistics (t, z and χ^2 values). For this reason, a direct comparison of these methods is not possible. Each method was merely examined based on a comparison in itself. In the MIIO method with a sample size of 100, the obtained statistical values were very close to zero. However, as the sample size increased, these values also increased. Test statistics varied between 0.00 and 5.87. An increase in the lowest violation coefficient had almost never effect on test statistics. The highest statistical values yielded by the MSCPM method was obtained in conditions where the sample size was 100 and the number of items was 5. It was observed that the higher the sample size and number of items were, the more stable the obtained values were. No pattern was observed in the findings yielded by the IT method. The value obtained with the increase in the lowest violation coefficient with the MSCPM method was very close to zero. However, in the IT method, especially in conditions where the sample size was 500, the number of items was 15, the item discrimination is high, and the response

categories were 5 and 7, χ^2 values were found to be very high even in conditions with the lowest violation coefficient of 0.45. Almost all the χ^2 values yielded by the IT method were at unexpected levels.

The findings regarding the H^T values are presented in Table 4. While the H^T values yielded by the MSCPM and IT methods were very close to each other, they were higher than those yielded by the MIIO method. However, the findings obtained from these two methods did not display any significant pattern. As the number of items increased, so did the H^T values yielded by all the methods. With a sample size of 250, higher H^T values were obtained in conditions where item discrimination was high. However, a similar pattern was not observed in the other simulation conditions. Consistent with the other findings, the MSCPM and IT methods were not affected by the lowest violation coefficient. The highest H^T values were yielded by the MSCPM and IT methods in conditions where the lowest violation method was 0.03. On the other hand, the lowest H^T values were obtained in conditions where the sample size was 500, the number of items was 15, the response category was 3 and the item discrimination was low.

When such is the case, it was observed in almost all the H^T values yielded by the MIIO method that the item ordering was not accurate. On the other hand, the MSCPM and IT methods can produce a moderate or high degree of accurate item ordering, especially in conditions where the lowest violation coefficient was 0.03. In conditions where the lowest violation coefficient was between 0.27 and 0.45, it was frequently observed, as in the MIIO method, that the item ordering used was not accurate.

DISCUSSION and CONCLUSION

This area of research initiated by Ligtvoet (2010) and Ligtvoet et al. (2011) with the methods they developed regarding invariant item ordering in polytomously categorized items is relatively new. Subsequent to these research studies in which methods were developed, even though some empirical studies are encountered in the literature, there are no technical or theoretical research studies. This implies that especially practitioners will be confused and will experience difficulties in deciding which method to use in which conditions and how to interpret the obtained coefficients. Especially in test administrations where items are ordered according to level of item difficulty – from easy to difficult, identification of the fixed item ordering is highly important for the interpretation of the test scores, especially in situations where items reflect the developmental traits of the measured cognitive stages or where item sets are clustered or hierarchical.

The most important findings obtained in the identification of invariant item ordering are the number of items violating the item ordering, the total number of item pairs causing violation, average test statistics, and the H^T values of the overall test (Ligtvoet, 2010). Hence, the present study focused on these values. The number of items violating ordering and the total number of item pairs causing violation yielded by the MSCPM and IT methods were higher than those yielded by the MIIO method. This finding is inconsistent with that reported in a study by Van der Ark (2012), where the MIIO and IT methods yielded a similar number of items to be removed. Moreover, Ligtvoet (2010) indicated that in a condition where the number of items was 20 and the response category was five, the IT method yielded 900 different violations in ordering. In the present study, the IT method yielded more than 1300 violations, much more than what the other methods identified. These two findings are in consistency.

While the MIIO method produced stable test statistics in all simulation conditions, the MSCPM method produced stable values in conditions where the sample size was 250 or above. However, the test statistics yielded by the IT method did not present any significant pattern. The fact that a condition where the lowest violation coefficient was 0.45 yields much higher values than those produced by a coefficient of 0.03 indicates that the values obtained via the IT method entails a high number of errors. While this is not consistent with the findings, the H^T values obtained via the MSCPM and IT methods were found to be higher. It was observed that the item ordering in almost all the H^T values obtained by means of the MIIO method was incorrect.

When the findings were considered in general, it was found that the MIIO method yielded the most stable values due to the fact that it was not affected by the lowest violation coefficient and was affected only slightly by simulation conditions. Especially in conditions where the violation coefficient is 0.03 (the default value in the Mokken package), it is recommended to use the MIIO method in identifying item ordering. Even though the MSCPM method yields similar findings to those of the IT method, it generates more stable findings in particularly high sample sizes. In conditions where sample size, number of items and item discrimination are high, the MSCPM is recommended to be used. However, further studies need to be conducted on the IT method. The use of the IT method is not recommended due to lack of theoretical information.

In this relatively new field of study, there is a need for further theoretical and empirical studies. Conducting further studies on obtaining error values as regards invariant item ordering, error type 1 and power analysis is recommended. There is also a need to conduct similar studies on real datasets. Especially MIIO method must be used as a scaling procedure for scale development, person ordering, item ordering and validity studies.

REFERENCES

Ahmadi, K., Reidpath, D. D., Allotey, P., & Hassali, M. A. A. (2016). A latent trait approach to measuring HIV/AIDS related stigma in healthcare professionals: Application of Mokken scaling technique. *BMC Medical Education*, *16*(1), 155-164. doi:10.1186/s12909-016-0676-3

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573. doi:10.1007/BF02293814

Desa, Z. N. (2012). *Bi-factor multidimensional Item Response Theory modeling for subscores estimation, reliability, and classification* (Doctoral dissertation, University of Kansas), ProQuest LLC.

Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement, 41*(3), 261-270. doi:10.1111/j.1745-3984.2004.tb01165.x

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, *13*(1), 77-90.

Gibbons, C. J., Small, N., Rick, J., Burt, J., Hann, M., & Bower, P. (2017). The patient assessment of chronic illness care produces measurements along a single dimension: Results from a Mokken analysis. *Health and Quality of Life Outcomes*, *15*(1), 61-69. doi:10.1186/s12955-017-0638-4

Lee, C. P., Chen, Y., Jiang, K. H., Chu, C. L., Chiu, Y. W., Chen, J. L., & Chen, C. Y. (2016). Development of a short version of the Aging Males' Symptoms scale: Mokken scaling analysis and Rasch analysis. *The Aging Male*, *19*(2), 117-123. doi:10.3109/13685538.2016.1157861

Ligtvoet, R. (2010). *Essays on invariant item ordering*. Unpublished doctoral dissertation, Tilburg University, the Netherlands, ProQuest LLC.

Ligtvoet, R., Van der Ark, L. A., Bergsma, W. P. & Sijtsma, K. (2011). Polytomous latent scales for the investigation of the ordering of items. *Psychometrika, 76*, 200-216. doi:10.1007/s11336-010-9199-8

Ligtvoet, R., Van der Ark, L. A., & Sijtsma, K. (2008). Selection of Alzheimer symptom items with manifest monotonicity and manifest invariant item ordering. *New Trends in Psychometrics*, *3*(1), 225-234.

Ligtvoet, R., Van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, *70*(4), 578-595. doi:10.1177/0013164409355697

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

Maydeu-Olivares, A., Kramp, U., García-Forero, C., Gallardo-Pujol, D., & Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior Research Methods*, *41*(2), 295-308. doi:10.3758/BRM.41.2.295

McGrory, S. (2015). *Non-parametric item response theory applications in the assessment of dementia*. Unpublished Doctoral Dissertation. University of Arizona, ProQuest LLC.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417-430. doi:10.1177/014662168200600404

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*(1), 59-71. doi:10.1177/014662169001400106

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen and Lydiche.

Saiepour, N., Najman, J. M., Clavarino, A., Baker, P. J., Ware, R. S., & Williams, G. (2014). Item ordering of personal disturbance scale (DSSI/sAD) in a longitudinal study; using Mokken scale analysis. *Personality and Individual Differences*, *58*, 37-42. doi:10.1016/j.paid.2013.09.030

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, *49*(1), 79-105. doi:10.1111/j.2044-8317.1996.tb01076.x

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*, 149-157. doi:10.1177/014662169201600204

Sijtsma, K., Meijer, R. R., & van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling procedures. *Personality and Individual Differences, 50*, 31-37. doi:10.1016/j.paid.2010.08.016

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage. Stewart, M. E., Allison, C., Baron-Cohen, S., & Watson, R. (2015). Investigating the structure of the autism-spectrum quotient using Mokken scaling. *Psychological assessment*, *27*(2), 596-604. doi:10.1037/pas0000058

Stochl, J., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: A non-parametric IRT method in empirical research for applied health researchers. *BMC Medical Research Methodology*, *12*(1), 74. doi:10.1186/1471-2288-12-74

Van Abswoude, A. A., van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*(1), 3-24. doi:10.1177/0146621603259277

Van Abswoude, A. A., Vermunt, J. K., Hemker, B. T., & van der Ark, L. A. (2004). Mokken scale analysis using hierarchical clustering procedures. *Applied Psychological Measurement*, *28*(5), 332-354. doi:10.1177/0146621604265510

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of statistical software*, *20*(11), 1-19.

Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1-27.

Van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: Dynamic assessment of the control of variables strategy. *Instructional Science*, *43*(3), 381-400. doi:10.1007/s11251-015-9344-y

Watson, R., Deary, I. J., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological medicine*, *38*(4), 575-579. doi:10.1017/S003329170800281X

Wechsler, D. (1999). *WISC-III: manual: Wechsler intelligence scale for children*. Psychological Corporation.

Yoon, S., Shaffer, J. A., & Bakken, S. (2015). Refining a self-assessment of informatics competency scale using Mokken scaling analysis. *Journal of Interprofessional Care*, *29*(6), 579-586. doi:10.3109/13561820.2015.1049340

Mokken Ölçekleme Analizleri Kullanılarak Çok Kategorili Puanlanan Maddelerde Değişmez Madde Sıralamasının İncelenmesi

Giriş

Testte yer alan maddelerin sıralaması geleneksel olarak madde güçlüğüne göre yapılmaktadır. Ancak bir maddenin diğerinden daha zor olması o maddenin teste ait tüm alt testlerde de aynı güçlük düzeyinde olduğu anlamına gelmez. Örneğin, bir test maddesi düşük yetenek gerektiren bir alt test için zor bir test maddesi olabilirken yüksek yetenek gerektiren bir alt test için tam tersi bir sıralama ortaya çıkabilir (Ligtvoet, 2010). Ancak ölçme uygulamalarında madde sıralaması, maddelerin zorluğuna ya da cazipliğine bağlı olarak tüm katılımcılar için aynı olmalıdır. Örneğin çocuklar için geliştirilen zekâ testlerinde sorular güçlük düzeyine göre sıralanmaktadır (Wechsler, 1999). Bu sıralamanın temel amacı, öğrencinin zor sorularla karşılaştığında panik olmasını engellemek ve performansını teste yansıtmasını sağlamaktır. Diğer amaç ise farklı yaş gruplarında yaş arttıkça alt testlerin güçlük düzeylerinin de artmasını sağlamaktır (Ligtvoet, 2010).

Test maddelerinin sadece madde güçlüğüne göre sıralanması ile ortaya çıkabilecek problemlere çözüm getirebilmek amacıyla *değişmez madde sıralaması* (DMS) (Sijstma ve Junker, 1996) geliştirilmiştir.

DMS, madde sıralamasının tüm katılımcılar için aynı olması durumudur ve kullanımının yararlı olduğu pek çok açıdan kanıtlanmıştır. DMS, madde tepki kuramı (MTK) çerçevesinde tanımlanmaktadır. Test maddelerinin DMS'sinin belirlenebilmesi için MTK modellerinin varsayımlarını sağlaması gerekmektedir. Sijtsma ve Junker (1996), DMS'nin yalnızca madde tepki fonksiyonunun (item response function – IRF) kesişmediği MTK modellerinde kullanılabileceğini göstermiştir. DMS, ikili puanlanan veri setlerinde yalnızca Rasch (1960) ve ikili monotonluk modeline (İMM) (Mokken ve Lewis, 1982) uygulanabilmektedir. Çok kategorili puanlanan veri setlerinde ise yalnızca dereceleme ölçeği modeli (Andrich, 1978) ve sınırlandırılmış dereceli tepki modeline (Muraki, 1990) DMS uygulanabilmektedir.

Bu araştırmanın amacı dereceli tepki modeli aracılığıyla elde edilen simülatif veri setlerinde üç farklı Mokken DMS yönteminden elde edilen sıralamayı ihlal eden madde sayısını, toplam ihlale neden olan madde çifti sayısını, test istatistiklerinin ortalamasını ve testin geneline ait H^T değerlerini belirlemek ve karşılaştırmaktır.

Yöntem

Çok kategorili puanlanan veri setlerinde yalnızca dereceleme ölçeği modeli (Andrich, 1978) ve sınırlandırılmış dereceli tepki modeli (Muraki, 1990) DMS gösterebilmektedir. Bu araştırmanın veri üretiminde dereceli tepki modeli kullanılmıştır. Her bir veri setine 20 tekrar uygulanmıştır. 2 (madde ayırt edicilik düzeyleri) x 3 (örneklem büyüklüğü) x 2 (madde sayısı) x 3 (yanıt kategorisi) olmak üzere 36 veri seti * 20 tekrar ile 720 veri kümesi elde edilmiştir. Araştırmanın bağımlı değişkenleri sıralamayı ihlal eden madde sayısı, toplam ihlale neden olan madde çifti sayısı, test istatistiklerinin ortalaması ve testin geneline ait H^T değerleridir. Veri üretimi WINGEN 2.0 programı ile yapılmıştır.

Tüm simülasyon koşulları 3 (en düşük ihlal katsayısı değerleri) x 2 (madde ayırt edicilik düzeyleri) x 3 (örneklem büyüklüğü) x 2 (madde sayısı) x 3 (yanıt kategorisi) olmak üzere 108 test koşulundan oluşmaktadır. Her bir hücre için Mokken ölçekleme analizleri çerçevesinde ele alınan MIIO, MSCPM ve IT yöntemleri uygulanarak elde edilen sıralamayı ihlal eden madde sayısı, toplam ihlal edilen madde çifti sayısı, test istatistiklerinin ortalaması (t, z ve χ^2 değerleri) ve testin geneline ait H^T değerlerini belirlenmiştir. Analizler R programındaki Mokken 2.8.10 (Van der ark, 2007) paketi ile gerçekleştirilmiştir.

İkili puanlanan veri setlerinde H^T katsayısını Sijstma ve Meijer (1992) geliştirmiştir. Çoklu puanlanan maddelerde, Ligtvoet vd. (2011) bu araştırmanın temel bağımlı değişkeni olan H^T katsayısını H ölçeklenebilirlik katsayısının yorumlanmasını genelleştirerek geliştirmiştir. MIIO, MSCPM ve IT yöntemlerinin aynı anda kullanıldığı araştırmalarda elde edilen ortak sıralamayı ihlal eden maddeler testten çıkartılması gereken maddelerdir. Bu ihlalin düzeyi en düşük ihlal katsayısı ile belirlenmekte ve bu değer varsayılan olarak 0.03 olarak ele alınmaktadır. Bu değerin azalması en küçük bir ihlalin bile kabul edilmesi anlamına gelmektedir. İhlalin düzeyi MIIO yönteminde t testi tekniği (t değerleri) ile, MSCPM yönteminde z testi tekniği (z değerleri) ile ve IT yönteminde ki-kare testi tekniği (χ^2 değerleri) ile ortaya koyulmaktadır. İstatistiksel olarak anlamlı olacak şekilde ihlale neden olan maddeler sırayla testten çıkartılmalı; eğer iki veya daha fazla madde yüksek düzeyde ihlale sahipse ölçeklenebilirlik katsayısı en düşük olan madde testten çıkartılır (Ligtvoet, 2010).

Sonuç ve Tartışma

Ligvoet (2010) ve Ligtvoet vd. (2011) çok kategorili maddelerde değişmez madde sıralamasına ait geliştirdiği yöntemler ile başlayan bu araştırma alanı oldukça yenidir. Yöntemlerin geliştirildiği bu araştırmalardan sonra bazı uygulama araştırmalarına rastlanmakla birlikte teknik ve kuramsal herhangi bir araştırma literatürde yer almamaktadır. Bu durum özellikle uygulayıcıların hangi yöntemi hangi durumda seçmeleri ve elde edilen katsayıların nasıl yorumlanacağı konusunda kafa karışıklığı yaşayarak zorlanacakları anlamına gelmektedir. Özellikle madde sıralamasının kolaydan zora doğru yapıldığı test uygulamalarında, maddelerin ölçtüğü bilişsel basamakların gelişim özelliklerini

yansıttığı veya madde setlerinin hiyerarşik ya da kümelenmiş olduğu durumlarda değişmez madde sıralamalarının belirlenmesi test puanlarının yorumlanması için oldukça büyük bir öneme sahiptir.

Değişmez madde sıralamasının belirlenmesinde elde edilen en önemli bulgular, sıralamayı ihlal eden madde sayısını, toplam ihlale neden olan madde çifti sayısını, test istatistiklerinin ortalamasını ve testin geneline ait H^T değerlerini belirlemek olduğu söylenebilir. Bu nedenle bu araştırma bu değişkenlere odaklanmıştır. MSCPM ve IT yöntemlerinin belirlediği sıralamayı ihlal eden madde sayısı ve toplam ihlale neden olan madde çifti sayısı MIIO yönteminden daha fazladır. Bu bulgu Van der Ark'ın (2012) MIIO ve IT yöntemlerinin benzer sayıda madde atılmasını önerdiğini belirttiği çalışması ile farklılık göstermektedir. Ayrıca Ligtvoet (2010) araştırmasında madde sayısının 20 ve cevap kategorisinin beş olduğu durumda IT yönteminin 900 farklı sıralama ihlali ürettiğini belirtmiştir. Bu araştırmada da IT yöntemi 1300'ün üzerinde ihlal üreterek diğer yöntemlerden çok daha fazla sayıda ihlal üretmiştir. Bu iki araştırma bulgusu benzerlik göstermektedir.

MIIO yöntemi tüm simülasyon koşullarında stabil test istatistiği değerleri elde ederken, MSCPM yöntemi örneklem büyüklüğünün 250 ve üstü olduğu durumlarda stabil değerler üretmiştir. Ancak IT yönteminden elde edilen test istatistikleri bir örüntü göstermemektedir. En düşük ihlal katsayısı 0.45 olduğu durumda, 0.03 olduğu duruma göre çok daha yüksek değerler elde edilmesi, IT yöntemi ile elde edilen değerlerin yüksek hata içerdiği hakkında ipucu vermektedir. Bu bulgularla örtüşmemekle birlikte, MSCPM ve IT yöntemlerinden elde edilen H^T değerlerinin daha yüksek olduğu belirlenmiştir. MIIO yönteminden elde edilen H^T değerlerinin neredeyse tamamında madde sıralamasının kullanımının doğru olmadığı görülmektedir.

Bulgulara genel olarak bakıldığında MIIO yönteminden elde edilen değerlerin en düşük ihlal katsayısından etkilenmemesi ve simülasyon koşullarından düşük düzeyde etkilenmesi gibi nedenlerden dolayı en stabil değerler ürettiği belirlenmiştir. Özellikle ihlal katsayısının 0.03 olduğu durumlarda (Mokken paketindeki varsayılan değer) MIIO yöntemi ile değişmez madde sıralamasının belirlenmesi önerilmektedir. MSCPM yöntemi IT yöntemine benzer bulgular üretmekle birlikte özellikle yüksek örneklem büyüklüklerinde daha stabil değerler üretmektedir. Örneklem büyüklüğü, madde sayısı ve madde ayırt ediciliğinin yüksek olduğu durumlarda kullanılması önerilebilir. Ancak IT yöntemi üzerinde daha fazla çalışma yapılması gerekmektedir. IT yönteminin kullanılması var olan kuramsal bilgi altında önerilmemektedir.

Çok yeni bir alan olan bu konuda kuramsal ve uygulamalı yeni araştırmalara ihtiyaç duyulmaktadır. Değişmez madde sıralamasına ait hata değerlerinin elde edilmesi ve I. tip hata ve güç oranlarının çalışılması önerilebilir. Gerçek veri setleri üzerinde de benzer araştırmaların yapılması gerekmektedir. Özellikle ADMS yöntemi ölçek geliştirme, madde ve kişi sıralama ve geçerlik çalışmaları gibi konularda bir ölçekleme yöntemi olarak kullanılabilir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(4);326-338

An Investigation of the Factors Affecting the Vertical Scaling of Multidimensional Mixed-Format Tests*

Akif AVCU ** Hülya KELECİOĞLU ***

Abstract

This study examined the effect of the structure of a common item set (only dichotomous common items – mixed-format common item sets), parameter estimation methods and scale shrinkage on vertical scaling results when multidimensional datasets were used within the context of Common Item Nonequivalent Group (CINEG) design. Interactions between these variables were also investigated. The study was performed using simulated data. Measurement error and bias indexes were used to evaluate the quality of vertical scaling. All the procedures used in the data analysis were replicated 50 times to increase the generalizability of the results. R program was used for the data generation, calibration of the parameters and vertical scaling procedures. Possible interactions were investigated with factorial analysis of variance by using SPSS. The results showed a consistent effect of the common item format in all conditions. In addition, some interactions between the variables were observed. These findings are discussed and some recommendations are provided.

Keywords: Vertical scaling, Multidimensional tests, Mixed format tests

INTRODUCTION

Test scores are among the primary sources of information that educators and educational institutions use in making important decisions about students. Thus, test scores must provide the accurate information to facilitate appropriate decisions (Kolen and Brennan, 2004). However, different forms of the same test are often used due to the reasons such as test safety and follow up student development. A functional link between these forms needs to be established so that the scores from different test forms are comparable. This process is called test linking. Test linking is the process of establishing a relationship between different test forms. There is no requirement that the content and difficulty levels between the test forms for test binding be the same. Test equating is a special form of linking in which the aim is to use the scores between the different test forms interchangeably. Hence, test forms should be similar in content and difficulty (Kolen and Brennan, 2004). Vertical scaling is similar to the equating because different test forms are linked to each other. However, test forms differ in content and difficulty because they reflect progression between classes or age groups. Therefore, while vertical scaling is used to compare different test forms, the scores at each level can not be used in place of each other. When the scores are put onto a common scale, students' grade-to-grade improvement can be seen. The main aim of vertical scaling is to observe student progress.

Dichotomous items were the most widely used item format in the 20th century (Koretz and Hamilton, 2006). Today, however, the use of mixed-format tests, which contain both dichotomous and polytomously scored items, is rapidly becoming widespread. Mixed-format tests offer many advantages. According to Livingston (2009), multiple-choice questions may be used to measure a test taker's ability with high reliability for a wide range of contents, in a short time and at low cost. On the other hand, open-ended questions measure higher-level cognitive skills more effectively, but tests

* This study is based on the dissertation "Çok boyutlu karma-format testlerin ölçeklenmesini etkileyen faktörlerin incelenmesi" advised by Prof. Dr. Hulya Kelecioglu in July, 2006.

** Phd, Marmara University, Atatürk Faculty of Education, İstanbul-Turkey, e-mail:avcuakif@gmail.com ORCID ID: 0003-1977-7592
*** Professor, Hacettepe University, Ankara-Turkey, e-mail:hulyakelecioglu@gmail.com, ORCID ID: 0002-0741-9934

To cite this article:
Avcu, A., & Kelecioğlu, H. (2018). An investigation of the factors affecting the vertical scaling of multidimensional mixed-format tests. *Journal of Measurement and Evaluation in Education and Psychology, 9*(4), 326-338. DOI: 10.21031/epod.394659

Received: 14.02.2018
Accepted:13.10.2018

made up of these items have narrower content, are costly to assess, and are likely to be subjective. Mixed-format tests eliminate the disadvantages of different formats and increase the psychometric qualities of the instruments.

One of the variables that this study examines is the effect of scale shrinkage which becomes relevant when measurement tools are applied at different time points to detect students' progress. Scale Shrinkage is the extent to which the variance and range of scores decrease in the second application compared to the first (Yen, 1985). As students continue any program, they become more homogenous in terms of their ability compared to beginning. This leads their score variances to shrink at the later test applications. The scale shrinkage corresponds to the homogeneity. So far, there has been a lack of research on how scale shrinkage affect the results of vertical scaling.

Another important variable examined in this study is the structure of the common item set. Mixed-format tests are scaled vertically through the use of only dichotomous items, mixed-format items (including at least two different items in the common item set and only polytomous items). In this study, only the dichotomous common item set and mixed-format common item set conditions were compared. Although, positive outcomes were obtained for the mixed common item set for vertical scaling applications (e.g., Kim and Lee, 2006), it could be valuable to see the results within the context of the current study where a different combination of variables is included.

Most software programs routinely carry out estimations using expectation - maximization (EM) algorithms (Bock and Aitkin, 1981). Another method that has recently started to be used is the Metropolis-Hastings Robbins-Monro (MHRM) method developed by Chai (2010). The performance of EM and MHRM have been compared in estimating multidimensional dichotomous models (Han and Paek, 2010) and multidimensional polytomous models (Kuo and Sheng, 2016). However, no study was found comparing their performances in the context of multidimension mixed-format scaling. A comparison of these estimation methods could contribute important insights to the literature. Thus, we also varied the estimation method across the study conditions.

Vertical Scaling of Mixed-Format Tests

Dichotomous Item Response (IRT) Models:

Dichotomous response models are based on a three-parameter logistic model developed by Birnbaum (1968). This model is expressed in formula 1 (Lord and Novick, 1968).

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}} \quad (1)$$

Here θ corresponds to the level of the individual's ability, a to the distinction parameter, b to the difficulty parameter, and c to the so-called chance parameter of the item. When this model was first introduced, it was used for one-dimensional tests, but since the 1980-s it has been used for multidimensional models as well. The generalization of Birnbaum' (1968) model for multidimensional tests is given in the following section.

For example, let us say that, i = 1,……, N are different participants, j = 1,……, and n are test items. Also, suppose m is a latent factor. θ = (θ_{i1},….,θ_{iN}). The slope parameters associated with the dimensions are a_j = (a_1,……, a_j). The likelihood of responding to a dichotomous item for multidimensional 3PLMs becomes as presented in formula 2:

$$\phi\,(x_{ij} = 1 \mid \theta_i, \alpha_j, d_j, \gamma_j) = \gamma_j + \frac{(1-\gamma_j)}{1+\exp[-D\left(\alpha_{j,}^{\mathrm{T}}\ \theta_i + d_j\right)]} \quad (2)$$

Here, d_j corresponds to the intersection parameter, γ_j corresponds to "*chance*" parameter, and D corresponds to the scaling constant. This value is generally taken as 1.702, and it is used to transform the logistic metric to the traditional normal ogive metric (Reckase, 2009).

Polytomous IRT Models

Although there are different models for polytomous items in the literature, the graded response model (GRM) is preferred in this study. In this model, developed by Samejima (1969, 1972), if we assume that the discrimination parameters are kept constant. $\tilde{P}_{ijk}$ corresponds to the cumulative probability that a person *i* with θ_i ability level can obtain a score beyond the category k of the item j. For the K_j categories, $\tilde{P}_{ijk}$ could be expressed as follows:

$$\tilde{P}_{ijk}=\tilde{P}_{jk}(\theta_i) = \tilde{P}(\theta_i)|\alpha_j, b_{jk}) = \begin{cases} 1 & \\ \dfrac{\exp[Da_j(\theta_i - b_{jk})]}{1+\exp[Da_j(\theta_i - b_{jk})]} & \begin{matrix} k = 1, \\ 2 \le k \le K_j, \\ K > K_j \end{matrix} \\ 0 & \end{cases} \quad (3)$$

Here, a_j corresponds to the discrimination parameter, b_{jk} corresponds to difficulty (or threshold parameter) from the second category to category K, and D corresponds to the scaling constant. The category response function, P_{ijk}, corresponds to the difference between two adjacent cumulative probabilities and is expressed as follows:

$$P_{ijk} = P_{jk}(\theta_i) = \tilde{P}_{ijk} - \tilde{P}_{ij(k+1)} \quad (4)$$

Samejima (1969) and Carlson (1995) used the GRM to generalize this to multidimensional situations.

In the model, the boundaries of the response categories for the C_j categories belonging to item *j* and the $d_j = d_1, \ldots\ldots d_{(c_{(j)}-1)}$ intersections are expressed as follows:

$$\phi(x_{ij} \ge 0|\theta_i, \alpha_j, d_j) = 1$$

$$\phi(x_{ij} \ge 1|\theta_i, \alpha_j, d_j = \frac{1}{1+\exp[-D(\alpha_j^T, \theta_i, d_j]}$$

$$\phi(x_{ij} \ge 2|\theta_i, \alpha_j, d_j = \frac{1}{1+\exp[-D(\alpha_j^T, \theta_i, d_j]}$$

.....

$$\phi(x_{ij} \ge 0|\theta_i, \alpha_j, d_j) = 0$$

(5)

Vertical Scaling

In the literature, moment and characteristic methods are the most commonly preferred methods used to apply vertical scaling. The moment methods, namely, mean / sigma (Marco 1977) and mean/mean (Loyd and Hoover, 1980), are the simplest methods, and only the parameter estimates need to be known in order to estimate linking constants. Alternative methods to the moment methods are the characteristic curve methods developed by Haebara (1980) and Stocking and Lord (1983). These methods based on minimization of the differences between characteristics curves of items. Comprehensive analysis and comparisons of these methods were provided by Kolen and Brennan (2004). These methods were extended to link mixed format tests. A detailed information can be found in Kim and Lee (2006)'s study. This study uses the Haebara method.

Purpose of the Study

Based on the literature presented above, the aim of the current study is to investgate the effect of common item structure, scale shrinkage, and estimation methods on the vertical scaling of multidimensional mixed-format tests.

METHOD

Data Simulation

Simulated datasets were used in the study. In addition, population parameters were also simulated considering that the values can be observable in real testing conditions. The dimensionality structure was prepared considering the two-tier model proposed by Cai (2010). In two-tier models, main dimensions and special dimensions are used as the source of the dimensionality. The terms "*main*" and "*special*" do not imply that main dimensions are theoretically more important or the variance/covariance structure between the dimensions is different. There is no theoretical relationship between main and special dimensions in a two-tier model. In addition, special factors are mutually orthogonal, and items have loading from only one dimension. On the other hand, the main factors may be related to each other. In the context of this study, content and item format were regarded as two sources of dimensionality in the data simulation. Accordingly, "*content*" was regarded as a special dimension source and "*item format*" as a main dimension source. Dual effect of content and item format on test dimensionality was investigated by Zhang (2016), but no study was found that take both factors into consideration when conducting scaling studies using simulated data. Figure 1 shows the model used for the data simulation in the current study. As seen, the three dimensions based on content and the two dimensions based on item format are intertwined in the model in compatible with two-tier models. The variance-covariance matrix used for the data simulation was established based on this model. The matrix is shown in Fügure 2. As seen in the figure, the relationship between the general factors is set to be 0.75 Among the special factors, this value is 0. Likewise, covariance values are assumed to be 0, showing no relationship between general and special items.

Simulation of Person and Item Parameters

In order to obtain accurate and stable parameter estimations in Multidimensional Item Response Theory (MIRT), as a sample size of 3000 was recomended (Yao and Boughton, 2009). Thus, in this study, the sample size consisted of 3000 simulated examinees. Theta scores were simulated from a normal distribution. The mean for the lower ability group was set to 0 and that for higher ability group was set to 1 with different scale shrinkage levels. One point ability difference between the groups was acceptable value that can be seen in real testing conditions (e. g., Kim, 2007). The theta vectors were simulated for each specific factor. Thus, final θ matrices with 3×3000 size were obtained as the population parameters for each group. In addition, variances of the population ability parameters were controlled. For the cases of scale shrinkage, variances were set to a shrinkage of 65% for the higher ability group. This amount of shrinkage was selected based on the study literature review provided by Yen (2005). This level of shrinkage was the case for half of the datasets, while for the rest of the datasets, variances were kept the same for both datasets used in the vertical scaling.

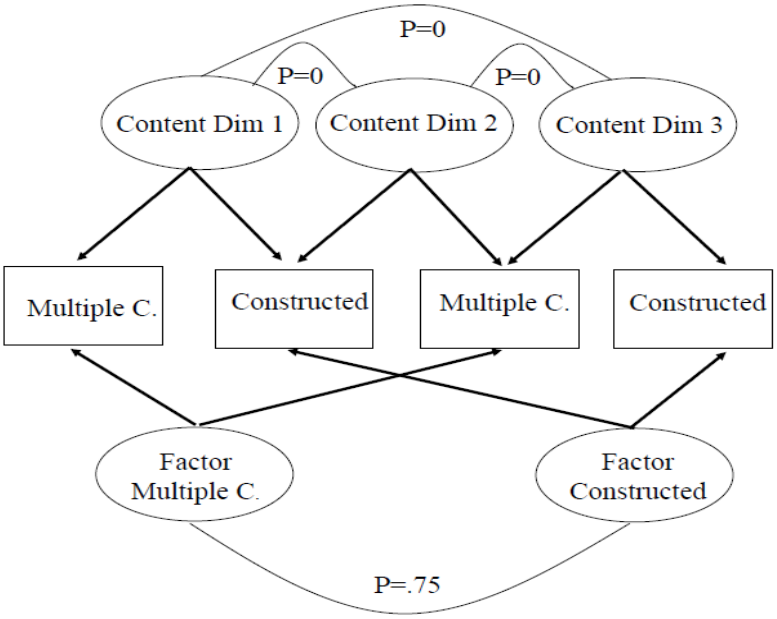


Figure 1. Two Tier Model

In addition, the datasets were created to be composed of 108 items (90 dichitomous and 18 polytomously scored items). In this scenario, there were 54 items (45 dichotomous and 9 polytomous) in each main factor and 36 items (30 dichotomous and 6 polytomous) in each factor.

Population *a* parameters for the generation of the data matrices were generated for each dimension (for each of the main and dimensions dimensions). Thus, a final matrix with a size of 5 × 108 was obtained for each dataset. For the main factors, if the item belonged to a dimension, the mean *a* value was determined as high discrimination power and fixed at 1 and the standard deviation at 0.15. If an item did not belong to a dimension, the mean value was fixed at 0.2 and the standard deviation at 0.03, because these items were not expected to have high level of discrimination. For special factors, if the item is included in that dimension, the mean value was fixed at 1 and the standard deviation at 0.15, while if the item was not included in that dimension, the all the *a* values were fixed to be 0 because of simple structure of spesific factors. All the simulated discrimination parameters were selected from the standard normal distribution.

$$\begin{bmatrix} 1 & 0.75 & 0 & 0 & 0 \\ 0.75 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 2. Variance-Covariances for Three Dimensional Data

The difficulty parameters (*b*) were produced as 1 × 108 vectors for dichotomous items. For the lower ability group, the mean was set to be 0 with a standard deviation of 1. For the higher ability group, the mean was set to be 1 with a standard deviation of 1. Polytomous items were configured as having a 5-point scoring format. For this reason, four intercept parameters for each item were simulated. The threshold values for the lower ability group are simulated with means that ranged from -1.5 to 1.5 with

a 1-point increase for every adjacent thresholds. For the higher ability group, same procedure was repeated except the range was set to -1 to 1. The distribution of the difficulty parameters was selected from a normal distribution with a standard deviation of 0.1. Data matrices were simulated as described above using parameter estimates and matrices produced for the calibration process.

Parameter Estimation

In this study, 3PLM and GRM were used to calibrate the mixed-format tests. This combination is preferred in many studies. Rosa, Swygert, Nelson, and Thissen (2001) pointed out that 3PLM is preferred for calibration because more parameters on the items are taken into account and the model therefore gives more information. In the literature, GRM and partial credit models are preferred in the calibration of polytomously response models (Kim and Cohen, 2002, Bastari, 2000, Tate, 2000). Dodd (1984) concluded that the two model types produce similar the results despite being conceptually and mathematically different. Cao, Yin and Gao (2007) also found that the two models yield similar results.

For theta estimation the MAP was preferred in this study. Each data set was calibrated separately so that the scaling process could be performed. In the analysis of each data set for EM cycles, the convergence value and the number of iterations were taken as 0.001 and 500, respectively. For the MH-RM estimation technique, the convergence value was set to 0.0001 and the number of iterations to 2000.

Evaluation Criteria

As the evaluation criteria of the results, root mean square error (RMSE) and bias were used in parallel with similar studies. RMSE shows the amount of random error fort he scaling process. The computation of RMSE is given in formula 6:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}} \tag{6}$$

The bias values provide information on the systematic error detected during scaling process and are calculated as described in formula 7:

$$\text{Bias} = \frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)}{n} \tag{7}$$

Data Analysis

The data analysis was performed using the R statistic program (R Core Team, 2015). Different R packages were loaded and the analyses were carried out. Firstly, the "*truncnorm*" package developed by Trautmann et al. (2014) was used when *d*-matrices were derived. This package is preferred for controlling the upper and lower bounds of derived values of population threshold parameters. In this way, it was ensured that the difference between successive threshold parameters did not fall below 0,3 and model-data mis-fit was prevented. Other population parameters were obtained by using the "*rnorm*" command in the R program.

Later, the "*mirt*" package developed by Chalmers (2012) was used. Response matrices is producesd with the command "*sim*", and calibration was performed with "*mirt*" command. Finally, the ability parameters were estimated using the "*fscores*" command. When the scaling was performed, the "*plink*" package developed by Weeks (2010) is utilized. Each analysis was replicated 50 times. Then, the error and bias values of the parameters obtained from 50 replications were used with analysis of variance (ANOVA) to compare the conditions tested, and in 2 × 2 × 2 factorial ANOVA to see the interactions among the conditions being tested.

RESULTS

This section discusses the amount of error (RMSE), bias (Bias) values, and results of the factorial ANOVA for each research question as the major findings of the study. Common item set was excluded in the calculation of the error and bias values. In addition, given values of the dichotomous and polytomous items were calculated separately. As a last caution to the reader, the error and bias values were calculated separately for each of the three dimensions. The values are presented in Table 1.

Table 1 shows that the common item structure had a significant effect on some estimates of the synchronization structure. The error values in the threshold parameters of the polytomous items for the first dimension were higher in cases where the mixed-format common item sets were used. Under all conditions, the *a* parameters of polytomous items were found to have higher in situations where mixed-format common item sets were used. In addition, for the threshold parameters with scale shrinkage and MHRM, the mixed-format common item structure elicited more errors. In the third dimension, the errors and bias values for mixed-format common items were lower except for in the *a* parameter of the polytomous items.

The scale shrinkage effect was examined as the next. For the first dimension, it was found that for the first dimension, the amount of error and bias obtained for the threshold parameter in the tests using dichotomous items was higher when the scale shrank. With the mixed-format common item structure, the MHRM estimation method and scale shrinkage, and the amount of error was lower for all the item parameters. For the second dimension, the bias amounts for the item parameters in the conditions using EM cycles and the only dichotomous common items were lower with no scale shrinkage. Similarly, the bias values of the ability parameters for datasets using mixed-format common items are also lower when the scale is not shrunk. In the third dimension, when EM cycles and dichotomous items were used, the error and bias amounts of the item parameters were generally lower in the cases of no scale shrinkage.

Regarding the estimation method, the error and bias values were lower with no scale shrinkage for parameters *a* and *b* of dichotomous items in the first dimension. On the other hand, with scale shrinkage, only the error values were lower in the EM estimation method. In addition, the error and bias values obtained from the *a* parameters of the polytomous items for the data in which only dichotomous items were included in the common item set were lower with the EM estimation method. These values for the second dimension showed similar changes to those in first dimension. Unlike for the first dimension, it was seen that, for this dimension, the bias values of the ability parameters were lower when the mixed-format item structure was used and there was no scale.Finally, the findings for the third dimension showed EM cycles produced lower error and bias values for all the item parameters with no scale shrinkage and a dichotomous common item structure was preferred.

Later, the 2 × 2 × 2 factorial ANOVA results were examined to see whether the observed differences in the bias and error values were significant, and whether there was an interaction between the conditions investigated. The results are presented in Table 2.

Regarding the interactions for the first dimension, there was a significant interaction between the CIF and SS conditions for the bias values of the ability parameters ($p < .05$). According to the analyses performed to test whether the levels of interaction of the CIF and EM conditions were meaningful, these two conditions interacted with the bias values of the threshold parameters of polytomous items ($p < .05$).

Table 1. Error and Bias Values Across the Conditions

		No Shrinkage				Shrinkage			
		EM		MHRM		EM		MHRM	
		Error	Bias	Error	Bias	Error	Bias	Error	Bias
	First Dimension								
Dich. common items	*θ*	0.068	-3.654	0.060	-3.294	0.060	-3.227	0.058	-3.431
	a param. (dich.)	0.074	-0.241	0.086	-0.340	0.069	-0.351	0.072	-0.337
	b param. (dich.)	0.624	2.911	0.641	2.992	0.604	2.817	0.669	3.119
	a param. (poly.)	0.132	-0.025	0.139	-0.036	0.136	-0.045	0.132	0.035
	b param. (poly.)	1.116	0.771	1.066	0.835	1.140	0.774	1.074	0.836
Mixed format common items	*θ*	0.034	-1,738	0.032	-1,736	0.032	-1.744	0.033	-1.964
	a param. (dich.)	0.050	0.158	0.050	0.104	0.045	0.158	0.052	0.101
	b param. (dich.)	0.457	2,226	0.475	2,315	0.433	2.111	0.489	2.384
	a param. (poly.)	0.123	0.034	0.129	0.023	0.121	0.035	0.130	0.019
	b param. (poly.)	1,308	0.512	1,189	0.498	1.246	0.547	1.233	0.496
	Second Dimension								
Dich. common items	*θ*	0.053	-2.858	0.049	-2.543	0.047	-2.568	0.050	-2.449
	a param. (dich.)	0.074	-0.244	0.084	-0.347	0.070	-0.351	0.087	-0.339
	b param. (dich.)	0.556	2.595	0.590	2.754	0.578	2.696	0.590	2.755
	a param. (poly.)	0.128	0.023	0.130	0.022	0.122	0.027	0.122	0.021
	b param. (poly.)	1.206	0.771	1.175	0.835	1.258	0.774	1.171	0.836
Mixed format common items	*θ*	0.030	-1,751	0.033	-1,798	0.034	-1.874	0.032	-1.909
	a param. (dich.)	0.050	0.174	0.052	0.121	0.046	0.172	0.052	0.116
	b param. (dich.)	0.528	2,576	0.543	2,644	0.491	2.392	0.552	2.689
	a param. (poly.)	0.139	-0.074	0.139	-0.052	0.142	-0.073	0.133	-0.053
	b param. (poly.)	1,242	0.512	1,175	0.498	1.194	0.547	1.188	0.496
	Third Dimension								
Dich. common items	*θ*	0.040	-2.344	0.039	-2.269	0.040	-2.168	0.034	-2.177
	a param. (dich.)	0.077	-0.250	0.087	-0.354	0.070	-0.361	0.080	-0.345
	b param. (dich.)	0.670	3.127	0.759	3.537	0.717	3.346	0.731	3.412
	a param. (poly.)	0.142	-0.037	0.150	-0.060	0.145	-0.059	0.140	-0.058
	b param. (poly.)	1.620	0.771	1.643	0.835	1.678	0.774	1.615	0.836
Mixed format common items	*θ*	0.016	-0.762	0.016	0.779	0.016	-0.701	0.016	-0.764
	a param. (dich.)	0.046	0.137	0.049	0.081	0.043	0.137	0.051	0.079
	b param. (dich.)	0.422	2,055	0.459	2,237	0.413	2.015	0.460	2.241
	a param. (poly.)	0.160	0.175	0.157	0.171	0.165	0.170	0.153	0.171
	b param. (poly.)	1,326	0.512	1,243	0.498	1.251	0.547	1.275	0.496

Finally, when the interactions between the three conditions were examined, it was found that there was a meaningful three-way interaction for the error values of the *a* parameters of the dichotomous items ($p < .05$). The second and third dimensions showed similar results. When all the results were considered en bloc, it could be seen that, in addition to a clear effect of the common item format, the estimation method had effect, at least, for some dimensions. Although, some interactions were observed, they did not come close to providing a meaningful picture.

Table 2: Factorial ANOVA Results

			Dichotomous			Polytomous	
	Conditions Being Tested		***θ***	***a***	***b***	***a***	***B***
First Dimension	Common Item Format (CIF)	RMSE	298.756**	99.88**	354.555**	9.947**	64.491**
		Bias	790.483**	312.124**	267.519**	113.801**	150.291**
	Scale Shrinkage (SS)	RMSE	3.963*	7.323**	0.001	0.103	0.039
		Bias	0.265	1.214	0.004	0.959	0.157
	Estimation Method (EM)	RMSE	1.620	2.783	18.726**	2.654	11.586**
		Bias	0.001	3.800	18.940**	1.542	0.412
	CIF*SS	RMSE	1.748	0.828	0.208	0.010	0.515
		Bias	4.150*	1.068	0.211	0.487	0.090
	CIF*EM	RMSE	2.633	1.714	0.056	1.046	0.049
		RMSE	1.953	0.121	0.016	1.170	4.051*
	SS*EM	Bias	0.841	0.054	5.577*	0.419	1.581
		RMSE	13.896**	1.167	5.617*	0.436	0.165
	CIF*SS*EM	Bias	0.443	4.341*	0.076	1.407	2.811
		RMSE	1.567	1.320	0.049	1.097	0.135
Second Dimension	Common Item Format (CIF)	RMSE	114.524**	98.007**	17.307**	11.491**	0.023
		Bias	197.916**	333.092**	4.709*	254.163	150.291**
	Scale Shrinkage (SS)	RMSE	0.955	1.476	0.019	1.244	0.032
		Bias	0.836	1.101	0.025	0.039	0.157
	Estimation Method (EM)	RMSE	3.629	4.941*	6.300*	0.140	6.532
		Bias	5.209*	3.818	6.417*	2.802	0.412
	CIF*SS	RMSE	0.178	0.666	1.066	0.503	1.256
		Bias	6.971**	0.840	1.079	0.034	0.090
	CIF*EM	RMSE	2.369	7.065**	0.345	0.492	0.366
		RMSE	3.020	0.027	0.401	4.888	4.051*
	SS*EM	Bias	3.555	3.778	0.272	0.377	0.007
		RMSE	2.282	1.206	0.315	0.041	0.165
	CIF*SS*EM	Bias	0.074	0.636	1.990	0.391	2.481
		RMSE	0.098	1.334	2.037	0.013	0.135
Third Dimension	Common Item Format (CIF)	RMSE	321.824**	115.478**	584.166**	12.644**	312.86**
		Bias	938.236**	302.728**	488.666**	2667.632**	150.291**
	Scale Shrinkage (SS)	RMSE	7.121**	3.017	0.079	0.179	0.023
		Bias	2.648	1.098	0.067	2.223	0.157
	Estimation Method (EM)	RMSE	6.009*	3.205	16.128**	0.343	1.419
		Bias	0.135	4.129*	16.029**	2.305	0.412
	CIF*SS	RMSE	1.225	0.566	0.360	0.157	0.795
		Bias	2.108	0.998	0.350	0.738	0.090
	CIF*EM	RMSE	2.401	3.281	0.159	1.428	0.050
		RMSE	0.515	0.062	0.096	1.164	4.051*
	SS*EM	Bias	0.049	1.816	1.964	1.742	0.066
		RMSE	0.510	1.410	1.857	2.787	0.165
	CIF*SS*EM	Bias	1.243	2.002	3.193	0.055	5.479*
		RMSE	0.240	1.448	3.074	1.255	0.135

p<0.05; p<0.01***

DISCUSSION

The findings showed that the common item structure significantly affected the amount of errors and bias obtained from the vertical scaling process. Specifically, for mixed-format tests, when the common item set only contained dichotomous items, it caused higher the amount of error, with few exceptions. As stated by Kolen and Brennan (2004), the common set of items needs to be a "mini version" of the total test in terms of content and statistical properties. This means that when polytomous items are placed in the common item set, the common item set becomes more similar to the total and this positively affects the scaling results.

Another finding suggests that the estimation methods provide similar amount of error in almost all conditions. The fact that parameter estimates performed with MHRM and EM cycles have no effect on the results of the scaling in many cases can be explained by the three dimensional data structure preferred in this study. Cai (2008, 2010a) reported that MHRM estimates yield better results for datasets with five or more dimensions, and that these two estimation methods give very similar results for datasets with a lower number of dimensional structures. As stated by Cai (2010a), when EM cycles are used, the number of quadrants required for estimates increases geometrically as the number of dimensions is linearly increased. There is no such requirement for MHRM estimates and it becomes more advantageous to use MHRM for higher dimensional data. As a result, these two estimation methods yielded similar results in the case of the three dimensional data scenario used in this study.

One interesting finding is that the scaling results yielded a high amount of bias that could be explained by the one-point θ difference between the groups which could rarely be observed in real life test applications. An effect of higher ability difference on bias value is reported in the literature (Kim and Lee, 2006). Indeed, many studies have confirmed that bias values gets higher when the differences real and estimated abilities between upper and lower groups increases (Brennan and Kolen, 2008; Kirkpatrick, 2005; Wang, Lee, Wu, Huang, Hu and Harris, 2009; Kim and Lee, 2006). On the other hand, Cao (2008) and Kirkpatrick (2005) stated that the effect of ability differences on measurement results is valid only in the context of the IRT. Furthermore, Hagge (2010) stated that when the difference of abilitiy among the groups is high, the bias value may be relatively low where the correlation between polytomous and dichotomous items is too high.

In the light of these findings, it is suggested that test developers should prefer that common item sets contain mixed-format items when vertical scaling is performed even if this involves some difficulties in practice, such as higher cost and a limited number of available polytomously scored items. Moreover, since this study was conducted by using simulation data, caution should be taken when making generalizations for testing applications. In future studies, it is suggested that the current study be replicated using real data.

REFERENCES

Baker, Frank B. (1992). *Item response theory: parameter estimation techniques*. New York: Marcel Dekker, Inc.

Bastari, B. (2000). *Linking multiple-choice and constructed-response items to a common proficiency scale* (Doctoral dissertation, University of Massachusetts Amherst). Retrieved from https://scholarworks.umass.edu/dissertations/AAI9960735

Birnbaum, A. (1968). Some latent trait models and their use in ınferring an examinee's ability. In FM Lord, MR Novick (eds.), *Statistical theories of mental test scores*, ss. 397-479. Addison-Wesley, Reading, MA.

Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459

Cai, L. (2008). *A Metropolis-Hastings Robbins-Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model (Doctoral dissertation*). Retrieved from https://cdr.lib.unc.edu.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, 75, 33-57.

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement*, 32, 79-96.

Cao, Y. (2008). *Mixed format test equating: Effects of test dimensionality and common-item sets* (Doctoral dissertation). Retrieved from https://drum.lib.umd.edu.

Cao, Y., Yin, P., & Gao, X. (2007). *Comparison of IRT and classical equating methods for tests consisting of polytomously-scored items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Chalmers, R. P.(2012). mirt: A multidimensional ıtem response theory package for the r environment. *Journal of Statistical Software, 48*(6), 1-29.

Dodd, B. G. (1984). *Attitude scaling: A comparison of the graded response and partial credit latent trait models* (Doctoral dissertation, University of Texas at Austin). Retrieved from https://elibrary.ru/item.asp?id=7426395.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*(3), 144–149.
Hagge, S. L. (2010). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups* (Doctoral dissertation), Retrieved from https://ir.uiowa.edu/etd/680/.
Han, K. T., & Paek I. (2014). A review of commercial software packages for multidimensional IRT modeling. *Appl. Psychol. Meas*. 38, 486–498.
Kim, J. (2007). "*A comparison of calibration methods and proficiency estimators for creating IRT vertical scales*." PhD (Doctor of Philosophy) thesis, University of Iowa.
Kim, S., & Lee W. (2006). An Extension of Four IRT Linking Methods for Mixed-Format Tests. *Journal of Educational Measurement*, *43*(1), 53-76.
Kim, S. H., & Cohen, A.S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, *26*(1), 25–41.
Kirkpatrick, R. K. (2005). *The effects of item format in common item equating* (Doctoral dissertation, University of Iowa). Available from ProQuest Dissertations and Theses database. (UMI No. 3184724)
Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. NY: Springer.
Koretz, D.M., & Hamilton, L.S. (2006). Testing for accountability in K-12. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., 531-578). Westport, CT: American Council on Education and Praeger Publishers.
Kuo T-C & Sheng Y (2016) A comparison of estimation methods for a multi-unidimensional graded response ırt model. *Front. Psychol*. 7, 880.
Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service.
Marco, G. L., (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139-160.
Muraki, E. & Carlson, E. B. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*. 19, 73-90.
R Core Team (2015). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria.
Reckase, Mark D. (2009). *Multidimensional item response theory*. New York: Springer.
Rosa, K., Swygert, K., Nelson, L. & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed response items: Scale scores for patterns of summed scores. D. Thissen and H. Wainer (Eds.), *Test scoring* (pp. 253-292). Hillsdale, NJ: Lawrence Erlbaum.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychomtrika Monograph*, No. 17.
Samejima, F. (1972). A general model for free-response data. *Psychometric Monograph*, No. 18.
Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement*. 37, 329-346.
Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design. *Applied Psychological Measurement, 32*(8), 632-651.
Weeks, J. P. (2010) plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software*, *35*(12), 1–33
Yao, L. ve Boughton, K. (2009). Multidimensional linking for tests with mixed ıtem types. *Journal of Educational Measurement*. *46*(2), 177–197.
Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika*, 50, 399-410.
Zhang, M. (2016). *Exploring dimensionality of scores for mixed-format tests* (Doctoral dissertation). Retrieved from https://ir.uiowa.edu/cgi/viewcontent.cgi?article=6628&context=etd

Çok Boyutlu Karışık Format Testlerinin Dikey Ölçeklemesini Etkileyen Faktörlerin İncelenmesi

Giriş

Testlerden elde edilen puanlar birçok başlık altında alınan önemli kararlar için temel bilgi kaynakları arasındadır. Alınacak önemli kararlardan bağımsız olarak, test puanlarının mümkün olan en kesin

bilgiyi sunması gerekmektedir. Daha kesin bilgi daha iyi kararların alınabilmesi için önemlidir. Bununla birlikte uygulamada test güvenliği ve öğrenci gelişiminin takip edilebilmesi gibi birtakım gerekçeler yüzünden aynı testin farklı formları kullanılmakta veya farklı zamanlarda uygulanan testlerde ortak maddeler kullanılarak testler ölçeklenmektedir. Farklı formlardan elde edilen puanlar daha sonrasında eşitlenmekte ya da ölçeklenmektedir. Bu işlemin hatasız olması gerçekleştirilen sınavların daha adil olması ve öğrencilerin geleceği ile ilgili doğru kararlar verebilmek için önemlidir. Buna göre, puanları önemli kararlar için kullanılan testlere uygulanan dikey ölçekleme yöntemlerinin psikometrik olarak savunulabilir olması önemlidir. Bu sebepten dolayı ölçekleme gerçekleştirilirken uygulayıcıların kararlarını dayandıracakları kuramsal çalışmalar büyük önem taşımaktadır. Bu sebepten dolayı farklı yöntemlerin karşılaştırılması ve farklı durumlar için en az hata veren yöntemlerin belirlenmesi gerekmektedir.

İki kategorili ve çok kategorili olarak puanlanan maddelerin birlikte yer aldığı karma format testlerin kullanımı gün geçtikçe artmaktadır. Benzer şekilde, büyük ölçekli ve öğrencilerle ilgili önemli kararların alındığı test uygulamalarında birden fazla formunun kullanımı da benzer şekilde yaygınlaşma eğilimindedir. Farklı test formlarından elde edilen puanların karşılaştırılabilir olabilmesi için bu formlar arasında fonksiyonel bir bağ oluşturulması gerekmektedir. Eğer kurulan bu bağ farklı sınıf (ya da test güçlüğü farklılaşan) formlar arasında gerçekleştirilirse, bu işlem dikey ölçekleme olarak adlandırılmaktadır. Dikey ölçeklemede farklı test formları birbirlerine bağlandığı için eşitleme ile benzerdir. Fakat test formları içerik ve güçlük olarak farklıdır çünkü formlar sınıflar arası ya da yaşa bağlı olarak ilerlemeyi yansıtmaktadırlar. Bundan dolayı, dikey ölçekleme farklı test formlarının karşılaştırılması için kullanılmakla birlikte her bir seviyedeki puanlar birbirlerinin yerine kullanılamazlar. Test ölçeklemesinde temel amaç farklı seviyelerdeki puanların karşılaştırılmasıdır. Seviye farklılığı bir öğrencinin bulunduğu sınıf, eğitim öğretim yılının bulunduğu aşama ya da yaştan kaynaklanabilir. Dikey ölçekleme genellikle aynı bireylerin farklı seviyelerde elde ettikleri puanların farklı zamanlara göre karşılaştırılabilmesi için kullanılmaktadır. Bu tür desenler ise DOGOM (Denk Olmayan Gruplarda Ortak Madde) deseni olarak adlandırılmaktadır.

Bu çalışma kapsamında karma format maddelerden oluşan boyutlu testler DOGOM deseni kullanılarak ölçeklendiğinde ortak madde setinin yapısı (yalnızca iki kategorili maddelerden oluşan ortak madde seti - iki ve çok kategorili maddelerin yer aldığı ortak madde seti), yetenek daralması (üst yetenek grubunda yetenek varyansının daralması - varyansın eşit kalması) ve parametre kestirim yöntemlerinin (EM - MHRM) ölçekleme sonuçları üzerindeki etkisi incelenmiştir. Ayrıca bu koşulların etkileşim içinde olup olmadığına bakılmıştır.

Yöntem

Çalışma, türetilmiş veriler kullanılarak gerçekleştirilmiştir. Ölçeklemenin niteliğinin değerlendirilmesinde ölçme hatası ve yanlılık değerleri kullanılmıştır. Veriler türetilirken yanıt matrisleri, içerisinde İKM (iki kategorili madde) ve ÇKM(çok kategorili madde)'ler yer alacak şekilde oluşturulmuştur. İKM'ler için parametre kestirimi 3 parametreli modele (3PLM) göre, ÇKM'ler için ise aşamalı tepki modeline (ATM) göre gerçekleştirilmiştir. Veri türetme ve analizi sürecinde gerçekleştirilen işlem 50 defa tekrarlanmıştır. Ayrıca, araştırmada gerçekleştirilen veri türetme, testlerin kalibrasyonu ve ölçekleme işlemleri için R programı kullanılmıştır. Etkileşimleri incelemek için kullanılan iki ve üç yönlü analizler SPSS ile gerçekleştirilmiştir.

Bulgular ve Tartışma

Araştırmada sonucunda ortak madde yapısının ölçekleme işlemi sonucunda ortaya çıkan hata ve yanlılık miktarını önemli ölçüde etkilediği görülmüştür. Buna göre karma format testlerde ortak madde setinin sadece iki kategorili puanlanan maddelerden oluşması ölçekleme hatasını bazı istisnalar haricinde arttırmaktadır. Elde edilen bu bulgu, diğer koşullardan bağımsız olarak tutarlı bir şekilde gözlenmiştir.

Varyans daralmasının etkisi incelendiğinde yetenek parametresi ve çok kategorili puanlanan maddelerie ait *a* parametreleri için farklılaşmalar olduğu görülmüştür. Gözlenen bu farklılaşmalar yanlılık değerlerine aittir. Çok kategorili puanlanan maddelere ait *a* parametreleri için ise hata değerlerinde farklılaşmalar olduğu bulunmuştur. Her iki parametre için varyansın azaldığı durumda daha iyi sonuçlar elde edildiği görülmüştür.

Kullanılan kestirim yönteminin etkisi incelendiğinde ise bazı boyutlar için yanlılık değerlerinin Metropolis–Hastings Robbins-Monro kestirim yöntemi için daha az olduğu görülmüştür. Ayrıca iki kategorili puanlanan maddelerin *a* ve *b* parametreleri ve çok kategorili puanlanan maddelerin eşik parametreleri için bazı durumlarda kestirim yönteminin hata ve yanlılık değerlerini etkilediği görülmüştür. Çok kategorili puanlanan maddelerin *a* parametresinin ise kestirim yönteminden etkilenmediği görülmüştür.

Son olarak, etkileşimler incelenmiştir. Buna göre, yetenek parametresi bazı koşullara göre yanlılık değerlerinin ikişerli ve üçerli etkileşimler gösterdiği bulunmuştur. İki kategorili maddelere ait *a* ve *b* parametreleri için bakıldığında *b* parametresine ait hata ve yanlılık değerlerinde testin bazı boyutlarında varyans daralması ve kestirim yönteminin etkileşim içinde oldukları görülmüştür. İki kategorili puanlanan maddelere ait *a* parametrelerine ait hata değerleri için birinci boyutunda üç koşulun etkileşim içinde olduğu bulunmuştur. Ayrıca, çok kategorili puanlanan maddelere ait *a* parametreleri ile eşik parametreleri için etkileşim gözlenmemiştir. Üç boyutun tamamı için ortak madde yapısı ve kestirim yöntemi koşulları arasında etkileşim olduğu görülmüştür.

Sonuç olarak, etkisi incelenen koşullar içinde ölçekleme sonuçları üzerinde en fazla etkisi olan koşulun ortak madde yapısı olduğu sonucuna varılmıştır.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(4);339-355

Self-regulated Learning Skills: Adaptation of Scale*

Şenol ŞEN ** Ayhan YILMAZ *** Ömer GEBAN ****

Abstract
The aim of this study is to adapt and examine the psychometric properties of Achievement Goal Scale (AGS) originally constructed by Elliot and McGregor (2001) and the Motivated Strategies for Learning Scale (MSLQ) originally constructed by Pintrich, Smith, Garcia and McKeachie (1991) in order to measure the self-regulated learning skills of high school students' in a chemistry course. The study group was comprised of 862 high school students attending a chemistry course in different public schools. The construct validity of the sub-scales included in the scales were tested by confirmatory factor analysis. For the reliability studies, the internal consistency coefficient Cronbach's alpha (α) values as well as McDonald's ω (omega) coefficients were calculated. In addition, item-total correlations were calculated for the reliability of each item in the scales. When the confirmatory factor analysis results were examined, it was accepted that the fit indices met the goodness of fit criteria for both the Achievement Goal Scale and Motivated Strategies for Learning Scale. Factor loadings of the items in both scales were statistically significant. These results showed that the Turkish forms of both scales have enough psychometric properties in terms of validity and reliability for a chemistry course.

Key Words: Adaptation of scale, chemistry, high school, self-regulation.

INTRODUCTION

Self-regulation is a cyclic process that individuals monitor their own behaviours; make a judgement by comparing based on their own criteria and regulate their behaviours. Self-regulated individuals affect, lead and control their own behaviours (Bandura; as cited in Senemoğlu, 2011). According to Zimmerman (2000) self-regulation is the thoughts, feelings and behaviours which individuals develop to achieve their goals and which emerge cyclically. Social cognitive theory contends that self-regulation develops in social environments and is internalised by individuals through time. According to the theory, self-regulation includes cognitive, metacognitive and motivational components in its structure (Zimmerman; as cited in Sakız & Yetkin Özdemir, 2014). Therefore, self-regulated students take on metacognitively, motivationally and behaviourally active roles in the process of learning, they set their own learning goals and they control this process (Zimmerman, 1989). Self-regulation is not defined as a mental ability or as an academic skill but rather as a self-directive process in which learners adapt their cognitive competencies in the form of academic abilities (Zimmerman, 2002).

Most of the learning models which have been developed by researchers in the field of education and which are based on self-regulation are based on Zimmerman's (1989) cyclical model. Pintrich's self-regulation model, one of those models, was developed in the context of Social Cognitive Theory. Motivational components play important roles in this model (Pintrich, 1999; 2000a; Pintrich & DeGroot, 1990; Pintrich et al., 1991; 1993). The feature of this model suggested by Pintrich is that it reflects a social cognitive perspective and that it includes motivational processes; because if students are not motivated to use their cognitive and metacognitive skills, these skills are not important

* The present study is a part of PhD Thesis entitled "Investigation of Students' Conceptual Understanding of Electrochemistry and Self-Regulated Learning Skills in Process Oriented Guided Inquiry Learning Environment" completed within Hacettepe University Graduate School of Educational Sciences. This study was supported by Research Fund of Hacettepe University.
** Assoc. Prof. Dr., Hacettepe University, Faculty of Education, Ankara, Turkey, e-mail: schenolschen@gmail.com, ORCID ID: http://orcid.org/0000-0003-3831-3953
*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara, Turkey, e-mail: ayhany@hacettepe.edu.tr, ORCID ID: http://orcid.org/0000-0003-4252-5510
**** Prof. Dr., Middle East Technical University, Faculty of Education, Ankara, Turkey, e-mail: geban@metu.edu.tr, ORCID ID: https://orcid.org/0000-0002-9433-0056

To cite this article:
Şen,, Ş., Yılmaz, A., & Geban, Ö. (2018). Self-regulated learning skills: adaptation of scale. *Journal of Measurement and Evaluation in Education and Psychology, 9*(4), 339-355. DOI: 10.21031/epod.439039

Received: 29.06.2018
Accepted: 13.10.2018

(McCoach & Siegle, 2003; Pintrich & DeGroot, 1990). According to Pintrich (2000a), self-regulation is a process in which learners set goals for themselves, follow them and try to organise their motivation, cognition and behaviours. This process is determined, organised and restricted by learners' goals and by the contextual properties of the environment they are in. Pintrich stresses that self-regulated learning is the learning actualised to develop self-efficacy and states that self-efficacy in addition to motivation is an important component of self-regulation (as cited in Sarı & Akınoğlu, 2009). Garcia and Pintrich also claim that motivation, an important component of self-regulation, is composed of individuals' beliefs about themselves such as personal goals, self-efficacy and value beliefs in addition to their perceptions about the classroom (as cited in Özturan, Sağırlı, Çiltaş, Azapağası & Zehir, 2010).

With the emergence of self-regulation models based on social cognitive theory, the notion of the importance of context in self-regulation processes has emerged. Context can be defined as the circumstances creating an environment for a situation, an idea or an event (Context, 2018). With the emergence of the idea that context can influence the validity of findings, measurements were made sensitive to the context. Thus, measurements for different domains of learning gained more and more importance (Pintrich; as cited in Özbay, 2008). Briefly, measurements sensitive to the context and directed to specific areas of learning and specific tasks instead of measurements based on generalisations became more important. Motivated Strategies for Learning Scale (MSLS) developed by Pintrich et. al (1991) is frequently used in the literature. Pintrich et al (1991) chose a course for university students as the unit of analysis in the scale (Özbay, 2008). MSLS was developed on the basis of the view that context had significant effects on the use of motivation and learning strategies and that different strategies should be used in different areas and tasks of learning (Özbay, 2008). MSLS contains five sub-dimensions as the indicators of students' cognitive regulation. They are labelled as rehearsal, elaboration, organisation, critical thinking and metacognitive self-regulation. There are some sub-dimensions on which cognition control activities and monitor measurements in the framework of self-regulated learning model suggested by Pintrich (2000a) and some performance control activities in the framework of the model suggested by Zimmerman (2000) are included. MSLS does not contain sub-dimensions for measuring motivational strategies related to organising motivation and feelings. Yet, there are sub-dimensions such as achievement goals containing performance and mastery, task value, self-efficacy for learning and performance and test anxiety at the forethought stage of Zimmerman's model. In relation to organising behaviours, MSLS includes three sub-dimensions. They are the sub-dimensions of effort regulation, time and study environment management and help seeking. Indeed, two self-regulation models which were developed by Zimmerman and Pintrich and which were based on social cognitive theory lay emphasis on such self-regulation strategies as performance control, time management, help seeking and environmental configuration. Lastly, MSLS contains two more sub-dimensions related to organising the context. They are called peer learning and time and study environment management. They are used to find how well students use their friends as sources of learning and how well they manage their study environment and time (Yumuşak, Sungur, & Çakıroğlu, 2007).

Achievement goals included in MSLS influence learners' task determination and problem-solving efforts in addition to their study behaviours and recalling. According to Bandura, individuals' setting goals can cause increase in their motivation (as cited in Driscoll, 2005). When individuals set their goals, they evaluate their performance and their level intrinsically and they decide on the basis of extrinsic criteria. If they cannot attain such a standard, they will insist on their efforts. However, all these goals will not maintain this insistence. Goals set should have certain properties for this. Motivated Strategies for Learning Scale has two types of achievement goals labelled as mastery goals and performance goals. Yet, performance goals are divided into two as performance approach and performance avoidance in the literature (Elliot & Church, 1997; Skaalvik; as cited in Şenler, 2011). In later studies, however, mastery goals are divided into two as mastery approach and mastery avoidance in a similar vein (Elliot, 1999; Pintrich, 2000b). While performance approach goals involve such goals as doing better than others do and being the best, performance avoidance goals involve such goals as avoiding being ordinary. Mastery approach goals aim to learn and understand in depth whereas mastery avoidance goals emphasise not learning and misunderstanding (Elliot & Church, 1997; Elliot & McGregor, 2001; Elliot & Reis, 2003). Therefore, the need for using sub-dimensions for mastery

goals and performance goals available in MSLS arises. Achievement goal Scale (AGS) can be used in analysing mastery goals in MSLS as mastery approach goals and mastery avoidance goals and performance goals as performance approach goals and performance avoidance goals in four parts. Thus, Achievement Goal Scale has four components: mastery approach goals, performance approach goals, mastery avoidance goals and performance avoidance goals (Elliot & McGregor, 2001). The other items in the scale are not the items for goal orientation. However, the researchers developing the scale recommend that these items be included and implemented in the scale although they are not used.

Purpose of the Study

The need for making measurements sensitive to the context emerges since context influences the validity of findings. For this reason, measurements directed to different areas of learning have been gaining more and more importance (Pintrich; as cited in Özbay, 2008). Yet, it was found in studies that there were no reliable and valid scales for determining high school students' self-regulated learning skills in different courses. Therefore, scales are needed for primarily use in assessment so as to develop students' self-regulated learning skills in chemistry course. Besides, the fact that achievement goals available in MSLS are limited to two goals in the literature made it necessary to use MSLS along with AGS. Therefore, the two scales should be adapted and validity and reliability of the scales should be examined. In line with this need, this study adapts MSLS and AGS into chemistry course and analyses the psychometric properties to determine high school students' self-regulated learning skills.

METHOD

Research model

This study employs survey model. Survey model is a research approach aiming to describe a situation existed in the past or existing at present as it is. When it is impossible to reach the population, study can be conducted with a small sample taken from the population in survey studies (Fraenkel & Wallen, 2000).

Participants

A total of 862 high school students who were the 9^{th}, 10^{th}, 11^{th} and 12^{th} graders in differing state schools in Ankara were included in the study. % 35.03 of the participants were female whereas 33.06% were male. In addition to that, 31.9% of the participants did not make any coding for gender. The participants' age ranged between 16 and 20.

Data Collection Instruments

Achievement Goal Scale (AGS)

Achievement Goal Scale (AGS) was developed by Elliot and McGregor (2001) and was adapted into Turkish by Şenler and Sungur (2007). The scale was adapted by Şenler and Sungur (2007) into science course and it was administered to primary school students. The 7-pointed Likert type scale was changed into 5-pointed Likert type. This study, on the other hand, adapts the scale into chemistry course for high school students using 7-pointed Likert type as in the original version by getting permission. The scale was administered to 862 students in total.

The scale has four sub-factors. The factor of mastery approach goals included items 1, 6 and 8; the factor of performance approach goals included items 4, 10 and 16; the factor of mastery avoidance goals included items 11, 14 and 17 and the factor of performance avoidance goals included items 2, 7, 13, 19, 20 and 21 (Elliot & McGregor, 2001). The other items included in the 21-item scale were not related to goal orientation. Yet, the researchers who had developed the scale recommended that these

items be included in the scale and be implemented although they were not used. Thus, items 15 and 18 available in the scale were in the factor of competence expectancies (Elliot & Church, 1997) and items 3, 5, 9 and 12 were in the factor of challenge and threat appraisals (Elliot & Reis, 2003).

Motivated Strategies for Learning Scale (MSLS)

Motivated Strategies for Learning Scale (MSLS) was developed by Pintrich, Smith and McKeachie (1991) so as to be informed of university students' motivation in classes and of the learning strategies they used in those classes. The scale was adapted into Turkish by Büyüköztürk, Akgün, Özkahveci and Demirel (2004) and Sungur (2004). It is a 7-pointed Likert type scale. It has two main components called motivation and learning strategies. The motivation component is composed of six sub-factors. These are intrinsic goal orientation, extrinsic goal orientation, task value, control of learning beliefs, self-efficacy for learning and performance and text anxiety. The learning strategies part is related to different cognitive and metacognitive strategies students use and contains 31 items. In addition to the 31 items, there are also 19 items related to different resource management strategies. Learning strategies part includes nine sub-factors labelled as rehearsal, organization, elaboration, critical thinking, metacognitive self-regulation, time and study environment, effort regulation, peer learning and help seeking. High scores received from any factor in MSLS indicate that students have high levels of the property related to the factor (Pintrich et al., 1991; Büyüköztürk et al., 2004). Having received the necessary permission, the scale was adapted for use with chemistry course with high school students, and thus it was administered to 862 students.

Data Analysis

Prior to analyzing the data, the items which were stated negatively in the original version of the scale were coded inversely. First order confirmatory factor analysis was conducted for construct validity to see whether or not Achievement goal scale and Motivated Strategies for Learning Scale measured the intended structure. Because factor loadings were not equal, both Cronbach's alpha (α) and McDonald Omega (α) reliability coefficients were calculated so as to determine reliability in the sense of internal consistency. In this study, LISREL software was used for confirmatory factor analysis and SPSS and Excel software packages were used for reliability analyses.

Language Validity

Turkish adaptations of MSLS from English made earlier (Büyüköztürk et al. 2004; Sungur, 2004; Taştan, 2009; Yalçınkaya, 2010) and AQS study in Turkey (Şenler and Sungur) were examined in this study and expert opinion was consulted for the translated items which were determined. Efforts were made to see whether or not the translated items were equivalent to the original items and to see the degree to which the items in the Turkish version were compatible with Turkish grammar and were intelligible. After expert opinion was obtained, modifications were made, the resultant form was administered to a group of high school students having similarities with the students with whom the application would be done. The items of the revised version were checked in terms of content, and the language of the form was modified based on students' feedback.

RESULTS

This section presents the findings obtained from the analyses done for validity and reliability of both scales. The results for confirmatory factor analysis conducted for structure validity of the scales are shown in Table 1.

Table 1. The Fit Indices for the Achievement Goal Scale

N	χ^2/df (<3.0)	RMSEA (<.08)	CFI (>.95)	IFI (>.90)	GFI (>.90)	NFI (>.90)	AGFI (>.85)	NNFI (>.95)	SRMR (<.1)
862	5.26	.070	.98	.98	.94	.97	.91	.97	.042

When the fit indices of the Achievement Goal Scale were examined in Table 1 and Figure 1, it was concluded that the values apart from Chi square/df (5.26)- which were fit indices- met the criterion for good fit (χ^2/df < 3.0; RMSEA<.08; CFI>.95; IFI >.90; GFI>.90; NFI >.90; AGFI >.85; NNFI >.95; SRMR <.1) (Çelik & Yılmaz, 2013; Schermelleh-Engel, Moosbrugger, & Müller, 2003).

Table 2. Reliability Analysis Results for the Achievement Goal Scale

Subscales	Item No	λx	δ	t	R^2	Item Total Cor.	α	ω
Mastery approach goals	m1	.74	.46	24.24	.55	.67	.85	.84
	m6	.84	.29	29.48	.71	.71		
	m8	.84	.29	29.45	.71	.70		
Mastery avoidance goals	m11	.72	.47	23.60	.52	.64	.79	.79
	m14	.87	.24	3.55	.76	.77		
	m17	.64	.59	2.03	.41	.59		
Performance approach goals	m4	.84	.29	29.30	.71	.77	.77	.78
	m10	.80	.36	27.44	.64	.73		
	m16	.55	.70	16.61	.30	.51		
Performance avoidance goals	m2	.38	.86	1.85	.14	.36	.67	.67
	m7	.48	.77	14.08	.23	.45		
	m13	.69	.53	21.70	.48	.61		
	m19	.71	.50	22.49	.50	.67		
	m20	.22	.95	6.32	.05	.22		
	m21	.49	.76	14.59	.24	.46		

Figure 1 and Table 2 show the variance values described with *t* values which were found to be significant for each item, the factor loadings (λx) and error variances (δ). Accordingly, it was found that factor loadings in the sub-factors were found to range between .22 and .87. to perform the reliability analysis for the scale, McDonald's coefficient (omega)- which is recommended when the factors loading in each factor were not equal- in addition to Cronbach's alpha values was also found (Zinbarg, Revelle, Yovel & Li, 2005). Cronbach's alpha reliability coefficients were found to range between .67 and .85. In addition to reliability analyses, total item correlation suggesting the consistency of each item with the whole factor in which the item belonged was also analysed. It was found in consequence that only the total correlation for item 20 was smaller than .30 yet, some studies in the literature (Briggs & Cheek, 1986, for instance) point out that item correlation coefficient in .15-.50 interval would be sufficient for scales measuring more comprehensive properties. Clark and Watson (1995), on the other hand, state that the values between .15 and .20 would be adequate for total item correlation in scales measuring more comprehensive properties. Since AGS measured the different properties of both mastery and performance, decision was made to include this item in the study.

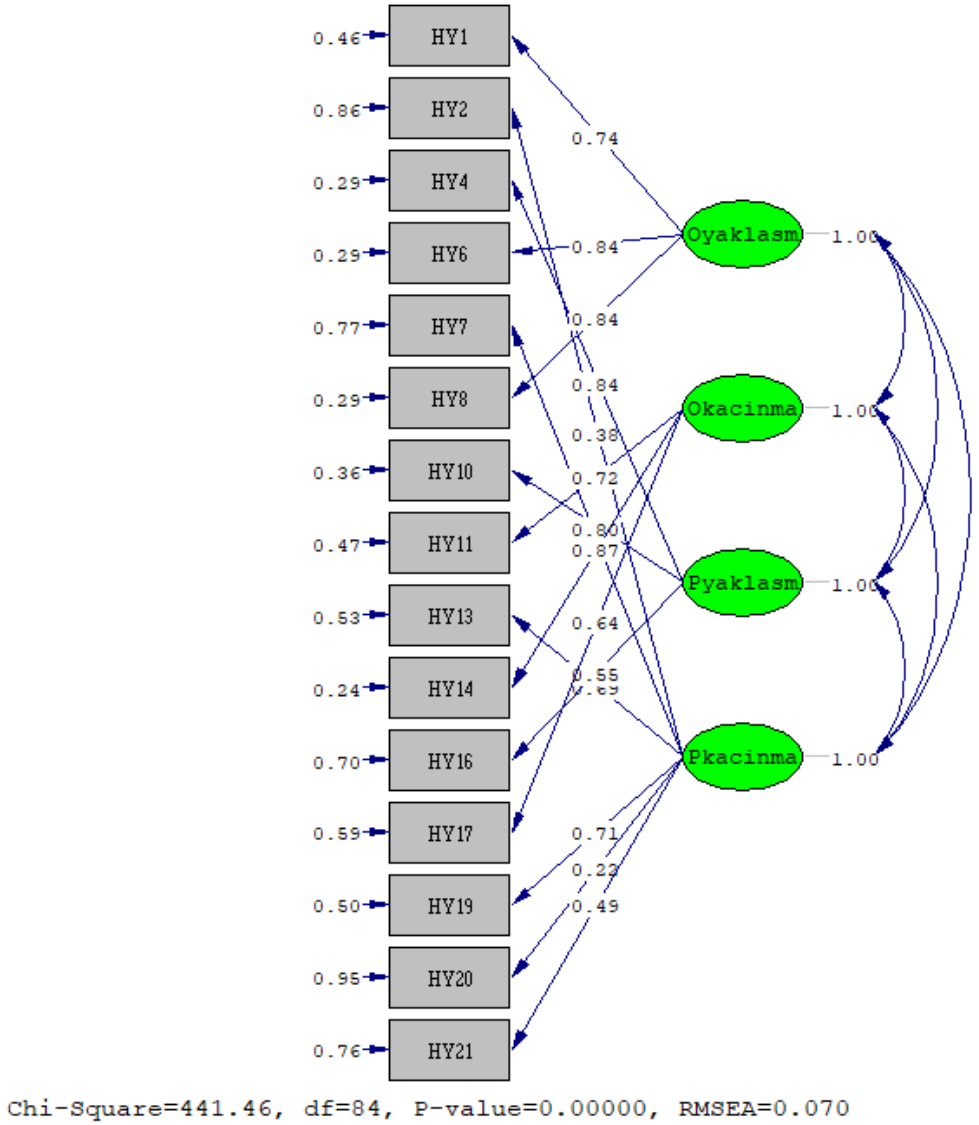


Figure 1. Path Diagram and Factor Loadings for the Achievement Goal Scale

Table 3. The Fit Indices for MSLS Motivation Section

N	χ^2/df (<3.0)	RMSEA (<.08)	GFI (>.90)	NNFI (>.95)	NFI (>.90)	CFI (>.95)	AGFI (>.85)	IFI (>.90)	SRMR (<.1)
862	5.22	.070	.89	.97	.97	.97	.87	.97	.044

Following the confirmatory factor analysis conducted for the motivation section of Motivated Strategies for Learning Scale, it was concluded that the values apart from Chi square/df (5.22)- which were fit indices- met the criterion for good fit (See Table 3 and Figure 2). (χ^2/df < 3.0; RMSEA<.08; CFI>.95; IFI >.90; GFI>.90; NFI >.90; AGFI >.85; NNFI >.95; SRMR <.1) (Çelik & Yılmaz, 2013; Schermelleh-Engel, Moosbrugger, & Müller, 2003).

Table 4. Reliability Analysis Results for MSLS Motivation Section

Subscales	Item No	λx	δ	t	R^2	Item Total Cor.	α	ω
Task Value	4	.58	.67	17.80	.34	.55	.85	.85
	10	.68	.53	21.99	.46	.65		
	17	.67	.55	21.58	.45	.62		
	23	.79	.38	26.98	.62	.72		
	26	.71	.50	23.12	.50	.65		
	27	.77	.40	26.08	.59	.71		
Control of Learning Beliefs	2	.67	.55	21.41	.45	.63	.73	.73
	9	.47	.78	13.90	.22	.46		
	18	.80	.37	26.67	.64	.73		
	25	.59	.65	18.17	.35	.56		
Self-efficacy for Learning and Performance	5	.53	.72	16.20	.28	.51	.87	.87
	6	.52	.73	15.74	.27	.50		
	12	.64	.59	2.53	.41	.61		
	15	.62	.61	19.66	.38	.57		
	20	.77	.41	26.04	.59	.71		
	21	.80	.36	27.56	.64	.73		
	29	.75	.43	25.31	.56	.71		
	31	.78	.40	26.56	.61	.74		
Text Anxiety	3	.42	.83	12.10	.18	.40	.61	.60
	8	.37	.86	1.75	.14	.35		
	14	.65	.58	19.64	.42	.58		
	19	.59	.65	17.71	.35	.56		
	28	.37	.86	1.62	.14	.35		

Figure 2 and Table 4 show the variance values described with t values which were found to be significant for each item, the factor loadings (λx) and error variances (δ). Accordingly, it was found that factor loadings in the sub-factors were found to range between .37 and .80. Cronbach's alpha reliability coefficients were found to range between .61 and .87.

Table 5. Fit Indices for MSLS Learning Strategies Section

N	χ^2/df	RMSEA	GFI	NFI	CFI	IFI	AGFI	NNFI	SRMR
862	3.99	0.059	.83	.94	.95	.95	.81	.95	.079

Following the confirmatory factor analysis conducted for the learning strategies section of Motivated Strategies for Learning Scale, it was concluded that the values apart from χ^2/df (3.99), GFI (.83) and AGFI (.81) - which were fit indices- met the criterion for good fit. (See Table 5 and Figure 3). χ^2/df (3.99), GFI (.83) and AGFI (.81) (Table 5 and Figure 3). (χ^2/df < 3.0; RMSEA<.08; CFI>.95; IFI >.90; GFI>.90; NFI >.90; AGFI >.85; NNFI >.95; SRMR <.1).

Figure 3 and Table 6 show the variance values described with t values which were found to be significant for each item, the factor loadings (λx) and error variances (δ) for MSLS learning strategies section. Accordingly, it was found that factor loadings in the sub-factors were found to range between .22 and .74. Cronbach's alpha reliability coefficients were found to range between .59 and .84.

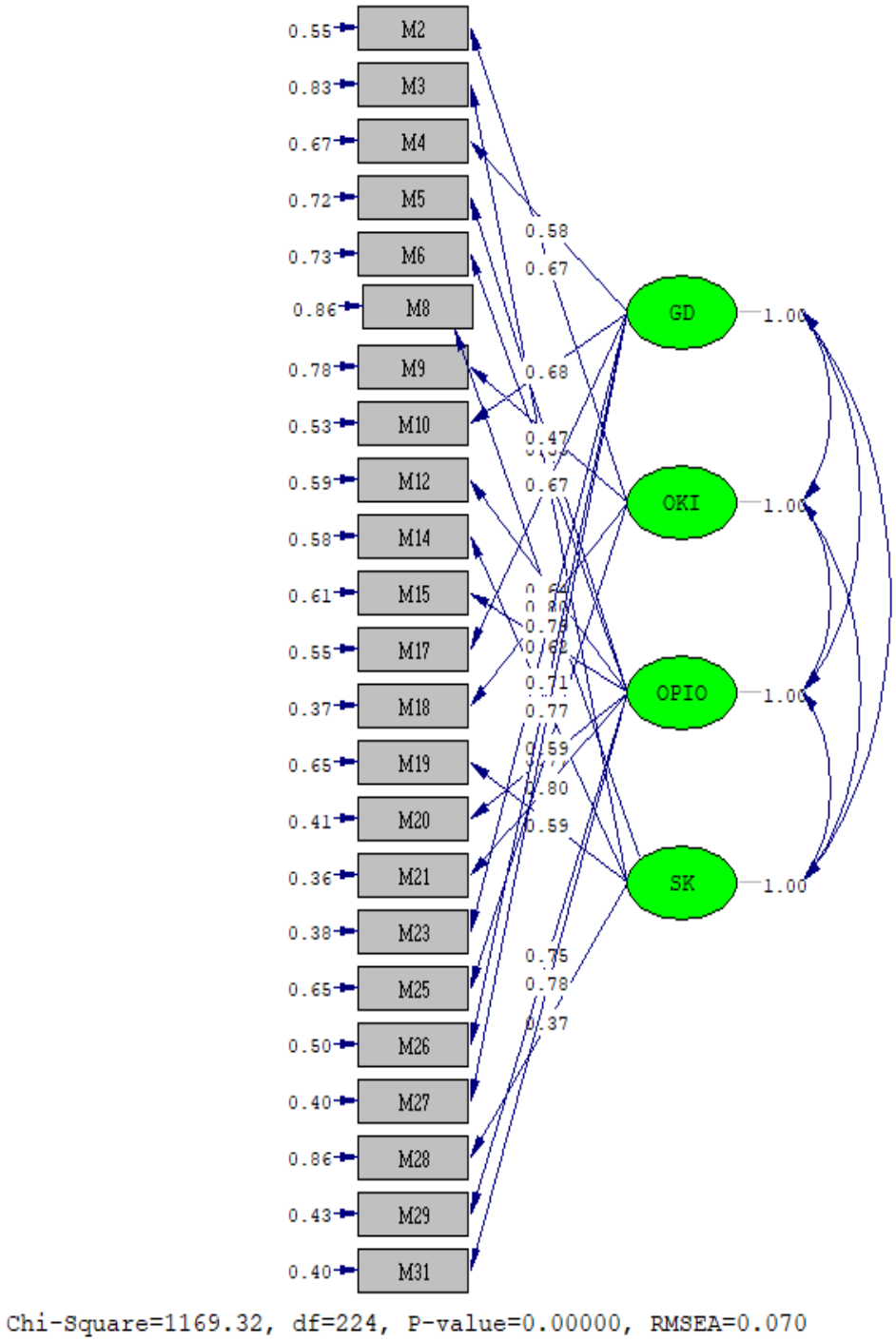


Figure 2. Path Diagram and Factor Loadings for MSLS Motivation Section

Table 6. Reliability Analysis Results for MSLS Learning Strategies Section

Subscales	Item No	λx	δ	t	R^2	Item Total Cor.	α	ω
Rehearsal	39	.67	.55	21.41	.45	.50	.76	.76
	46	.71	.49	23.31	.50	.53		
	59	.66	.56	21.24	.44	.57		
	72	.62	.62	19.35	.38	.53		
Organization	32	.59	.65	16.21	.35	.34	.68	.68
	42	.65	.57	18.23	.42	.43		
	49	.56	.69	15.25	.31	.46		
	63	.56	.69	15.15	.31	.39		
Elaboration	53	.69	.53	22.10	.48	.57	.78	.78
	62	.67	.55	21.49	.45	.57		
	64	.55	.69	16.96	.30	.44		
	67	.70	.52	22.56	.49	.51		
	69	.74	.45	24.68	.55	.55		
	81	.29	.92	8.26	.08	.24		
Critical Thinking	38	.49	.76	14.42	.24	.42	.76	.76
	47	.64	.59	19.62	.41	.52		
	51	.64	.59	19.79	.41	.50		
	66	.68	.54	21.20	.46	.51		
	71	.66	.56	2.67	.44	.48		
Help Seeking	40	.22	.95	5.80	.05	.38	.59	.59
	58	.65	.58	18.07	.42	.28		
	68	.60	.65	16.53	.36	.27		
	75	.57	.67	15.84	.32	.18		
Peer Learning	34	.66	.56	21.31	.44	.53	.71	.71
	45	.65	.58	2.63	.42	.49		
	50	.69	.52	22.37	.48	.57		
Metacognitive Self-regulation	33	.25	.94	7.11	.06	.43	.84	.85
	36	.62	.61	19.43	.38	.47		
	41	.56	.68	17.19	.31	.42		
	44	.57	.67	17.54	.32	.47		
	54	.59	.65	18.34	.35	.46		
	55	.65	.58	2.55	.42	.45		
	56	.59	.65	18.34	.35	.40		
	57	.30	.91	8.54	.09	.45		
	61	.57	.68	17.24	.32	.52		
	76	.69	.53	22.13	.48	.46		
	78	.69	.52	22.38	.48	.59		
	79	.66	.57	2.80	.44	.32		
Effort Regulation	37	.59	.65	16.46	.35	.39	.69	.69
	48	.65	.58	18.50	.42	.09		
	60	.56	.68	15.71	.31	.13		
	74	.57	.67	15.94	.32	.12		
Time and Study Environment	35	.67	.55	21.67	.45	.50	.75	.76
	43	.60	.65	18.66	.36	.53		
	52	.33	.89	9.59	.11	.57		
	65	.55	.70	16.90	.30	.53		
	70	.59	.65	18.57	.35	.34		
	73	.53	.72	16.36	.28	.43		
	77	.69	.52	22.66	.48	.46		
	80	.27	.93	7.79	.07	.39		

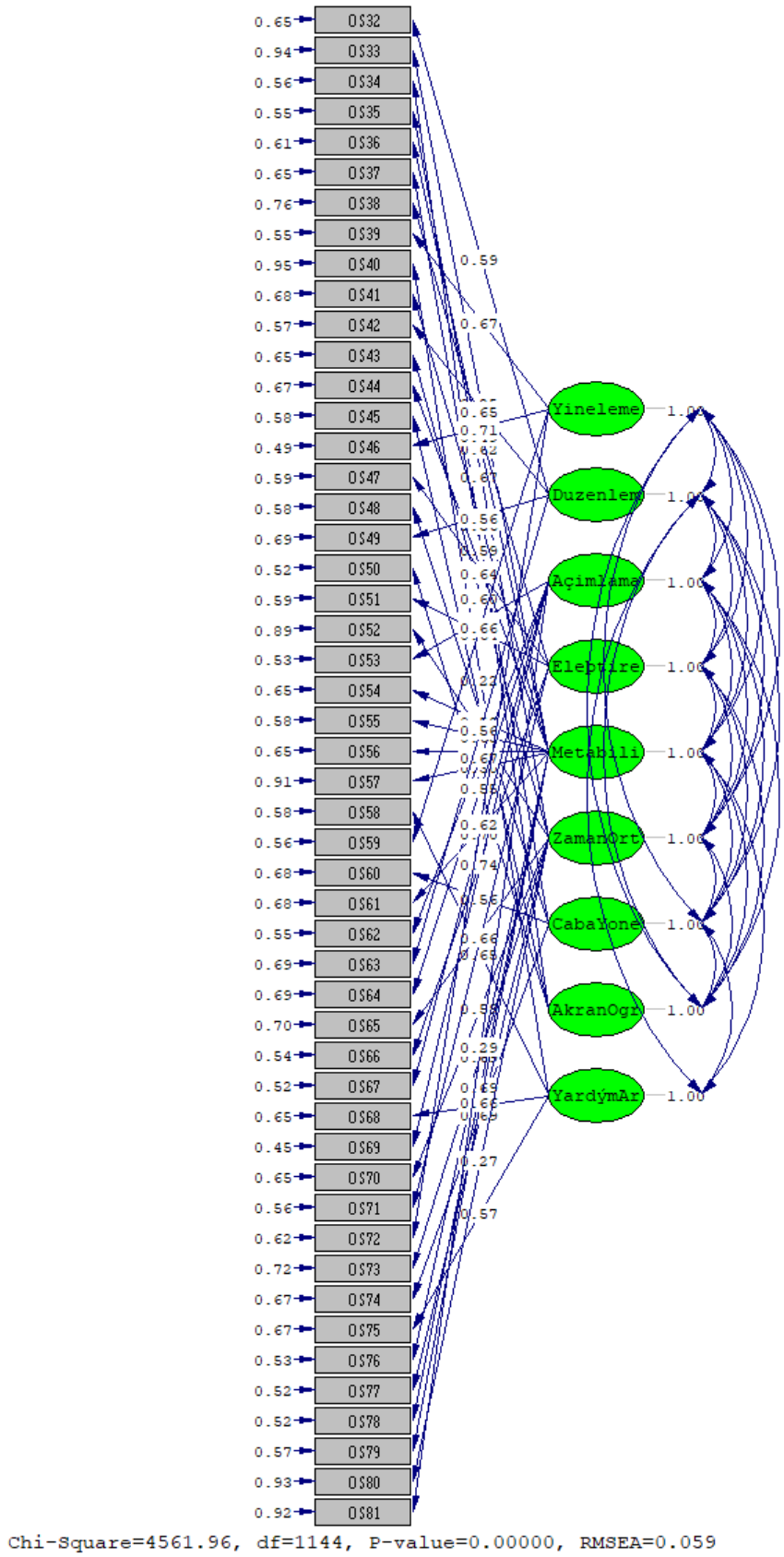


Figure 3. Path Diagram and Factor Loadings for MSLS Learning Strategies Section

DISCUSSION and CONCLUSION

Students' individual differences are the properties that should be taken into consideration in teaching-learning process. This is because the teaching-learning approaches students choose and their responses to teaching change according to the difference in their individual properties. Their individual properties can be divided into cognitive, affective, social and physiological categories. Several factors which can be described as individual differences such as having different levels of motivation, difference in perceptual preferences, intelligence level and psychological factors are influential in individuals' teaching-learning processes (Kuzgun-Deryakulu, 2004). One of those individual differences is students' self-regulated learning skills. Therefore, the validity and reliability of the scales were deemed adequate to reveal students' self-regulated learning skills in teaching-learning environments. The emergence of the view that the importance of contexts in self-regulation processes could not be ignored with the arise of self-regulation models based on social cognitive theory made us feel the necessity for scales which could be used with differing courses. For this reason, this study adapted Achievement Goal Scale and Motivated strategies for Learning Scale for chemistry course and analysed the psychometric properties so as to measure high school students' self-regulated learning skills. The sub-factors in the scales were analysed by means of confirmatory factor analysis. In addition to Cronbach's alpha- which was an internal consistency coefficient- McDonald's Omega coefficient was also calculated. Moreover, total item correlations were also analysed for the reliability of each item in the scales.

On examining the results for confirmatory factor analysis performed for Achievement goal Scale, the fit indices for the scale were found as RMSEA= .07; GFI=.94; NFI= .97; AGFI=.91, NNFI=.97; CFI= .98 and SRMR=.042. An examination of fit indices makes it clear that only chi-square/df ratio is below 3. Yet, on considering the other fit indices, it can be concluded that there is good fit. Garver and Mentzer (1999) state that NNFI, CFI and RMSEA can be used in determining model-data fit. Considering the acceptability of RMSEA below 0.8 and having RMSEA of 0.7 in this study along with the other fit indices, it was regarded that the model had good fit (Schermelleh-Engel, Moosbrugger, & Müller, 2003). Besides, due to the fact that NNFI and CFI (>.90) had acceptable values in this study, the scale was assumed to have construct validity. It can be said that the fit indices obtained in this study yields results similar to the ones in the original scale and the ones in other adaptations. On examining the results of confirmatory factor analysis performed for Achievement Goal Scale developed by Elliot and McGregor (2001), it was found that Chi-square (48, N=148) = 60.49, p=.11; RMSEA= .042, Tucker-Lewis Index (TLI) = .99 and CFI= .99. Another adaptation made by Pamuk (2014) found, on examining the results of confirmatory factor analysis performed for each sub-factor, that fit indices were perfect for three sub-factors apart from the sub-factor of performance avoidance. The fit indices for performance avoidance was reported as Chi-square/df=22.55, NFI=.97, CFI=.97, SRMR=.04 and GFI=.98.

The results of reliability analyses done for Achievement Goal Scale indicated that the Cronbach's Alpha (α) found for mastery approach was .85, it was .79 for mastery avoidance, .77 for performance approach and .67 for performance avoidance. Nunnally (1978) suggested that reliability coefficient be .70 as a general rule. But O'Rourke, Hatcher and Stepanski (2005) pointed out that values below .70 were also adequate and that social scientists even reported values below 60 occasionally (for example Dekovic, Janssens & Gerris, 1991; Holden, Fekken & Cotton, 1991). Therefore, when considered along with all other results for the scale, it was concluded that the factors of the scale satisfied the reliability criteria. Additionally, it was found that the other adaptations of this scale made in Turkey had also calculated similar reliability indices. Cronbach's Alpha- which was the internal consistency coefficient- calculated for Achievement Goal Scale developed by Elliot and McGregor (2001) ranged between .83 and .87. Şenler and Sungur (2007), on the other hand, found that Cronbach's Alpha took on values between .64 and .84. Cronbach's Alpha coefficients ranged between .65 and .76 in Pamuk (2014). Examining the results for the scale and the adaptations made in Turkey, it can be said that achievement objectives, a component of self-regulated learning skills, can be measured more comprehensively with this scale (Şen, 2015).

The fit indices for the motivation part of Motivated Strategies for Learning Scale (MSLS) were found as RMSEA= .07; GFI=.89; NFI= .97; AGFI=.87, NNFI=.97; CFI= .97 and SRMR=. 044. On examining the adaptations in the literature and the original version, it can be said that the fit indices found in this study are higher. The results of confirmatory factor analysis conducted for the motivation part of the scale developed by Pintrich et al (1991) were found as Chi square/sd = 3.49; RMR=.07; GFI=.77. In an adaptation made by Sungur (2004) the results for the motivation part were as in the following: Chi-square/sd = 5.3, GFI = .77, and RMR = .11. Adaptation made by Büyüköztürk et al (2004), however, reported results for the motivation part as: Chi-square/sd =4.47, RMSEA=.06, GFI=.88, AGFI=.85, CFI=.82, NNFI=.80, RMR=.18 and SRMR=.06. It was found that the fit indices for the motivation part of the model in the scale prepared by Pintrich et al (1991) and the fit indices of the adaptations made in Turkey did not have enough model-data fit. Considering the adaptations made by Büyüköztürk et al. (2004), Sungur (2004), Taştan (2009) and Yalçınkaya (20109 and the fit indices for the original version of the scale, it was regarded that the motivation part met the criteria for fit indices. Besides, Pintrich et al (1991) stated that motivational attitudes could change according to the properties of a course, teachers' demands and students' individual properties although the fit indices they had obtained were not within the desired interval; and they claimed that the values they had found were adequate.

In consequence of the reliability analyses performed for MSLS, the Cronbach's Alpha was found as .85, it was found as .73 for the factor of control of learning beliefs, .87 for the factor of self-efficacy for learning and performance and .61 for the factor of test anxiety. On reviewing the adaptations and original versions in the literature, this study can be said to have higher reliability indices. Only the reliability coefficient found for test anxiety was below .70 in this study. But because O'Rourke, Hatcher and Stepanski (2005) state that the values below .70 are also adequate; it was regarded that Cronbach's Alpha- which was calculated for the sub-factors of the motivation part of MSLS and which was also an internal consistency coefficient, McDonald's omega coefficients and total item correlations met the criteria for reliability. Cronbach's Alpha values found for MSLS following the reliability analyses reported in the literature were found as .62-.93 by Pintrich et al (1991), as .54-.89 by Sungur (2004) and as .52-.86 by Büyüköztürk et al (2004).

The fit indices found for the learning strategies part of MSLS were as in the following: RMSEA= .059; NFI= .94; GFI=.83; NNFI=.95; AGFI=.81, CFI= .95 and SRMR=.079. Reviewing the adaptations in the literature and the original version, it can be said that the fit indices found in this study are higher. The results of confirmatory factor analysis conducted for the learning strategies part of the scale developed by Pintrich et al (1991) were as in the following: Chi-square/sd = 2.26; RMR=.08; GFI=.78. Sungur (2004) found the values in an adaptation for biology course as: Chi-square/sd = 4.5, GFI = .71, and RMR = .08. Büyüköztürk et al (2004) found the fit indices for the learning strategies part as: Chi-square/sd =4.73, GFI=.80, AGFI=.77, CFI=.70 NNFI=.67 RMR=.22, SRMR=.06 and RMSEA=.07.It was found that the fit indices in the adaptation made in Turkey did not meet the model-data fit values as in the fit indices for the learning strategies part of the model in the scale prepared by Pintrich et al (1991). Considering the original scale and the fit indices of the adaptations in Turkey, it was regarded that the fit indices for the scale met the indices for good fit. Besides, Pintrich et al (1991) state that students' use of strategies differs according to students' individual differences, teachers' properties and the structure of courses; and that therefore researchers consider the values they find as acceptable. For this reason, considering the adaptations made for MSLS (Büyüköztürk et al., 2004; Pintrich et al., 1991, Sungur, 2004) it may be said that the reliability indices found are acceptable.

Cronbach's Alpha found for the factor of rehearsal of MSLS was .76, it was .68 for the factor of organisation, .78 for the factor of elaboration, .76 for the factor of critical thinking, .84 for the factor of metacognitive self-regulation, .75 for the factor of time and study environment, .69 for the factor of effort regulation, .71 for the factor of peer learning and .59 for the factor of help seeking (Şen, 2015). On examining the adaptations in the literature and the original version, it can be stated that the reliability indices found in this study are higher. Cronbach's Alpha was found as .52-.80 by Pintrich et al (1991), as .57-.81 by Sungur (2004) and as .41-.75 by Büyüköztürk et al (2004). On examining the Cronbach's alpha values, McDonald's Omega coefficients and total item correlations, it was regarded

that the sub-factors in the learning strategies part of the scale met the criteria for reliability. Considering all the figures for the questionnaire it was concluded that the questionnaires had met the reliability criteria. In consequence, having done validity and reliability analyses, both questionnaires can contribute to the literature as questionnaires which are capable of serving to the purpose of determining self-regulated learning skills. Besides, educators can also analyse the results by using each sub-factor available in the questionnaires separately.

REFERENCES

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 54*, 106–148. doi: 10.1111/j.1467-6494.1986.tb00391.x

Büyüköztürk, Ş., Akgün, Ö. E., Demirel, F. ve Özkahveci, Ö. (2004). Güdülenme ve Öğrenme Stratejileri Ölçeği'nin Türkçe formunun geçerlik ve güvenirlik çalışması. *Kuram ve Uygulamada Eğitim Bilimleri, 4*(2), 207-239.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319.

Context. (2018). *In Oxford dictionaries*, Retrieved from https://en.oxforddictionaries.com/definition/context.

Çelik, H. E. ve Yılmaz, V. (2013). *Lisrel 9.1 ile yapısal eşitlik modellemesi* (Yenilenmiş 2. Baskı). Ankara: Anı Yayıncılık.

Dekovic, M., Janssens, J. M. A. M., & Gerris, J. R. M. (1991). Factor structure and construct validity of the Block Child Rearing Practices Report (CRPR). *Psychological Assessment, 3*, 182–187. doi: 10.1037/1040-3590.3.2.182

Driscoll, M. P. (2005). *The psychology of learning for instruction* (3rd Edition). Boston, MA: Pearson.

Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational psychologist, 34*(3), 169-189. doi: 10.1207/s15326985ep3403_3

Elliot, A.J., & Church, M.A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 72*(1), 218-232. doi: 10.1037/0022-3514.72.1.218

Elliot, A.J., & McGregor, H.A. (2001). A 2x2 achievement goal framework. *Journal of Personality and Social Psychology, 80*, 501-519. doi: 10.1037/0022-3514.80.3.501

Elliot, A.J., & Reis, H.T. (2003). Attachment and exploration in adulthood. *Journal of Personality and Social Psychology, 85*, 317–331. doi: 10.1037/0022-3514.85.2.317

Fraenkel, J.R., & Wallen, N.E. (2000). *How to design and evaluate research in education* (4th ed). Boston. McGraw Hill.

Garver, M. S., & Mentzer, J.T. (1999). Logistics research methods: Employing structural equation modeling to test for construct validity. *Journal of Business Logistics, 20*(1), 33-57.

Holden, R. R., Fekken, G. C., & Cotton, D. H. G. (1991). Assessing psychopathology using structured test item response latencies. *Psychological Assessment, 3*, 111–118. doi: 10.1037/1040-3590.3.1.111

Kuzgun, Y. ve Deryakulu, D. (2004). Bireysel farklılıklar ve eğitime yansımaları. Y. Kuzgun (Ed.) ve D. Deryakulu (Ed.), *Eğitimde bireysel farklılıklar* (s.95–136). Ankara: Nobel Yayın Dağıtım.

McCoach, D. B., & Siegle, D. (2003). Factors that differentiate underachieving gifted students from high achieving gifted students. *Gifted Child Quarterly, 47*, 144 – 154.

Nunnally, J. (1978). *Psychometric theory.* New York: McGraw-Hill.

O'Rourke, N., Hatcher, L., & Stepanski E. J. (2005). *A Step-by-Step Approach to Using SAS for Univariate and Multivariate Statistics* (Second Edition). Cary, NC: SAS Institute Inc.

Özbay, A. (2008). *Yabancı dilde bilgilendirici yazma alanında öz düzenleme becerilerinin kullanımı ve başarı arasındaki ilişki.* Yayımlanmamış Doktora Tezi, Hacettepe Üniversitesi, Ankara, Türkiye. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Özturan Sağırlı, M., Çiltaş, A., Azapağası, E. ve Zehir, K. (2010). Yükseköğretimin öz-düzenlemeyi öğrenme becerilerine etkisi (Atatürk Üniversitesi örneği). *Kastamonu Eğitim Dergisi, 18*(2), 587-596.

Pamuk, S. (2014). Multilevel analysis of students science achievement in relation to constructivist learning environment perceptions, epistemological beliefs, self-regulation and science teachers characteristics. *Unpublished doctoral dissertation. Middle East Technical University, Ankara, Turkey*. Retrieved from http://etd.lib.metu.edu.tr/upload/12617892/index.pdf

Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research*, 31(6), 459-470. doi: 10.1016/S0883-0355(99)00015-4

Pintrich, P. R. (2000a). *The role of goal orientation in self-regulated learning.* In M., Boekaerts & P. R., Pintrich (Eds.), *Handbook of self-regulation (pp.13-39).* San Diego, CA: Academic Press.

Pintrich, P. R. (2000b). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology, 92*,544–555. doi: 10.1037/0022-0663.92.3.544

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33-40. doi: 10.1037/0022-0663.82.1.33

Pintrich, P. R., Smith, D. A. F., Garcia, T. & McKeachie, W. J. (1991). A manual for the use of the motivated strategies for learning questionnaire (MSLQ). National Center for Research to Improve Postsecondary Teaching and Learning. Ann Arbor: Michigan. ED 338 122. Retrieved from https://files.eric.ed.gov/fulltext/ED338122.pdf

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement, 53*(3), 801-813.doi: 10.1177/0013164493053003024

Sakız, G. ve Yetkin Özdemir, İ. E. (2014). Özdüzenleme ve özdüzenlemeli öğrenme: Kuramsal bakış. G. Sakız (Ed.), *Özdüzenleme* (s. 29-47). Ankara: Nobel Akademik Yayıncılık.

Sarı, A. ve Akınoğlu, O. (2009). Öz-düzenlemeli öğrenme: Modeller ve uygulamalar. *M.Ü. Atatürk Eğitim Fakültesi Eğitim Bilimleri Dergisi, 29*, 139-154.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research-Online, 8* (2), 23-74.

Senemoğlu, N. (2011). *Gelişim, öğrenme ve öğretim* (20. Baskı). Ankara: Pegem Akademi.

Sungur, S. (2004). *An implementation of problem based learning in high school biology courses.* Unpublished Dissertation, Middle East Technical University, Ankara, Turkey. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Şen, Ş. (2015). *Investigation of Students' Conceptual Understanding of Electrochemistry and Self-Regulated Learning Skills in Process Oriented Guided Inquiry Learning Environment*. Unpublished Dissertation, Hacettepe University, Ankara, Turkey. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Şenler, B. (2011). *Pre-service science teachers'self-efficacy in relation to personality traits and academic self-regulation*. Unpublished Dissertation, Middle East Technical University, Ankara, Türkiye. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Şenler, B., & Sungur, S. (2007). Hedef yönelimi anketinin Türkçe 'ye çevrilmesi ve adaptasyonu. *1. Ulusal İlköğretim Kongresi,* Ankara.

Taştan, O. (2009). *Effect of cooperative learning based on conceptual change conditions on motivation and understanding of reaction rate*. Unpublished Dissertation, Middle East Technical University, Ankara, Türkiye. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Yalçınkaya, E. (2010). *Effect of case based learning on 10th grade students' understanding of gas concepts, their attitude and motivation*. Unpublished Dissertation, Middle East Technical University, Ankara, Türkiye. Retrieved from http://etd.lib.metu.edu.tr/upload/3/12611523/index.pdf

Yumuşak, N., Sungur, S., & Çakıroğlu, J. (2007). Turkish high school students' biology achievement in relation to academic self-regulation. *Educational Research and Evaluation, 13*(1), 53 – 69. doi: 10.1080/13803610600853749

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81*(3), 329-339. doi:10.1037/0022-0663.81.3.329

Zimmerman, B. J. (2000). Attaining Self-Regulation: A Social Cognitive Perspective. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Self-Regulation: Theory, Research, and Applications* (13-39). San Diego, CA: Academic Press.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into practice, 41*(2), 64-70. Retrieved from https://www.jstor.org/stable/1477457

Zinbarg, R. E., Revelle, W., Yovel, I. & Li, W. (2005). Cronbach's α, Revelle's, β and McDonalds ω: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70(1),* 1-11. doi: 10.1007/s11336-003-0974-7

Öz-düzenleyici Öğrenme Becerileri: Ölçek Uyarlama

Giriş

Öz-düzenleme, bireylerin davranışlarını gözlemlemesi ve kendi ölçütleriyle karşılaştırmalar yaparak yargıda bulunması ve gerektiğinde davranışlarını kendi ölçütlerine göre yeniden düzenlemesidir. Öz-düzenleyici bireyler kendi davranışlarını etkilerler, yönlendirirler ve kontrol ederler (Bandura; aktaran, Senemoğlu, 2011). Zimmerman (2000)'a göre öz-düzenleme, bireylerin bireysel hedeflerine ulaşmak

adına geliştirdikleri, planlı ve döngüsel olarak ortaya çıkan düşünceler, duygular ve davranışlardır. Sosyal bilişsel kurama göre öz-düzenleme, sosyal ortamda gelişir ve zamanla bireyler tarafından içselleştirilir. Bu kurama göre öz-düzenlemenin yapısında bilişsel, metabilişsel ve motivasyonel bileşenler bulunmaktadır (Zimmerman; aktaran, Sakız & Yetkin Özdemir, 2014). Bundan dolayı öz-düzenleyici öğrenciler öğrenme sürecinde metabilişsel, motivasyonel ve davranışsal olarak etkin bir rol alırlar, kendi öğrenme hedeflerini oluştururlar ve bu süreci kontrol ederler (Zimmerman, 1989). Bu tanımlara göre, öz-düzenleme zihinsel bir beceri ya da akademik bir yetenek olarak tanımlanmayıp, öğrenenin sahip olduğu bilişsel yeterliklerini akademik yetenekler şeklinde adapte ettiği ve bunu da kendisi tarafından yönettiği bir süreç olarak özetlenebilir (Zimmerman, 2002).

Sosyal bilişsel kurama dayalı öz-düzenleme modellerinin ortaya çıkmasıyla, bağlamın öz-düzenleme süreçlerindeki öneminin göz ardı edilemeyeceği fikri ortaya çıkmıştır. Bağlam (kontekst); bir durum, bir fikir veya bir olay için çevreyi oluşturan koşullar şeklinde tanımlanabilir ("Context", 2018). Bağlamın, bulguların geçerliliğini etkileyebileceği fikrinin ortaya çıkması ile yapılan ölçümler bağlama duyarlı hale getirilmiştir. Böylece farklı öğrenme alanlarına yönelik ölçümler giderek daha fazla önem kazanmıştır (Pintrich; aktaran, Özbay,2008). Kısacası, durumlar arası genellemelere dayalı ölçümler yerine bağlama duyarlı, özel öğrenme alanlarına ve görevlerine yönelik ölçümler daha fazla önem kazanmıştır. Literatürde yapılan çalışmalar arasında sosyal bilişsel kurama dayalı olarak Pintrich vd. (1991) tarafından geliştirilen "Öğrenmede Güdüsel Stratejiler Ölçeği" (ÖGSÖ) sıklıkla kullanılmaktadır. Pintrich vd. (1991) ölçekte üniversite öğrencilerine yönelik bir dersi analiz birimi olarak belirlemişlerdir (Özbay, 2008). ÖGSÖ, motivasyon ve öğrenme stratejilerinin kullanımında bağlamın önemli bir etkisinin olduğu, farklı öğrenme alanlarında ve görevlerinde farklı stratejilerinin kullanımının gerektiği görüşüne dayalı olarak geliştirilmiştir (Özbay, 2008). ÖGSÖ, öğrencilerin bilişsel düzenleme göstergeleri olarak beş alt boyut içerir. Bu alt boyutlar; yineleme, açımlama, düzenleme, eleştirel düşünme ve metabilişsel özdüzenleme boyutlarıdır. Pintrich (2000a) tarafından önerilen öz-düzenleyici öğrenme modeli çerçevesinde bazı biliş kontrol aktiviteleri ve izleme ölçümleri ile Zimmerman (2000) tarafından önerilen model çerçevesinde bazı performans kontrol aktivitelerinin yer aldığı alt boyutlar vardır. Motivasyonun ve duyuların düzenlenmesi ile ilgili olarak ÖGSÖ motivasyonel stratejilerin ölçülmesine yönelik alt boyutlar içermemektedir. Fakat performans ve öğrenme hedefleri olmak üzere başarı hedefleri (achievement goals), görev değeri, öğrenme ve performansla ilişkili öz-yeterlik ve sınav kaygısı Zimmermann'ın modelinde yer alan önsezi aşamasında vurgulanan öğrencilerin motivasyonel inançlarının yer aldığı alt boyutlar bulunmaktadır. Davranışın düzenlenmesi ile ilgili olarak ise ÖGSÖ'da üç alt boyut mevcuttur. Bunlar; öğrencilerin zor ve ilgi çekmeyen görevlerle karşılaştıklarında kendi çabalarını düzenlemeleri, zaman ve çalışma ortamı yönetimi ve yardım almak için birini belirlemeye yönelik alt boyutlardır. Aslında hem Zimmerman hem de Pintrich tarafından geliştirilen ve sosyal bilişsel teoriye dayalı olan öz-düzenleyici öğrenme modellerinde performans kontrolü, zaman yönetimi, yardım arama ve çevresel yapılandırma gibi öz-düzenleyici stratejilerin vurgusu yapılmaktadır. Son olarak ÖGSÖ bağlamın düzenlenmesi ile ilişkili iki alt boyut daha içermektedir. Bu alt boyutlar akran öğrenimi ile zaman ve çalışma ortamı yönetimidir. Bu alt boyutlar öğrencilerin öğrenme kaynağı olarak arkadaşlarını ne kadar iyi kullandıklarını ve çalışma ortamı ile zamanlarını ne kadar iyi yönettiklerini belirlemek için kullanılır (Yumuşak, Sungur & Çakıroğlu, 2007).

Bağlamın, bulguların geçerliliğini etkilemesinden dolayı bağlama duyarlı ölçümlerin yapılması ihtiyacı ortaya çıkmaktadır. Bu sebeple farklı öğrenme alanlarına yönelik ölçümler giderek daha fazla önem kazanmıştır (Pintrich; aktaran, Özbay, 2008). Fakat yapılan çalışmalarda lise öğrencilerinin farklı derslerdeki öz-düzenleyici öğrenme becerilerinin belirlenmesi için geçerli ve güvenilir ölçeklerin olmadığı belirlenmiştir. Bundan dolayı öğrencilerin kimya dersindeki öz-düzenleyici öğrenme becerilerini geliştirmek için öncelikle değerlendirmede kullanılacak ölçeklere ihtiyaç duyulmaktadır. Ayrıca alanyazında kullanılan ÖGSÖ'da yer alan başarı hedeflerinin iki genel başarı hedefleri şeklinde sınırlandırılmış olması Öğrenmede Güdüsel Stratejiler Ölçeği ve Hedef Yönelimi Ölçeklerinin birlikte kullanımı gerekliliğini ortaya çıkarmıştır. Bundan dolayı bu iki ölçeğin uyarlanarak geçerlik ve güvenirlik çalışmaları yapılmalıdır. Bu doğrultuda, bu çalışmada; lise öğrencilerinin öz-düzenleyici öğrenme becerilerini belirlemek için ÖGSÖ ve HYÖ kimya dersi için uyarlanmış ve psikometrik özellikleri incelenmiştir.

Yöntem

Çalışmaya 9., 10., 11., ve 12. sınıflara devam etmekte olan toplam 862 lise öğrencisi katılmıştır. Öğrencilerin %35.03'ü kız, %33.06'sı erkek öğrencilerden ve %31.9'u da herhangi bir kodlama yapmamıştır. Öğrencilerin, yaşları 16-20 arasında değişmektedir.

Veri Toplama aracı olarak HYÖ ve ÖGSÖ ölçekleri kullanılmıştır. Elliot ve McGregor (2001) tarafından üniversite öğrencileri için geliştirilmiş olan Hedef Yönelimi Ölçeği (HYÖ) Şenler ve Sungur (2007) tarafından Türkçeye adaptasyonu yapılmıştır. Şenler ve Sungur tarafından ölçek fen dersleri için uyarlanmış olup ilköğretim öğrencilerine uygulanmıştır. 7'li likert tipi olan ölçek araştırmacılar tarafından 5'li likert şeklinde uyarlanmıştır. Bu çalışmada ise, ölçek orijinal versiyonunda olduğu gibi 7'li likert şeklinde ve lise öğrencilerine yönelik kimya dersleri için izin alınarak uyarlanmıştır. ÖGSÖ, Pintrich, Smith, Garcia ve McKeachie (1991) tarafından üniversite öğrencilerinin derslerdeki motivasyonları ve bu derslerde kullandıkları öğrenme stratejileri hakkında bilgi elde etmek için geliştirilmiştir. Ölçek, Büyüköztürk, Akgün, Özkahveci ve Demirel (2004) ve Sungur (2004) tarafından Türkçe'ye uyarlanmıştır. 7'li Likert tipi bir ölçektir. ÖGSÖ'nın motivasyon ve öğrenme stratejileri olmak üzere iki ana bileşeni bulunmaktadır.

Ölçeklerde yer alan alt boyutların yapı geçerliği için doğrulayıcı faktör analizi yapılarak analiz edilmiştir. Ölçeklere ilişkin güvenirlik değerlerini elde etmek için ise bir iç tutarlılık katsayısı olan Cronbach Alfa değerlerinin yanı sıra McDonalds'ın Omega (ω) katsayısı hesaplanmıştır. Ayrıca ölçeklerde yer alan her bir maddenin güvenirliği için madde toplam korelasyon değerleri incelenmiştir.

Sonuç ve Tartışma

Öğrencilerin sahip oldukları bireysel farklılıklar, öğretme-öğrenme sürecinde dikkate alınması gereken önemli özelliklerdir. Çünkü öğrencilerin tercih ettikleri öğretme-öğrenme yaklaşımları, öğretim uygulamalarına verdikleri tepkiler sahip oldukları bu bireysel özelliklerindeki farklılıklara göre değişmektedir. Bu bireysel özellikler, bilişsel, duyuşsal, toplumsal ve fizyolojik kategoriler altında sınıflandırılabilir. Farklı motivasyon düzeylerine sahip olmak, algısal tercihlerdeki farklılıklar, zeka düzeyi ve psikolojik faktörler gibi bireysel farklılıklar olarak tanımlanabilecek bir çok faktör bireylerin öğretme-öğrenme süreçlerini etkiler (Kuzgun & Deryakulu, 2004). Bu bireysel farklılıklardan bir tanesi de öğrencilerin öz-düzenleyici öğrenme becerileridir. Dolayısıyla öğrencilerin öğretme-öğrenme ortamlarındaki öz-düzenleyici öğrenme becerilerini belirlemek için geçerli ve güvenilir ölçeklere ihtiyaç duyulmaktadır. Sosyal bilişsel kurama dayalı öz-düzenleme modellerinin ortaya çıkmasıyla, bağlamın öz-düzenleme süreçlerindeki öneminin göz ardı edilemeyeceği fikrinin ortaya çıkması farklı derslerde kullanılacak olan ölçeklere gereksinim duyulmaktadır. Bu sebeple bu çalışmada lise öğrencilerinin öz-düzenleyici öğrenme becerilerinin ölçülmesi amacıyla Hedef Yönelimi Ölçeği ile birlikte Öğrenmede Güdüsel Stratejiler Ölçeği kimya dersi için uyarlanarak psikometrik özellikleri incelenmiştir. Ölçeklerde yer alan alt boyutların yapı geçerliği için doğrulayıcı faktör analizi yapılarak analiz edilmiştir. Ölçeklere ilişkin güvenirlik değerlerini elde etmek için ise bir iç tutarlılık katsayısı olan Cronbach Alfa değerlerinin yanı sıra McDonalds'ın Omega (ω) katsayısı hesaplanmıştır. Ayrıca ölçeklerde yer alan her bir maddenin güvenirliği için madde toplam korelasyon değerleri incelenmiştir.

Hedef Yönelimi Ölçeği için yapılan doğrulayıcı faktör analizi sonuçları incelendiğinde; ölçeğe ait uyum değerleri RMSEA= .07; GFI=.94; NFI= .97; AGFI=.91, NNFI=.97; CFI= .98 ve SRMR=.042 şeklindedir. Çalışmada öğrenme yaklaşma boyutu için tespit edilen Cronbach Alfa (α) değeri .85, öğrenme kaçınma boyutu için .79; performans yaklaşma boyutu için .77 ve performans kaçınma boyutu için ise bu değer .67 olarak hesaplanmıştır. Ölçeğe ait sonuçlar incelendiğinde özdüzenleyici öğrenme becerilerinin bir bileşeni olan başarı hedeflerinin daha detaylı bir şekilde bu ölçekle ölçülebileceği söylenebilir (Şen, 2015).

Öğrenmede Güdüsel Stratejiler Ölçeğinin (ÖGSÖ) motivasyon boyutuna ait uyum değerleri (fit indices) ise RMSEA= .07; GFI=.89; NFI= .97; AGFI= .87, NNFI= .97; CFI= .97 ve SRMR= .044

şeklindedir. Çalışmada Görev Değeri boyutu için hesaplanan Cronbach Alfa (α) değeri .85, Öğrenmeye İlişkin Kontrol İnancı boyutu için .73; Öğrenme ve Performansla ilgili Özyeterlik boyutu için .87 ve Sınav Kaygısı boyutu için ise bu değer .61 olarak hesaplanmıştır. Ölçeğin öğrenme stratejileri boyutu için hesaplanan uyum değerleri; RMSEA= .059; NFI= .94; GFI=.83; NNFI= .95; AGFI=.81, CFI= .95 ve SRMR= .079 şeklindedir. Çalışmada yineleme boyutu için belirlenen Cronbach Alfa (α) değeri .76, düzenleme boyutu için .68, açımlama boyutu için .78, eleştirel düşünme boyutu için .76, metabilişsel özdüzenleme boyutu için .84, zaman ve çalışma alanı yönetimi boyutu için .75, çaba yönetimi boyutu için .69, akran öğrenimi boyutu için .71, yardım arama boyutu için .59 olarak hesaplanmıştır (Şen,2015). Ölçeklere ait tüm değerler göz önünde bulundurulduğu zaman ölçeklerin geçerlik ve güvenirlik açısından psikometrik özellikleri karşıladığına karar verilmiştir. Sonuç olarak geçerlik ile güvenirliği sağlanmış olan her iki ölçek, eğitimcilerin öz-düzenleyici öğrenme becerilerini belirleme amacına hizmet edebilecek ölçekler olarak literatüre katkı sağlayabilir. Ayrıca eğitimciler ölçeklerde yer alan her bir alt boyutu ayrı ayrı olarak da kullanarak sonuçları inceleyebilirler.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(4);356-375

The Effect of Cooperative Learning on Students' Anxiety and Achievement in Musical Ear Training Lessons*

Gökhan ÖZTÜRK** Nesrin KALYONCU***

Abstract
Anxiety that can be experienced in musical ear training lessons is a significant psychological factor that can negatively affect the acquisition of aural skills, their effective application and evaluation process as well. This study looks into the influence of 'Cooperative Learning Method' on anxiety in musical ear training classes, on student achievement and exam anxiety/state anxiety. The study was designed, and conducted as an experiment with a pre-test and post-test control group, and the experimental procedures were completed in eight weeks. The study was carried out with thirty seven students [(n_e=19), (n_c=18)], who were taking Musical Ear Training-IV (MET-IV) lesson in the spring semester of 2010-2011 academic year in the Music Teacher Training Bachelor Program of a university in West Black Sea Region in Turkey. Experiment and control groups were equilibrated by taking into account the grade means in the MET courses that students had taken for the previous three semesters, and the gender aspect. The lessons in the experimental group have been taught mainly through 'Cooperative Learning Method', while mainly through 'Expository Teaching Approach' in the control group. Research data were collected by means of 'MET Lesson Anxiety Scale', 'Music Theory Test', and 'Musical Writing (Dictation) Test' tools, which have been developed by the researchers, and by means of 'State Anxiety Scale' tool developed by Spielberger et al. The collected data were analysed using dependent samples *t*-test and independent samples *t*-test. According to the results, Cooperative Learning has shown no significant effect on the achievement in music theory and in musical writing, and likewise on exam anxiety/state anxiety. However, the results show a significant positive effect of Cooperative Learning on lesson anxiety in musical ear training classes.

Key Words: Musical ear training lesson, cooperative learning method, achievement, lesson anxiety, exam anxiety/state anxiety

INTRODUCTION

Musical Ear Training (MET), which is one of the basic branches of music education, has a complicated structure. Musical ear training, the general aims of which can be defined as gaining student the skills of musical perception and musical memorization, musical imagination, musical sensitivity, transforming the perceived into musical notation/symbols, and musical reading (Brink, 1980; De Larminat, 2008; Harrison, 1990; Kalyoncu, 2005; Paney, 2007; Scheele, 1993; Sevgi, 1982; Shanefield, 2011), encompasses a range of theoretical and practical content. In the theoretical dimension of the lesson, the students are given the knowledge of musical elements to enable them to comprehend and analyse the language of music, and the knowledge that explains inter-elemental relations (notation, rhythm, meter, intervals, musical scale, chords, alteration, modulation etc.). The practical dimension, which is taught as woven from theoretical knowledge, includes the practices such as listening, perceiving/recognizing/identifying single and multiple tones, hearing and identifying metric, rhythmic, melodic, and harmonic structures separately or together, reading notation through

* This article is based on PhD thesis titled "The Effect of Cooperative Learning Method on Students' Anxiety and Achievement in Ear Training" completed by the first author, under the consultancy of second author. The results were presented at the 23rd EAS Conference/5th ISME European Regional Conference in Rostock, Germany.

** Asst. Prof. Dr., Tokat Gaziosmanpaşa University, Faculty of Education, Tokat-Turkey, gokhan.ozturk@gop.edu.tr, ORCID ID: https://orcid.org/0000-0002-1667-3758
*** Prof. Dr., Bolu Abant İzzet Baysal University, Faculty of Fine Arts, Bolu-Turkey, kly00nega@gmail.com, ORCID ID: https://orcid.org/0000-0002-2083-7487

To cite this article
Öztürk, G., & Kalyoncu, N. (2018). The effect of cooperative learning on students' anxiety and achievement in musical ear training lessons. *Journal of Measurement and Evaluation in Education and Psychology, 9(4),* 356-375. DOI: 10.21031/epod.411010

Received: 30.03.2018
Accepted: 11.09.2018

solfege or other solmization methods, sight-reading, dictating the tones and musical structures heard, and to put forth the internally imagined. Through this multifaceted and comprehensive structure, MET lessons constitute the basic for many theory-focused and performance-focused lessons in the musical training process and underpin music-specific courses (Aydoğan, 1998; Dunlap, 1989; Ferrante, 2010; Gates, 2001; Karkın, 2007; Potts, 2009; Sevgi, 1982).

There are various factors that affect achievement in musical ear training lessons, like in any lesson, such as student ability, teacher competency, curriculum, readiness, pre-learning level, the teaching method or the attitude of students towards the relevant lesson. Notwithstanding, anxiety experienced in the lesson or related to the lesson is considered as a factor with an effect on student achievement, and it is reported that anxiety experienced particularly in situations like examination is a variable that negatively affect learning and academic achievement throughout student life (Bacanlı, 2011; Hasselberg, 2010; Morgan, 2000; Slavin, 2006; Wine, 1971; Woolfolk, 1993; Zanden and Pace, 1984). Such views that are underlining the impacts of anxiety on learning make us consider anxiety as a factor to be counted in for musical ear training lessons as well.

In a broad sense, anxiety, as an important study subject of educational psychology as it is a basic sense of the mankind, "is an emotion of fear, discontent, and worry that is unconscious in its basis […] but identified consciously by the individual" (Öner, 1972, p.152). Anxiety is also qualified as a mood "with no clear reason" (Hançerlioğlu, 1988, p.223) or "that lost its source or object" (Dağ, 1999, p.181). Anxiety, which is defined by Spielberger (1983) as an emotion of "subjective feeling of tension, apprehension, nervousness, and worry associated with an arousal of the autonomic nervous system" (as cited in Horwitz, 2001, p.113), can be accompanied by various physical reactions as well (Carlson and Buskist, 1997; Hasselberg, 2010; Plotnik, 2009).

Öner (1977) states that the anxiety process comprises a series of complex cognitive, emotional, physiologic and behavioural activities. For this reason, anxiety can lead to different changes in individuals. While the symptoms of anxiety is categorised under two as subjective (spiritual) and objective (physical) complaints (Köknel, 1983; as cited in Hançerlioğlu, 1988, p.225), the changes are recognised as cognitive, emotional, behavioural, and physiologic changes as well (Ceyhan and Namlu, 2000; Dürü, 1999; Hasselberg, 2010). The *cognitive aspect* of anxiety is the self-criticism of the individual and feeling concerned for his/her performance; the *emotional aspect* is feeling nervousness and unease; and the *behavioural aspect* includes reactions such as clumsiness, silence, reticence, withdrawal (Geen, 1985; as cited in Kapıkıran, 2006, p.2). As for the *physiological aspect* of anxiety, it reveals itself through the symptoms like palmar sweating, pulse and respiration increase, stomach complaints etc. (Carlson and Buskist, 1997; Doğan and Baş, 2003; Öner, 1972; Plotnik, 2009).

The anxiety phenomenon, which is identified with regard to a variety of lessons (Awan, Azher, Anwar and Naz, 2010; Batton, 2010; Daneshamooz and Alamolhodaei, 2012; Elkhafaifi, 2005; Suwantarathip and Wichadee, 2010), also exist in the context of MET lessons (Hannon, 2015; Karpinski, 2000b; Mishra, 1998; Rifkin and Urista, 2006). As a general trend in the existing sources, the content of lesson, approaches in teaching, class environment, the students' perception towards itself and towards teacher, teaching methods and exam applications are considered as significant factors with a potential to lead to anxiety in musical ear training. MET lessons are most times perceived as 'hard' due to their comprehensive content. The studies conducted by Wunsch (1973), Covington (1992), Karpinski (2000a), Sevgi (2000), Sisley (2008) and Hannon (2015) calls attention to this challenge. In this context, the nature of MET lesson has the characteristics that can cause anxiety come out or develop further in students. However, the comprehensiveness of contents is not the sole reason that cause anxiety. Özgür and Aydoğan (1999) remarks that learning, itself, is an intangible process and contends, along with this fact, that the abstract character of the sound phenomenon makes learning harder in the process of musical ear training. Most of the activities performed in MET lessons involve the concurrent utilization of the skills related to perception and comprehension, in other words, the intricate procedures of sensing, perception, coding, decoding, and recognition/identification. MET lessons include practices that have predominantly psychological dimension, since the coordination of perception, memory, attention, and the use of knowledge step forward during these procedures

(Spencer, 1947). Aural skills can be affected by lack of perception (Wunsch, 1973), psychological barriers, nervousness, and lack of attention (Hannon, 2015; Karpinski, 2000b). When viewed from this perspective, it is possible to consider anxiety as a secret agent that has the potential to block or undermine achievement, which can be reached in MET, through leading to the malfunctioning of the steps of the musical hearing process and some research results at hand support this opinion (Mishra, 1998; Öztürk and Kalyoncu, 2017).

Various pedagogical measures can be taken to reduce the negative effects of the factors that are expected to lead to anxiety in musical ear training. The discussions on the compatibility of the conventional teaching methods with the nature of ear training have brought forward the necessity of pursuing different approaches. For example, Gates (2001) indicates that the teaching strategies followed by the music theorists, who ignore the researches on learning strategies and base on self-experience, harm students, and points out that more clear decisions could be provided on what and how to teach through the synthesis of the researches and practical experiences. On the other hand, Scandrett (2005) states that most MET classes consist of 15-20 individuals and that students show divergence with regard to their backgrounds and abilities. He stresses that some students are able to perform a task easily while others have more difficulty; and for this reason, that it would be beneficial giving individual instruction to every student according to their needs, pre-learning level, and abilities. Likewise, Rifkin and Urista (2006) reports that dictation causes to the debilitating anxiety particularly in the students who have weak hearing skills, that polarization is observed in the classes in which strong and weak students co-exist in the same environment, and that successful students get bored while weak students experience anxiety and even umbrage. The authors advice the employment of methods that are based on interaction and cooperation, encourage active learning and enhance the feeling of confidence. Berry (2008) indicates that expository method is a conventional approach used to perform teaching in music theory lessons, and that instruction could be seen as the best choice since it ensures the conveyance of excessive knowledge and due to diversity in student pre-learning. Arguing that the compatibility of the method does not squarely mean that it is effective, Berry points out the deprivation of instruction from the effectiveness of active learning approach, and states, based on the researches, that there is a tendency of transition from the conventional methods to active methods. In another research, the lecturers linked student failure in MET lesson to the difficulty of the curriculum, while the students showed 'the deficiency of the teaching method' as the primary reason for failure (Aydoğan, 1998). According to the same study, the students, who found the lecturers running MET classes incompetent with regard to teaching methods, thought that they could not be inspired due to the monotonously instructed lessons. Parker (2007), however, reports that teachers continue teaching in the way they were educated, and that innovative ideas, social interaction, and teaching applications that ensure complete learning and that are directed to problem solving, are ignored during the implementation of music curricula. Having observed the destructive impacts of anxiety in MET lessons, Hannon (2015) emphasizes the role of the teacher and argues that teachers can help students to overcome anxiety through methods, which are creative, interaction-focused and multi-activity-oriented.

The developments in the 20th century has shown that learning is not only a stimulus-bound reactive process; on the contrary, it is a process of construction based on cognitive and social interaction (Ergün and Özsüer, 2006; Laney, 1999; Seifert and Sutton, 2009; Senemoğlu, 2005; Slavin, 2006). This fact, which is in congruence with the nature of musical ear training, has contributed to the development of various methods, which puts the active participation of students in the centre. As one of these, Cooperative Learning Method is accepted as an alternative approach that can meet the above-mentioned expectation. In this context, along with the conventional teaching methods in MET lessons, participative, interactive and cooperative classroom environments can be built instead of the competition-based classroom environment particularly to avoid time pressure, reduce anxiety, and achieve different results in practice (Hannon, 2015; Rifkin and Urista, 2006; Slavin, 2006; Woolfolk, 1993).

Cooperative Learning Method is a teaching strategy in which students assist each other in learning that is aimed at a common goal/learning objective by forming small mixed groups or teams, are rewarded

as a group (Johnson and Johnson, 1988, 1994; Kagan, 1994; Slavin, 1980, 1987), and "work together, sharing ideas, information, and resources, as they progress toward identified goals" (Kaplan and Stauffer, 1994, p.1). Cooperative learning* is a method that can be applied to different lessons and groups, that bases the working methods of small groups on specific theoretical foundations, and that incorporates various techniques such as 'Teams-Games-Tournament', 'Academic Controversy', 'Jigsaw 1-2', 'Co-op Co-op', 'Group Investigation', and 'Student Teams-Achievement Divisions'.

Academic achievement being in the first place, Cooperative Learning Method has important impacts on cognitive, social, and emotional learning outputs. Particularly by the activities based on the principle of social interaction, it contributes to students' psychic-emotional development by supporting self-confidence, attitude, and motivation, while gaining them the emotion of sharing and responsibility along with the skills of critical thinking and problem solving (Johnson, Johnson and Smith, 1991; Slavin, 1987, 1990). With these contributions, Cooperative Learning Method has an important effect on anxiety as well. It is supposed that anxiety, which is considered as one of the chief factors effecting productivity and the construction of positive relations, is rarely observed in cooperative classroom environments (Johnson et al. 1991; Kagan, 1994). Results of the researches conducted on various lessons showed the positive effect of Cooperative Learning Method on class anxiety (Batton, 2010; Courtney, Courtney and Nicholson, 1992; Edelbrock, 1990; Lavasani and Khandan, 2011; Mehdizadeh, Nojabaee and Asgari, 2013; Okebukola, 1986; Suwantarathip and Wichadee, 2010; Valentino, 1988).

The use of Cooperative Learning Method was highlighted in some theoretical studies conducted in early 1990s in the field of music education as well (Baloche and DeLorenzo, 1994; Di Natale and Russel, 1995; Friedmann, 1989; Kaplan and Stauffer, 1994), and various application samples were presented. Several experimental studies presented its outlook in practice. Some of these researches highlight the positive impacts of Cooperative Learning Method both on the learning of musical content and on the student's psychic process. Its effects on listening skills (Hosterman, 1992), musical knowledge and singing skills (Bilen, 1995; Güven, 2011; Kocabaş, 1995; Parker, 2007; Söker, 1998; Uysal, 2004), attitude towards instrumental course and performance, confidence, and also motivation in technical applications and style works (Fisher, 2010; Sözen, 2012) are to name but a few. Moreover, it is also reported that student participation in musical activities improve and that they develop positive attitude, as well, in cooperative learning environments (Bilen, 1995; Djordjevic, 2007; Fisher, 2010; Goliger, 1995; Güven, 2011; Hwong, Caswell, Johnson and Johnson, 1993; Kocabaş, 1995).

There are other studies, along with the above ones, which emphasize the effect of cooperative learning on the gaining of aural skills. For example, Therrien (1997) found in his study in which he compared the effect of computer-aided individual teaching and cooperative learning strategy on the achievements of music theory and musical ear training that the both teaching methods were effective in the acquisition of musical knowledge and aural skills, and in the improvement of course success. Similarly, Nacakçı (2011) found that the Cooperative Learning Method was more effective compared to the conventional methods. Along with these studies, there are diverse studies showing that Cooperative Learning Method has effects on the gaining and/or improvement of the skills of reading tonal and rhythmic phrase (Inzenga, 1999); analysis of melody, meter, and timbre (Holloway, 2001); analysis of musical texture, genre, and style (Smialek and Boburka, 2006); attitude towards music theory lesson, perception of self-confidence and success attributions (Canakay, 2007); rhythmic counting and sight-reading (Parker, 2007); being informed on sight-reading, harmonising and playing the given tune (Fisher, 2010); polyphonic solfege reading and musical hearing-writing (Gürpınar, 2014).

* It is known that the views and suggestions on the use of cooperation in education dates late back in history and the approach of teaching-learning in small groups have existed for centuries (Fisher, 2010). It is also cited that social psychologists worked on subjects that compare cooperation and competition well before the cooperative learning programs developed for application in classrooms (Slavin, 1987). However, Cooperative Learning Method is a highly structured approach compared to many group learning techniques being used (Kaplan and Stauffer, 1994).

Despite these researches, the authors of this article have not yet come across with a study that looked into the effect of the Cooperative Learning Method on anxiety in MET lessons. Additionally, it is difficult to mean that there is a large number of researches concentrated on anxiety and musical ear training. One of a few researches that we spotted during the literature review in the context of our topic is Mishra's (1998) study, which is one of the pioneering studies about the anxiety in MET. Mishra addressed the anxiety and variables of ear training, and reports that the students experiencing high level of anxiety get lower scores from MET exams in comparison to the students experiencing low level of anxiety. Further, Mishra indicates that the students who do not feel confident with their aural skills get remarkably lower scores from MET exam compared to the students who feel confident with their abilities/skills. In their study with the students of music teacher training bachelor program, Öztürk and Kalyoncu (2017) found that the students experience anxiety in MET classes and exams, that -similar to Mishra's findings- there is a significant correlation in the negative direction between lesson achievement and anxiety towards lesson and exam. In the context of researches, a survey that was conducted to identify 141 music major students' preconceived ideas about music theory and aural skills can be referred here (Hannon, 2015). In the survey, 41,8% of the students answered questions with a negative response, while 28,4% had a general fear and expressed the difficulty of aural skills. The limited number of researches in the music education literature complicates stable arguments about the place and impacts of anxiety in MET lessons, and crystalizes the need for further studies that will look into the causes of anxiety experienced in MET classes, its relations with different variables, how its impact on success can be reduced or what preventions to be taken against anxiety during teaching. In this context, the purpose of this research is to examine whether or not the Cooperative Learning Method, which enables the student to learn in group by teacher and peer support in a social environment, have an effect on *a) lesson anxiety, b) exam anxiety/state anxiety, c) achievement* in MET lesson. However, the fact that anxiety, which is a popular study subject in educational research, has not yet been widely dealt with in the national and international literature in the context of musical ear training is considered as the challenging but stimulating aspect of our research.

METHOD

Research Model

Experimental model with pre-test and post-test control group is used in this research.

Table 1. Overview of the Research Process

Before Experiment	Group	Pre-test	Experimental Process	Time	Post-test
Assignment of the groups	Experimental group	- Personal information form - Music theory test - Musical writing test - State anxiety scale - MET lesson anxiety scale	Cooperative learning method (Student teams-achievement divisions - STAD technique)	8 weeks	- Music theory test - Musical writing test - State anxiety scale - MET lesson anxiety scale
	Control group		Expository methods (Instruction, Q&A, and Discussion techniques)		

Study Group

The study group was formed by 40 students who were taking Musical Ear Training-IV (MET-IV) lesson in the spring semester of 2010-2011 academic year in the Music Teacher Training Bachelor Program of a university in West Black Sea Region in Turkey. In experimental studies with pre-test and post-test control group, where the number of participants are fewer, it is preferred to match the participants rather than random selection (Açıkgöz, 1992). The students' success grades in the MET-I, II, III courses in the previous semesters were taken as the key criterion to equilibrate the experimental and control groups. A score of success was obtained for each student by taking the mean of the grades belonging to the previous three semesters of the students who were studying in the fourth semester.

The class was divided into two, the experimental and control groups were equalized according to these achievement scores and 20 students were assigned to each group (see Table 2). In addition, there were no students with absolute pitch ability in the study group, and all participants had relative pitch ability.

Table 2. Means of Success Grade in MET-I, II, III Courses

Groups	N	$\bar{X}$	SD	t	df
Experimental	20	69,93	9,94	,67	38
Control	20	67,38	13,80		

The groups were balanced in terms of gender in order to achieve heterogeneous group structure, which is one of the basic principles of the Cooperative Learning Method (see Table 3). In the process, 3 (three) students in total who were not regularly attending classes in both groups were excluded from measurements; hence, the number of students in the study group was set as 37.

Table 3. Distribution of Students according to Gender

Groups	Female		Male		Total	
	f	%	f	%	N	%
Experimental	13	35,13	6	16,22	19	51,4
Control	12	32,43	6	16,22	18	48,6
Total	25	67,56	12	32,44	37	100

Table 4 shows that 59,46% of the students who participated in the research exercised MET lesson for 1-4 hours per week on a regular basis; however, it is seen that 27,03% of the students worked only before the exam and 10,81% of the students did not work at all. The development of aural skills depends on regular and continuous work. However, a certain number of students state that they regularly study MET lesson, while the statements of more than one third of the research group show that they do not study regularly for MET lessons.

Table 4. Students' Weekly Exercise Hours for MET Lesson

Exercise type		Experimental	Control	Total
Never	f	3	1	4
	%	8,11	2,7	10,81
Working 1-4 hours per week regularly	f	9	13	22
	%	24,32	35,14	59,46
Studying on a daily basis	f	1	0	1
	%	2,7	0	2,7
Studying just before exams	f	6	4	10
	%	16,22	10,81	27,03
Total	f	19	18	37
	%	51,4	48,6	100

Experimental Process

Planning of experimental application

At the planning stage of the experiment, the course contents to be taught during the experimental period were selected primarily from the framework program of the MET-IV lesson in the current Music Teacher Training Bachelor Program (YÖK, 2006). These contents are structured to be the same in both experimental and control groups (see Table 5). Again, the exercises, dictation and solfege pieces, and home assignments were prepared the same for both groups without making tonal, maqamic or rhythmical alterations to the structure. For the under-mentioned contents, the process of learning-

teaching for each lesson was planned clearly and through cascading in the context of Expository Teaching Approach for the control group and Cooperative Learning Method for the experimental group by putting the 'Student Teams-Achievement Divisions' (STAD) in the centre. Lesson blocks were scheduled for two hours.

Table 5. MET-IV Lesson Content and Learning-Teaching Methods Used During the Study

Week	Lesson blocks	Lesson content	Learning-teaching methods	
			Experimental group	Control group
1	1. block	Application of pre-tests	---	---
	2. block	Seventh chords and the dominant seventh chord	Jigsaw I	Instruction/ Q&A/ Discussion
2	1. block	Seventh chords and the dominant seventh chord	STAD	
	2. block	Modulation	STAD	
3	1. block	Four-sharp major tonality	STAD	
	2. block	Four-sharp major tonality	STAD	
4	1. block	Hearing and writing in E major tonality	STAD	
	2. block	Four-sharp minor tonality	STAD	
5	1. block	Hearing and Writing in C# minor tonality	STAD	
	2. block	Hicaz and Zirgüleli Hicaz maqams	STAD	
6	1. block	Nikriz maqam	Instruction/Q&A	
	2. block	Hearing-reading-writing-application in maqamic structures	STAD	
7	1. block	Hearing-reading-writing-application in maqamic structures	Instruction/Q&A	
	2. block	Four-flat major tonality	STAD	
8	1. block	Four-flat minor tonality and Ornaments	Instruction/Q&A	
	2. block	Overall evaluation	Instruction/Q&A	

The conduct of the experimental application

All the scheduled lessons were taught in the experimental application process. The lessons of the experimental group were taught using the STAD technique due to its availability for the effective use of time. The STAD technique, which also allows for the use of conventional teaching methods at the stage of lecturing, is considered available for benefiting both from the impact of the conventional teaching methods such as instruction, and from smooth transition to cooperative works after starting with the expository teaching that the students are familiar with. These advantages were influential on the choice of the STAD technique. The implementation of the STAD technique consisted of five stages below:

1. '*Presentation'*, the delivery of the content of learning by the teacher through instruction and discussion as a first step in the classroom,
2. '*Teams'* of four formed heterogeneously in consideration of the students' characteristics such as academic achievement and gender,
3. '*Individual exams*' taken by the students in short intervals or at the end of each work in addition to group works in cooperative activities,
4. '*Individual progress scores*', to determine whether the student showed better success compared to the previous test scores,
5. '*Team reward*', awarded to the team members as per the pre-determined criteria in order to motivate the students in the team.

'Discussion' technique was used in two lessons, as well, along with the 'instruction' and 'Q&A' techniques in the context of expository teaching approach in the lessons done with the control group. Throughout the research period, no group work that required cooperation with the control group was performed, and the control group was provided with no information in order to prevent the creation of a competitive environment with the experimental group. The traditional seating order of the class, which was peculiar to common lessons, remained the same, and no change was made to the place

during the study period. The basic differences that separated the lessons done in the experimental and control groups are presented in Table 6.

Table 6. The Basic Differences of the Lessons Done with the Experimental and Control Groups

Control group	Experimental group
Students act individually	Students act with the group
Presentation of the subject through expository methods	Delivering hand papers in addition to instruction through exposition
No intervention in the student's individual learning	Teacher steers group works
Students study individually	Students work in interaction with the members of the group which they belong
Examination only between and at the end of the semesters	Small quizzes or other measurement applications upon the completion of each topic
No shared distribution of tasks, and individual responsibility of students from the works as a whole	The tasks of the members are clearly defined in group works and task sharing is performed
Individual reward	Individual or group awarding according to the contribution of the students to the group

Data Collection Tools

Research data is collected by means of 'Music Theory Test', 'Musical Writing (Dictation) Test', 'MET Lesson Anxiety Scale', and 'State-Trait Anxiety Inventory'. The characteristics of the study group were identified by using 'Personal Information Form'. Data gathering tools are explained below:

Personal Information Form: Data on the gender of the students, the type of school they graduated and their exercise habits on MET lesson were collected by this form developed by the researchers.

Music Theory Test: Developed by the researchers, this test consists of 48 items covering the subjects of interval, chord, rhythm, meter, tonality, and maqam in the program of MET-IV course. For the content validity of the test, the opinions of four academics[*] working in the musical ear training field of the music education departments of different universities were taken. Upon corrections in some questions in the light of expert opinions, the test was applied to a total of 185 students who were attending the 2nd, 3rd and 4th classes of the Music Teacher Training Undergraduate Programs of four universities for reliability procedures. The data obtained from 163 students who completed the test were analysed. Iteman 3.50 program was used in the analysis of the data. As a result of the reliability procedures of the test, KR-20 reliability coefficient value was found to be 0.89, mean difficulty index was found to be 0.47, and mean distinctiveness value was found to be 0.52. Since the reliability of the tests prepared for use in group comparisons is expected to be 0.60-0.80 (Tavşancıl, 2006), it can be said that the Music Theory Test developed is reliable.

Musical Writing (Dictation) Test: The test is developed by the researchers, and its content validity is ensured by receiving the opinions of four academics employed in different universities. There are 61 items in the test. The test consists of three basic dimensions that aim to measure the skills of perceiving, converting into musical notation, and analysing harmonic, melodic, and rhythmic structures. There are questions that aim at measuring: recognising/analysing and writing intervals and chords in *the first dimension*; writing rhythmical phrases in simple, compound, and irregular meters in *the second dimension*; and in *the third dimension*, they are aimed at recognising maqamic and tonal scales, and identifying, writing, and creating maqamic and tonal tunes in different meters. Each question is based on the items of the 'Music Theory Test'. The mean scores obtained by the application of the test to the students were compared with their mean success scores in MET-I, II, III lessons. The capacity of the prepared test to measure students' achievement in the MET lesson was

[*] We would like to thank Lecturer MA Adnan Atalay, Prof. Ali Sevgi, Dr. Özcan Özbek, and Asst. Prof. Dr. Salih Aydoğan for their help, sharing, and guidance in the development process of 'Music Theory Test' and 'Musical Writing (Dictation) Test'.

determined by this way. The concurrent validity coefficient between the mean score of MET-I, II, III lessons and the mean scores of the developed test was found to be r=,71.

MET Lesson Anxiety Scale: The draft of this scale, which is developed by the researchers, was applied to 272 students attending the 2nd, 3rd and 4th classes of music teacher training bachelor programs of five universities. A four-dimension scale of 28 items was obtained by the analysis of the gathered data. There are 5 positive expressions in the scale along with the negative expressions. The respondent student is asked to choose the most suitable option for themselves from five-item choices. The highest anxiety score that can be obtained from the scale is 140, and the lowest score is 28. A high score indicates that the anxiety is high towards the MET lesson, while a low score means anxiety is also low. The construct validity of the scale was tested by factor analysis. The adequacy of data set for factor analysis was assessed by the Kaiser–Meyer–Olkin (KMO) Measure of Sample Adequacy, and Bartlett's Test of Sphericity. The Kaiser-Meyer-Olkin (KMO) compliance measurement value was found to be 0,92. Bartlett's Test of Sphericity value was significant at the level of 0,001 (X^2_{378}=3415,93). According to the results of the factor analysis performed by using basic components analysis and orthogonal rotation, there are 4 factors with eigenvalues greater than 1 on the scale. These four factors account for 56,19% of the total variance. The total variance explained by each factor is 19,14; 13,58; 12,48 and 11,00. According to the initial eigenvalues, the fact that the eigenvalue (9,98) of the first factor is too high from the eigenvalue (2,50) of the second factor indicates that the scale has a general factor as a whole. According to the item analysis results based on item scale correlation, the correlation values ranged from r=,37 to r=,75 and found to be significant at the level of p<,01. The correlation values showed that the characteristic to be measured by the overall scale were the same with the characteristic to be measured with each item, showing that all the items had the quality to be included in the scale. The whole t value, which was obtained as a result of the analysis made by the comparison of the responses to each item by the participants who were in the bottom-top 27% (n=65) slice, was found to be significant at the level of p<,01. This result shows that all items have the quality to determine whether they have, or not, the quality to measure the characteristic intended by that item. The Cronbach Alpha internal consistency coefficients for each dimension of the scale and for the test as a whole were: α=0,90 for the 1st factor; α=0,85 for the 2nd factor; α=0,85 for the 3rd factor; α=0,73 for the 4th factor, and α=0,93 for the overall scale.

State-Trait Anxiety Inventory: The Inventory 'State Anxiety Subscale' was developed by Spielberger et al. in 1970, and Turkish adaptation and standardization were done by Öner and Le Compte (1985). In adaptation studies, the coefficient of invariance calculated by the Pearson product-moment correlation coefficient was found to vary between .26 and .68 according to the test-retest test. The KR-20 reliability coefficient was between .94 and .96. The scale, which has Likert type 20 items, requires the individual to describe how he or she feels at a certain time, under certain conditions, and about a particular situation, to respond by taking into account their feelings about the situation they are in. The emotions or behaviours expressed in the scale items are responded according to the severity of the above-mentioned experiences by marking one of the options of never, some, very, and completely. The total score obtained from the scale varies between 20 and 80. The higher the score, the higher the level of anxiety (Öner and Le Compte, 1985). The scale was used by the researchers upon prior permission from N. Öner.

Data Analysis

The data were analysed using the SPSS software. Shapiro-Wilk (S-W, N<30) test values and skewness-kurtosis values of each data group were calculated to determine the normal distribution characteristics of the data. As a result of the S-W test, it was determined that all score types except the Musical Writing (Dictation) pre-test scores of the experimental group showed normal distribution and the skewness-kurtosis coefficients were within ±1,5. It is observed in the literature that acceptable skewness-kurtosis values for normality can be between ±1 and ±3. In this study, the ±1,5 approach to normality distribution (Tabachnick and Fidell, 2007) is taken as a basis. 'Independent samples t-test' and 'dependent samples t-test' from parametric tests were used to compare pre-test and post-test scores of experimental and control groups. The level of significance was set to p<,05. The effect size

was examined in cases where the difference between the groups appeared significant. The widely preferred Cohen's *d* formula is used in the calculation of the effect size. Cohen's *d* value was interpreted as .20=small, .50=medium, and .80=large (Cohen, 1988).

RESULTS

The Effect of Cooperative Learning Method on Lesson Anxiety

Both the inter-group and intra-group lesson anxiety scores were compared in order to determine the effect of Cooperative Learning Method on general anxiety towards the MET lesson.

Table 7. The Results of the Dependent Samples t-test Belonging to the Pre-test and Post-test Scores of 'MET Lesson Anxiety Scale'

Groups	Application	N	$\bar{X}$	SD	t	df
Experimental	Pre-test	19	83,42	21,28	6,05*	18
	Post-test	19	56,95	17,42		
Control	Pre-test	18	90,33	20,66	4,62*	17
	Post-test	18	70,06	14,11		

*p<,05

As seen from Table 7, there is a significant difference ($t_{(18)}$=6,05, p<,05) between the pre-test ($\bar{X}$=83,42) and post-test ($\bar{X}$=56,95) mean scores of the experimental group's 'MET Lesson Anxiety Scale'. Similarly, there was a significant difference ($t_{(17)}$=4,62, p<,05) between the pre-test ($\bar{X}$=90,33) and post-test ($\bar{X}$=70,06) mean scores of the control group. It can be argued according to these findings that the two applied teaching-learning approaches are effective in moderating students' anxiety towards the MET lesson.

Table 8. The Results of the Independent Samples t-test Belonging to the Pre-test and Post-test Scores of 'MET Lesson Anxiety Scale'

Application	Groups	N	$\bar{X}$	SD	t	df	Cohen's d
Pre-test	Experimental	19	83,42	21,28	-1,00	35	---
	Control	18	90,33	20,66			
Post-test	Experimental	19	56,95	17,42	-2,51*	35	-.83
	Control	18	70,06	14,11			

*p<,05

As seen from Table 8, there is no statistically significant difference ($t_{(35)}$=-1,00, p>,05) between the experimental group ($\bar{X}$=83,42) and the control group ($\bar{X}$=90,33) in terms of 'MET Lesson Anxiety Scale' pre-test mean scores. In the post-test, a significant difference ($t_{(35)}$=-2,51, p<,05) was found between the mean scores of the experimental group ($\bar{X}$=56,95) and control group ($\bar{X}$=70,06), and the effect size was also found to be large (d=-.83>.80). Although the mean of anxiety scores for both groups declined, the difference is in favour of the experimental group in which the Cooperative Learning Method is centered and the fall in anxiety scores is greater.

The Effect of Cooperative Learning Method on Exam Anxiety/State Anxiety

Both the inter-group and intra-group state anxiety scores were compared in order to determine the effect of Cooperative Learning Method on state anxiety towards the MET exam.

Table 9. Dependent Samples t-test Results Belonging to the Pre-test and Post-test Scores of 'Music Theory Exam State Anxiety Scale'

Groups	Application	N	$\bar{X}$	SD	t	df
Experimental	Pre-test	19	37,11	9,03	0,22	18
	Post-test	19	36,63	10,72		
Control	Pre-test	18	48,33	11,10	2,61*	17
	Post-test	18	42,67	10,78		

*p<,05

According to Table 9, there is no significant difference ($t_{(18)}$=0,22, p>,05) between the mean scores of the experimental group's 'Music Theory Exam State Anxiety Scale' pre-test ($\bar{X}$=37,11) and post-test ($\bar{X}$=36,63). On the contrary, a significant difference ($t_{(17)}$=2,61, p<,05) was found between the control group's pre-test ($\bar{X}$=48,33) and post-test ($\bar{X}$=42,67) mean scores. The finding at hand shows that the Music Theory exam state anxiety scores of the students in the control group, for whom the expository teaching approach was centered, declined significantly to make a difference.

Table 10. Dependent Samples t-test Results Belonging to the Pre-test and Post-test Scores of 'Musical Writing (Dictation) Exam State Anxiety Scale'

Groups	Application	N	$\bar{X}$	SD	t	df
Experimental	Pre-test	19	43,52	14,97	0,00	18
	Post-test	19	43,52	14,29		
Control	Pre-test	18	50,44	12,43	-0,69	17
	Post-test	18	51,72	15,18		

As seen from Table 10, the mean scores of the experimental group's 'Musical Writing (Dictation) Exam State Anxiety Scale' pre-test ($\bar{X}$=43,52) and post-test ($\bar{X}$=43,52) did not change and there is no difference ($t_{(18)}$=0,00, p>,05) between the means of two measurements. There is no significant difference ($t_{(17)}$=-0,69, p>,05) between the pre-test ($\bar{X}$=50,44) and post-test ($\bar{X}$=51,72) mean scores of the control group and there is also a small increase in their post-test mean. These findings show, at the end of the study that there is no decline in the state anxiety of the students in both groups towards Musical Writing (Dictation) exam.

Table 11. Independent Samples t-test Results Belonging to the Pre-test and Post-test Scores of Both Exams related State Anxiety

Statae anxiety	Application	Groups	N	$\bar{X}$	SD	T	df	Cohen's d
Music theory exam anxiety	Pre-test	Experimental	19	37,10	9,03	-3,36*	35	-1.11
		Control	18	48,33	11,10			
	Post-test	Experimental	19	36,63	10,71	-1,70	35	---
		Control	18	42,66	10,78			
Musical writing (Dictation) exam anxiety	Pre-test	Experimental	19	43,52	14,97	1,52	35	---
		Control	18	50,44	12,43			
	Post-test	Experimental	19	43,52	14,29	1,69	35	---
		Control	18	51,72	15,18			

*p<,05

In Table 11, it is seen that there is a significant difference ($t_{(35)}$=-3,36, p<,05) in terms of state anxiety towards Music Theory pre-test mean scores between the experimental group ($\bar{X}$=37,10) and the control group ($\bar{X}$=48,33), and that the effect size is large (d=-1.11>.80). This result indicates that the control group had a higher level of exam anxiety than the experimental group before the study started. No significant difference ($t_{(35)}$=-1,70, p>,05) was found between the experimental group ($\bar{X}$=36,63) and control group ($\bar{X}$=42,66) in terms of state anxiety post-test mean scores for the same exam; the

significant difference between the two groups was eliminated as a result of the decrease in the control group scores. Based on the findings at hand, it can be said that teaching through expository methods is more effective in reducing state anxiety towards the Music Theory exam.

According to the same table, there is no significant difference ($t_{(35)}$=1,52, p>,05) in the state anxiety pre-test scores for the Musical Writing (Dictation) exam between the experimental group ($\overline{X}$=43,52) and the control group ($\overline{X}$=50,44). No significant difference ($t_{(35)}$=1,69, p>,05) was found in the post-test mean scores for the same exam between the experimental group ($\overline{X}$=43,52) and control group ($\overline{X}$=51,72). Neither of the two methods applied in the classes were effective in reducing state anxiety towards Musical Writing (Dictation) exam.

The Effect of Cooperative Learning Method on Achievement

Both inter-group and intra-group achievement scores were compared in order to determine the effect of the Cooperative Learning Method on achievement in MET lesson.

Table 12. Dependent Samples t-test Results Belonging to the Pre-test and Post-test Scores of 'Music Theory Test'

Groups	Application	N	$\overline{X}$	SD	t	df
Experimental	Pre-test	19	33,11	5,88	-8,72*	18
	Post-test	19	39,95	5,20		
Control	Pre-test	18	30,94	5,64	-6,51*	17
	Post-test	18	38,17	7,10		

*p<,05

As it is seen from Table 12, there is a significant difference ($t_{(18)}$=-8,72, p<,05) between the 'Music Theory Test' pre-test ($\overline{X}$=33,11) and post-test ($\overline{X}$=39,95) mean scores of the experimental group. Similarly, there is a significant difference ($t_{(17)}$=-6,51, p<,05) between the pre-test ($\overline{X}$=30,94) and post-test ($\overline{X}$=38,17) mean scores of the control group. According to these findings, it can be argued that the two learning-teaching approaches are effective in improving students' Music Theory achievement.

Table 13. Dependent Samples t-test Results Belonging to the Pre-test and Post-test Scores of 'Musical Writing (Dictation) Test'

Groups	Application	N	$\overline{X}$	SD	t	df
Experimental	Pre-test	19	27,00	10,89	-4,50*	18
	Post-test	19	33,57	8,87		
Control	Pre-test	18	26,05	11,52	-1,80	17
	Post-test	18	28,61	12,07		

*p<,05

According to Table 13, there seems to be a significant difference ($t_{(18)}$=-4,50, p<,05) between the mean scores of the experimental group's 'Musical Writing (Dictation) Test' pre-test ($\overline{X}$=27,00) and post-test ($\overline{X}$=33,57). However, there is no significant difference between the pre-test ($\overline{X}$=26,05) and post-test ($\overline{X}$=28,61) mean scores of the control group ($t_{(17)}$=-1,80, p>,05). Findings suggest that Musical Writing (Dictation) scores of the students in both groups are improved as a result of the study; however, it shows that the scores of the experimental group increased in a statistically significant manner. In addition, Musical Writing (Dictation) exam mean scores of both groups are lower than the means of the Music Theory exam as well.

Table 14. Independent Samples t-test Results Belonging to the Pre-test and Post-test Scores of Both Tests

MET Test	Application	Groups	N	$\bar{X}$	SD	t	df
Music theory	Pre-test	Experimental	19	33,11	5,88	1,14	35
		Control	18	30,94	5,64		
	Post-test	Experimental	19	39,95	5,20	0,87	35
		Control	18	38,17	7,11		
Musical writing (Dictation)	Pre-test	Experimental	19	27,00	10,89	0,25	35
		Control	18	26,05	11,52		
	Post-test	Experimental	19	33,57	8,87	1,43	35
		Control	18	28,61	12,07		

According to Table 14, there is no significant difference ($t_{(35)}$=1,14, p>,05) between the experimental group ($\bar{X}$=33,11) and the control group ($\bar{X}$=30,94) in terms of Music Theory pre-test achievement mean scores, and knowledge levels have been close to each other prior to the beginning of the study. There is also no significant difference ($t_{(35)}$=0,87, p>,05) between the post-test achievement mean scores in the experimental group ($\bar{X}$=39,95) and control group ($\bar{X}$=38,17). Although there was an increase in the achievement scores of Music Theory in both groups, the difference is not statistically significant.

As it is seen from the same table, there was no significant difference ($t_{(35)}$=0,25, p>,05) between the experimental group ($\bar{X}$=27,00) and the control group ($\bar{X}$=26,05) with regard to the Musical Writing (Dictation) pre-test achievement mean scores, and dictating skills have been close to each other before the beginning of the study. There is also no significant difference ($t_{(35)}$=1,43, p>,05) between the post-test achievement means for the experiment ($\bar{X}$=33,57) and control group ($\bar{X}$=28,61). Despite the improvement in Musical Writing (Dictation) achievement scores in both groups, the improvement is not statistically significant.

CONCLUSION and DISCUSSION

Based on the results related to *MET lesson anxiety*, it can be argued that Cooperative Learning Method have proved to be more effective in moderating general anxiety towards the lesson compared to the expository methods. The main reason for the effectiveness of the method is considered to be the 'classroom environment' provided for the student. Cooperative classroom allow for more opportunities for the student to improve his/her relationships with the environment and to develop social behaviours, and thus, increase attendance to lesson (Cornacchio, 2008; Kassner, 2002). Activities are mostly student-centered in these classes, as the learning process is more important than the learning outcomes (Lavasani and Khandan, 2011). This environment, on the one hand, supports the student to actively take responsibility, while on the other hand, it is encouraging even those with weak social relationships to communicate with other learners and to aim at a common goal together in "positive interdependence" (Johnson et al. 1991, p.17-18). We argue that the 'interdependency' in the classroom provided by the Cooperative Learning Method is effective in reducing the lesson anxiety of students'.

Another reason may be the contribution that cooperative learning makes to the improvement of 'self-esteem' and 'self-confidence' feelings. Self-esteem and self-confidence are important factors effecting learners' classroom experiences, motivation for lessons, and course achievement. According to the literature, it is argued that students with low self-confidence are more anxious or anxiety generates lack of self-confidence (Alkan, 2011; Başoğlu, 2007; Lawal, Idemudia and Adewale, 2017). Some students who participated in ear training lesson may also experience lack of self-esteem: Pratt and Henson (1987), for example, indicate that most learners do not care about musical ear training because of lack of confidence in their competence to conduct studies on ear training, and as a consequence, they feel like they are being threatened by such studies. Vygotsky (1978) and Bandura (1986) emphasize that the emotional and cognitive support to the learners by more knowledgeable people around them is influential in increasing their self-esteem and moderating their anxiety (as cited in

Alkan, 2011, p.95-96). Moreover, Rifkin and Urista indicate that students work at their own pace in MET classes in which game-like interaction based methods are pursued, that they learn "from each other's ideas and mistakes in a stimulating, cooperative atmosphere that builds confidence" (2006, p.76). It is considered that the cooperative learning activities in the experimental group might have provided the student with the above-mentioned support and solidarity, and enabled the student to demonstrate what they can do in the direction of their capacity, and this might have consequently been effective in moderating lesson anxiety by making a positive contribution to self-esteem and self-confidence in the research process.

Based on the results related to *MET exam anxiety/state anxiety*, it is concluded that Cooperative Learning Method is not effective in reducing state anxiety. Many reasons may underlie this fact. That there was no fall in the state anxiety of the experiment group, which worked through Cooperative Learning Method, despite the decline in their general lesson anxiety can be attributed to their act with the group in the lesson and obligation to exert individual performance in the exam. The withdrawal of the social support and solidarity from the environment in exam, which was provided during the lesson, might have made the students feel anxious. The decline in the Music Theory test anxiety of the students in the control group, whose assessment and evaluation procedures were based on individual performance, may underpin this conclusion: The control group students, who had to cope with exam stress individually during and at the end of the process, might have developed their own individual coping methods against anxiety. However, there was not a significant decline Musical Writing (Dictation) test to make a difference in the control group; in contrast, there was even a minor increase -though statistically insignificant- in the mean scores. The dictation exam anxiety scores of the experimental group also did not decrease, and the means remained the same. That there was no decline in Musical Writing (Dictation) exam anxiety for both groups can be attributed to the considerable difference in the procedures demanded by the two exams: While the students answer the questions in the Music Theory test in an order and time they wish, each question in Musical Writing (Dictation) exam is responded in a certain time and through reactions consisting of complex procedures (perception, memorisation, use of knowledge, identification, analyse, transformation and writing) to be given instantly. There is scarcely any chance of retrospective compensation by the end of the time given for response. The procedural difference in the Musical Writing (Dictation) exam that required mental flexibility and time dependence might have caused high anxiety towards the exam, since time pressure in examinations is one of the important factors that generates exam anxiety (Bekdemir, 2007; Birenbaum and Pinku, 1997).

Another reason may be the students' lack of study. Regular study contributes to the consolidation of knowledge and skills and the development of cognitive competences. Hence, the development of aural skills can be achieved not only by the activities performed in the lesson, but also by regular practices out of the classroom. However, more than one third of the students in our research group do not have regular working habits: For the MET-IV lesson, they either 'never study' (%10,81) or 'study just before the exam' (%27,03) (see Table 4). Some of the research on the sources of exam anxiety emphasizes 'the absence of study skills' (Bozanoğlu, 2004; Culler and Holahan, 1980). In this context, being not sure about what they learned as a result of the absence of regular study habit and the inadequacy of knowledge and skills level may have increased exam anxiety.

Depending on the results related to *MET lesson achievement*, it is concluded that the Cooperative Learning Method does not have a significant effect on course achievement or it has an effect similar to that of expository methods. There were improvements in the achievement scores of the both groups, which was an expected result. But there is no statistically significant difference between the experiment and control groups. Findings related to state anxiety showed that the Cooperative Learning Method was ineffective in moderating either Music Theory or Musical Writing (Dictation) exam anxiety, and that expository method was ineffective in moderating Musical Writing (Dictation) exam anxiety. We know that exam anxiety is an obstacle before achievement in ear training (Karpinski, 2000b; Mishra, 1998; Wunsch, 1973). Musical Writing (Dictation) is a complicated activity and it involves more ability, attention, effort, and skills compared to the obtaining and use of Music Theory knowledge. In exams, while the Music Theory dimension is closely related to the use of long-term

memory, the Musical Writing (Dictation) dimension involves the use of complex procedural steps in coordination, such as attention, perception, memorisation, analysis, comprehension, and transformation. This opinion is supported by the fact that the mean scores of the students in Musical Writing (Dictation) exam are lower than the mean of Music Theory. Considering that the bio-physiological changes caused by high anxiety during examination inhibit the formation of protein chains necessary for learning in the brain and disturb mental activities such as reasoning and thinking (Kaya and Varol, 2004; Özer, 2005), it is thought that the anxiety that arise during exam negatively effects the vital procedural steps of the musical hearing process. In this research, we are dwelling on the possibility that the likely difference in achievement, which could arise from cooperative learning at the end of the experiment, might not have realised due to exam anxiety.

The study was conducted by separating a class into two. Despite the fact that both groups are informed about the principles of study and learning content, the control group was not given any information about the details of the study with the experimental group and about the different teaching method. However, the activities conducted with the experimental group may have evoked curiosity in the control group. The members of the both groups may have exchanged views about the research in other common classes, which may have let the control group enter into competition with the experimental group. This likely sharing is explained as 'Hawthorne Effect[*]' in social psychology and can be faced in experimental studies in the field of education. The selection of the study group from the same institution to ensure the equivalence of the control and experimental groups should be taken into account as a reason that might have led to the appearance of the mentioned effect during the research period.

In conclusion, it can be argued with regard to exam anxiety/state anxiety and course achievement that the Cooperative Learning Method applied in the MET lessons has a similar effect with the Expository Teaching Approach; and in contrast, it is effective in moderating general anxiety developed by the students against the lesson. The following suggestions are presented relying on these results:

1. We suggest more inclusion of Cooperative Learning Method in lessons, since the cooperative classroom environment in the MET lessons can moderate lesson anxiety.
2. We know that measure of emotional (affective) characteristics and processes are more difficult than the measure of the cognitive and psycho-motor characteristics (Kalyoncu, 2002; Nartgün, 2008; Turgut, 1984). We suggest the determination of the anxiety experienced in the MET lessons and its underlying reasons through scales.
3. We suggest that the students keep a MET diary in which they note the states and experiences that make them feel anxious prior to and after each lesson throughout the semester. Facing the tangible states in these diaries and seeking solutions by sharing these states with friends or teachers is considered to help them in coping with anxiety.
4. We suggest the preference of other methods that are available for paired assessment and/or group assessment by, from time to time, getting out of the objective methods that are used for assessment in MET lessons.
5. To overcome the anxiety in MET lessons and exams, it is proposed to distance oneself from clichés in teaching and to develop creative approaches -as demonstrated by Hannon's (2015) numerous examples-, which provide diverse support to students both inside and outside the classroom and help to realize their full potential.

[*] Hawthorne effect "is also encountered in the situations in which there is no difference between the control and experimental groups. This attracts attention rather particularly when the researcher gives pre-test to the experimental and control groups prior to the experimental study. This test may serve as a hint for the students to sense the aim of the research. The students in the control group my initiate some improving activities to score better in the second test. For example, the students may learn from their friends in the experimental group about the studies they are performing, do the same things, and learn them. It can be observed at the end of the experience that there is no difference between the achievement of the control and experimental groups, although the experiment has been effective in reality" (Kaptan, 1998, p.157).

6. Stage anxiety is frequently addressed in musicology literature; however, there is not a considerable tendency toward researching the anxiety experienced in musical ear training or in other performance-focused courses. For this reason, we suggest the conduct of musical ear training and anxiety-related in-depth research by considering different variables.

REFERENCES

Açıkgöz, K. Ü. (1992). *İşbirlikli öğrenme: Kuram araştırma uygulama*. Malatya: Uğurel Matbaası.

Alkan, V. (2011). Etkili matematik öğretiminin gerçekleştirilmesindeki engellerden biri: Kaygı ve nedenleri. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 29, 89-107.

Awan, R. N., Azher, M., Anwar, M. N. & Naz, A. (2010). An investigation of foreign language classroom anxiety and its relationship with students' achievement. *Journal of College Teaching and Learning*, *7*(11), 33-40.

Aydoğan, S. (1998). *Müzik öğretmeni yetiştiren kurumlarda müziksel işitme okuma öğretimi* (Doctoral Dissertation, Gazi University, Ankara).

Bacanlı, H. (2011). *Eğitim psikolojisi* (16th Edition). Ankara: Pegem Akademi.

Baloche, L. & DeLorenzo, L. C. (1994). Cooperative learning making music together. *General Music Today*, *8*(1), 9-12. DOI: 10.1177/104837139400800103

Başoğlu S. T. (2007) Sınav kaygısı ile özgüven arasındaki ilişkinin erinlik döneminde incelenmesi (Master Thesis, Maltepe University, İstanbul). Retrieved from http://tez2.yok.gov.tr/

Batton, M. (2010). *The effect of cooperative groups on math anxiety* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 3427021).

Bekdemir, B. (2007). İlköğretim matematik öğretmen adaylarındaki matematik kaygısının nedenleri ve azaltılması için öneriler (Erzincan eğitim fakültesi örneği). *Erzincan Eğitim Fakültesi Dergisi*, *9*(2), 131-144.

Berry, W. (2008). Surviving lecture: A pedagogical alternative. *College Teaching*, *56*(3), 149-153.

Bilen, S. (1995). *İşbirlikli öğrenmenin müzik öğretimi ve güdüsel süreçler üzerindeki etkileri* (Doctoral Dissertation, Dokuz Eylül University, İzmir). Retrieved from http://tez2.yok.gov.tr/

Birenbaum, M. & Pinku, P. (1997). Effects of test anxiety, information organization, and testing situation on performance on two test formats. *Contemporary Educational Psychology*, *22*(1), 23-38.

Bozanoğlu, İ. (2004). *Bilişsel davranışçı yaklaşıma dayalı grup rehberliğinin akademik risk altındaki öğrencilerin akademik alandaki güdülenme, benlik saygısı, başarı ve sınav kaygısı düzeylerine etkisi* (Doctoral Dissertation, Ankara University, Ankara). Retrieved from http://tez2.yok.gov.tr/

Brink, E. R. (1980). *A cognitive approach to the teaching of aural skills viewed as applied music theory* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 8026770).

Canakay, E. U. (2007). *Aktif öğrenmenin müzik teorisi dersine ilişkin akademik başarı, tutum, özyeterlik algısı ve yüklemeler üzerindeki etkileri* (Doctoral Dissertation, Dokuz Eylül University, İzmir). Retrieved from http://tez2.yok.gov.tr/

Carlson, N. R. & Buskist, W. (1997). *Psychology: The science of behavior* (5th Edition). Boston: Allyn and Bacon.

Ceyhan, E. & Namlu A. G. (2000). Bilgisayar kaygısı ölçeği (BKÖ): Geçerlik ve güvenirlik. *Anadolu Üniversitesi Eğitim Fakültesi Dergisi*, *10*(2), 77-93.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cornacchio, R. A. (2008). *Effect of cooperative learning on music composition, interactions, and acceptance in elementary school music classrooms* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 3325655).

Courtney, D. P., Courtney, M. & Nicholson, D. (1992, November). *The effect of cooperative learning as an instructional practice at the college level*. Paper presented at the annual meeting of The Mid-South Educational Research Association, Knoxville, Tennessee.

Covington, K. (1992). An alternative approach to aural training. *Journal of Music Theory Pedagogy*, 6, 5-18.

Culler, R. E. & Holahan, C. J. (1980). Test anxiety and academic performance: The effects of study-related behaviors. *Journal of Educational Psychology*, *72*(1), 16-20.

Dağ, İ. (1999). Psikolojinin ışığında kaygı. *Doğu Batı Düşünce Dergisi* (4th Edition), 6, 181-190.

Daneshamooz, S. & Alamolhodaei, H. (2012). Cooperative learning and academic hardiness on students' mathematical performance with different levels of mathematics anxiety. *Educational Research*, *3*(3), 270-276.

De Larminat, V. (2008). Gehörbildung zwischen französischer und deutscher tradition. Versuch einer synthese. *Zeitschrift der Gesellschaft für Musiktheorie, 5*(1), 121-162. DOI:10.31751/358

Di Natale, J. J. & Russel, G. S. (1995). Cooperative learning for better performance. *Music Educators Journal*, *82*(2), 26-28. DOI:10.2307/3398865

Djordjevic, S. A. (2007). *Student perceptions of cooperative learning in instrumental music* (Master Thesis). Available from ProOuest Dissertations and Theses database (UMI No. 1456226).

Doğan, A. A. & Baş, M. (2003). Beden eğitimi ve spor bölümü öğrencilerinin durumluk kaygı düzeyleri ile başarıları arasındaki ilişki. *Atatürk Üniversitesi Beden Eğitimi ve Spor Bilimleri Dergisi, 5*(3), 1-5.

Dunlap, M. P. (1989). *The effects of singing and solmization training on the musical achievement of beginning fifth-grade instrumental students* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 9013890).

Dürü, Ç. (1999). Kaygı ve depresyon: Psikopatolojik bir bakış. *Doğu Batı Düşünce Dergisi* (4th Edition), 6, 191-196.

Edelbrock, R. C. (1990). *Computer anxiety reduction: The effect of cooperative learning* (Doctoral Dissertation Abstract). Available from ProOuest Dissertations and Theses database (UMI No. 9020173).

Elkhafaifi, H. (2005). Listening comprehension and anxiety in the arabic language classroom. *The Modern Language Journal*, *89*(2), 206-220.

Ergün, M. & Özsüer, S. (2006). Vygotsky'nin yeniden değerlendirilmesi. *Afyon Karahisar Üniversitesi Sosyal Bilimler Dergisi, 8*(2), 269-292.

Ferrante, J. D. (2010). *An investigation of the effects of regularly employed melodic dictation tasks on the sight-singing skills of high school choral* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 3399484).

Fisher, C. (2010). *Teaching piano in groups*. New York: Oxford University Press.

Friedmann, M. (1989). Stimulating classroom learning with small groups. *Music Educators Journal*, *76*(2), 53-56.

Gates, L. S. (2001). *The effects of ear training on beginning horn students: A qualitative case study design* (Master Thesis). Available from ProOuest Dissertations and Theses database (UMI No.1405354).

Goliger, J. M. (1995). *Implementation of a program of cooperative learning ın an urban secondary piano laboratory* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 9539810).

Gürpınar, E. (2014). *İşbirlikli öğrenme yöntemine dayalı çoksesli solfej uygulamalarının müziksel işitme-okuma-yazma ve koro ders başarılarına etkisi* (Doctoral Dissertation, İnönü University, Malatya). Retrieved from http://tez2.yok.gov.tr/

Güven, E. (2011). *Kaynaştırma uygulamasının yapıldığı sınıflarda işbirlikli öğrenmenin müzik öğretimi üzerindeki etkileri* (Doctoral Dissertation, Gazi University, Ankara). Retrieved from http://tez2.yok.gov.tr/

Hançerlioğlu, O. (1988). *Ruhbilim sözlüğü*. İstanbul: Remzi Kitabevi.

Hannon, A. (2015). Helping non-singers overcome fear and anxiety in aural skills. *Music Theory Pedagogy Online, 3*(5), 1-15. Retrieved from https://drive.google.com/file/d/0B0ZI8di-pEDvZmVaMzk2MC1RR28/view

Harrison, C. S. (1990). Relationships between graded in the components of freshman music theory and selected background variables. *Journal of Research in Music Education, 38*(3), 175-186. DOI:10.2307/3345181

Hasselberg, S. (2010). *Leistungsangst in der schule. Praeventions- und interventionsmöglichkeiten der schulsozialarbeit* (Diplomarbeit, Hochschule Neubrandenburg, Neubrandenburg). Retrieved from http://digibib.hs-nb.de/

Holloway, M. S. (2001). *The use of collaborative action learning to increase music appreciation students' listening skills* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 9998928).

Horwitz, E. K. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics*, 21, 112-126.

Hosterman, G. L. (1992). *Cooperative learning and traditional lecture/demonstration in an undergraduate music appreciation course* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 9226705).

Hwong, N., Caswell, A., Johnson, D. W. & Johnson, R. T. (1993). Effects of cooperative and individualistic learning on prospective elementary teachers' music achievement and attitudes. *The Journal of Social Psychology*, *133*(1), 53-64.

Inzenga, A. (1999). *Learning to read music cooperatively in a choral setting: A case study* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 9943999).

Johnson, D. W. & Johnson, R. T. (1988). *Cooperation in the classroom*. Minnesota: Interaction Book Company.

Johnson, D. W. & Johnson, R. T. (1994). An overview of cooperative learning. In J. Thousand, A. Villa & A. Nevin (Eds.) *Creativity and collaborative Learning*. Baltimore: Brookes Press.

Johnson, D. W., Johnson, R. T. & Smith, K. A. (1991). *Cooperative learning: Increasing college faculty instructional productivity*. Washington: ASHE-ERIC Higher Educational Report No.4, The George Washington University.

Kagan, S. (1994). *Cooperative learning*. San Clemente: Kagan Publications.

Kalyoncu, N. (2002). *Musikunterricht in der deutschen und türkischen grundschule. Eine vergleichende didaktische analyse*. Frankfurt am Main: Peter Lang Verlag.

Kalyoncu, N. (2005). Eğitim fakültelerinde uygulanan müzik öğretmenliği lisans programının revizyon gerekçeleriyle tutarlılığı. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 25*(3), 207-220.

Kapıkıran, N. A. (2006). Başarı kaygısı ölçeğinin geçerliği ve güvenirliği. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 19, 1-6.

Kaplan, P. R. & Stauffer, S. L. (1994). *Cooperative learning in music*. Reston, Virginia: MENC: Music Educators National Conference.

Kaptan, S. (1998). *Bilimsel araştırma ve istatistik teknikleri*. Ankara: Tekışık Ofset.

Karkın, A. M. (2007). Müzik teorisi ve işitme eğitimi dersinin piyano eğitimi üzerindeki etkileri, karşılaşılan sorunlar ve çözüm önerileri. *Kastamonu Eğitim Dergisi*, *15*(1), 411-422.

Karpinski, G. S. (2000a). *Lessons from the past: Music theory pedagogy and the future.* Music Theory Online, *6*(3). Retrieved from http://www.mtosmt.org/issues/mto.00.6.3/mto.00.6.3.karpinski_frames.html

Karpinski, G. S. (2000b). *Aural skills acquisition: The development of listening, reading, and performing skills in college-level musicians*. New York: Oxford University Press.

Kassner, K. (2002). Cooperative learning revisited: A way to address the standards. *Music Educators Journal*, *88*(4), 17-23. DOI:10.2307/3399786

Kaya, M. & Varol, K. (2004). İlahiyat fakültesi öğrencilerinin durumluk-sürekli kaygı düzeyleri ve kaygı nedenleri (Samsun örneği). *19 Mayıs Üniversitesi İlahiyat Fakültesi Dergisi*, 17, 31-63.

Kocabaş, A. (1995). *İşbirlikli öğrenmenin blokflüt öğretimi ve öğrenme stratejileri üzerindeki etkileri* (Doctoral Dissertation, Dokuz Eylül University, İzmir). Retrieved from http://tez2.yok.gov.tr/

Laney, J. D. (1999). A sample lesson in economics for primary students: How cooperative and mastery learning methods can enhance social studies teaching. *The Social Studies*, *90*(4), 152-158.

Lavasani, M. G. & Khandan, F. (2011). The effect of cooperative learning on mathematics anxiety and help seeking behavior. *Procedia Social and Behavioral Sciences*, 15, 271-276.

Lawal, A. M., Idemudia, A. S. & Adewale, O. P. (2017). Academic self-confidence effects on test anxiety among Nigerian university students. *Journal of Psychology in Africa*, *27*(6), 507-510. DOI:10.1080/14330237.2017.1375203

Mehdizadeh, S., Nojabaee, S. S. & Asgari, M. H. (2013). The effect of cooperative learning on math anxiety, help seeking behavior. *Journal of Basic and Applied Scientific Research*, *3*(3), 1185-1190.

Mishra, J. (1998, January). *The effects of anxiety on ear training test scores*. Paper presented at the Ohio Music Education Association/Music Educators National Conference North Central Division Professional Conference, Columbus, Ohio.

Morgan, C. T. (2000). *Psikolojiye giriş* (Trans. H. Arıcı et al. - 14th Edition). Ankara: Hacettepe Üniversitesi Psikoloji Bölümü Yayınları.

Nacakcı, Z. (2011). Müziksel işitme okuma ve yazma dersinde işbirliğine dayalı öğrenmenin öğrencilerin başarılarına etkisi. *E-Journal of New World Sciences Academy*, *6*(2), 180-186.

Nartgün, Z. (2008). Duyuşsal nitelikler ve ölçülmesi. In S. Erkan and M. Gömleksiz (Eds.) *Eğitimde ölçme ve değerlendirme* (p. 143-196). Ankara: Nobel Yayınevi.

Okebukola, P. A. (1986). Reducing anxiety in science classes: An experiment involving some models of class interaction. *Educational Research*, *28*(2), 146-149. DOI: 10.1080/0013188860280211

Öner, N. (1972). Kaygı ve başarı. *Hacettepe Sosyal ve Beşeri Bilimler Dergisi*, 2, 151-163.

Öner, N. (1977). *Durumluk-sürekli kaygı envanterinin türk toplumunda geçerliği* (Thesis for Associate Professorship, Hacettepe University, Ankara).

Öner, N. & Le Compte, A. (1985). *Süreksiz durumluk-sürekli kaygı envanteri el kitabı*. İstanbul: Boğaziçi Üniversitesi Yayınları.

Özer, M. A. (2005). Etkin öğrenmede yeni arayışlar: İşbirliğine dayalı öğrenme ve buluş yoluyla öğrenme. *Bilig*, 35, 105-131.

Özgür, Ü. & Aydoğan, S. (1999). *Müziksel işitme okuma*. Ankara: Sözkesen Matbaası.

Özgüven, İ. E. (2003). *Psikolojik testler* (5th Edition). Ankara: PDREM Yayınları.

Öztürk, G. & Kalyoncu, N. (2017, December). *Müziksel işitme eğitiminde kaygı ve çeşitli değişkenlerle ilişkisi.* Paper presented at the 3rd Cyprus International Congress of Educational Research, Gazimağusa, KKTC.

Paney, A. S. (2007). *Directing attention in melodic dictation* (Doctoral Dissertation). Retrieved from https://ttu-ir.tdl.org.

Parker, N. R. (2007). *A team-based learning model to improve sight-singing in the choral music classroom* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 3256882).

Plotnik, R. (2009). *Psikolojiye giriş* (Trans. T. Geniş). İstanbul: Kaknüs Yayınları.

Potts, S. D. (2009). *Choral sight-singing instruction: An aural-based ensemble method for developing individual sight-reading skills compared to a non-aural-based sing-singing method* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 3387446).

Pratt, G. & Henson, M. (1987). Aural teaching in the first year of tertiary education: An outline for a course. *British Journal of Music Education*, *4*(2), 115-138.

Rifkin, D. & Urista, D. (2006). Developing aural skills: It's not just a game. *Journal of Music Theory Pedagogy,* 20, 57-78.

Scandrett, J. F. (2005). *The efficacy of concept mapping in aural skills training* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 3188996).

Scheele, J. (1993). *Hochschul-Gehörbildung*. Hamburg: Diplomica Verlag.

Seifert, K. & Sutton, R. (2009). *Educational psychology* (2nd Edition). Zurich: A Global Text.

Senemoğlu, N. (2005). *Gelişim, öğrenme ve öğretim: Kuramdan uygulamaya* (12th Edition). Ankara: Gazi Kitabevi.

Sevgi, A. (1982). *Gazi yüksek öğretmen okulu müzik bölümü müziksel işitme okuma yazma eğitimi I. ve II. yarıyıllarında kullanılacak kaynak yöntem ve araç gereçler üzerine bir araştırma* (Thesis for Assistantship, Gazi Yüksek Öğretmen Okulu, Ankara).

Sevgi, A. (2000). *Çoksesli dikte ve okuma parçaları*. Ankara: Yurtrenkleri Yayınevi.

Shanefield, A. (2011). *A qualitative investigation of the attitudes and self-perceptions of music theory faculty not trained in teaching pedagogy on their classroom effectiveness* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No.3452449).

Sisley, B. A. (2008). *A comparative study of approaches to teaching melodic dictation* (Master Thesis). Retrieved from https://etd.ohiolink.edu/

Slavin, R. E. (1980). Cooperative learning. *Review of Educational Research*, *50*(2), 315-342. DOI:10.3102/00346543050002315

Slavin, R. E. (1987). *Cooperative learning: Student teams* (2nd Edition). Washington: National Education Association Publication.

Slavin, R. E. (1990). *Cooperative learning: Theory, research, and practice*. California: Prentice Hall.

Slavin, R. E. (2006). *Educational psychology: Theory and practice* (8th Edition). Boston: Pearson Education, Inc., Allyn and Bacon.

Smialek, T. & Boburka, R. R. (2006). The effect of cooperative listening exercises on the critical listening skills of college music-appreciation students. *Journal of Research in Music Education*, *54*(1), 57-72.

Söker, S. (1998). *İşbirlikli (ortak çalışma yoluyla) öğretmenin şarkı öğretimine etkileri* (Abstract of Master Thesis, Marmara University, İstanbul). Retrieved from http://tez2.yok.gov.tr/

Sözen, İ. (2012). *İşbirlikli öğrenme yaklaşımı ile yapılan toplu bağlama öğretiminin performans ve tutuma etkisi* (Doctoral Dissertation, Abant İzzet Baysal University, Bolu). Retrieved from http://tez2.yok.gov.tr/

Spencer, H. S. (1947). Ear training in music education. *Music Educators Journal*, *33*(4), 44+46.

Suwantarathip, O. & Wichadee, S. (2010). The impacts of cooperative learning on anxiety and proficiency in an EFL class. *Journal of College Teaching and Learning*, *7*(11), 51-57.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics* (5th Edition). Boston: Pearson/Allyn&Bacon.

Tavşancıl, E. (2006). *Tutumların ölçülmesi ve SPSS ile veri analizi* (3rd Edition). Ankara: Nobel Yayınevi.

Therrien, M. C. (1997). *Guidelines for the instructional design of technological and cooperative applications in a music program* (Master Thesis). Available from ProOuest Dissertations and Theses database (UMI No. MQ40234).

Turgut, M. F. (1984). *Eğitimde ölçme ve değerlendirme metotları* (3rd Edition). Ankara: Saydam Matbaacılık.

Uysal, G. (2004). *İlköğretimde işbirlikli öğrenmenin müzik öğretiminde sınıf atmosferi ve şarkı söyleme becerileri üzerindeki etkisi* (Master Thesis, Dokuz Eylül University, İzmir). Retrieved from http://tez2.yok.gov.tr/

Valentino, V. R. (1988). *A study of achievement, anxiety, and attitude toward mathematics in college algebra students using small group interaction methods* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 8905132).

Wine, J. (1971). Test anxiety and direction of attention. *Psychological Bulletin*, *76*(2), 92-104.

Woolfolk, A. E. (1993). *Educational psychology* (3rd Edition). Boston: Allyn and Bacon.

Wunsch, I. G. (1973). Brainwriting in the theory class: The importance of perception in taking dictation. *Music Educators Journal*, *60*(1), 55-59.

Yükseköğretim Kurulu [YÖK]. (2006). *Eğitim fakültesi öğretmen yetiştirme lisans programları: Müzik öğretmenliği lisans programı.* Ankara: YÖK Yayınları.

Zanden, J. W. V. & Pace, A. J. (1984). *Educational psychology: In theory and practice* (2nd Edition). New York: Random House.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(4);376-392

A Content Analysis Study on the Use of Analytic Hierarchy Process in Educational Studies

Muhittin ŞAHİN* Halil YURDUGÜL**

Abstract
In this study, it is aimed to examine the studies based on the AHP (Analytic Hierarchy Process) method in the field of education and to present the researcher's perspective on how to use the AHP method in the field of education. Within the scope of this aim, firstly the AHP method was introduced with a sample application and then the results were interpreted. The other aim of the research; studies which based on the AHP methods in the field of education in the last five years have been examined through content analysis. AHP; is one of the *"Multiple Criteria Decision Making (MCDC)"* methods that can determine the priority or weights among the criteria and alternatives based on comparative judgments. The content analysis conducted within the scope of the research was carried out in the context of eight criteria determined by the researchers. According to the results of the analysis; the AHP method has shown an increasing tendency compared to the years, and usually is used for determining and prioritizing teaching priorities. Especially in Asia Pacific countries, the AHP method is used much more intensive. Another result is that the AHP method is used to make group decisions rather than individual decisions. It has been seen that the research has been done especially with undergraduate students. In addition to these, there are lots of studies with academicians and experts.

Key Words: Analytic hierarchy process, educational studies, content analysis, decision making algorithms

INTRODUCTION

Decision making is one of the indispensable component of human life. Because we need to make a decision at every stage of our lives and in any situation. However, the first condition for decision making involves multiple alternatives, the person tries to determine the most appropriate alternative for him / her based on more than one criterion. If this process is tried to explain this process via a simple example; when the individual wants to choose any university or department for higher education the individual pass through a decision-making process. There are many criteria that influence this decision-making process: location of the university, facilities, education quality etc. As you can see, many criteria are influencing the decision to choose the university. One of the most critical points in making a decision is determining the important criteria that influence decision making (Saaty, 1990). In this context, the decision-making process can be based on the individual's perceptions, predictions and also can explained via a mathematical model. Multiple criteria decision making is an analytical method used to rank, classify, or select alternatives according to the criteria specified when there are multiple criteria. Especially this method, which is widely used in business, politics, engineering, agriculture and economics (and nowadays decision-support systems), unfortunately does not seem to have much use in the educational field. Within the scope of this research, the studies which is in educational field via AHP have been examined.

The Purpose of the Study

In this study, it is aimed to examine the studies made using the AHS method in the field of education and to present the perspective of the researchers about how to use the AHS method in the field of

*Res. Ass. Dr., Ege University, Faculty of Education, Izmir-TURKEY, muhittin.sahin@ege.edu.tr, ORCID ID: https://orcid.org/0000-0002-9462-1953
**Prof. Dr., Hacettepe University, Faculty of Education, Ankara-TURKEY, yurdugul@hacettepe.edu.tr, ORCID ID: https://orcid.org/0000-0001-7856-4664

To cite this article:
Şahin, M., & Yurdugül, H. (2018). A content analysis study on the use of analytic hierarchy process in educational studies. *Journal of Measurement and Evaluation in Education and Psychology, 9*(4), 376-392. DOI: 10.21031/epod.373784

Received: 02.01.2018
Accepted: 08.08.2018

education. For this purpose; a) AHS method was introduced, b) AHP method was elaborated by a sample application, c) the AHP studies in the field of education were examined by content analysis and a perspective was established with related researchers.

Analytic Hierarchy Process-AHP

The AHP is one of the multiple decision-making methods that model decision-making processes mathematically and are used to solve complex problems (Saaty, 1980). Although the AHP using since 1980, decision-making processes were already known with comparative judgment and similar scaling techniques. In particular, it is possible to say that the law of comparative judgment was first put forward by Thurstone in 1927. Alternatives in comparative judgment; are compared in the form of larger, better, more negative, better-looking, and the alternatives are shown on a number line as a result of the analyzes (Details: Turgut & Baykul, 1992). In essence AHP is also based on comparative judgments. But, it seems that the scaling techniques and first order decision making techniques do not include the influence of the criteria that are effective in the decision making process. The AHP aims that solving the hierarchical model by including in the model the criteria that are effective in the decision making process by adding second or higher order layers to the scaling techniques.

Unlike multi-criteria decision making algorithms (TOPSIS, ELECTRE, UTA, PROMETHE, etc.); The AHP aims to combine qualitative and quantitative factors and arrive at a single judgment (Alsamaray, 2017). Advantages of the AHS method; a) use of hierarchical and ratio scales, b) comparisons of intuitive, qualitative, quantitative and rational factors, c) comparison of both criteria and alternatives according to criteria and d) solving decision problems which have objective and subjective criterias (Bhutta & Huq, 2002).

Another advantage of the AHP is that can be used both individually (to be applied to one person) and in group decisions. In the process of obtaining individual decisions, there are some algorithmic operations on comparison matrices, but in making group decisions there are some differences. Because there are more than one individuals when group decision is made, there are naturally more than one comparison matrix. These comparison matrices are reduced to a single matrix. Geometric averaging is often used when this reduction is done (Saaty, 2008).

Because of inherently AHP based on comparative judgment all alternatives and criteria are compared with each other in pairs. It's decided to according to the eigen values which is the result of decomposing the obtained matrices.

The AHP method also has a conceptual hierarchical structure. This structure was constructed by Saaty (1990); a) establishment of hierarchical structure of problem, b) determination of comparative judgments decisions and c) determination of priorities. Zahedi's structure (1986) is very similar to Saaty as; a) setting up a decision hierarchy, b) collecting data with comparative judgment, c) using eigen values to calculate relative weights and d) obtaining a range of ratings for alternatives. If the AHP method will be used, it must be followed this process. The steps of the AHP are much more detailed by Timor (2011) and Esen (2008);

- Identification of the decision problem and determination of the goal,
- Determination of appropriate decision criteria,
- Determination of the alternatives,
- Constructing of the hierarchical structure of the decision problem,
- Comparison of criteria for each level of the hierarchy and determination of importance levels,
- Comparative judgment of alternatives according to the criteria and calculation of priorities,
- Calculation of corresponding index,
- Sorting alternatives according to relative priorities,
- And as a last step consistence analysis.

Within the scope of the research AHP method was used as; a) the identification of the purpose of the decision and the constructing of the hierarchical structure, b) making comparative judgments and c) determination of priorities.

Step 1

At this step, the problem is identified and the hierarchical structure is constructed. The most important and priority step in the AHP is constructing the hierarchical structure (Zahedi, 1986). Information about how to construct the hierarchy is given in Figure 1.

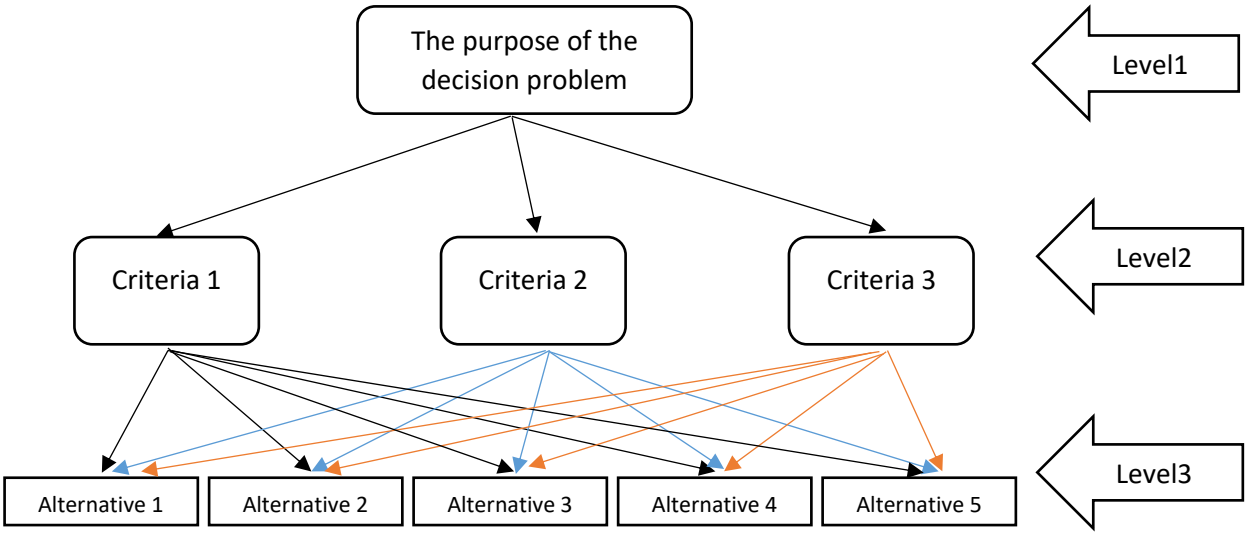


Figure 1: The Hierarchical Structure of AHP

As shown in Figure 1, there are several levels in the AHP process. At the top of the hierarchy is the purpose of the decision problem. In the second level, the criteria set for this purpose and in the third level, alternatives which will be determined the priority order according to these criteria. The AHP structure can be manipulated as desired by the investigator or researchers. For example, a hierarchical structure can be created by adding sub-criteria under this criterion.

Step 2

At this step, the data is collected via the data collection tool which is created in accordance with the hierarchical structure. Comparative judgment of criteria and alternatives in the hierarchical structure are made. In order to make these comparative judgments, a scale with 17 bipolar and equally spaced units is used. This scale is a similarity to semantic differential scale. Descriptions of units of this scale are referred to as " *Intensity of Importance*" and these intensity of importance are given in Table 1

Table 1. Intensity of Importance Table (Saaty, 1990)

Intencity of importance	**Definition**	**Explanation**
1	Equal Importance	Both factors have the same importance
3	Moderate Importance	According to experience and judgment is more one factor important than the other.
5	Strong Importance	One factor is strongly more important than the other.
7	Very Strong Importance	One factor is strongly preferred at a higher level than the other.
9	Extreme Importance	One of the factors is very important at a very high rate.
2,4,6,8	Intermediate Values	These are the intermediate rates, they use when compromise is needed.

One of the points to be noted in the comparison is that the alternatives are repeatedly compared according to each criterion, not just once. If we give an example through our model; Alternative1, Alternative2, Alternative3, Alternative4 and Alternative5 are repeatedly compared according to the first criterion, the second criterion and the third criterion.

Step 3

The third step is to determination of the priorities. For this firstly, comparison matrices are created. For example, if there are 3 different criteria (fuel consumption, performance, comfort) in deciding which of the 5 different car models (alternatives) will be taken, 3 * 3 comparison matrix for the criteria, and a 5 * 5 comparison matrix for the alternatives have been created. Then priority calculations are made based on these matrices. At the last step, consistency analysis is performed to obtain the validity of the results and then the results are reported. The comparison matrices for the criteria are given in Table 2.

Table 2. Comparison Matrix Structure

Criterias	Criteria1 (C1)	Criteria2 (C2)	Criteria3 (C3)
Criteria1 (C1)	-	C1-C2 comparison	C1-C3 comparison
Criteria2 (C2)	C2-C1 comparison	-	C2-C3 comparison
Criteria3 (C3)	C3-C1 comparison	C3-C2 comparison	-

The package programs can be used AHP method analyze, and also can be made manually step by step.

Sample application: AHP used for university selection

In this example, the AHP process and approaches of students in university selections are examined. Hence, the decision problem of the AHP method is the university selection and university selection is placed at the top of the hierarchical structure (Figure 2). The AHP method can be applied to only one person, event or situation and also it can be applied multiple persons, event or situationWithin the scope of this study, AHP method has been employed in 3 steps as a) determination of decision problem and establishment of hierarchical structure, b) comparative judgment and c) determination of priorities.

Step 1

The hierarchical structure of the research as shown in Figure 2.

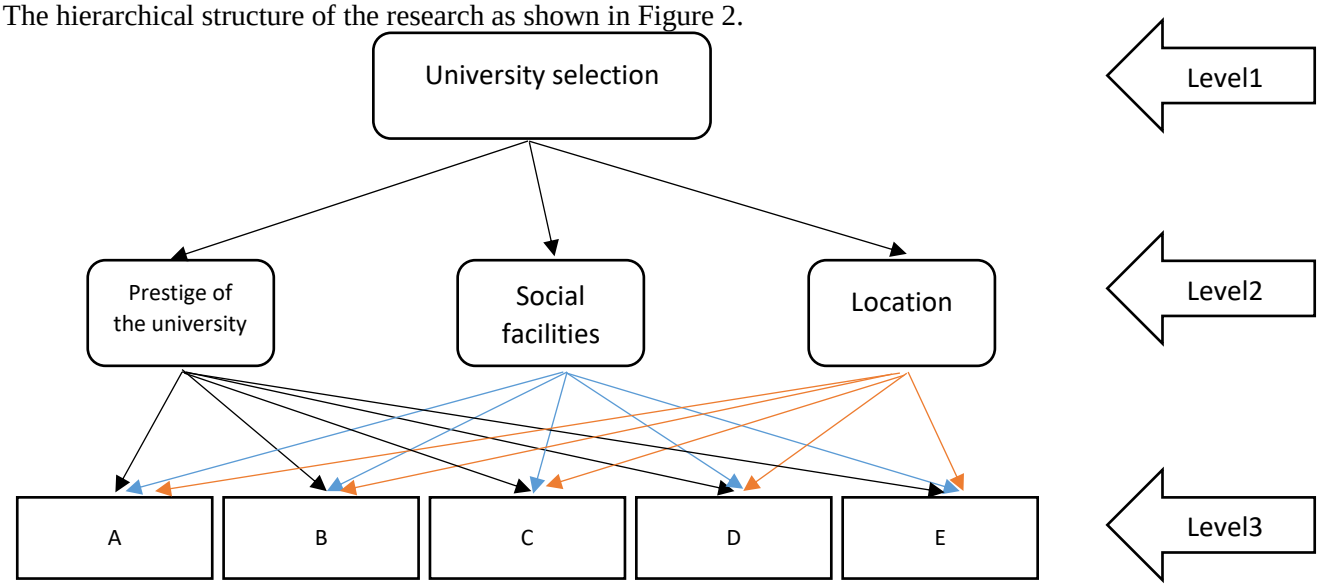


Figure 2. AHS Hierarchical Structure of University Selection

As shown in Figure 2, the university selection was identified as the main problem and placed at the top of the hierarchical structure. Then the prestige of the university, the social facilities of the university and the location of the university were determined as criteria. Finally, according to these criteria; Universities A, B, C, D and E are determined as alternative. After the determination of the first step of decision-making and the establishment of the hierarchical structure have been completed, the second step, comparative judgment, has been passed.

Step 2

In the second step, the data obtained from the comparative judgment is placed in the comparison matrices. An example is given in Figure 3 to show how comparative judgment are made.

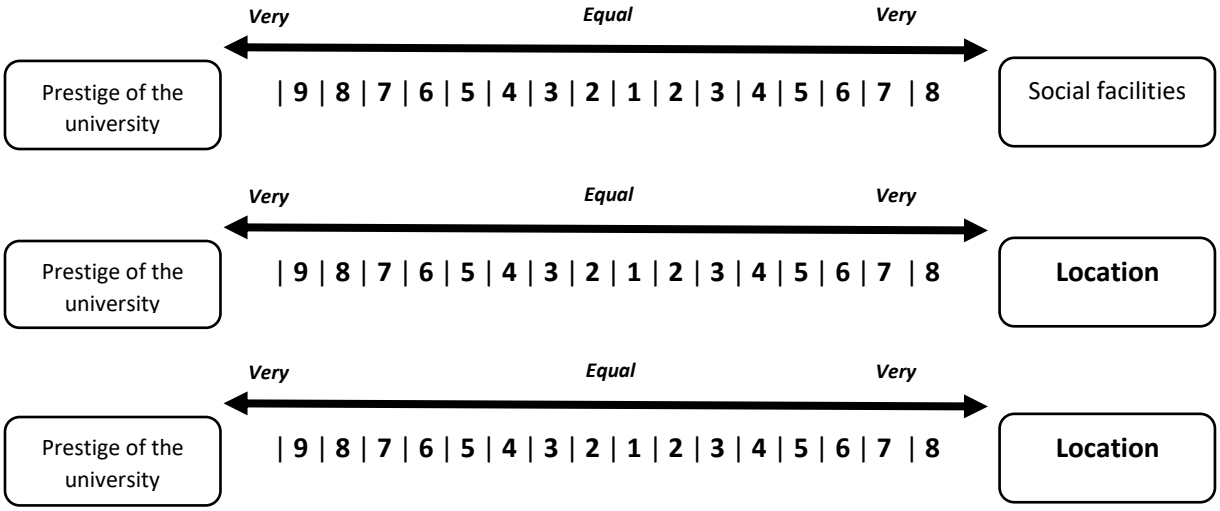


Figure 3. Comparative Judgment Structure of Criteria

Comparative judgments are read from left to right. For example, if the answer of the individual is to the left of the expression "1" which expresses equality, it is "9,8,7,6,5,4,3,2" and f it is on the right side of this expression it is "1 / 2,1 / 4/5, 1/6, 1/7, 1/8, 1/9 ". After the comparative judgment were made, a matrix was created which the criteria were compared with each other. This matrix is given in Table 3.

Table 3. Comparison Matrix of Criteria

Criteria	Prestige of the university (C1)	Social facilities (C2)	Location (C3)
C1	1	9	4
C2	0,111	1	0,20
C3	0,25	5	1

The comparison matrix is a symmetric matrix. The diagonal values are "1". The comparison of the C1 criterion with C2 is "9" and the comparison of the C2 criterion with C1 is “0,111 (1/9)”. After the criteria matrix, the alternative matrix is constructed which include comparative judgments of the alternatives based on each criterion. The matrix is shown in Table 4.

Table 4. Comparison Matrix of Alternatives Based on Prestige of the University

Criteria	A (A1)	B (A2)	C (A3)	D (A4)	E (A5)
A1	1,00	2,00	5,00	7,00	9,00
A2	0,50	1,00	3,00	1,00	9,00
A3	0,20	0,33	1,00	0,20	5,00
A4	0,14	1,00	2,00	1,00	8,00
A5	0,11	0,11	0,20	0,13	1,00

The comparison matrix for the alternatives was also established for social facilities and the university's location criteria. These values which are on the tables represent the real data obtained from the implementation of the example problem situation.

Step 3:

In the third step, mathematical operations were performed on the comparison matrices and eigen values were determined. Then consistency analysis was performed. As a final step, priorities have been determined.

- First of all, normalized matrix is calculated.
- For this purpose, first the column values are sum and then the normalized matrix is calculated by dividing each element in the column by the column sum.
- The vector of priorities is calculated by taking the average of each line in this matrix. At this stage, all the priorities matrices have to be obtained.
- The priorities vector is multiplied by the comparison matrix given at the beginning and the matrix of all priorities is calculated.

The normalized matrix for the criteria, the priorities vector and all priorities matrix for are given in Table 5.

Tablo 5. Normalized Matrix, Priority Vector and All Priorities Matrix

	C1	C2	C3	Priority Vector	All Priorities Matrix
Prestige of the university	0,735	0,6	0,769	0,701	3,1475
Social facilities	0,082	0,07	0,038	0,062	3,0113
Location	0,184	0,33	0,192	0,236	3,058

After the matrix calculations is completion, the consistency index is calculated. Equation 1 is used to calculate the consistency.

$$\mathrm{CR} = \frac{\mathrm{CI}}{\mathrm{RI}} \quad \text{Equation 1}$$

CI: Consistency index

RI: Random Consistency Index

Equation 2 is used for the consistency index calculation.

$$\mathrm{CI} = \frac{(\lambda_{\max} - \mathrm{n})}{\mathrm{n}-1} \quad \text{Equation 2}$$

λ_{max} refers to maximum eigen value and n refers to number of criteria or alternatives. The random consistency index is given in Table 6.

Table 6. Random Consistency Index Table (Saaty, 1980)

n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Random Consistency Index	0,00	0,58	0,90	1,12	1,24	1,32	1,41	1,45	1,49	1,51	1,48	1,56	1,57	1,59

The consistency index, the random consistency index and the consistency ratio which is obtained according to these calculations are given in Table 7.

Table 7. Consistency Index, Random Consistency Index ve Consistency Ratio

CI	0,0361
RI	0,5800
CR	0,0623

The consistency ratio should be less than 0,1, otherwise an attempt should be made to increase consistency (Saaty, 1990). It can be said that the consistency rate which is obtained (0.0623) is below the desired value and the consistency is acceptable.

In the analysis phase, the calculation has been made just for the criteria. These calculations are made in the same way for all alternatives based on each criterion. The priorities which is identified for the criteria are presented in Fig.

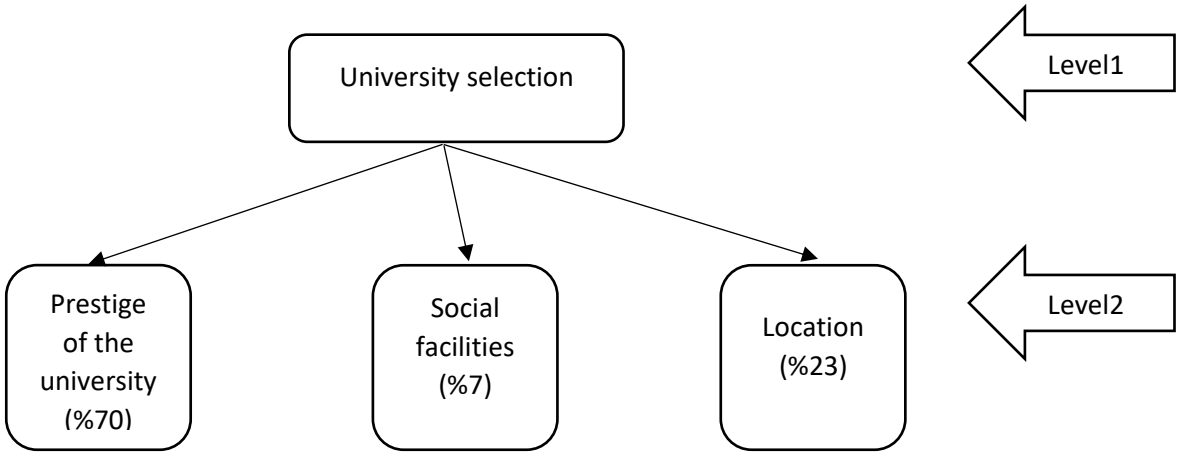


Figure 4. The Priorities for the Criteria

As shown in Figure 4, this person's first priority for university selection was determined theprestige of the university (%70), the second is location of the university (%23), and the last priority is college social facilities (%7). These results are the second level results. The third level results will be evaluated according to three different criteria. The priorities of the student based on the prestige of the university for the five universities are given in Figure 5.

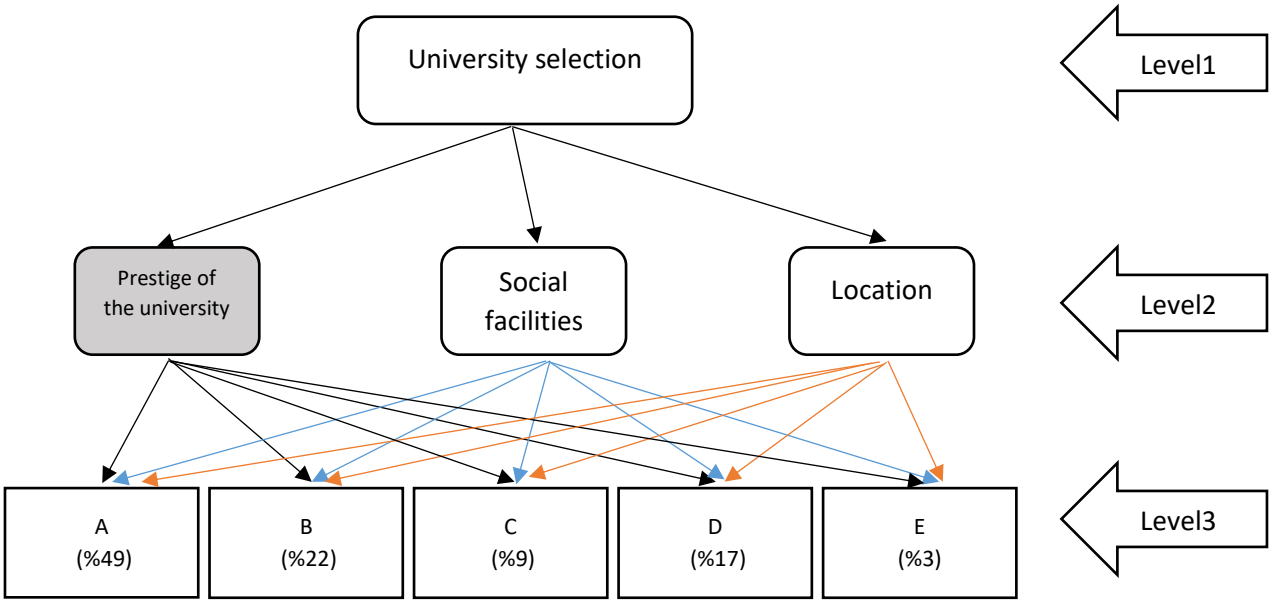


Figure 5. Priorities of University's Based on Prestige of the Univesity Criterian

As shown in Figure 5, this person's first priority for university selection based on prestige of the university criterion was determined A (%49), the second is B (%22), the third is D (%17), the fourth is C (%9) and the last priority is E university (%3). The priorities of the student based on the social facilities criterion for the five universities are given in Figure 6.

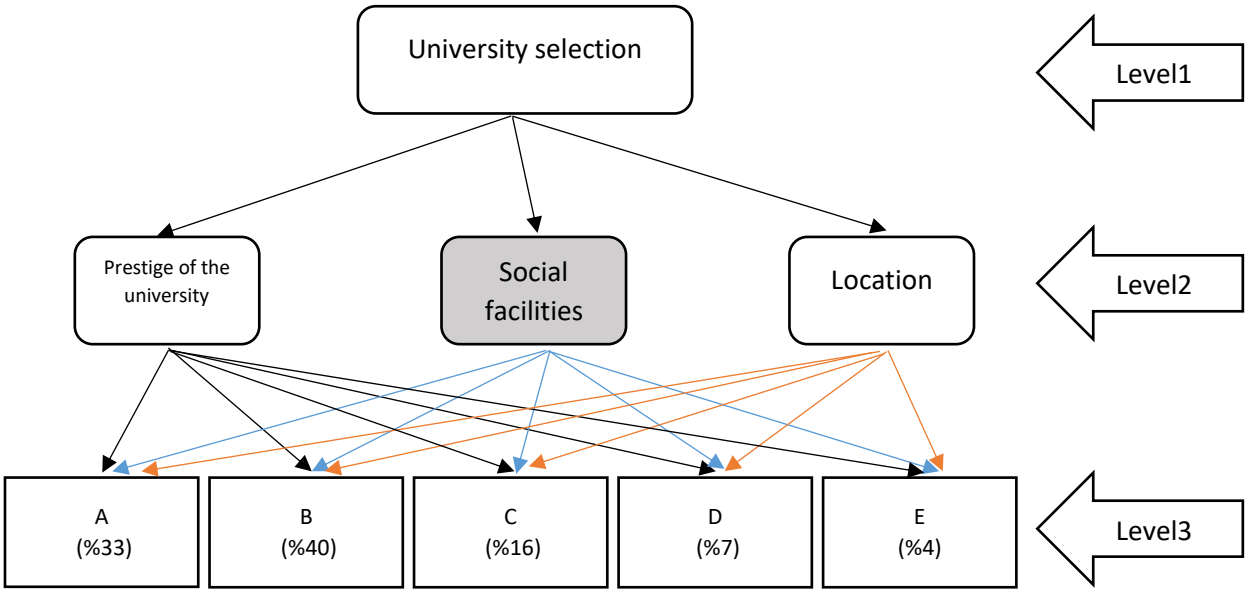


Figure 6. Priorities of University's Based on Social Facilities Criterian

As shown in Figure 6, this person's first priority for university selection based on social facilities was determined B (%40), the second is B (%33), the third is A (%16), the fourth is C (%7) and the last priority is E university (%4). Finally, the priorities of the student based on the location criterian for the five universities are given in Figure 7.

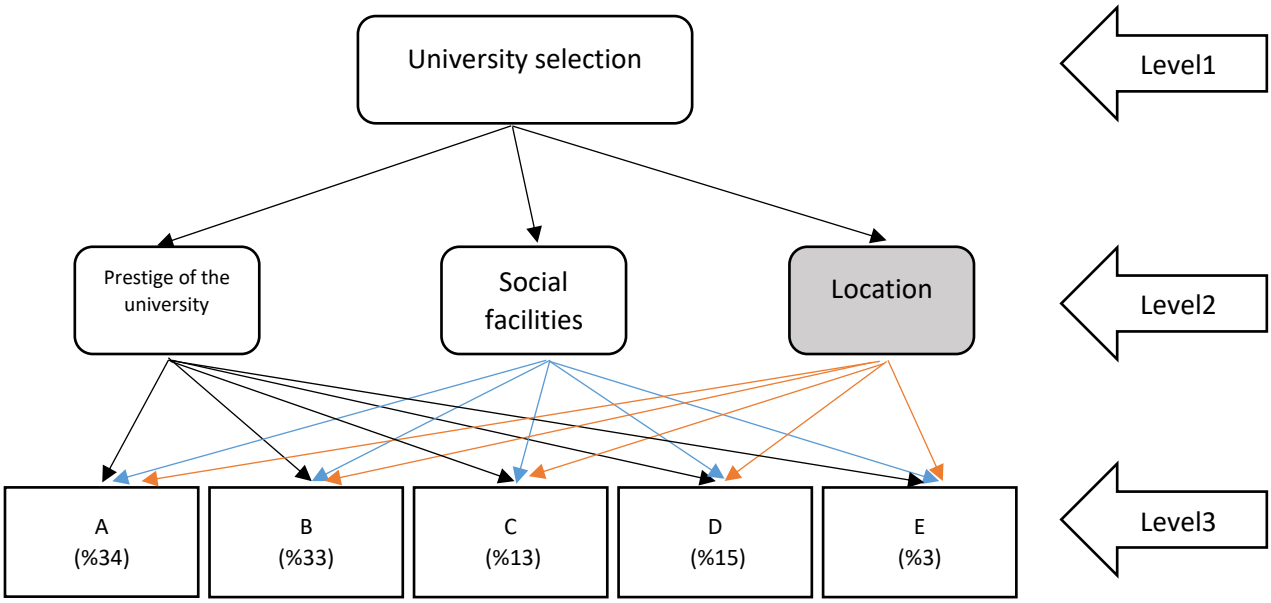


Figure 7. Priorities of University's Based on Location Criterian

As shown in Figure 7, this person's first priority for university selection based on location of the university criterian was determined A (%34), the second is B (%33), the third is D (%15), the fourth is C (%13) and the last priority is E university (%3). The priorities of a person or a group can be determined via AHP method. In addition to this, the AHP method can be used to determine the weights of the criteria or to sort them. In order to determine the situation of AHP studies in the field of education, the studies were examined with content analysis

METHOD

Content analysis was used as a method in this part of the research. Content analysis classifies texts by reducing them into interrelated and manageable data sets (Weber, 1990). Content analysis can be performed in four steps ; a) collecting data, b) data coding, c) finding themes and d) arrangement of codes and findings identification (Yıldırım ve Şimşek, 2004). Within the scope of the research, first of all literature rewiev was made.

- The properties which used for the literature review are as follows. “Analytic Hierarchy Process + Education” in web of science database,
- “Analitik Hiyerarşi Süreci + Eğitim” in Google Scholar.Last five years’ studies beetwen 2013-2017,
- Studies in education fields,
- Research that can be accessed from the databases provided by the university,
- The publication language is Turkish and English.

As a result of the literature rewiev, 42 articles were included in the content analysis and examined. While content analysis is performed, coding is performed according to previously determined criteria. These criteria;

- The purpose of the study
- The purpose of the AHP method
- Group / Individiual decision
- Year of the study
- The region where the study was conducted
- The level of the AHP

- Sample size
- Study group level

The results of the content analysis is presented in findings section.

FINDINGS

In this section findings which based on content analysis are presented.

Findings of the Purpose of the Studies

The aims of the studies used AHP method is system development, evaluation, selection and prioritization. Findings for the purpose of the studies are given in Table 8 in detail.

Table 8. Findings for the Purpose of the Studies

Purpose of the Study	Product of the Study	Frequency
System development	Mathematical model	1
	Decision support system	3
	Quality evaluation system	1
	University effectiveness system	1
Evaluation	Environment and material evaluation	3
	Evaluation of the instruction	9
Selection	Statistical software selection	1
	City selection for appointment	1
	Student selection	2
	Course selection	3
	University selection	2
Determination of the priorities	For instruction	12
	Career	1
	Infrastructure	2
Total		42

As seen in Table 8, the AHP method firstly was used for determination of priority (15), secondly for evaluation (12), thirdly for selection select (9) and finally for the system developing (6).Environment and material evaluation includes product, evalution of distance learning and gamification. Evaluation of the instruction includes evaluation of universities, students, academics instruction performance, method, etc.

Findings of the Purpose of the AHP Method

In some studies, a different multi-criteria decision making method has been used in addition to the AHP method. Therefore, a title for the purposes of use of the AHP method has been included. Findings of the purpose of using the AHP method are presented in Table 9.

Table 9. The Purpose of the Using AHP Method

Purpose of the AHP	Frequency	Percent
Selection	4	%9,52
Ranking	18	%42,86
Weight determination	6	%14,29
Evaluation	10	%23,81
System development	4	%9,52
Total	42	%100

As shown in Table 9, the AHP method was used in order to rank the criteria, sub-criteria or alternatives (42.86%) in the most studies. Secondly for evaluation (23,81%), thirdly determination of weight

(14,29%), and lastly for selection and system development (9,52%). Especially the studies which used AHP method for determination of weights, after this procedure another method is utilized and the study is carried out in this way.

Findings of Group / Individual Decision

AHP method is used for determination of group or individual decision. For this reason, there's a sub title about group or individual decision in this study. Using this method, it is possible to determine the priorities or trends of a person or a group or a university. Findings about this is given in Table 10.

Table 10. Findings of Group/Individual Decision

Individual/Group	Frequency	Percent
Individual	6	14,29
Group	36	85,71
Total	42	100,00

As shown in Table 10, most of the studies which is used the AHP method have been used to determine group priorities. This method was used in 14.29% of the studies to determine the individual and 85.71% to determine the group priorities.

Findings About Year of the Study

The studies which are publicated between 2013 and 2017 are examined. The distribution of the studeis by years is given in Figure 8.

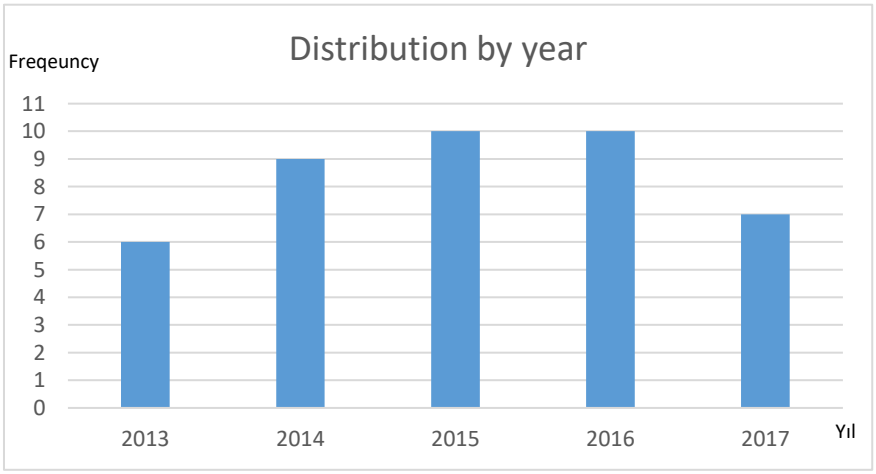


Figure 8. Findings About Studies' Years

As shown in Figure 8, six studies were conducted in 2013, nine studies in 2014, ten studies in 2015 and 2016, and seven studies in 2017 (Studies in 2017 is limited to research conducted until the beginning of December 2017. Because the literature rewiev was done at the beginning of December). When we look at the distribution of studies by years, the studies are increasing year by year. This can be interpreted as AHP method is increasingly being used more and more in educational research.

Findings About Studies Region

The regions where the research was conducted were examined under four regions. The Europan region includes; Italy, Spain, Swiss, Greece and Turkey. The America region includes; United States of America, Chile, Canada and Mexico. The Asia region includes; South Korea, China, Taiwan,

Kazakhistan and Pakistan. The African region includes; Saudi Arabia, United Arab Emirates, India and Malezia. Detailed information about the studies which carried out in these regions is given in Table 11.

Table 11. Findings About the Studies Region

Region	Frequency	Percent
Europe	16	0,38
America	3	0,07
Asia	17	0,41
Africa	6	0,14
Total	42	100,00

As shown in Table 11, it is seen that the most studies in Asia (%41), secondly Europe (%38), thirdly Africa (%14) and lastly America (%7) are the most investigated. %88 of studies in the Asian region (15) were conducted in Asia-Pacific countries (South Korea, China and Taiwan). %63 of the study in Europe region (10) were conducted in Turkey. Because, the literature rewiev was conducted via not only English but also Turkish keywords.

Findings About Level of the Studies

The first step of the AHP method is to determine the decision problem and to create a hierarchical structure. According to the problem situation, the number of levels of the structure can be completely determined by the researcher(s) and manipulated. So this criteria determined by the researchers in this study. Detailed information on the level of studies conducted is given in Table 12.

Table 12. Findings About Level of the Studies

Number of Level	Frequncy	Percent
2	4	%9,52
3	22	%52,39
4	14	%33,33
5	2	%4,76
Total	42	%100

As it is seen in Table 12, it is seen that the most of the studies have 3 levels (52.38%) and there are not many studies which have 5 levels. It has been found that studies consisting of four levels generally consist of criteria, sub-criteria and alternatives. Studies consisting of three levels include studies on criteria and sub-criteria. There are criteria in two level studies and findings about these criteria.

Findings About Sample Size

In the AHP method, it is possible to perform both group and individual calculations and decisions. The study group intervals were determined by the researchers. Detailed information about the sample size is presented in Table 13.

Table 13. Sample Size of the Studies

Sample Size Interval	Frequecy	Percent
1	3	%7,14
Between 2 and 100	23	%54,76
Between 101 and 1000	7	%16,67
1000 and more	1	%2,38
No information	8	%19,05
Total	42	%100,00

As shown in Table 13, the size of the study group varies. The group size were determined by the researchers as four interval. Another group is the ones that do not mention how many people they study with in their research (19,05%). Individual calculations are possible via AHP method, so there is also a single person studies (7,14%). Findings show that the preferred group size is usually 2-100 participants (54.46%) in the research. It is seen that the ratio of the studies' group size between 101-1000 is %16,67. The studies which have 1000 and over participants are very few (%2,38).

Findings About Study Group Level

It is possible to conduct research with all the stakeholders of a problem situation via AHP method. Detailed information about level of the group is presented in Figure 9.

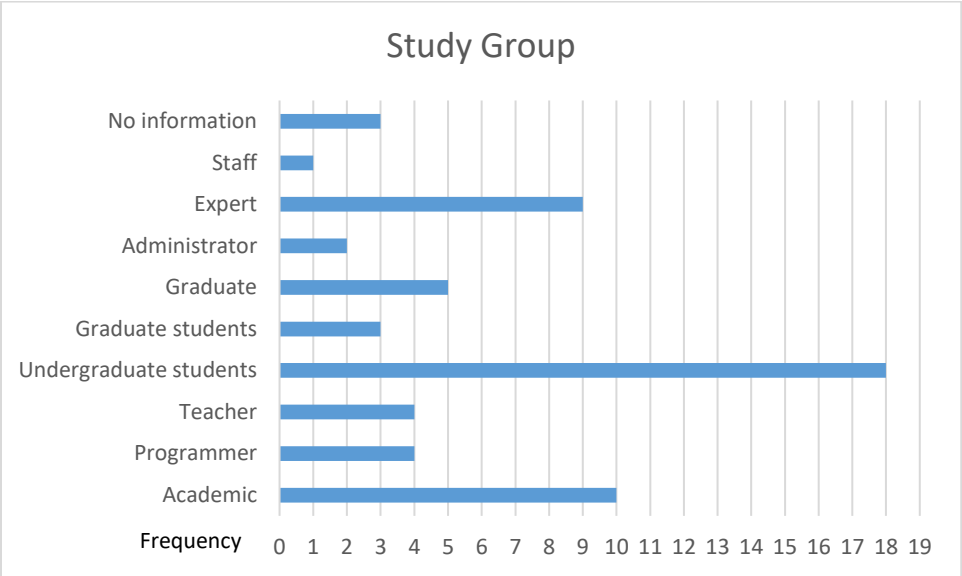


Figure 9. Level of the Study Group

As can be seen in Figure 9, studies using the AHP method were frequently carried out with undergraduate students. The academics and experts are also seen as very preferred groups. One of the reasons for this is that the literature rewiev is done with the "education" keyword. There were not so much studies which conducted with administrator and staff in the education field. Some studies were conducted with different and multiple study groups. There is no information about study group level in some studies.

CONCLUSION and DISCUSSION

Within the context of the research, AHP that one of the most multi-criteria decision-making techniques, has been introduced, applied through an authentic sample and presented in the content analysis of research conducted in the field of education between the years 2013-2017. AHP method is one of multi-criteria decision making methods and has a hierarchical structure. Information about this method is presented in the introduction section of this study. Then, this method was explained via an authentic sample. The university selection for the authentic sample is considered as a decision problem. For solving this decision problem firstly, a) determined the purpose of the decision and established a hierarchical structure, b) made comparative judgment and c) determined priorities.

When the results of the content analysis are examined, it is found that the studies which using the AHP method for system development, evaluation, selection and determination of priorities. System development studies includes decision support systems and quality evaluation systems. Especially in the development of decision support systems, the AHP method offers researchers opportunities. Because

there are multiple criteria and alternatives in decision making and this situation make the process complicated. Decision support systems can be developed, which can help individuals make decisions, via AHP method. The evaluation studies consist of product and material evaluation and evaluation of instruction. The selection studies consist of student selection, course selection and university selection. It has also been observed that these studies are usually done on paper, but not via an electronic system. It's suggested to the researchers firstly development an electronic system based on AHP. It seems that there are lots of studies which used AHP method about sociology, politics, auatomation profit and loss analysis, budget planning etc (Zahedi, 1986). There are also studies used in education field (Wang, 2014; Weng, Zhang & Liu, 2014; Thanassoulis, Dey, Petridis, Goniadis & Georgiou, 2017) but it is seen that these studies are limited. Educational studies refer to the teaching and learning process. It is possible to say that the studies which carried out in faculty of education are much more limited. is suggested studies which is used AHP should be conducted in order to determine the learners' need.

The purpose of using the AHP method in studies may differ from the purpose of the study. For example, while the purpose of the research may be to choose the best method, but AHP can be used with the purpose of determining the weights criteria. In the literatuere, the AHP method is used for the purpose of respectively ranking, evaluation, weight determining, selection and system development. There are lots of studies ranking the criteria or alternatives. In studies which using weight determination purposes, the AHP is the first leg of the study; weights of criteria or alternatives are determined via AHP and in the second step a ranking was obtained by the TOPSIS method (Kecek & Söylemez, 2016; Lokare & Jadhav, 2016).

Both individual and group decisions can be made by the the AHP method studies. Comparison matrix uses for individual decision, but for the group decision the matrices have to reduce just one comparison matrix. Geometric mean is used for reducing the matrices. It is generally seen that the AHP method is used for group decision and even if it is a small number AHP is used for individual decision.

It is seen that the distribution of the studies using the AHP method is increasing year by year. The studies which have done in 2017 is limited to the beginning of December 2017. Because the literature rewiev was done by the beginning of December. The most studies were conducted Asia region. And the frequency of the studies that conducted especially in the Asia-Pacific countries is remarkable. Europen region is the second order. The reason of this is the keywords. For the literature rewiev both English and Turkish keywords were used. In order to reveal the situation in our country Turkish keywords were used.

The AHP is an approach that adds a second order level to the scaling techniques and adds criteria to the model and resolves the resulting hierarchical model. Hence the AHP includes more than one level. In the literature it is seen that the studies consisting of three levels are the majority. These studies include usually the criteria and the sub-criteria. In addition, there are three levels of study in which the criteria and alternatives are included. The level of the hierarchical structure can be determined and manipulated by the researcher(s). Thus providing a very flexible structure to the researcher(s). Levels can be determined appropriately for the purpose of the study. In the literature, it is also seen that qualitative studies have been conducted for determining criteria and alternatives (Ertuğ & Girginer, 2014; Chiu, Kao, Pu, Lo & Huang, 2015). The criteria, sub-criteria or alternatives are situated in a hierarchical structure based on findings of the qualitative studies.

The findings about the sample size, it is concluded that the maximum frequency is within the range of 2-100. Besides this, there are also studies carried out with one person. There is also a study which the size of the study group is 1000 and more. Studies which used the AHP method were usually carried out with small groups. The groups priorities or preferences can be determined with the small groups studies. However, it is thought that it is necessary to work with wider working groups in order to reach a general judgment. Most of AHP studies carried out with undergraduate students. Other than this the study groups were comprised of academics and experts. Students and academics are the most studied group because keywords include "education". Especially for the system development studies, AHP can be used for the need analysis which is the first step of the studies. All stakeholders' priority or preferences, differences and similarities can be determined via the AHP method. According to these findings, designs can be

configured. In addition to these, the AHP method is different from the known likert-like scales and is enjoyable by participants. However, it is recommended that it must be implemented with a moderator.

REFERENCES

Ahmad, S. Z., & Hussain, M. (2017). An investigation of the factors determining student destination choice for higher education in the United Arab Emirates. *Studies in Higher Education, 42*(7), 1324-1343. Doi: 10.1080/03075079.2015.1099622

Alsamaray, H. S. (2017). AHP as multi-criteria decision making technique, empirical study in cooperative learning at Gulf University. *European Scientific Journal, ESJ, 13*(13), 272-289. Doi: 10.19044/esj.2017.v13n13p272

Altamirano-Corro, A., & Peniche-Vera, R. (2014). Measuring the institutional efficiency using dea and ahp: The case of a mexican university. *Journal of Applied Research and Technology, 12*(1), 63-71.

Bhutta, K. S., & Huq, F. (2002). Supplier selection problem: A comparison of the total cost of ownership and analytic hierarchy process approaches. *Supply Chain Management: An International Journal, 7*(3), 126-135.

Blanco, M., Gonzalez, C., Sanchez-Lite, A., & Sebastian, M. A. (2017). A practical evaluation of a collaborative learning method for engineering project subjects. *IEEE Access, 5*, 19363-19372.

Certa, A., Enea, M., & Hopps, F. (2015). A multi-criteria approach for the group assessment of an academic course: A case study. *Studies in Educational Evaluation, 44*, 16-22.

Chiu, P. S., & Huang, Y. M. (2016). The development of a decision support system for mobile learning: A case study in Taiwan. *Innovations in Education and Teaching International, 53*(5), 532-544.

Chiu, P. S., Kao, C. C., Pu, Y. H., Lo, P. F., & Huang, Y. M. (2015, July). *The development of a decision support system for successful mobile learning*. In Advanced Learning Technologies (ICALT), 2015 IEEE 15th International Conference on (pp. 114-115), China.

Çiçekli, U. G., & Karaçizmeli, A. (2013). Bulanık analitik hiyerarşi süreci ile başarılı öğrenci seçimi: Ege üniversitesi iktisadi ve idari bilimler fakültesi örneği. *Ege Stratejik Araştırmalar Dergisi, 4*(1), 71-94.

Dai, L., Guo, J., & Zhao, J. (2013). *Application of analytical hierarchy pro-cess on evaluation of teaching quality in farmer distance education platform*. In Proceedings of the 2013 International Conference on Information, Business and Education Technology (ICIBET 2013). Atlantis Press.

Dündar, S. (2008). Ders seçiminde analitik hiyerarşi proses uygulaması. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 13*(2), 217-226.

Ertuğ, Z. K., & Girginer, N. (2014). A multi criteria approach for statistical software selection in education. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 29*(2), 129-143.

Esen, Ö. (2008). *Uygulamalı yöneylem araştırması, yöneticiler için bilgisayar destekli karar modelleri: Excel ile modelleme ve çözüm teknikleri*. İstanbul: Çağlayan.

Fardinpour, A., Pedram, M. M., & Burkle, M. (2014). Intelligent learning management systems: Definition, features and measurement of intelligence. *International Journal of Distance Education Technologies (IJDET), 12*(4), 19-31.

Farid, S., Ahmad, R., Niaz, I. A., Arif, M., Shamshirband, S., & Khattak, M. D. (2015). Identification and prioritization of critical issues for the promotion of e-learning in Pakistan. *Computers in Human Behavior, 51*, 161-171.

Frangos, C. C., Fragkos, K. C., Sotiropoulos, I., Manolopoulos, I., & Gkika, E. (2014). *Student preferences of teachers and course ımportance using the analytic hierarchy process model*. In Proceedings of the World Congress on Engineering (Vol. 2), United Kingdom.

Han, S., Li, Z., & Tang, X. (2014). Study o*f the relationship between tutors and master graduates based on analytic hierarchy process*. 2nd International Conference on Advances in Social Science, Humanities, and Management (ASSHM 2014), China.

Ho, S. Y., Chen, W. T., & Hsu, W. L. (2017). Assessment system for junior high schools in taiwan to select environmental education facilities and sites. *EURASIA Journal of Mathematics, Science & Technology Education, 13*(5), 1485-1499.

Huang, D. F., & Singh, M. (2014). Critical perspectives on testing teaching: Reframing teacher education for English medium instruction. *Asia-Pacific Journal of Teacher Education, 42*(4), 363-378.

Huang, Y., & Shi, Y. (2013, June). *College teachers teaching evaluation model based on ahp-dfs.* In 2013 the International Conference on Education Technology and Information System (ICETIS 2013), China.

Ishizaka, A., & Nemery, P. (2013). *Multi-criteria decision analysis: Methods and software*. John Wiley & Sons. New Jersey.

Kahraman, C., Suder, A., & Cebi, S. (2013). Fuzzy multi-criteria and multi-experts evaluation of government investments in higher education: The case of Turkey. *Technological and Economic Development of Economy, 19*(4), 549-569.

Karaarslan, M. H., & Özbakır, L. (2017). Mühendislik öğrencilerinin kariyer tercihlerinin belirlenmesi. *Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 19*(1), 83-103.

Kecek, G., & Söylemez, C. (2016). Course selection in postgraduate studies through analytic hierarchy process and topsis methods. *British Journal of Economics, Finance and Management Sciences, 11*(1), 142-157.

Kim, N., Park, J., & Choi, J. J. (2017). Perceptual differences in core competencies between tourism industry practitioners and students using analytic hierarchy process (AHP). *Journal of Hospitality, Leisure, Sport & Tourism Education, 20*, 76-86.

Kim, S. (2014). *Decision support model for introduction of gamification solution using ahp*. The Scientific World Journal, 2014, 1-7.

Köksal, G., & Eğitman, A. (1998). Planning and design of industrial engineering education quality. *Computers & Industrial Engineering, 35*(3-4), 639-642.

Lokare, V. T., & Jadhav, P. M. (2016, January). *Using the AHP and TOPSIS methods for decision making in best course selection after HSC*. In Computer Communication and Informatics (ICCCI), 2016 International Conference on (pp. 1-6). India.

Lu, Y. L., Lian, I. B., & Lien, C. J. (2015). The application of the analytic hierarchy process for evaluating creative products in science class and its modification for educational evaluation. *International Journal of Science and Mathematics Education, 13*(2), 413-435.

Madbouly, A. I., Noaman, A. Y., Ragab, A. H. M., Khedra, A. M., & Fayoumi, A. G. (2016). Assessment model of classroom acoustics criteria for enhancing speech intelligibility and learning quality. *Applied Acoustics, 114*, 147-158.

Noaman, A. Y., Ragab, A. H. M., Madbouly, A. I., Khedra, A. M., & Fayoumi, A. G. (2017). Higher education quality assessment model: Towards achieving educational quality standard. *Studies in Higher Education, 42*(1), 23-46.

Oddershede, A., Donoso, J., Farias, F., & Jarufe, P. (2015). ICT support assessment in primary school teaching and learning through AHP. *Procedia Computer Science, 55*, 149-158.

Ognjanovic, I., Gasevic, D., & Dawson, S. (2016). Using institutional data to predict student course selections in higher education. *The Internet and Higher Education, 29*, 49-62.

Pellicer, E., Sierra, L. A., & Yepes, V. (2016). Appraisal of infrastructure sustainability by graduate students using an active-learning method. *Journal of Cleaner Production, 113*, 884-896.

Rombe, E., Allo, P.L.D., Tolla, M.A. & KusumaDewi, S. (2016). *What are the current quality ıssues in higher education*? Proceedings of the 2016 International Conference on Education, Management Science and Economics. Singapore.

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences, 1*(1), 83-98.

Saaty, T. L. (1980). *The analytic hierarchy process.* McGraw-Hill: New York.

Saaty, T. L. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research, 48*, 9-26.

Samut, P. K. (2014). İki aşamalı çok kriterli karar verme ile performans değerlendirmesi: AHP ve TOPSIS yöntemlerinin entegrasyonu. *Anadolu Üniversitesi Sosyal Bilimler Dergisi, 14*(4), 57-67.

Sinem, A., & Arslan, M. (2015). Yabancılara Türkçe öğretiminde dilsel becerilerin gelişimine etkisi bakımından ders materyallerinin önem derecelerinin analitik hiyerarşi süreci (AHS) ile belirlenmesi. *Bartın Üniversitesi Eğitim Fakültesi Dergisi, 4*(2), 711-726. Doi: 10.14686/buefad.v4i2.5000138861

Soba, M., Şimşek, A., Erdin, E., & Can, A. (2016). Ahp temelli vikor yöntemi ile doktora öğrenci seçimi. *Dumlupinar University Journal of Social Science/Dumlupinar Üniversitesi Sosyal Bilimler Dergisi, 50.*

Thanassoulis, E., Dey, P. K., Petridis, K., Goniadis, I., & Georgiou, A. C. (2017). Evaluating higher education teaching performance using combined analytic hierarchy process and data envelopment analysis. *Journal of the Operational Research Society, 68*(4), 431-445.

Thurstone, L.L. (1927). A low of comperative judgement. *Psychological Review, 34*, 273-286.

Tian, Y., Yang, P., Zhang, N., & Yang, G. (2013). *Teaching quality evaluation of a new university mathematics teaching mode-an empirical research.* In Conference: International Conference on Education Technology and Information System (ICETIS). Sanya.

Timor, M. (2011). *Analitik hiyerarşi prosesi*. İstanbul: Türkmen.

Turgut, M. F., & Baykul, Y. (1992). *Ölçekleme teknikleri*. Ankara: ÖSYM Yayınları.

Türkmen, E. G., Güngör, İ., & Erinci, F. (2015). Öğretmenlerin tayin yeri seçiminde analitik hiyerarşi proses uygulaması. *Uluslararası Alanya İşletme Fakültesi Dergisi, 7*(3), 35-49.

Uvalieva, I., Garifullina, Z., Utegenova, A., Toibayeva, S., & Issin, B. (2015). *Development of intelligent system to support management decision-making in education*. In Modeling, Simulation, and Applied Optimization (ICMSAO), 2015 6th International Conference on (pp. 1-7). Turkey.

Venkadasalam, S. (2015). An analytic hierarchy process (AHP) approach to training typology selection based on student perspective: Empirical evidence from Malaysian Maritime Academy. *Asia-Pacific Journal of Business Administration, 7*(2), 140-146.

Wang, L. Y. (2014). Research on evaluation system for comprehensive quality of college and university students based on analytic hierarchy process model. *In Applied Mechanics and Materials*, *678*, 648-652. Trans Tech Publications.

Wang, Y., Li, J., Li, D., & Chen, G. (2015, May). *Analysis of influencing factors on graduate students' achievements in scientific research*. In Control and Decision Conference (CCDC), 2015 27th Chinese (pp. 3188-3191). China.

Weber, R. P. (1990). *Basic content analysis* (No. 49). Sage.

Weng, Y., Zhang, C., & Liu, Y. (2014, May). *Evaluation of teaching quality system designing based on AHP*. In Electronics, Computer and Applications, 2014 IEEE Workshop on (pp. 438-440). IEEE.

Xingfeng, L. I. U. (2017). Performance evaluation of engineering teachers in universities based AHP and fuzzy mathematical methods. *Revista de la Facultad de Ingeniería*, *32*(5), 141-149.

Xu, L. (2013, June). Teaching quality about application of multimedia in higher education. *In 2013 Conference on Education Technology and Management Science (ICETMS 2013).* China.

Yacan, İ. (2016). *Eğitim kalitesinin belirlenmesinde etkili olan faktörlerin bulanık AHP ve Bulanık Topsıs yöntemi ile değerlendirilmesi* (Yüksek lisans tezi, Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü, Denizli).

Yıldırım, A., & Şimşek, H. (2004). Sosyal *bilimlerde nitel araştırma yöntemleri*. Ankara: Seçkin.

Zahedi, F. (1986). The analytic hierarchy process-A survey of the method and its applications. *Interfaces*, *16*(4), 96-108.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(4);393-413

Turkish Prospective Teachers' Attitudes towards the Teaching Profession: A Meta-Analysis Study*

Erkan Hasan ATALMIŞ** Akif KÖSE***

Abstract
This research aims to explore whether prospective teachers' attitudes towards the teaching profession vary across demographic characteristics. A meta-analysis has been conducted for the related studies on prospective teachers' attitudes towards the teaching profession in Turkey. The effect sizes for random effects model have been employed over 103 studies in terms of gender, 26 regarding grade level, 18 for the presence of a teacher in the family, and 11 for the graduated faculties by using Hedges' *g* coefficient. Various methods have been utilized in an attempt to examine publication bias in the meta-analysis, such as the funnel plots, Duval and Tweedie's trim and fill method, and Egger's regression test. The findings revealed that prospective teachers' attitudes towards the teaching profession significantly varied depending on the gender in favor of the females with medium effect while the variables; grade level, the presence of a teacher in the family, and type of faculty, did not significantly change prospective teachers' attitudes towards teaching profession. This indicates that only the gender variable from the demographic characteristics changes prospective teachers' attitude towards the teaching profession.

Key Words: Effect size, meta-analysis, prospective teachers, teaching profession, attitude

INTRODUCTION

Considering the studies conducted within the scope of educational sciences discipline, the number of studies regarding attitude towards teaching profession has increased considerably since 2000's. These studies are generally empirical researches that aim to reveal how the attitude towards teaching profession varies across demographic characteristics. When the results of these studies are examined, it is observed that the findings are different and inconsistent from one another; hence it is hard to obtain generalizable knowledge. In this regard, meta-analysis studies are at the forefront. This research aims to examine whether prospective teachers' attitudes towards teaching profession differ across their demographic characteristics through use of meta-analysis method.

With a view to understanding the significance of the attitude towards the teaching profession, it is essential to scan the definitions made about the concept of attitude from past to present. Fishbein and Ajben (1975) have defined attitude as positive or negative pre-disposition to respond to a stimulus object. On the other hand, Pratkanis and Greenwald (1989) have described attitude not only as a function of the stimulus object, but also as a function of the personality variables and the roles as well as tasks that one must perform in a particular situation. Eagly and Chaiken (1993) identify attitude as a psychological tendency that evaluates a certain entity positively or negatively. These definitions have suggested that attitude is not a behavior but a tendency that prepares for behaviors (Tuncer & Bahadır, 2016). Within the framework of these definitions, the attitude towards the teaching profession can also be defined as the thoughts and feelings that an individual holds in mind regarding the teaching profession (Camadan & Duysak, 2010). In this context, the attitudes of the teachers towards the teaching profession may lead to the emergence of their behaviors necessary for teaching (Emre &

* This study was supported by Kahramanmaras Sütçü Imam University Scientific Research Projects Coordination Unit (Project number: 2017 / 5-16M) and a part of the study was presented at the 1st International Turkish World Strategic Research Congress (TUDSAK).
** Assistant professor, Kahramanmaraş Sütçü Imam University, Faculty of Education, Kahramanmaraş-Türkiye, e-posta: eatalmis@ksu.edu.tr, ORCID ID: orcid.org/0000-001-9610-491X
*** Assistant professor, Kahramanmaraş Sütçü Imam University, Faculty of Education, Kahramanmaraş-Türkiye, e-posta: akifkose@ksu.edu.tr, ORCID ID: orcid.org/0000-002-6961-6052

To cite this article:
Atalmış, E., H., & Köse, A. (2018). Turkish prospective teachers' attitudes towards the teaching profession: a meta-analysis study. *Journal of Measurement and Evaluation in Education and Psychology, 9*(4), 393-413. DOI: 10.21031/epod.410287

Received: 28.03.2018
Accepted: 11.09.2018

Ünsal, 2017; Kartal & Afacan, 2012). The relevant literature has revealed that the individuals who possess a positive attitude towards the teaching profession will make an effort for their professional competency, they will be keen on their profession, they will have real communication with the students and they will create different learning environments (Çeliköz & Çetin, 2004; Demirtaş, Cömert & Özer, 2011; Semerci & Semerci, 2004).

Since attitude is an affective behavior, attitude scales developed by different researchers have been used since 1980s in order to measure attitudes towards the teaching profession, and it appears that the number of these scales have increased in recent years. The attitude towards teaching profession is generally found to have one factor in these scales and the number of items varies between 10 and 34 (Aşkar & Erden, 1987; Başbay, Ünver & Bümen, 2009; Bulut, 2009; Erkuş, Sanlı, Bağlı & Güven, 2000; Semerci, 1999; Üstüner, 2006). More than that, there are also scales that include more than one factor related to the teaching attitude. The attitude scale developed by Ünlü (2011) consists of 23 items and 2 dimensions-"Affection for Profession" and "Concern about the Profession"; the scale developed by Çapa and Çil (2000) has 32 items and three dimensions including "Affection for Profession", "Self-Confidence in Profession" and "Respect for Profession"; the scale developed by Çetin (2006) possesses 34 items and three dimensions-"Affection for Profession", "Value for Profession" and "Compatibility with Profession". Considering the dimensions in these studies, the dimension of "Affection for Profession" is common and corresponds to the factor having the highest percentage of the total variance explained in all three studies. This offers an insight into the fact that the attitude towards the teaching profession is highly related to the professional affection that plays a significant role in shaping professional behavior.

Numerous studies have been conducted on the relationship between teachers' or prospective teachers' attitudes towards teaching profession and different variables. The studies have analyzed the relationship between teachers'/prospective teachers' attitudes towards teaching profession and their pedagogical competency perceptions (Adıgüzel, 2017), self-efficacy perceptions (Bakaç & Özen, 2017; Dadandı, Kalyon & Yazıcı, 2016), their personalities based on adjectives (Aslan & Yalçın, 2013), teaching motivation (Ayık & Ataş, 2014), academic motivation (Bedel, 2015), personal values (Bektaş & Nalçacı, 2012), professional alienation levels (Çağlar, 2013), life-long learning levels (Çam & Üstün, 2016), job satisfaction (Çetin, 2016), professional competency levels (Çetinkaya, 2007), learning styles (Çiğdem & Memiş, 2010), communication skills (Çimen, 2016; Tümkaya, 2016), professional motivation (Çimen, 2016), professional field knowledge (Dikmenli & Çifçi, 2015), occupational self-esteem (Dilmaç, Çıkılı, Işık & Sungur, 2009; Girgin, Akamca, Ellez & Oğuz, 2010), professional anxiety (Doğan & Çoban, 2009), academic dishonesty tendency (Hançer, 2017), contemporary teaching perceptions (İlğan, Sevinç & Arı, 2013), learning styles (Kahyaoğlu, Tan & Kaya, 2013; Saracaloğlu & Dursun, 2011), emotional intelligence levels (Kayserili, 2009), critical thinking and creativity skills (Kesicioğlu & Deniz, 2014), life satisfaction levels (Kiralp & Bolkan, 2016; Recepoğlu, 2013), liking of children (Kuşcu, Erbay, Acar & Gülnar, 2015), social skills (Kozağaç, 2015), academic procrastination behaviors (Kutlu, Gökdere & Çakır, 2015), classroom management approaches (Süral, 2015), vocational motivation levels (Ömür & Nartgün, 2013), job satisfaction levels (Orhan, 2013), attitude towards cheating (Özyurt & Altay, 2014), personal values (Parlar & Cansoy, 2016), attitudes towards school (Baykara Pehlivan, 2004), professional concerns (Serin, Güneş & Değirmenci, 2015), technopedagogical field knowledge (Tuncer & Bahadır, 2016) ve reflective thinking skills (Yumuşak, 2015).

In addition to the examination of the relationship between the attitude toward the teaching profession and related variables, a number of studies have analyzed how attitude towards teaching profession varies across individuals' demographic characteristics. The related studies mostly involve the demographic characteristics such as gender (Camadan & Duysak, 2010; Çiğdem & Memiş, 2011), high school type (Can, 2010), education status (İlğan, Sevinç & Arı, 2013; Tok, 2012), grade level (Kaplan & İpek, 2002; Tümkaya, 2011), faculty type (Kozağaç, 2015; Ömür & Nartgün, 2013), religious status (Parvez & Shakır, 2014) ve the presence of a teacher in the family (Kutlu, Gökdere & Çakır, 2015; Recepoğlu, 2013). It seems difficult to determine the demographic characteristics that influence the attitude towards the teaching profession due to the use of different sample sizes and

inconsistencies between the results in the studies. Thus, a meta-analysis is required to determine whether the demographic variables are real determinants of the attitude towards teaching profession.

The Purpose and Significance of Study

Upon examining the relevant literature, it has been determined that meta-analysis studies on the attitude toward the teaching profession are limited. A meta-analysis study conducted by Erdemar, Aytaç, Türk and Arseven (2016) and including 35 studies carried out between 2004 and 2015 has only examined whether the attitude towards teaching profession differs across gender. In their survey study, Eren, Çelik and Oğuz (2014) have found 109 studies conducted between 1984 and 2013 related to the attitudes towards the teaching profession. The limited number of demographic variables used in a few number of meta-analyses has raised need for conducting such a study. In this meta-analysis, both the number of studies related to the subject and included in this research have been increased, and the demographic characteristics that are related to attitude have been augmented. Within this scope, answers to the following questions have been sought:

1. Do prospective teachers' attitudes towards teaching profession significantly vary across their gender?
2. Do prospective teachers' attitudes towards teaching profession significantly differ across their grade level (freshman and senior)?
3. Do prospective teachers' attitudes towards teaching profession significantly vary across the presence of a teacher in the family?
4. Do prospective teachers' attitudes towards teaching profession significantly vary across being education faculty graduate/studying in education faculty or graduates of other faculties/ studying in the other faculties (faculty type)?

METHOD

Research Design

This study has been designed through use of the meta-analysis method. Meta-analysis is a quantitative research that takes place by statistically combining the results from multiple studies on the topic in the related literature (Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper, Hedges, & Valentine, 2009).

This research has scanned databases such as ULAKBİM, Google Academic, Web of Science and ERIC in order to explore whether prospective teachers' attitudes towards teaching profession vary across what factors; moreover, 249 studies have been achieved through the search made by using such keywords as "teacher candidates", "the teaching profession", "teaching attitude" and "attitudes towards the teaching profession". Of all the studies, those that meet the specified criteria are included in the meta-analysis study. The following criteria have been used in determining the studies to be included in this research: 1) Research questions include variables such as gender, grade level, being a teacher in the family, and faculty type. 2) The parametric tests (t-test and ANOVA) have been used in during data analysis, sample size, group mean and standard deviation values have been presented. 3) The reliability coefficients of the teachers' attitude scales used in the quantitative studies have been provided and the values are greater than .70. Taking the criteria into consideration, 249 studies meeting the first criterion have been included in the study. 129 of 249 studies have been determined to be available according to the second criterion which signifies that both parametric tests are used and the sample size, group mean and standard deviation values are provided. Lastly, 113 studies have been included in the meta-analysis considering the third criterion referring to the fact that the reliability coefficient of the teaching attitude scale is greater than .70.

Sample

The number of theses and articles have been initially determined for each research question, and the samples used in these studies have been examined. Table 1 presents the number of the studies, their demographic characteristics and sample sizes.

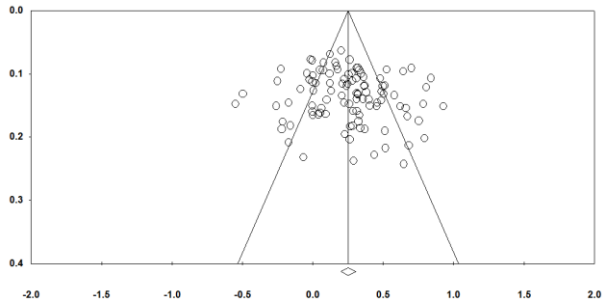
Table 1. Descriptive Tables Regarding the Studies

Variables	Number of studies	Category	Sample size
Gender	103	Female	18252
		Male	13661
Grade level	26	1st grade	2880
		4th grade	2185
The presence of a teacher in the family	18	Yes	1854
		No	3061
Faculty type	11	Education Faculty	1423
		Other Faculties	1118

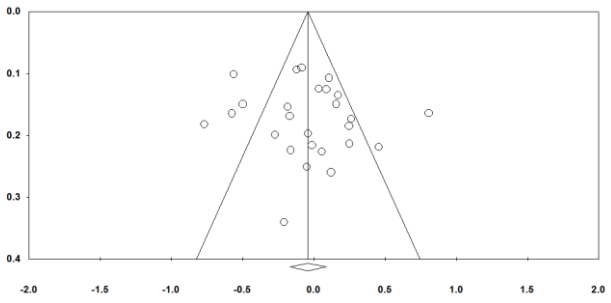
Table 1 displays that 103 of the studies include "gender" variable and that the sample of these studies holds a total of 31913 prospective teachers. Besides, the "grade level" variable available in the second research question is included in 26 studies as 1st and 4th grade and the total sample in these studies is composed of 5065 prospective teachers. The reason for the selection of these classes is that the first grade represents the first year of faculty and the fourth year represents the last year / years. As for the third research question, the variable of "the presence of a teacher in the family" takes place in 18 studies and the total number of samples in these studies is 4915. There are 11 studies about the last research question, "faculty type", and the sample consists of 2541 individuals.

Publication Bias

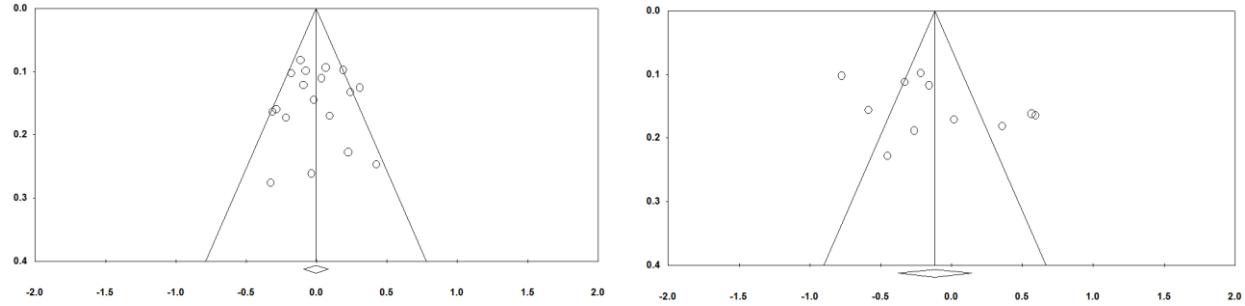
Publication bias refers to the likelihood that a group of studies selected from published studies on a particular topic may not represent all studies (Rothstein, Sutton, & Borenstein, 2005). If the studies that are statistically significant are mostly examined in a meta-analysis, it is likely that this analysis has publication bias (Borenstein et al., 2009). In this regard, several methods are used for detecting publication bias. The most commonly used of these methods are the Funnel Plots, Duval and Tweedie's trim and fill method and Egger's Linear Regression Test. This research used Funnel plot so as to test the publication bias. Figure 1 presents The Funnel Plots showing the publication bias.



Gender



Grade level



The presence of a teacher in the family | Faculty type

Figure 1. Funnel Plots

Sterne et al. (2011) have stated that if the points calculated for each study and their effect values are scattered symmetrically around the vertical line in the funnel, it will not result in publication bias.

Figure 1 suggests that four funnel plots is distributed symmetrically. However, the results of Duval and Tweedie's trim and fill method and Egger's Linear Regression Test are shown in Table 2 since the funnel plots do not provide a statistically clear result.

Table 2. Test Results Regarding Publication Bias

	Duval And Tweedie's Trim and Fill Method		Egger's Regression Test (*p*)
Variable	Trimmed Study	Observed/Filled	
Gender	17	0.25 (0.20, 030) / 0.17 (0.12, 0.23)	0.178
Grade level	5	-0.04 (-0.17, 0.09) /-0.14 (-0.28, -0.01)	0.429
The presence of a teacher in the family	0	0.00 (-0.09, 0.09)/ 0.00 (-0.09, 0.09)	0.950
Faculty type	3	-0.12 (-0.38, 0.14)/ -0.12 (-0.38, 0.14)	0.159

Table 2 shows the results of Duval and Tweedie's trim and fill method conducted to determine the publication bias. This method recalculates the size effect in the case of trimming of this number, showing the number of studies that need to be trimmed in order to correct the asymmetric state of the funnel plot. If the difference between the two results (observed/corrected value) is not statistically significant, it is understood that the analysis does not result in publication bias (Pamuk, Atli & Kış, 2015). The difference between the observed and corrected values in the 95% confidence interval of the four variables in this study is not statistically significant, indicating that the study is free from publication bias.

Egger's regression test is another method used to test publication bias. The insignificant result indicates that there is no publication bias (Klassen & Tze, 2014). The values obtained in this study support the absence of publication bias for four variables.

Heterogeneity Test

Heterogeneity test has been used before data analysis as this test is especially one of the assumptions underlying in the use of the random-effects model (Başak, Aşkın and Gelbal, 2016; Üstün and Eryılmaz, 2014). Hence, the Q-test showing the heterogeneity of the effect sizes of the studies in the meta-analysis and the results obtained by calculating I^2 value indicating the level of heterogeneity are presented in Table 3.

Table 3. Heterogeneity Test Results

	K	Q	I^2	Degree of Freedom	p
Gender	103	477.94	78.658	102	.000
Grade level	26	113.194	77.914	25	.000
The presence of a teacher in the family	18	34.055	50.080	17	.008
Faculty type	11	97.034	89.694	10	.000

The level of heterogeneity varies across I^2 value. This value indicates that the variance among the studies' results from heterogeneity rather than chance. If the value of I^2 is less than 25%, the level of heterogeneity is low, if it is 50%, the level of heterogeneity is medium and if it is greater than 75%, the level is classified as high. According to Table 2, the heterogeneity levels for the variables of gender, grade and faculty type have been identified to be high since I^2 is greater than 75%, while medium level heterogeneity has been found in terms of the presence of a teacher in the family variable.

Data analysis

The effect sizes can be calculated through the fixed effects model and the random effects model in meta-analysis. In the fixed effect model, the same effect level is calculated for all studies and weighting is performed based on the number of observations in the samples of the studies (Borenstein et al., 2009). As for the random effects model, the effect size differs due to the demographic and other characteristics of the sample (Cooper, Hedges, & Valentine, 2009; Üstün & Eryılmaz, 2014) and presents more generalizable results (Card, 2011). In addition, Borenstein et al. (2009) have recommended that the random effects model be used in meta-analysis of published studies. Thus, this research used random effects model.

The effect sizes related to the population and each study have been calculated by using the "Hedges' *g*" method through the "Comprehensive Meta-Analysis" program. The other methods, "Cohen's *d*" and "Glass Δ" methods have mild bias for small samples in the calculation of the effect sizes of population (Üstün and Eryılmaz, 2014). The "Hedges' *g*" method proposes a solution to this situation with the J correction factor as shown below.

$$g = J.d$$

Here, *J* is the correction factor and *d* is Cohen's formula. These two factors are presented as a formula below.

$$J = 1 - \frac{3}{4d_f - 1}$$

$$Cohen\ d = \frac{x_e - x_c}{S_p}$$

Here, d_f in *J* correction factor represents degree of freedom; while x_e, x_c and S_p in "Cohen's *d*" refer to the mean of the experimental group, the mean of the control group and combined standard deviation of the two groups, respectively.

RESULTS

This research has analyzed 103 studies consisting of 31913 individuals with the aim of determining whether prospective teachers' attitudes towards teaching profession differ across their gender, which is the first research question. According to the random effects model, the *z* score has been found to be statistically significant (z=9.494, p<0.05), while the standardized mean difference between the 18252 female and 13661 male prospective teachers has been identified to be 0.252 at the 95% confidence interval, meaning that female prospective teachers have a statistically more positive attitude towards

the teaching profession compared to males. Appendix-2 shows the forest plot of the effect sizes of 103 studies including "gender" variable.

As for the second research question, 26 studies consisting of 5065 individuals have been examined in order to reveal whether prospective teachers' attitudes towards teaching profession vary across their "grade level". According to the random effects model, no significant difference has been determined in *z* score (z=-0.617, p>0.05), while the standardized mean difference between the 2880 1st grade and 2185 4th grade prospective teachers has been found to be -0.041 at the 95% confidence interval. This indicates that prospective teachers' attitudes towards teaching profession do not significantly differ across their grade level. Appendix-3 displays the forest plot of the effect sizes of 26 studies including "grade level" variable.

When it comes to the third research question, 18 studies conducted with 4915 individuals have been analyzed so as to explore whether prospective teachers' attitudes towards teaching profession vary across "the presence of a teacher in the family". According to the random effects model, the standardized mean difference between the 1854 prospective teachers who have teachers in their families and 3061 who do not have teachers in their families has been determined to be -0.003 at the 95% confidence interval, and no significant difference has been found in terms of *z* score (z=-0.074, p>0.05). This supports the view that the presence of a teacher in the family does not significantly change prospective teachers' attitudes towards teaching profession. Appendix-4 presents the forest plot of the effect sizes of 18 studies including "the presence of a teacher in the family" variable.

Considering the last research question, 11 studies composed of 2541 individuals have been analyzed in an attempt to determine whether prospective teachers' attitudes towards teaching profession vary across "faculty type". According to the random effects model, the standardized mean difference between the 1423 prospective teachers from education faculty and 1118 teachers from the other faculties has been identified to be -0.119 at the 95% confidence interval, and no significant difference has been found in terms of *z* score (z= -0.895, p>0.05). This sheds light onto the fact that faculty type does not significantly change prospective teachers' attitudes towards teaching profession. Appendix-5 shows the forest plot of the effect sizes of 11 studies including "faculty type" variable.

DISCUSSION and CONCLUSION

The present study has explored as to whether the attitudes of the prospective teachers towards the teaching profession vary across gender, grade level, the presence of a teacher in the family and faculty type. In this regard, meta-analysis method has been used to obtain generalizable information from the related studies that were previously made, that are different and inconsistent.

After satisfying the specified criteria, 106, 26, 18 and 11 studies have been accessed for the variables such as gender, grade level, the presence of a teacher in the family and faculty type, respectively. Taking the sample sizes into account, a total of 31913 prospective teachers-18252 female and 13661 male-; a total of 5065 prospective teachers-2880 are 1st grade and 2185 are 4th grade-; 4915 prospective teachers-1854 having teachers in their families and 3061 no teacher in their families-; a total of 2541 teachers-1423 from education faculty and 1118 from other faculties- have been determined as the research sample.

Statistical analyzes pave the way for the fact that female prospective teachers have a more positive attitude regarding the teaching profession compared to the males. This result is especially true supporting the same meta-analysis study conducted by Erdemar et al. (2016) with 35 studies. The previous studies lead forth the reason for this difference as such. Teaching profession is much more compatible with the women's perceptions as compared to male teachers (Terzi & Tezci, 2007), hence it is conceivable that women are more willing to prefer the teaching profession and that they have plans to devote their whole lives to children.

No significant difference has been observed among prospective teachers' attitudes towards the teaching profession in terms of their grade level. Similar results have emerged in the studies conducted

by Dalkıran & Yıldız (2016) and Pehlivan (2008). However, several studies have put forwards that 1st grade prospective teachers have higher level of perceptions towards teaching profession compared to 4th graders (Yildizer, Ozboke, Tascioglu & Yilmaz, 2017). On the other hand, some studies have shown that 4th grade prospective teachers' attitudes towards the teaching profession are more positive than 1st graders (Aydın & Tekneci, 2013; Çelen & Eskicioğlu, 2015). However, the underlying reasons for the 1st and 4th grade prospective teachers' attitudes towards teaching profession are prospective teachers and faculty. In particular, the reasons why prospective teachers prefer the related teaching programs are that teaching is a job-guaranteed profession or the will of the family (Ekici, 2014; Kartal & Afacan, 2012), leading to the fact that they may see the teaching profession as a "profession" rather than "sanctity", and their attitudes towards the profession may not decrease even if they receive 4-year undergraduate education. It is hotly-debated that the teaching profession courses are generally taught theoretically rather than practically by the faculty members and this situation is compensated by the teaching practice course (Eraslan, 2009). In addition, Paker (2008) has emphasized the problems experienced by prospective teachers in their teaching practice lessons, particularly clarified the fact that teachers have not received enough feedback from the observations and presentations they have made. This may cause prospective teachers' failure in internalizing their profession even at the end of 4 year-undergraduate education, and therefore there may not be any change in their attitudes towards the teaching profession. For this reason, it is necessary to increase the teaching application hours and this course should be processed with the principle of accountability in an attempt to improve the attitudes of the prospective teachers towards teaching profession.

Research results have also revealed that prospective teachers' attitudes towards teaching profession do not significantly differ across the presence of a teacher in the family. This result is in line with that of all meta-analysis studies except for the one conducted by Akpınar, Yıldız & Ergin (2006). The attitude towards the teaching profession can be explained as a situation that is not expected to be changed by the external factor, "the presence of a teacher in the family", when it is considered to be defined internally (Çapa & Çil, 2000; Çetin, 2006; Eagly & Chaiken, 1993; Fishbein & Ajben, 1975; Ünlü, 2011).

Last but not least, no significant difference has been determined among the attitudes of the prospective teachers towards the teaching profession in terms of the faculty type. Upon examining the studies included in the scope of the meta-analysis, different findings have been found in the present study. Several studies have determined that prospective teachers from education faculty hold more positive attitudes towards teaching profession (Kaplan & İpek, 2002; Uyulgan & Kartal, 2012); whereas in other studies, the difference has been found in favor of those who study in the other faculties and who receive pedagogical formation education (Bağçeci, Yildirim, Kara & Keskinpalta, 2015; Ömür & Nartgün, 2013; Polat, 2013). This may be explained by the fact that the attitudes of the prospective teachers in the education faculty towards the teaching profession do not vary across grade level. In other words, whether it is teaching profession education taken in the education faculty, whether it is a pedagogical formation education program given in a short period of time about 1 year, what is significant in this process is to put the activities into practice and to determine the internal reasons of prospective teachers for teaching profession. Thus, not the quantity but the quality of the education must be revised by the universities and YÖK.

This research has been carried out through employing the meta-analysis method to obtain generalized information regarding the change of the attitude towards the teaching profession depending on the demographic characteristics. However, just as all studies, this study also has various limitations. First, the studies related to the attitude towards teaching profession have been generally considered to be in Turkey even though both national and international literature review has been performed during the meta-analysis. Second, only four demographic characteristics-gender, grade level, the presence of a teacher in the family and faculty type- have been used in the meta-analysis study. Other demographic and personal characteristics as well as reasons for being a teacher may be added in the further studies. Once for all, the current meta-analysis study aims to reveal how the attitude towards the teaching profession differs across each variable separately; nevertheless, this particular purpose ignores the

mutual or overlapping effect of the variables altogether. In this context, meta-regression studies may be included in the subsequent studies.

REFERENCES

Adıgüzel, A. (2017). The relationship between teacher candidates' pedagogical competence perceptions and their attitudes about teaching profession. *Turkish Journal of Education, 6*(3), 113-127. DOI: 10.19128/turje.296481

Akpınar, E., Yıldız, E., & Ergin, Ö. (2006). Prospective teachers' professional knowledge and their attitudes toward their profession affect their success of teaching. *Buca Journal of Education Faculty*, *19*(1), 56-62.

Aslan, S., & Yalçın, M. (2013). The prediction of attitude towards to profession of teacher through five factor personality dimensions. *National Education, 42*(197), 169-178.

Aşkar, P., & Erden, M. (1987). Attitude scale towards teaching profession. *Contemporary Education, 121*(12), 8-11.

Aydın, A., & Tekneci, E. (2013). Attitudes towards profession and anxiety levels of education of mentally handicapped students. *Pegem Journal of Education & Instruction, 3*(2), 1-12. DOI: https://doi.org/10.14527/V3N2M1

Ayık, A., & Ataş, Ö. (2014). The relationship between pre-service teachers' attitudes towards the teaching profession and their motivation to teach. *Journal of Educational Sciences Research, 4*(1), 25-43. DOI: http://dx.doi.org/10.12973/jesr.2014.41.2

Bağçeci, B., Yıldırım, İ., Kara, K., & Keskinpalta, D. (2013). A comparative study on the attitudes of students from education faculties and science faculties towards being a teacher. *Erzincan University Journal of Education Faculty*, *17*(1), 307-324. DOI: 10.17556/jef.52416

Bakaç, E., & Özen, R. (2017). Relationship between pedagocical certificate program students' attitudes and self-efficacy beliefs towards teacher profession. *Kastamonu Journal of Education, 25*(4), 1389-1404.

Başar, T., Aşkın, İ., & Gelbal, S. (2016). The effect of mastery learning model on students academic achievement: A meta-analysis study. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(2), 355-377. DOI: 10.21031/epod.277891

Başbay, M., Ünver, G., & Bümen, N. T. (2009). A longitudinal study on secondary education teacher candidates' attitudes towards teaching profession. *Educational Administration: Theory and Practice*, *59*(1), 345-366.

Baykara Pehlivan, K. (2004). The relationship between classroom teacher candidates' attitudes towards the teaching profession and school attitudes. *Journal of Educational Researches, 14*(4), 211-218.

Bedel, E. F. (2015). Exploring academic motivation, academic self-efficacy and attitudes toward teaching in pre-service early childhood education teachers. *Journal of Education and Training Studies, 4*(1), 142-149. DOI: 10.11114/jets.v4i1.561

Bektas, F., & Nalcaci, A. (2012). The relationship between personal values and attitude towards teaching profession. *Educational Sciences: Theory and Practice, 12*(2), 1244-1248.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* UK: Wiley.

Bulut, İ. (2009). Evaluation of teacher candidates' attitudes concerning teaching profession (Dicle and Firat University sample). *Dicle University Ziya Gökalp Education Faculty Journal, 14*(1), 13-24.

Camadan, F., & Duysak, A. (2010). Comparing pre-service teachers' attitudes in the different programs toward teaching profession in terms of different variables: Example of Rize University. *Sakarya University Journal of Education Faculty*, *20*(1), 30-42.

Can, Ş. (2010). Attitudes of the students who attend the non-thesis graduated education program towards the teaching profession. *Muğla University Journal of Social Sciences Institute, 24*(1), 13-28.

Card, N. A. (2011). *Applied meta-analysis for social science research: Methodology in the social sciences*. New York: Guilford.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage.

Çağlar, C. (2013). The relationship between the levels of alienation of the education faculty students and their attitudes towards the teaching profession. *Educational Sciences: Theory and Practice*, *13*(3), 1507-1513. DOI: 10.12738/estp.2013.3.1577

Çam, E., & Üstün, A. (2016). The relation between professional attitude and life long learning tendency of teachers. *Hitit University Journal of Social Sciences Institute, 9*(1), 461-477. DOI: http://dx.doi.org/10.17218/husbed.58800

Çapa, Y., & Çil, N. (2000). Investigation of teacher candidates' attitudes towards teaching profession in terms of several variables. *Hacettepe University Journal of Education Faculty*, *18*(1). 69-73.

Çelen, A., & Eskicioğlu, Y. (2015). Analysis of attitude toward teachig profession and state-trait anxiety level of students in teaching departments accepting students through special aptitude tests. *Route Educational and Social Science Journal*, *2(*3), 1-18.

Çeliköz, N., & Çetin, F. (2004). Factors affecting the attitudes of Anatolian teacher education students towards the teaching profession. *Journal of National Education*, *162*(1), 139-157.

Çetin, F. (2016). The relationship between the classroom management competence of teachers and their attitudes towards the profession of teaching and job satisfaction. *Electronic Turkish Studies*, *11*(3), 791-808. DOI: 10.7827/TurkishStudies.9285

Çetin, Ş. (2006). Reliability and validity study of an attitude scale of teaching profession. *The Journal of Industrial Arts Education Faculty of Gazi University*, *18*(1), 28-37.

Çetinkaya, R. (2007). *Teacher candidates' perceptions of proficiency and their attitudes towards teaching profession* (Master's Thesis, Selçuk University Social Sciences Institute, Konya, Turkey). Retrieved from http://tez2.yok.gov.tr/.

Çiğdem, G., & Memiş, A. (2011). Investigation of attitudes of elementary school prospective teachers towards learning styles and teaching profession in terms of various variables. *Çukurova University Journal of Education Faculty*, *3*(40), 57-77.

Çimen, L. K. (2016). A study on the prediction of the teaching profession attitudes by communication skills and professional motivation. *Journal of Education and Training Studies*, *4*(11), 21-38. DOI: https://doi.org/10.11114/jets.v4i11.1842

Dadandı, İ., Kalyon, A., & Yazıcı, H. (2016). Teacher self-efficacy beliefs, concerns and attitudes towards teaching profession of faculty of education and pedagogical formation students. *Bayburt Education Faculty Journal*, *11*(1), 253-259.

Dalkıran, E., & Yıldız, G. (2016). Investigation of teaching profession attitudes of the music education department students. *Fine Arts*, *11*(4), 153-160. DOI: http://dx.doi.org/10.12739/NWSA.2016.11.4.D0180

Demirtaş, H., Cömert, M., & Özer, N. (2011). Pre-Service teachers' self-efficacy beliefs and attitudes towards profession. *Education and Science*, *36*(159), 96-111.

Dikmenli, Y., & Çifçi, T. (2015). Attitudes to teaching profession and field knowledge levels of the geography teacher candidates taking pedagogical formation education. *Cumhuriyet University Journal of Social Sciences*, *39*(2), 155-172.

Dilmaç, B., Çıkılı, Y., Işık, H., & Sungur, C. (2009). Technical teacher candidates' vocational self-esteem as predictor of attitudes related teaching prefessions. *Journal of Selçuk-Technical*, *8*(2), 127-143.

Doğan, T., & Çoban, A. E. (2009). The investigation of the relations between students' attitude toward teaching profession and anxiety level in faculty of education. *Education and Science*, *34*(153), 157-159.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Orlando, FL: Harcourt Brace Jovanovich College Publishers.

Ekici, F. Y. (2014). Examining prospective teachers' attitudes towards teaching profession in terms of various variables (Istanbul Sabahattin Zaim University sample). *Journal of International Social Research*, *7*(35), 658-665.

Emre, Ş. C., & Ünsal, S. (2017). The investigation of the relationship between secondary school teachers' self efficacy beliefs and attitude towards teaching. *European Journal of Education Studies*, *3*(6), 94-111. DOI: 10.5281/zenodo.572344

Eraslan, A. (2009). Prospective mathematics teachers' opinions on 'teaching practice'. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education*, *3*(1), 207-221.

Erdamar, G., Aytaç, T., Türk, N., & Arseven, Z. (2016). The effects of gender on attitudes of preservice teachers towards the teaching profession: A meta-analysis study. *Universal Journal of Educational Research*, *4*(2), 445-456. DOI: 10.13189/ujer.2016.040219

Eren, B., Çelik, M., & Oğuz, A. (2014). Investigation of dissertations and articles in Turkey about the attitudes towards teaching profession. *Dumlupinar University Journal of Social Sciences*, *42(1)*, 359-370.

Erkuş, A., Sanlı, N., Bağlı, M. T., & Güven, K. (2000). Developing an attitude scale toward teaching as a profession. *Education and Science*, *25*(116), 27-33.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Girgin, G., Akamca, G. Ö., Ellez, A. M., & Oğuz, E. (2010). Preschool teacher candidates' attitudes towards profession, self efficacy beliefs and professional self respects. *Buca Education Faculty Journal*, *28*(1), 1-15.

Hançer, A. H. (2017). Effects of science teacher candidates' attitudes towards the teaching profession on academic fraud inclinations. *Electronic Turkish Studies*, *12*(6), 387-402. DOI: 10.7827/TurkishStudies.11418

İlğan, A., Sevinç, Ö. S., & Arı, E. (2013). The perceptions of teachers towards professional attitude contemporary teachers qualifications. *Ondokuz Mayıs University Journal of Education Faculty*, *32*(2), 175-195.

Kahyaoğlu, M., Tan, Ç., & Kaya, M. F. (2013). Attitudes of elementary school teacher candidates towards learning styles and teaching profession. *Mustafa Kemal University Journal of Social Sciences Institute*, *10*(21), 225-236.

Kaplan, A., & İpek, A. S. (2002). Examining the attitudes of mathematics teacher candidates towards the teaching profession. *Education and Science*, *27*(125), 69-73.

Kartal, T., & Afacan, Ö. (2012). Examining attitudes of prospective teachers who took pedagogical formation education towards teaching profession. *Mehmet Akif Ersoy University Journal of Education Faculty*, *24*(1), 76-96.

Kayserili, T. (2009). *Examination of the attitudes of pre-school teachers and teacher candidates regarding their emotional intelligence and pre-school teachership* (Master Thesis, Atatürk University Institute of Social Sciences, Erzurum, Turkey). Retrieved from http://tez2.yok.gov.tr/.

Kesicioğlu, O. S., & Deniz, Ü. (2014). Investigation of the relationship between pre-service preschool teachers' attidues towards teaching profession and their skills of crticial thinking and creativeness. *Turkish Studies-International Periodical for The Languages, Literature and History of Turkish or Turkic*, *9*(8), 651-659. DOI: http://dx.doi.org/10.7827/TurkishStudies.7206

Kiralp, F. S. S., & Bolkan, A. (2016). Relationship between candidate teacher's attitude towards teaching profession and their life satisfaction levels. *The Anthropologist*, *23*(1-2), 11-20. DOI: 10.1080/09720073.2016.11891919

Klassen, R. M., & Tze, V. M. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review*, *12*(1), 59-76. DOI: https://doi.org/10.1016/j.edurev.2014.06.001

Kozağaç, Z. B. (2015). *The determination of multiple intelligence area of preservice teacher of the department of mathematics and examination of attitude with regard to social abilities with teaching profession* (Master Thesis, Adnan Menderes University Institute of Social Sciences, Aydın, Turkey). Retrieved from http://tez2.yok.gov.tr/.

Kuşcu, Ö., Erbay, F., Acar, Ş., & Gülnar, E. (2015). Examination of attitudes of pre-school prospective teachers towards teaching profession in terms of liking of children. *International Journal of Educational Sciences, 2(*3), 155-122.

Kutlu, N., Gökdere, M., & Çakır, R. (2015). A comparative study of the attitudes of the prospective teachers towards the academic procrastination behavior and the teaching profession. *Kastamonu Education Journal*, *23*(3), 1311-1330.

Orhan, N. (2013). *Job satisfaction and occupational attitude levels of trainee teachers* (Master Thesis, Dokuz Eylül University Institute of Educational Sciences, İzmir, Turkey). Retrieved from http://tez2.yok.gov.tr/.

Ömür, Y. E., & Nartgün, Ş. S. (2013). The relationship between teacher candidates' attitudes towards teaching profession and motivational levels. *Journal of Policy Analysis in Education, 2*(2), 41-55.

Özyurt, Y., & Altay, E. (2014). The appearance of science teacher candidates' attitudes towards teaching profession and cheating. *Bartın University Education Faculty Journal, 3(*1), 78-101. DOI: 10.14686/BUEFAD.201416208

Paker, T. (2008). Problems of student teachers regarding the feedback of university supervisors and mentors during teaching practice. *Pamukkale University Journal of Education Faculty*, *23*(23), 132-139.

Pamuk, M., Atli, A., & Kış, A. (2015). Investigation of theses in Turkey on loneliness in terms of gender: A meta-analytic study. *Journal of Theory & Practice in Education (JTPE), 11*(4), 1392-1414.

Parlar, H., & Cansoy, R. (2016). Individual values as a predictor of teachers' attitudes towards the teaching profession. *Journal of Educational Sciences*, *44*(1), 125-142. DOI: 10.15285/maruaebd.286490

Parvez, M., & Shakir, M. (2016). A comparative study of the attitudes of muslim and non-muslim prospective teachers towards teaching profession. *International Journal for Educational Studies, 7*(1), 67-74.

Pehlivan, K. B. (2008). A study on the socio-cultural characteristics of classroom teacher candidates and their attitudes towards teaching profession. *Mersin University Education Faculty Journal*, *4*(2), 151-168.

Polat, S. (2013). Examination of the pedagogical formation certificate program and education faculty students' attitudes towards teaching profession. *e-Journal of International Educational Researches*, *4*(2), 48-60.

Pratkanis, A. R., & Greenwald, A. G. (1989). A sociocognitive model of attitude structure and function. In *Advances in experimental social psychology*, *22*(1), 245-285. DOI: https://doi.org/10.1016/S0065-2601(08)60310-X

Recepoğlu, E. (2013). Analyzing the relationship between prospective teachers' life satisfaction and attitudes concerning teaching profession. *H. U. Journal of Education, Özel Sayı*(1), 311-326.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. England: John Wiley & Sons.

Saracaloğlu, A. S., & Dursun, F. (2011). The relationship between classroom teacher candidates' attitudes towards teaching profession and learning strategies. *Education Sciences, 6*(3), 2328-2346.

Semerci, Ç. (1999). The attitude scale of the students' attitudes towards the teaching profession. *Education and Science, 23*(111), 51-55.

Semerci, N., & Semerci, Ç. (2004). Teaching profession attitudes in Turkey. *Journal of Social Science, 14*(1), 137-146.

Serin, M. K., Güneş, A. M., & Değirmenci, H. (2015). The relationship between classroom teachers' attitudes towards the teaching profession and the level of anxiety for the profession. *Cumhuriyet International Education Journal, 4*(1), 21-34. DOI: 10.30703/cije.321360

Sterne, J. A., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., … Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ, 343*, 1-8. . DOI: 10.1136/bmj.d4002

Süral, S. (2015). The relationship between elementary school teachers' teaching styles, classroom management approaches and attitudes towards teaching profession. *Journal of International Social Research, 8*(41), 1027-1038.

Terzi, A. R., & Tezci, E. (2007). Necatibey Education Faculty students' attitudes towards teaching profession. *Educational Management in Theory and Practice, 52*(1), 593-614.

Tok, T. N. (2012). Teacher candidates' attitudes towards the teaching profession in Turkey. *Alberta Journal of Educational Research, 58*(3), 381-403.

Tuncer, M., & Bahadır, F. (2016). Evaluation of teacher candidates in terms of attitudes towards technopedagogical field competencies and teaching profession. *Electronic Turkish Studies, 11*(9), 839-858. DOI: http://dx.doi.org/10.7827/TurkishStudies.9635

Tümkaya, S. (2011). Investigation of communication skills and teaching attitudes of students in classroom teaching. *Çukurova University Journal of Social Sciences Institute, 20*(2), 49-62.

Uyulgan, M. A., & Kartal, M. (2012). Evaluation of the senior students' chemistry subject matter knowledge and attitudes towards teaching profession in faculty of science chemistry department and faculty of education chemistry education department. *Buca Education Faculty Journal, 32*(1), 104-118.

Ünlü, H. (2011). Developing an attitude scale for the profession of physical education Teaching (ASPPET). *Educational Sciences: Theory & Practice, 11(4)*, 2005-2020.

Üstün, U., & Eryılmaz, A. (2014). A research method for effective research synthesis: Meta-analysis. *Education and Science, 39*(174), 1-32. DOI: 10.15390/EB.2014.3379

Üstüner, M. (2006). Validity and reliability study of attitude scale towards teaching profession. *Educational Sciences: Theory & Practice, 12*(1), 109-127.

Yildizer, G., Ozboke, C., Tascioglu, R., & Yilmaz, I. (2017). Examining attitudes of physical education teacher education program students toward the teaching profession. *Montenegrin Journal of Sports Science and Medicine, 6*(2), 27-33.

Yumuşak, G. K. (2015). Teacher candidates' reflective thinking tendencies and attitudes towards the profession. *Bartın University Education Faculty Journal, 4*(2), 466-481. DOI: 10.14686/buefad.v4i2.1082000206

Appendices

Appendix-1: Studies included in Meta-Analysis

Adıgüzel, A. (2017). The relationship between teacher candidates' pedagogical competence perceptions and their attitudes about teaching profession. *Turkish Journal of Education, 6*(3), 113-127. DOI: 10.19128/turje.296481

Akgün, F., & Özgür, H., (2014). Examination of the anxiety levels and attitudes of the information technology pre-service teachers towards the teaching profession. *Journal of Theory and Practice in Education, 10(*5), 1206-1223.

Akkaya, N. (2009). An investigetion of prospective teachers' attitudes regarding teaching various in terms of variables. *Buca Faculty of Education Journal, 25*(1), 35-42.

Akpınar, E., Yıldız, E., & Ergin, Ö. (2006). Prospective teachers' professional knowledge and their attitudes toward their profession affect their success of teaching. *Buca Journal of Education Faculty*, *19*(1), 56-62.

Aktop, A., & Beyazgül, G. (2014). Pre-service physical education teacher's attitudes towards teaching professionals. *Procedia-Social and Behavioral Sciences, 116,* 3194-3197. DOI: 10.1016/j.sbspro.2014.01.733

Alci, B., Karatas, H., Yurtseven, N., & Alci, E. (2013). The correlation between teacher candidates' attitudes towards teaching profession and their school practicum achievement. *Journal of Teaching and Education, 2*(3), 281-287.

Altunkeser, F., & Ünal, E. (2015). Predicting attitudes of elementary preservice teachers towards teaching as a profession regarding various variable. *Ahi Evran University Journal of Institute of Social Sciences, 2*(1), 1-15.

Arastaman, G. (2013). Examination of education and arts and sciences faculty students' self-efficacy beliefs and their attitudes toward teaching profession. *Journal of Kirsehir Education Faculty, 14*(2), 205-217.

Aydın, R., & Sağlam, G. (2012). Teacher applicants' views toward teaching profession (example from Mehmet Akif Ersoy University). *J. Turk. Educ. Sci*, *10*(2), 291-294.

Aydın, A., & Tekneci, E. (2013). Attitudes towards profession and anxiety levels of education of mentally handicapped students. *Pegem Journal of Education & Instruction, 3*(2), 1-12.

Bağçeci, B., Yıldırım, İ., Kara, K., & Keskinpalta, D. (2015). A comparative study on the attitudes of students from education faculties and science faculties towards being a teacher. *Erzincan University Journal of Education Faculty*, *17*(1), 307-324. DOI: 10.17556/jef.52416

Bakaç, E., & Özen, R. (2017). Relationship between pedagogical certificate program students' attitudes and self-efficacy beliefs towards teacher profession. *Kastamonu Education Journal, 25*(4), 1389-1404.

Bal, A. P. (2016). The effect of pedagogic formation training on vocational attitudes of mathematics teacher candidates. *International Journal of Social Sciences and Education Research*, *3*(1), 58-69.

Başbay, M., Ünver, G., & Bümen, N. T. (2009). A longitudinal study on secondary education teacher candidates' attitudes towards teaching profession. *Educational Administration: Theory and Practice*, *59*(1), 345-366.

Bozdoğan, A. E., Aydın, D., & Yıldırım, K. (2007). Attitudes of teacher candidates towards teaching profession. *Journal of Kirsehir Education Faculty*, *8*(2), 83-97.

Bozkirli, K. Ç., & Er, O. (2011). The examination of Turkish / Turkish language and literature teacher candidates' attitudes toward teacher profession according to various variables (Kafkas University sample). *Electronic Turkish Studies, 6*(4), 457-466. DOI: 10.7827/TurkishStudies.2826

Bulut, İ. (2009). Evaluation of teacher candidates' attitutes concerning teaching profession (Dicle and Firat University sample). *Dicle University Journal of Ziya Gokalp Faculty of Education, 14*(1), 13-24.

Bulut, D. (2011). Attitudes of music teacher candidates towards the profession of teaching. *Gazi University Journal of Gazi Educational Faculty*, *31*(3), 651-674.

Bulut, H., & Doğar, Ç. (2006). The investigation of student teachers' attitudes towards their occupations. *Erzincan University Journal of Education Faculty*, *8*(2), 13-27.

Bümen, N. T., & Özaydın, T. E. (2013). Changes on teacher self-efficacy and attitudes towards teaching profession from candidacy to induction. *Education and Science*, *38*(169), 109-125.

Camadan, F., & Duysak, A. (2010). Comparing pre-service teachers' attitudes in the different programs toward teaching profession in terms of different variables: Example of Rize University. *The Journal of Sakarya University Education Faculty, 20*(1), 30-42.

Can, Ş. (2010). Attitudes of the students who attend the non-thesis graduated education program towards the teaching profession. Journal of *Mugla University Social Science Institute, 24*(1), 13-28.

Cinpolat, T., Alıncak, F., & Abakay, U. (2016). Examination of the attitudes of physical education and sports college students towards teaching profession. *Gaziantep University Journal of Sport Sciences, 1*(1), 38-47.

Cüceoğlu-Önder, G. (2014). Attitudes of pre-service music teachers towards the teaching profession in Turkey. *Educational Research and Reviews, 9*(18), 703-710. DOI: https://doi.org/10.5897/ERR2014.1770

Çapri, B., & Çelikkaleli, Ö. (2008). Investigation of preservice teachers' attitudes towards teaching and professional self-efficacy beliefs according to their gender, programs, and faculties. *Inonu University Journal of the Faculty of Education, 9*(15), 33-53.

Çelen, A., & Eskicioğlu, Y. (2015). Analysis of attitude toward teachig profession and state-trait anxiety level of students in teaching departments accepting students through special aptitude tests. *Route Educational and Social Science Journal, 2(*3), 1-18.

Çeliköz, M., & Çağdaş, M. (2012). Giyim öğretmen adaylarının öğretmenlik mesleğine yönelik tutumlarının bazı değişkenler açısından incelenmesi. *Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi, 29*(1), 14-28.

Çetinkaya, R. (2007). *Qualification perception of the Turkish teacher candidates and attitudes to teaching occupation* (Master Thesis, Selçuk University Institute of Social Science, Konya, Turkey). Retrieved from http://tez2.yok.gov.tr.

Çetinkaya, Z. (2009). Identifying Turkish pre-service teachers' attitudes toward teaching profession. *Elementary Education Online, 8*(2), 298-305.

Çiğdem, G., & Memiş, A. (2011). Sınıf öğretmenliği adaylarının öğrenme stilleri ve öğretmenlik mesleğine yönelik tutumlarının çeşitli değişkenler açısından incelenmesi. *Çukurova Üniversitesi Eğitim Fakültesi Dergisi, 3*(40), 57-77.

Çimen, L. K. (2016). A study on the prediction of the teaching profession attitudes by communication skills and professional motivation. *Journal of Education and Training Studies, 4*(11), 21-38. DOI: https://doi.org/10.11114/jets.v4i11.1842

Dalkıran, E., & Yıldız, G. (2016). Investigation of teaching profession attitudes of the music education department students. *Fine Arts, 11*(4), 153-160. DOI: http://dx.doi.org/10.12739/NWSA.2016.11.4.D0180

Demircioğlu, E., & Özdemir, M. (2014). Analyzing attitudes of students studying at the faculty of arts and sciences towards teaching profession according to various variables. *Mersin University Journal of the Faculty of Education, 10*(3), 110-122.

Demirtaş, Z., & Aksoy, G. P. (2016). Investigation of pedagogical formation certification program students' attitudes towards teaching profession in terms of some variables. *International Journal of Educational Research Review, 1*(1), 21-28.

Demirtaş, H., Cömert, M., & Özer, N. (2011). Pre-service teachers' self-efficacy beliefs and attitudes towards profession. *Education and Science, 36*(159), 96-111.

Derman, A. (2007). *Chemistry student teachers' self efficacy beliefs and attitudes toward teaching profession*. (Master Thesis, Selçuk University Institute of Science Institute, Konya, Turkey). Retrieved from http://tez2.yok.gov.tr.

Dikmenli, Y., & Çifçi, T. (2015). Pedagogical formation trainees' attitudes towards teaching profession and their field knowledge. Cumhuriyet University Faculty of Literature Journal of Social Sciences, *39*(2), 155-172.

Doğan, T., & Çoban, A. E. (2009). The investigation of the relations between students' attitude toward teaching profession and anxiety level in faculty of education. *Education and Science, 34*(153), 157-159.

Dönmez, C., & Uslu, S. (2013). The attitudes of social studies teacher candidates' towards teaching profession. *The Journal of Turkish Educational Sciences, 11*(1), 42-63.

Durmuşoğlu, M. C., Yanık, C., & Akkoyunlu, B. (2009). Turkish and Azerbaijani prospective teachers' attitudes to their profession. *Hacettepe University Journal of Education, 36*(1), 76-86.

Engin, G., & Koç, G. Ç. (2014). The attitudes of prospective teachers towards teaching (the case of Ege University, Faculty of Education). *The Journal of Turkish Social Research, 182*(2), 153-168.

Eraslan, L., & Çakıcı, D. (2011). Pedagogical formation program students 'attitudes towards teaching profession. *Kastamonu Education Journal, 19*(2), 427-438.

Erbas, M. K. (2014). The relationship between alienation levels of physical education teacher candidates and their attitudes towards the teaching profession. *Australian Journal of Teacher Education, 39*(8), 37-52. DOI: 10.14221/ajte.2014v39n8.3

Erdoğan, D. G., & Güneş, D. Z. (2012). The attitudes of the first grade students' towards teaching profession at education faculty of Sakarya University. *Journal of Uludağ University Faculty of Education, 25*(1), 51-62.

Ergen, Y., & Töman, U. (2014). Research of classroom teacher 4. grade students 'attitudes towards teaching profession (Bayburt University Faculty of Education sample). *Journal of Research in Education and Teaching, 3*(1), 375-383.

Eroglu, C., & Unlu, H. (2015). Self-efficacy: Its effects on physical education teacher candidates' attitudes toward the teaching profession. *Educational Sciences: Theory and Practice, 15*(1), 201-212. DOI 10.12738/estp.2015.1.2282

Ertem, S., & Kete, R. (2015, Mayıs). *Formasyon öğretmen adaylarının mesleki tutum ve beklentilerinin farklı değişkenlere göre karşılaştırılması.* VII. Ulusal Lisansüstü Eğitim Sempozyumu, Sakaya, Türkiye.

Fadlelmula, F. K. (2013). Pre-service teachers' learning styles and attitudes toward teaching profession. *Turkish Journal of Education, 2*(4), 55-63.

Gökçe, F., & Sezer, G. O. (2012). The attitudes of student teachers towards teaching profession: Uludag University sample. *Journal of Uludag University Faculty of Education, 25*(1), 1-23.

Göktaş, Z. (2017). Pre-service teachers' attitudes towards teaching profession in the school of physical education and sports at Balikesir University. *The Journal of International Social Research, 10*(51), 1288-1295. DOI: http://dx.doi.org/10.17719/jisr.2017.1856

Güneyli, A., & Aslan, C. (2009). Evaluation of Turkish prospective teachers' attitudes towards teaching profession (Near East University case). *Procedia-Social and Behavioral Sciences, 1*(1), 313-319.

Hançer, A. H. (2017). Effects of science teacher candidates' attitudes towards the teaching profession on academic fraud inclinations. *Electronic Turkish Studies, 12*(6), 387-402. DOI: 10.7827/TurkishStudies.11418

İlğan, A., Sevinç, Ö. S., & Arı, E. (2013). The perceptions of teachers towards professional attitude contemporary teachers qualifications. *Ondokuz Mayıs University Journal of Education Faculty, 32*(2), 175-195.

İpek, C., & Camadan, F. (2012). Primary teachers' and primary pre-service teachers' self-efficacy beliefs and attitudes toward teaching profession. *Journal of Human Sciences, 9*(2), 1206-1216.

İpek, C., Kahveci, G., & Camadan, F. (2015). The attitudes of pre-service class teachers towards teaching profession and school principalship. *Kastamonu Education Journal, 23*(1), 211-226.

Kahyaoğlu, M., Tan, Ç., & Kaya, M. F. (2013). Attitudes of elementary school teacher candidates towards learning styles and teaching profession. *Mustafa Kemal University Journal of Social Sciences Institute, 10*(21), 225-236.

Kalemoğlu-Varol, Y., Erbaş, M. K., & Ünlü, H. (2014). Beden eğitimi öğretmen adaylarının mesleki kaygı düzeylerinin öğretmenlik mesleğine yönelik tutumlarını yordama gücü. *Ankara Üniversitesi Spor Bilimleri Fakültesi Dergisi, 12*(2), 113-123.

Kaplan, A., & İpek, A. S. (2002). Examining the attitudes of mathematics teacher candidates towards the teaching profession. *Education and Science*, *27*(125), 69-73.

Kartal, T., & Afacan, Ö. (2012). Examining attitudes of prospective teachers who took pedagogical formation education towards teaching profession. *Mehmet Akif Ersoy University Journal of Education Faculty, 24*(1), 76-96. DOI: 10.1016/j.sbspro.2012.05.561

Kartal, T., Kaya, V. H., Öztürk, N., & Ekici, G. (2012). The exploration of the attitudes of the science teacher candidates towards teaching profession. *Procedia-Social and Behavioral Sciences, 46*(2012), 2759-2764. DOI: http://dx.doi.org/10.12973/jesr. 2015.51.2

Kaya, Ç., & Kaya, S. (2015). Relationship between prospective teachers' dysfunctional attitudes and their attitudes towards the teaching profession. *Journal of Educational Sciences Research*, *5*(1), 23-40.

Keskin, Y. (2017). Attitude and concern levels of geography teacher candidates towards the profession of teaching (Erzurum example). *E-Kafkas Journal Of Educational Research, 4*(2), 43-57.

Kılıç, D., & Bektaş, F. (2008). Evaluating of the attitudes of the class teacher candidates oriented to the teaching job. *Journal of Kazım Karabekir Education Faculty, 18*(1), 15-25.

Kılıç, S. K., Cihan, H., & Öncü, E. (2015). Metacognitive learning strategies and academic self-efficacy of pre-service physical education teachers and their attitudes towards the profession of teaching. *Hacettepe Journal of Sport Sciences, 26*(3), 77-89.

Kiralp, F. S. S., & Bolkan, A. (2016). Relationship between candidate teacher's attitude towards teaching profession and their life satisfaction levels. *The Anthropologist, 23*(1-2), 11-20. DOI: 10.1080/09720073.2016.11891919

Kızıltaş, E., Halmatov, M., & Sarıçam, H. (2012). Preschool department students' attitudes about profession of teaching (the case of Agri Ibrahim Cecen University). *Mehmet Akif Ersoy University Journal of Education Faculty, 12*(23), 173-189.

Kocaarslan, M. (2014). Analysis of prospective teachers' attitudes towards teaching as a profession. *Asian Journal of Instruction*, *2*(1), 46-55.

Küçükosmanoğlu, H., & Can, M. (2013). Attitudes towards teaching of N.E.U A.K.E.F. music education department students. Journal of Research in Education and Teaching, *2(*4), 338-343.

Kutlu, N., Gökdere, M., & Çakır, R. (2015). A comparative study of the attitudes of the prospective teachers towards the academic procrastination behavior and the teaching profession. *Kastamonu Education Journal*, *23*(3), 1311-1330.

Nakip, C., & Özcan, G. (2016). The relation between preservice teachers' sense of self efficacy and attitudes toward teaching profession. *Mersin University Journal of the Faculty of Education, 12*(3), 783-795. DOI: http://dx.doi.org/10.17860/mersinefd.282380

Oğuz, E., & Kalkan, M. (2011). Examining teacher candidates' attitudes towards teaching profession and pupil control ideology. *International Online Journal of Educational Sciences, 3*(3), 903-917.

Okçu, V., & Çelik, C. (2011). Effect of candidate teachers' opinions to public personnel selection examination (PPSE) on attitudes of teaching. *The International Journal of Research in Teacher Education, 2*(1), 30-54.

Orhan, N. (2013). *Job satisfaction and occupational attitude levels of trainee teachers* (Master Thesis, Dokuz Eylül University Institute of Educational Sciences, İzmir, Turkey). Retrieved from http://tez2.yok.gov.tr/.

Orhan, E. E., & Ok, A. (2014). Who prefer teacher education programs? Candidates' entry characteristics and attitude towards teaching. *Hacettepe University Journal of Education, 29*(4), 75-92.

Ömür, Y. E., & Nartgün, Ş. S. (2013). The relationship between teacher candidates' attitudes towards teaching profession and motivational levels. *Journal of Policy Analysis in Education, 2*(2), 41-55.

Özdemir, V. (2014). A research on the attitudes of English language teacher candidates in terms of demographic variables and public personnel selection exam. *International Journal of Eurasia Social Sciences, 5*(15), 16-28. DOI: 10.5430/ijhe.v6n3p57

Özdemir, Y., & Gungor, S. (2017). Attitudes of students enrolled in the pedagogical formation programs towards the teaching profession. *International Journal of Higher Education, 6*(3), 57-69. DOI: 10.5430/ijhe.v6n3p57

Özder, H., Konedralı, G., & Zeki, C. P. (2010). Examining the attitudes towards the teaching profession and academic achievements of prospective teachers. *Educational Administration: Theory and Practice, 16*(2), 253-275.

Pehlivan, H. (2017). An analysis of general high school teachers' attitudes towards teaching profession. *Journal of Human Sciences, 14*(3), 2244-2258. DOI: 10.14687/jhs.v14i3.4527

Pehlivan, K. B. (2008). A study on the socio-cultural characteristics of classroom teacher candidates and their attitudes towards teaching profession. *Mersin University Education Faculty Journal, 4*(2), 151-168.

Pektaş, M., & Kamer, S. T. (2011). The attitudes of science teacher trainees for teaching profession. *The Journal of Turkish Educational Sciences, 9*(4), 829-850.

Polat, S. (2013). Examination of the pedagogical formation certificate program and education faculty students' attitudes towards teaching profession. *e-Journal of International Educational Researches*, *4*(2), 48-60.

Recepoğlu, E. (2013). Analyzing the relationship between prospective teachers' life satisfaction and attitudes concerning teaching profession. *H. U. Journal of Education, Special Issue*(1), 311-326.

Sağlam, A. Ç. (2008). The attitudes of the branch of music students toward the teaching profession. *YYU Journal of Education Faculty*, *5*(1), 59-69.

Saracaloğlu, A. S., & Dursun, F. (2011). The relationship between classroom teacher candidates' attitudes towards teaching profession and learning strategies. *Education Sciences, 6*(3), 2328-2346.

Sezer, A., Pınar, A., & Yıldırım, T. (2010). An investigation of geography student teachers' profiles and attitudes toward teaching profession. *Journal of Marmara Geography, 22*, 43-69.

Süral, S. (2015). The relationship between elementary school teachers' teaching styles, classroom management approaches and attitudes towards teaching profession. *Journal of International Social Research, 8*(41), 1027-1038.

Şahin, F. S. (2010). Teacher candidates' attitudes towards teaching profession and life satisfaction levels. *Procedia-Social and Behavioral Sciences, 2(*2), 5196-5201. DOI: 10.1016/j.sbspro.2010.03.845

Şendil, C. (2017). Pedagogical formation program students' opinions towards teaching profession. *Journal of Kirsehir Education Faculty*, *18*(1), 595-611.

Şentürk, Z., Zahal, O., Yurga, C., Gürpınar, E., & Altun, F. (2016). Investigation of music teachers' attitudes toward teaching profession according to their profile properties. *Journal of Human Sciences, 13*(3), 5032-5052. DOI: 10.14687/jhs.v13i3.4024

Tanel, R., Şengören, S. K., & Tanel, Z. (2007). Investigating attitudes of prospective physics teachers towards teaching as a profession regarding various variables. *Pamukkale University Journal of Education, 22*, 1-9.

Tanrıöğen, A. (1997). The attitudes of the students at Buca Faculty of Education towards teaching profession. *Pamukkale University Journal of Education, 3*(3), 55-67.

Taşdemir, C. (2014). İlköğretim matematik öğretmen adaylarının öğretmenlik mesleğine yönelik tutumlarının incelenmesi. *Bilgisayar ve Eğitim Araştırmaları Dergisi, 2*(3), 91-113.

Taşkın, Ç. Ş., & Hacıömeroğlu, G. (2010). Examining preservice teachers' attitudes towards teaching profession in elementary education: A combination of quantitative and qualitative methods. *Elementary Education Online, 9*(3), 922-933,

Terzi, A. R., & Tezci, E. (2007). Necatibey Education Faculty students' attitudes towards teaching profession. *Educational Management in Theory and Practice, 52*(1), 593-614.

Timur, B., & İmer-Çetin, N. (2017). Examining self-efficacy beliefs and attitudes of pre-service science teachers' and pedogogical proficiency students' towards science teaching profession. *International Journal of Active Learning, 2*(2), 15-27.

Tok, T. N. (2012). Teacher candidates' attitudes towards the teaching profession in Turkey. *Alberta Journal of Educational Research, 58*(3), 381-403.

Toprak, N., & Saraç, L. (2014). An examination of attitudes toward teaching profession among female and male physical education and sports department entrance examination applicants. *Pamukkale Journal of Sport Sciences, 5*(2), 35-47.

Tunçeli, H. İ. (2013). The relationship between candidate teachers' communication skills and their attitudes towards teaching profession (Sakarya University sample). *Pegem Journal of Education & Instruction, 3*(3), 51-58.

Tuncer, M., & Bahadır, F. (2016). Evaluation of teacher candidates in terms of attitudes towards technopedagogical field competencies and teaching profession. *Electronic Turkish Studies*, *11*(9), 839-858. DOI: http://dx.doi.org/10.7827/TurkishStudies.9635

Tüfekçi, A., & Kocabatmaz, H. (2015). Evaluation of prospective information technology teachers' attitudes towards teaching profession. *Gazi University Journal of Educational Faculty*, *35*(3), 523-555.

Tüfekçi-Aslim, S. (2013). Evaluation of the attitudes of candidate elementary teachers to the profession of teaching. *The Journal of Industrial Arts Education Faculty of Gazi University*, *32*(1), 65-81.

Tümkaya, S. (2011). Investigation of communication skills and teaching attitudes of students in classroom teaching. *Çukurova University Journal of Social Sciences Institute*, *20*(2), 49-62.

Uğurlu, C. T. (2012). The attitudes of Anatolian teacher training high school students towards the profession of teaching (Adıyaman province case). *Journal of Uludag University Faculty of Education, 25*(1), 217-232.

Uygun, M., & Kunt, H. (2017). An analysis of the relationship between prospective teachers' thinking styles and their attitudes to teaching profession according to various variables. *International Electronic Journal of Elementary Education, 6*(2), 357-370.

Uyulgan, M. A., & Kartal, M. (2012). Evaluation of the senior students' chemistry subject matter knowledge and attitudes towards teaching profession in faculty of science chemistry department and faculty of education chemistry education department. *Buca Education Faculty Journal, 32*(1), 104-118.

Üstüner, M., Demirtaş, H., & Cömert, M. (2010). The attitudes of prospective teachers towards the profession of teaching (the case of Inonu University, Faculty of Education). *Education and Science*, *34(*151), 140-155.

Yalız, D. (2010). Comparison attitudes towards teaching profession of students in the department of physical education and sports teaching at Anadolu University. *CBU Journal of Physical Education and Sport*, Sciences, *5*(1), 7-14.

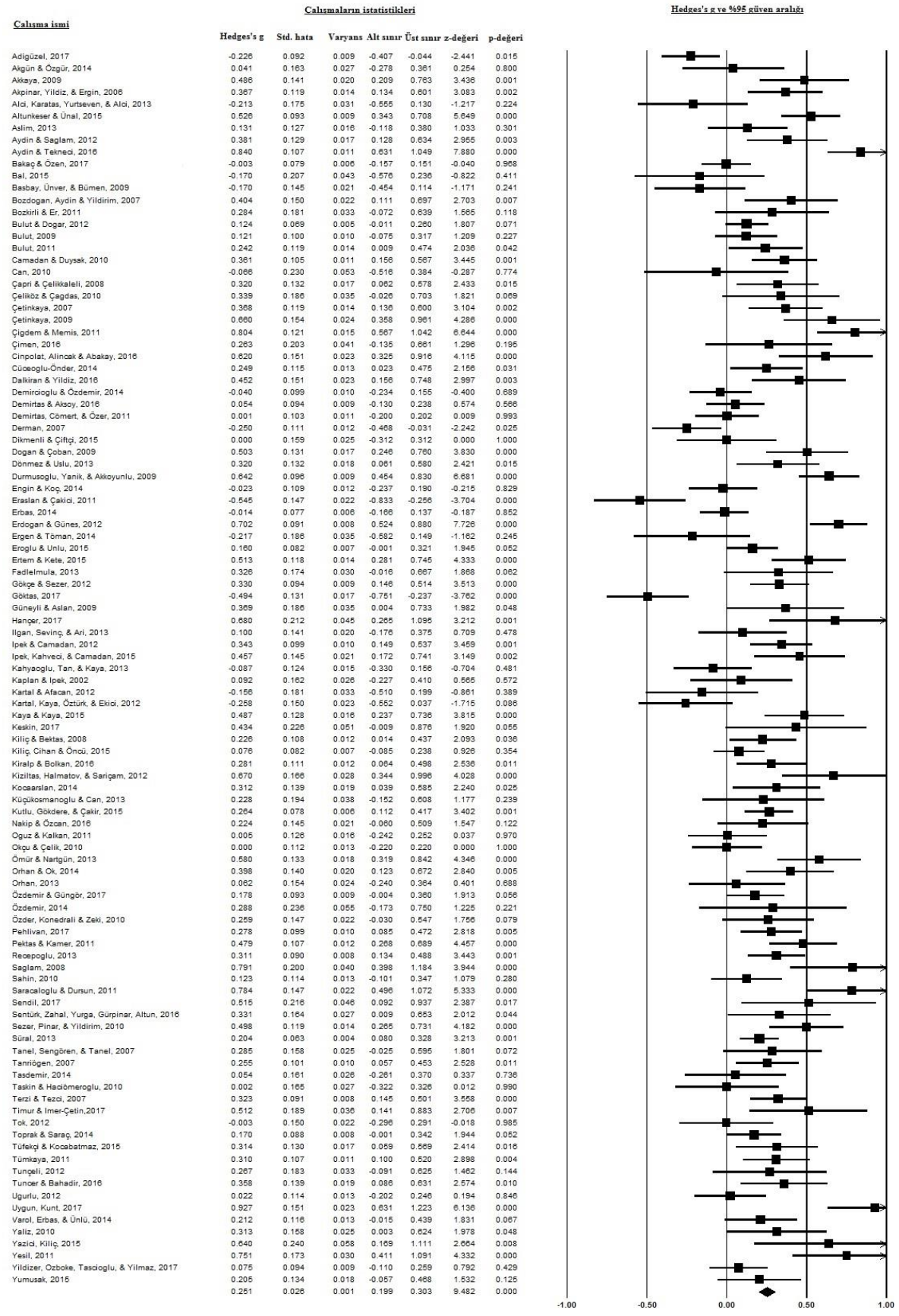
Yazıcı, T., & Kılıç, I. (2015). Attitude of students in faculties of fine arts and design and the conservatories to profession of music teaching. *İnönü University Journal of Art and Design, 5*(11), 79-88. DOI: 10.16950/std.30346

Yeşil, H. (2011). Turkish language teaching students' attitudes towards teaching profession. *International Online Journal of Educational Sciences, 3*(1), 200-219.

Yildizer, G., Ozboke, C., Tascioglu, R. ve Yilmaz, I. (2017). Examining attitudes of physical education teacher education program students toward the teaching profession. *Montenegrin Journal of Sports Science and Medicine*, *6*(2), 27-33.

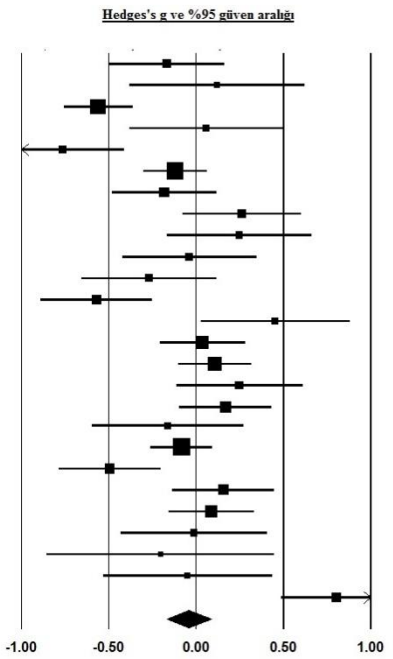
Yumuşak, G. K. (2015). Teacher candidates' reflective thinking tendencies and attitudes towards the profession. *Bartın University Education Faculty Journal, 4*(2), 466-481. DOI: 10.14686/buefad.v4i2.1082000206

Appendix-2: Forest Diagram of Effect Sizes of Studies Including "Gender" Variable



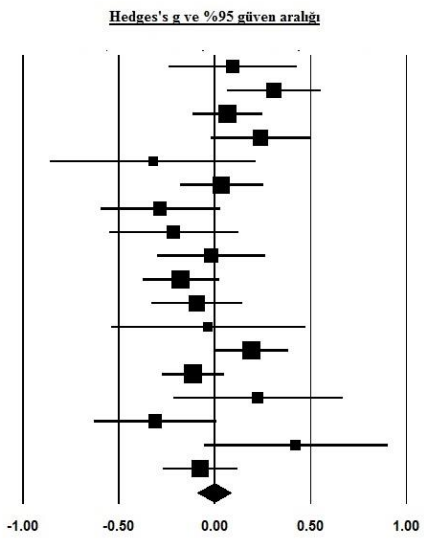
Appendix-3: Forest Diagram of Effect Sizes of Studies Including "Grade Level" Variable

Çalışma ismi	Çalışmaların istatistikleri						
	Hedges's g	Std. hata	Varyans	Alt sınır	Üst sınır	z-değeri	p-değeri
Akpinar, Yildiz, & Ergin, 2006	-0.167	0.168	0.028	-0.496	0.161	-0.998	0.318
Aktop & Beyazgül, 2014	0.121	0.256	0.066	-0.382	0.623	0.470	0.638
Aydin & Tekneci, 2016	-0.561	0.101	0.010	-0.758	-0.363	-5.561	0.000
Bümen & Özaydin, 2013	0.056	0.224	0.050	-0.384	0.495	0.248	0.804
Çelen & Eskicioglu, 2015	-0.764	0.181	0.033	-1.118	-0.410	-4.229	0.000
Çeliköz & Çagdas, 2010	-0.119	0.093	0.009	-0.302	0.063	-1.280	0.201
Çigdem & Memis, 2011	-0.182	0.153	0.023	-0.482	0.118	-1.186	0.236
Cüceoglu & Önder, 2014	0.262	0.172	0.030	-0.076	0.600	1.521	0.128
Dalkiran & Yildiz, 2016	0.248	0.212	0.045	-0.167	0.662	1.170	0.242
Durmusoglu, Yanik, & Akkoyunlu, 2009	-0.039	0.196	0.039	-0.424	0.346	-0.198	0.843
Göktas, 2017	-0.270	0.197	0.039	-0.657	0.117	-1.368	0.171
Kaplan & Ipek, 2002	-0.571	0.164	0.027	-0.892	-0.251	-3.492	0.000
Kaya & Kaya, 2015	0.453	0.217	0.047	0.027	0.878	2.084	0.037
Kiliç, Cihan, & Öncü, 2015	0.036	0.124	0.016	-0.208	0.280	0.291	0.771
Kutlu, Gökdere, & Çakir, 2015	0.106	0.107	0.011	-0.103	0.316	0.997	0.319
Pehlivan, 2008	0.248	0.184	0.034	-0.113	0.609	1.346	0.178
Saracaloglu & Dursun, 2011	0.168	0.135	0.018	-0.096	0.432	1.250	0.211
Tasdemir, 2014	-0.161	0.222	0.049	-0.596	0.274	-0.727	0.467
Terzi & Tezci, 2007	-0.084	0.090	0.008	-0.261	0.093	-0.931	0.352
Tümkaya, 2011	-0.494	0.149	0.022	-0.786	-0.203	-3.322	0.001
Tunçeli, 2013	0.156	0.149	0.022	-0.136	0.448	1.048	0.295
Üstüner, Demirtas, & Cömert, 2010	0.087	0.125	0.016	-0.158	0.333	0.696	0.487
Yaliz, 2010	-0.013	0.214	0.046	-0.433	0.407	-0.060	0.952
Yazici & Kiliç, 2015	-0.205	0.332	0.111	-0.856	0.447	-0.615	0.538
Yesil, 2011	-0.048	0.248	0.061	-0.533	0.437	-0.195	0.845
Yildizer et al., 2017	0.805	0.164	0.027	0.484	1.125	4.915	0.000
	-0.041	0.066	0.004	-0.170	0.088	-0.617	0.537



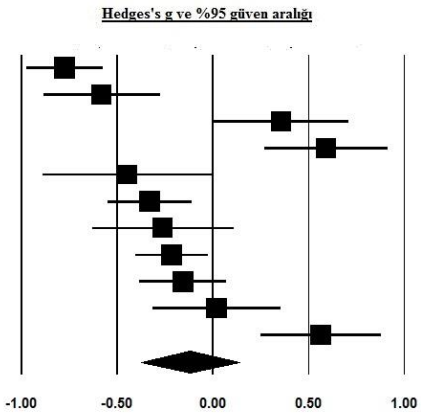
Appendix-4: Forest Diagram of Effect Sizes of Studies Including "The Presence of a Teacher in the Family" Variable

Çalışma ismi	Çalışmaların istatistikleri						
	Hedges's g	Std. hata	Varyans	Alt sınır	Üst sınır	z-değeri	p-değeri
Akgün & Özgür, 2014	0.094	0.169	0.029	-0.238	0.426	0.553	0.580
Akpinar, Yildiz, & Ergin, 2006	0.309	0.125	0.016	0.063	0.554	2.466	0.014
Aslim, 2013	0.067	0.094	0.009	-0.117	0.250	0.711	0.477
Aydin & Saglam, 2012	0.240	0.132	0.017	-0.019	0.498	1.817	0.069
Can, 2010	-0.321	0.274	0.075	-0.858	0.215	-1.174	0.240
Çeliköz & Çagdas, 2010	0.035	0.110	0.012	-0.181	0.251	0.317	0.751
Çetinkaya, 2009	-0.285	0.159	0.025	-0.597	0.027	-1.791	0.073
Kartal & Afacan, 2012	-0.215	0.172	0.029	-0.552	0.121	-1.255	0.210
Kocaarslan, 2014	-0.018	0.144	0.021	-0.301	0.265	-0.125	0.900
Kutlu, Gökdere, & Çakir, 2015	-0.178	0.102	0.010	-0.378	0.023	-1.739	0.082
Orhan & Ok, 2014	-0.092	0.121	0.015	-0.330	0.145	-0.762	0.446
Özdemir,2014	-0.035	0.259	0.067	-0.542	0.472	-0.135	0.892
Pehlivan, 2017	0.191	0.097	0.009	0.001	0.382	1.966	0.049
Recepoglu, 2013	-0.113	0.082	0.007	-0.273	0.048	-1.378	0.168
Sendil, 2017	0.225	0.226	0.051	-0.218	0.668	0.996	0.319
Tanel, Sengören, & Tanel, 2007	-0.311	0.162	0.026	-0.629	0.007	-1.914	0.056
Yazici & Kiliç, 2015	0.422	0.244	0.060	-0.057	0.901	1.727	0.084
Yildizer, Ozboke, Tascioglu, & Yilmaz, 2017	0.075	0.099	0.010	-0.269	0.118	-0.763	0.445
	-0.003	0.045	0.002	-0.092	0.085	-0.073	0.942



Appendix-5: Forest Diagram of Effect Sizes of Studies Including "Faculty Type" Variable

Çalışma ismi	Çalışmaların istatistikleri						
	Hedges's g	Std. hata	Varyans	Alt sınır	Üst sınır	z-değeri	p-değeri
Arastaman, 2013	-0.773	0.102	0.010	-0.974	-0.573	-7.574	0.000
Bagçeci, Yildirim, Kara, & Keskinpalta, 2015	-0.583	0.156	0.024	-0.888	-0.278	-3.746	0.000
Bozkirli & Er, 2011	0.356	0.180	0.032	0.004	0.709	1.980	0.048
Kaplan & Ipek, 2002	0.590	0.164	0.027	0.269	0.911	3.599	0.000
Keskin, 2017	-0.446	0.226	0.051	-0.889	-0.003	-1.975	0.048
Ömür & Nartgün, 2013	-0.329	0.112	0.012	-0.548	-0.110	-2.946	0.003
Orhan, 2013	-0.261	0.188	0.035	-0.629	0.106	-1.393	0.164
Polat, 2013	-0.216	0.098	0.010	-0.407	-0.025	-2.216	0.027
Sezer, Pinar, & Yildirim, 2010	-0.157	0.117	0.014	-0.386	0.071	-1.348	0.178
Timur & Imer-Çetin,2017	0.019	0.170	0.029	-0.315	0.353	0.112	0.911
Uyulgan & Kartal, 2012	0.563	0.161	0.026	0.246	0.879	3.486	0.000
	-0.119	0.132	0.018	-0.378	0.141	-0.896	0.370



ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(4);414-421

Exploratory and Confirmatory Factor Analysis: Which One to Use First?*

Fatih ORÇAN**

Abstract
There exist differences between the use of Exploratory and Confirmatory Factor analysis at scale adaptation or development studies. The order of factor analysis used would cause the discrepancy in the results. Besides, multiple confirmatory factor analysis would fit well on a single data set. In this study simulated data sets were fitted to three different models. Based on the results 64% of the data sets fit well on all three models. Also, a different data set was fit both on a confirmatory and an exploratory factor analysis. The result showed that confirmatory factor analyses were not sufficient to detect the best fitting model.

Keywords: Scale adaptation, scale development, confirmatory factor analysis, exploratory factor analysis

INTRODUCTION

Confirmatory Factor Analysis (CFA) and Exploratory Factor Analysis (EFA) are two common techniques used in scale development and scale adaptation studies. If the relationship among the items is not known it is recommended to use EFA, but if the relationship is tested and the factors and related items are known, CFA is recommended to be used (Bandalos & Finney, 2010; Büyüköztürk, 2002; Kline, 2011). For scale adaptation studies the use of these methods and their orders of use showed diversity form one study to another. Güvendir and Özkan (2015) explored scale adaptation and development studies published in Turkey between 2006 and 2014. Based on their results, total of 25 studies used EFA out of 26 scale development studies and 16 of them used CFA. Moreover, 22 scale development studies started with EFA to analyze their data while 11 stared with CFA.

Experts may guess how items will be structured beforehand; however, a statistical technique is required to decide about the structure of the items and number of latent factors. Thus, the items which works (explains variation) could be determined easily. Therefore, for a scale development study first an EFA should be used in order to discover underlying latent structure (Brown, 2006; Schumacker & Lomax, 2010). In fact, 96% of the studies in Turkey used EFA (Güvendir & Özkan, 2015). Besides, in a scale development process, CFA should be run using a data set different from the EFA data set (Schumacker & Lomax, 2010). Thus, the validity of the EFA structure found as a result of EFA will be shown by using CFA with a different data set. Two different ways can be discussed in the creation of the data set to be used for factor analysis. First, after a sufficient number of samples are collected in a single run to make both EFA and CFA, some of them (eg 50%) can be randomly selected for EFA and the rest for CFA. Another way is to collect two different data sets and analyze one for EFA and the other for CFA.

In adaptation studies, the use of EFA and CFA varies. For example, the process of translating the items from the original language to a new language is an important step for scale adaptation studies. Failure of transferring the original item meanings may cause a variation called scale error in scale scores. As a result of this meaning shift, it is possible to create a structure different from the original scale structure. Therefore, in an adaptation study, it is necessary to make sure that the translation of the item is done correctly before starting the analysis. A coherent translation process is very important for the elimination of structural differences. Sousa and Rojjanasrirat (2011) defined a step by step process for

* A part of this paper was presented at VII. International Congress of Research in Education.
** Dr.Öğr. Üyesi, Trabzon Üniversitesi, Fatih Eğitim Fakültesi, Trabzon-Türkiye, e-posta: fatihorcan@ktu.edu.tr, ORCID ID: https://orcid.org/0000-0003-17270456

To cite this article:
Orçan, F., (2018). Exploratory and confirmatory factor analysis: which one to use first?. *Journal of Measurement and Evaluation in Education and Psychology, 9*(4), 414-421. DOI: 10.21031/epod.394323

Received: 13.02.2018
Accepted: 28.09.2018

translating a scale into another language. According to Souse and Rojjanasrirat (2011), at least two people should translate the article first (forward translation). Then the work of these two independent translators should be reviewed by a third expert and the translation should be finalized. In the third step, the translated materials should be translated back into the original language at least by two different experts (back translation) and the final version of the scale should be obtained after these translations are examined by a third independent expert. In subsequent steps, the pilot and actual implementation stages are presented in detail (Sousa & Rojjanasrirat, 2011). A similar process was made by Sperber (2004). Sperber (2004) also stressed that word to word translation may not be accurate and the items translation should be culturally adaptive in order to prevent meaning shifts.

It is obvious that the translation error will affect the validity and reliability of the adapted scale. Therefore, in this study psychometric tests, as mentioned by Sousa and Rojjanasrirat's (2011) the seventh step of the process, used to test the validity and reliability (EFA and CFA) were considered. In this study the reasons and the order of use of these techniques were examined via simulated data in order to explore the possible differences in the results.

EFA is a statistical technique used in the social sciences for determining underling latent variables (factors). In other words, EFA stands out as a technique used in scale development. It is used where there is no knowledge among the items of the scale, that is, how many factors there are between the items and which factors are determined by which items. As the name suggests, EFA helps *explain* the structure that exists (Hayton, Allen, & Scarpello, 2004; Hurley, Scandura, Schriesheim, Brannick, Seers, et al., 1997). Some critical decisions need to be made during EFA, such as which method of estimation will be used, whether rotation will be made or by which criteria the number of factors will be determined. There are many studies in the literature about these (Costello & Osborne, 2005; Hanson & Roberts, 2006; Schmitt, 2011). Therefore, these factors were kept constant in the study. More detailed information about these concepts (transformation, sample size or number of factors) can be found in Büyüköztürk (2002), Costello and Osborne (2005).

Unlike EFA, CFA is used when there is a strong model assumption. With CFA, the existence of a previously proven structure is investigated with a new data set. In scale development studies, CFA should be used to test the validity of the structure obtained after EFA (Worthington & Whittaker, 2006). However, the use of CFA in scale adaptation studies differs in practice. In some adaptation studies, it is seen that both EFA and CFA are used, while in others only CFA is used. Use of CFA only in adaptation studies may cause some problems. For example, if a translation error occurred in an adaptation study, using the CFA only might result in a different situation than would actually occur, and the model could be misleading. In addition, a data set may fit with more than one CFA model, so it would be more appropriate to conduct an EFA first to introduce possible cultural differences in the adaptation. In such a case, if an EFA is not performed, a researcher will not test a second model since the first tested model fit to the data. Thus, it is important to run an EFA first to recognize the possible error.

Purpose of Study

The main purpose of this study was to determine how a data set can fit to more than one CFA model, and also how the use of EFA or CFA first may differ in the outcome model. In this study, the models were compared in two respects. Firstly, data were generated according to the model shown in Figure 1 in the R-cran program and these data were tested according to three different CFA models. Second, a simulated data set was evaluated on the basis of item and the possible differences that could occur as a result of using EFA or CFA first (as an example of scale adaptation or development procedures) were revealed. Thus, it is aimed to show what the different scale development procedures can produce.

METHOD

For the simulation part of the study, 100 data sets were simulated via R-cran for the sample size of 300 with model shown at Figure 1. Since the sample size is not a design factor in this study, as indicated

in Orçan and Yang (2016), 300 samples will be sufficient for this study. However, as the model complexity increases, the need for sample size will increase. The model consisted of two factors and eight observed variables. For four items loaded on each factor, factor loadings were set at .40, .50, .60 and .70 to provide diversity. In addition, the correlation between factors was determined as .70. This model was preferred in order not to increase the model complexity. The aim of the study is not to compare the behavior of CFA and EFA under different conditions. But it is aimed to show that the same data set can fit more than one model. Therefore, it would be sufficient to examine a single case to show that a data set might fit well more than one CFA model.

To produce data according to the model, first, the factor scores were randomly generated with the mean of 0 and the standard deviation of 1. Then, Cholesky method has been applied to ensure that the item data is multinomial in accordance with the specified factor loadings. Observed values were obtained as a linear combination of factor scores (Orçan & Yang, 2016). Finally, the simulated continuous variables converted into five categories in order to reflect the five-point Likert item properties. For the second aim of the study, a data set with sample size of 300 was tested with EFA and CFA models.

Analysis

Mplus 5.1 (Muthén & Muthén, 1998-2008) was used to analyze the data. 100 data sets were categorically generated based on the model (Model 1) as in Figure 1 that is correctly specified model. The data sets were analyzed according to model 1and two misspecified models; Model 2 where item 5 is loaded on the first factor and Model 3 where item 4 is loaded on the second factor See figure 2). These models are shown in figure 2. Factor loadings of items 4 and 5 are lower than others. In case of high factor loadings, the item-factor correlation will be higher and misspecification will be more prominent. However, for the purpose of the study lower factor loadings will be sufficient. In another study, factor loadings can be taken as research design.

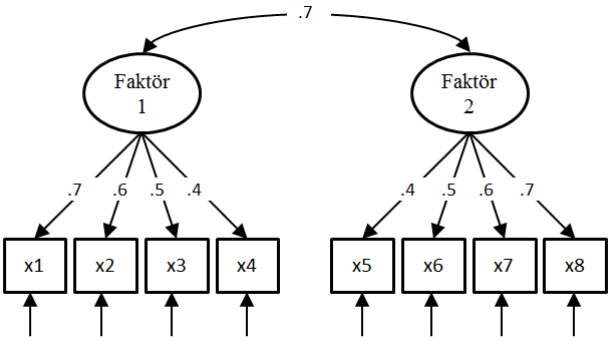
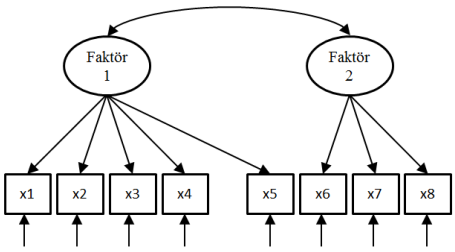
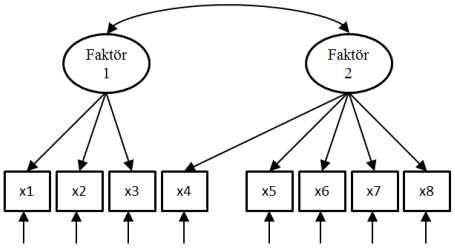


Figure 1. Data Generation Model (Model 1)

Although the data were generated according to normal distribution, during the categorization process the data were distanced from normality. Therefore Maximum Likelihood (MLR) estimation method was used for the models. For each of the models the p-value of the chi-square test, the comparative fit indices (CFI) , the root mean square error of approximation (RMSEA) and the standardized root mean square residual (SRMR) values were compared with Hu and Bentler's (1999) criteria. Besides, the correlation between the factors was examined via the descriptive statistics and the root mean square error (RMSE).



Model 2



Model 3

Figure 2. Misspecified Models

RESULTS

First, the models were evaluated based on chi-square, CFI, RMSEA and SRMR values. The number of data where model fit indices indicated good fit was shown at table 1. For example, 87% of the data fitted well for model 1 in terms of chi-square. However, 64% of the same data sets fit models 2 and 3. Similarly, for the RMSEA value, all the data fit to model 1 (100%), whereas for model 2 and model 3 these values were 91% and 89% respectively. Therefore, misspecified models model 2 and model 3 were considered to be true at 91% and 89%, respectively.

Table 1. Model-Data Fit Counts

	Model-Data Fit: Yes		
	Model 1	Model 2	Model 3
Ki-Kare	87	64	64
CFI	98	80	70
RMSEA	100	91	89
SRMR	100	100	100

The CFA models are generally evaluated based on the four fit indices. That is, a CFA model said to show a good model-data fit if the p-value of chi-square test is higher than .05, the CFI is higher than .95, the RMSEA and SRMR values are less than .06 and .08, respectively (Hu & Bentler, 1999). Table 2 shows the number of fit indices which indicates good model-data fit. For example, under model 1, 87% of the data-sets indicated good fit for four indices at the same time. This value is 64% for model 2 and 3. In detail, out of these 87 data-sets 63 of them also showed good model data fit for four indices under model 2 and 3. That is, even for misspecified model all four fit indices indicated good model data fit.

Table 2. Number of Fit Indices

		Model 1				
	Number of Indices	1	2	3	4	Total
Model 2	1	0	1	6	2	9
	2	0	1	2	8	11
	3	0	0	1	14	16
	4	0	0	1	63	64
Model 3	1	0	2	5	4	11
	2	0	0	5	14	19
	3	0	0	0	6	6
	4	0	0	1	63	64
	Total	0	2	10	87	100

The correlation coefficient between the factors is an important research question for many studies. In this study, root mean square error (RMSE) was calculated for the correlation coefficients obtained from the models by using .70 for the true correlation coefficient. Table 3 shows the descriptive values

of the correlation coefficients obtained from each model. Considering the table, the correlation coefficients for models 2 and 3 seem to have been greater than the true value. As shown in Table 3, the mean correlation coefficient for these models was .74 and .75, respectively. Besides, the RMSE values of models 2 and 3 were also higher than for the value of Model 1.

Table 3. Estimated Correlation Coefficients Statistics

	Min	Max	Mean	S. Deviation	RMSE
Model 1	.56	.89	.71	.07	.069
Model 2	.58	.97	.74	.07	.086
Model 3	.60	.94	.75	.08	.092

An example for Application

How the use of EFA or CFA first may changes the results of scale adaptation or development study was investigated in this section. Table 4 shows the correlation coefficients between the items and the mean and standard deviation values of the items in a data set with sample size of 300.

Table 4. Item Correlation and Descriptive Statistics

Items	M1	M2	M3	M4	M5	M6	M7	M8
M2	.34							
M3	.38	.32						
M4	.21	.21	.22					
M5	.19	.22	.19	.02*				
M6	.26	.20	.18	.15	.17			
M7	.23	.23	.24	.18	.12	.37		
M8	.32	.24	.22	.13	.21	.34	.37	1.00
Mean	3.00	3.01	3.01	2.97	3.04	3.01	3.03	3.06
Standard Deviation	.93	.92	.89	.91	.84	.94	.94	.85

* p>.05

This data set was first tested with each CFA models in the Mplus 5.1 program. MLR was used for the estimation. The model results were shown in table 5. Based on the results each model indicated good model-data fit in terms of all the fit indices. Mplus modification indexes had also given no warning. In the light of these results, a researcher who had started to research with any of the model will not need to try a second model because he/she had a good model-data fit already. Due to the nature of CFA, there is also no need for such a search when the model was *confirmed*. Therefore, the model set as default will be presented as the result. However, as it was seen, a data set fit well with all three models at the same time.

Table 5. The Result of CFA Models

	Chi-Square	*Sd*	*p*-value	CFI	RMSEA	SRMR
Model 1	14.13	19	.78	1.00	.00	.03
Model 2	12.25	19	.87	1.00	.00	.03
Model 3	21.52	19	.31	.99	.02	.04

In the second case, an EFA was run on SPSS with the same data to answer the question of what kind of a situation would occur. Thus, how the results can be changed when a researcher runs the same data set with EFA or CFA can be pointed. Since the PCA was not a factor analysis (Brown, 2006; Schmitt, 2011), principal axis factoring (PAF) was used as the estimation method. Since it was expected to have correlation between possible factors *promax* rotation was used. According to the results of this factor analysis (EFA1), the KMO value was .80 and Bartlet's test was significant ($\chi^2 = 319.08,\ p < .01$). As a result, the data set was suitable for an EFA. Table 6 shows EFA1 results. According to the factor analysis, a two-factor structure was formed.

Factor loadings in a factor analysis are expected to be higher than .30 (Martin & Newell, 2004; Seçer, Halmatov & Gençdoğan, 2013). The fact that the factor loadings of item 5 were less than this value in both factors which indicated inadequacy of the item. In addition, the internal consistency (Cronbach alpha) of the five items loaded on the first factor increases slightly, in the case the item 5 was removed. Accordingly, it was decided to remove item 5 from the analysis.

The result of new factor analysis (EFA2) was given in table 6. KMO and Bartlett test indicate that the data was suitable for factor analysis. According to the result, there were four items in the first factor and three items in the second factor. Internal consistency of the factors were .61 and .63 respectively. Each item was only loaded on one factor and these loadings were higher than .30. To conclude, the model without item 5 (EFA2) gave better result for the given data set.

Table 6. Results of Exploratory Factor Analysis

	EFA 1		FFA 2	
	Factor 1	Factor 2	Factor 1	Factor 2
Item 1	.58		.59	
Item 2	.55		.53	
Item 3	.63		.63	
Item 4	.31		.34	
Item 5	.23	.12	-	-
Item 6		.61		.61
Item 7		.61		.63
Item 8		.54		.53
KMO	.80		.79	
Bartlett's test	319.08		289.70	
p-value	.00		.00	
Correlation between factors	.64		.64	
Eigenvalues	2.66	1.04	2.55	1.04
% Variation	24.91	4.71	27.01	5.36

CONCLUSION and DISCUSSION

There is not a common way of using EFA or CFA first for scale adaptation studies. For an adaptation study some studies started with EFA while others started with CFA (Güvendir and Özkan, 2015). EFA is used when it is not known how many factors there are between the items and which factors are determined by which items while CFA is used if there is a strong theory about the structure. In this study, a data set is examined to fit to more than one CFA model via a simulation study. In addition, a data set was investigated to show how the use of EFA or CFA first might affect the results a scale adaptation. The simulated data generated according to a model specified in the R program were analyzed in the Mplus and SPSS programs.

Firstly, three different CFA models were evaluated for the same data set. The results clearly showed that more than one CFA model can fit well to a data set. For example, 63 of the 87 data sets that fit to model 1 also fit to model 2. This situation creates an ambiguity. *Which model shows the actual factor structure? Should all possible factor combinations be tried to determine the actual factor structure?* Using CFA for exploratory purposes may be limiting and even misleading the results (Schmitt, 2011). For this reason, as the result of the study showed, having a good fitting CFA model for a data set does not indicate that this model is actually the best model. In a scale adaptation studies, there may be changes in the structure resulting from cultural differences, as well as changes that may result from the item translation. Translating a scale into a new language requires not only translating language, but also language, culture and psychology as a whole (van de Vijver & Tanzer, 2004).

The possible models that may occur can be clearly defined in EFA. Structures not recognizable in CFA can easily be discovered through EFA (Bandalos & Finney, 2010). In other words, the possible changes in the structure in adaptation studies can be easily understood with the help of EFA. It is normal to have a change in the structure when a scale is translated into another language. It may even be possible to remove an item from the scale in some cases. Based on the results of this study, in order to achieve a consistent result and to establish a standard in scale adaptation studies, it is suggested to start with

an EFA to notice possible differences across cultures and languages. Then, a CFA will be a good step to verify the structure of adapted scale by using a different data set.

The use of different approaches in scale adaptation studies results in quite different conclusions. In this study, points to be considered in scale adaptation or development studies were indicated. How the results may change was examined via a simulation study. Also, differences were highlighted on a data set. As a result, in adaptation studies as well as in scale development studies, it is recommended to run an EFA and then a CFA to show the validity of the structure. Let us assume that the structure of the adapted scale is the same for both the adaptation and the original language. In this case, it will not be a problem to start with FFA and the same result will be achieved in every way. Otherwise, if there is a change in the structure, as it is seen in this study, we may not be able to detect it only with CFA. Therefore, it would be more beneficial to run EFA first in adaptation studies. Changing the design factors can change the results.

In this study, simulation design factors were limited because the aim was to show that a data set can fit different models. In other words, the sample size was 300, the correlation was fixed to .70 between the factors and the factor loadings had fixed values. It can be said that changing the design factors of the simulation may change the results. But this will not affect the conclusion that a data set can fit more than one model. Therefore, this was sufficient, even for one case (constant correlation and sample size). However, by changing the design factors, the simulation studies can be repeated and be examined in terms of the fit indices.

REFERENCES

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93-114). New York, NY: Routledge.

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, W. S., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research & Evaluation, 18*(6), 1-13.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Büyüköztürk, Ş. (2002). Faktör analizi: Temel kavramlar ve ölçek geliştirmede kullanımı. *Kuram ve Uygulamada Eğitim Yöntemleri, 32*, 470-483.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10*(7), 1-8.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of explanatory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.

Güvendir, M. A. & Özkan, Y. Ö. (2015). Türkiye'deki eğitim alanında yayımlanan bilimsel dergilerde ölçek geliştirme ve uyarlama konulu makalelerin incelenmesi. *Elektronik Sosyal Bilimler Dergisi*, *14*(52), 23-33.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191-205. doi: 10.1177/1094428104263675

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393-416. doi: 10.1177/0013164405282485

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

Hurley A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vanderberg, R. J., & Williams L. J. (1997). Exploratory and confirmatory factor analysis: Guidelines, issues, and alternatives. *Journal of Organizational Behavior, 18*, 667-683.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Martin, C. R., & Newell, R. J. (2004). Factor structure of the hospital anxiety and depression scale in individuals with facial disfigurement. *Psychology, Health & Medicine, 9*(3), 327-336. doi:10.1080/13548500410001721891.

Muthén, B., & Muthén, L. (1998-2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Orçan, F., & Yang, Y. (2016). A note on the use of item parceling in structural equation modeling with missing data. *Journal of Measurement and Evaluation in Education and Psychology, 7*(1), 59-72. doi:10.21031/epod.88204.

Schmitt R. S. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*(4), 304-321. doi:10.1177/0734282911406653.

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). New York, NY: Routledge.

Seçer, İ., Halmatov, S., & Gençdoğan (2013). Duygusal tepkisellik ölçeğinin Türkçeye uyarlanması: Güvenirlik ve geçerlilik çalışması. *Sakarya University Journal of Education, 3*(1), 77-89.

Sousa, V. D., & Rojjanasrirat, W. (2011). Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: A clear and user-friendly guideline. *Journal of Evaluation in Clinical Practice, 17*, 268-274. doi:10.1111/j.1365-2753.2010.01434.x

Sperber, A. D. (2004). Translation and validation of study instruments for cross-cultural research. *Gastroenterology, 126*, 124–128. doi:10.1053/j.gastro.2003.10.016.

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée, 54*, 119-135. doi:10.1016/j.erap.2003.12.004

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(4);422-435

Can TIMSS Mathematics Assessments be Implemented as Computerized Adaptive Test?

Semirhan GÖKÇE* Cees A.W. GLAS**

Abstract
In recent years, there has been a growing interest and extensive use of computerized adaptive testing (CAT) especially in large-scale assessments. Numerous simulation studies have been conducted on both real and simulated data sets to determine the optimum conditions and develop CAT versions. Being one of the most popular large-scale assessment programs, Trends in International Mathematics and Science Study (TIMSS) has been implemented as paper and pencil tests to monitor student achievement in mathematics and science at fourth and eighth grade levels since 1995. The purpose of this study is to investigate the optimum CAT algorithm for TIMSS eighth grade mathematics assessments. Since Turkey and USA participated in 2007, 2011 and 2015 administrations, their data were combined and then 393 items were calibrated on the same scale by using marginal maximum likelihood estimation method. With this item pool, several scenarios were proposed and tested to determine not only the optimum starting rule, ability estimation method, test termination rule but also the efficiency of exposure control method. The results of the study indicated that estimating abilities with expected a posteriori method after 6 random items, terminating the fixed-length test after 20 items seemed to be the optimum algorithm for TIMSS eighth grade mathematics assessments. Also, it was found that using item exposure control had a prior importance for the effective use of the item pool. This study has some implications for both national and international large-scale test developers in determining the optimum CAT algorithm and its consequences compared with paper and pencil versions.

Key Words: computerized adaptive testing, item response theory, mathematics assessment, simulation study, TIMSS.

INTRODUCTION

Educational testing has mainly been focused on traditional paper and pencil tests until the technological developments have supported the emergence of computers. At first, computers were responsible for displaying items and collecting responses, but since then they have also supported innovative item formats (Zenisky & Sireci, 2002) and fast score reporting. Then, instead of administering same set of items to the participants, different test forms have been assembled in computer-based testing. Eventually, this becomes meaningful when the participant's cumulative performance on earlier items determines the selection of newer items (Davey & Pitoniak, 2006). Actually, this is the main idea behind computerized adaptive testing (CAT). The intuitive principle underlying CAT is to maximize the item information. Statistically, each item gives information about the participants in terms of the trait being measured, but when the item parameters fit to their interim ability estimations, the amount of information maximizes. Therefore, the correct response of a participant is followed by more difficult item and the incorrect response is followed by an easier item (Hambleton, Swaminathan, & Rogers, 1991; Luecht & Sireci, 2012; van der Linden, 2010). This optimization process continues until the test administrators have enough certainty about the sufficiency of information about participant's ability level. Unlike traditional tests in which all participants take a single form, the CAT algorithm tailors the items according to the response patterns (Sireci, Baldwin, Martone, Kaira, Lam, & Hambleton, 2008) and finitely many test forms can be created during test

* Asst. Prof. Dr., Niğde Ömer Halisdemir University, Niğde-Turkey, e-mail: semirhan@gmail.com,
ORCID ID: 0000-0002-4752-5598
** Prof. Dr., University of Twente, Enschede-The Netherlands, e-mail: c.a.w.glas@utwente.nl,
ORCID ID: 0000-0001-6531-5503

To cite this article:
Gökçe, S., & Glas, C. A. W. (2018). Can TIMSS mathematics assessments be implemented as computerized adaptive test?. *Journal of Measurement and Evaluation in Education and Psychology, 9*(4), 422-435. DOI: 10.21031/epod.487351
Received: 25.11.2018
Accepted: 21.12.2018

administration. In this manner, different types of computer based tests range in a wide spectrum, from linear tests to adaptive tests.

Compared with linear tests in which fixed test forms are used, CAT has many advantages, such as testing on demand (Glas & Geerlings, 2009; Hambleton et al., 1991; van der Linden, 2001; Wainer, 2000), shortening tests without loss of measurement precision (Eggen, 2007; Hambleton et al., 1991; Meijer & Nering, 1999; Mills & Stocking, 1996; Verschoor & Straetmans, 2010), enabling immediate test scoring (Eggen, 2007; Wainer, 2000) and minimizing test frustration (Hambleton et al., 1991, Mills & Stocking, 1996). On the other hand, CAT has also some disadvantages such as reducing the control over tests and requiring a large calibrated item bank (Meijer & Nering, 1999).

The theoretical framework of CAT is based on Item Response Theory (IRT) framework in which the probability of a correct response to an item can be written as a mathematical function of participant's ability and item parameters. With IRT, the ability estimations of the participants can be obtained by independent set of items administered with a standard error. Hambleton et al. (1991) states that IRT provides a framework for comparing the ability estimations of differrent participants even if they have different set of items. Therefore, in order to match the item parameters with the ability levels of the participants, a large set of items (it is called item pool or item bank) is required whose statistical characteristics are obtained. van der Linden (1995) lists four steps of developing an iterative CAT algorithm as (1) defining the starting rule, (2) deciding on the item selection criteria, (3) choosing the ability estimation method, and (4) determining the termination rule. While determining the optimum starting rule, the difficulty of the first few items is important. Many testing programs have been using easier items at the beginning of a test in order to provide an initial success experience or motivation of the participants (Mills & Stocking, 1996). As the item selection methods are concerned, mainly there are two approaches such as Fisher's maximum information and Bayesian methods. Although Wainer (2000) states that both of the item selection methods give good results, Bayesian criteria needs more demand on the computer capabilities (Eggen, 2004). As the ability estimation methods are discussed, there are four ability estimation methods: maximum likelihood (ML), weighted maximum likelihood (WML), maximum a posteriori (MAP) estimation and expected a posteriori (EAP) estimation. According to Gu and Reckase (2007), MAP and EAP produce smaller standard errors compared to MLE and WML for the same number of items but they may produce biased estimations for inappropriate prior distributions. In test termination, there are mainly two options either to use fixed-length test or variable-length test. The former guarantees the implementation of a specified number of items to each participant but ends up with different standard error values for ability estimations. On the other hand, the latter stops the algorithm either obtaining sufficiently accurate ability estimation by comparing the standard error with a reference value or looking at the difference between consecutive ability estimations. At this point, the test developers should decide on test termination rule either to use a fixed-length test or a variable-length test depending on the purpose of the test and the content validity as well.

Due to the development of information and communication technologies and the widespread use of computers, many large-scale tests have been implemented as computer based test or even CAT such as Graduate Record Examinations (GRE), Graduate Management Admission Test (GMAT), Armed Services Vocational Aptitude Battery (ASVAB), and United States Medical Licensing Examination (USMLE). GRE, which was developed by Educational Testing Service (ETS), was implemented as a CAT as of 1992, Graduate Management Admission Council's GMAT was implemented as a CAT as of 1997 (Luecht & Sireci, 2012).

Trends in Mathematics and Science Study (TIMSS) is also a large scale assessment program aimed to monitor student achievement in mathematics and science at fourth and eighth grade levels in four-year-cycle since 1995 (Mullis, Martin, & Loveless, 2016). TIMSS assessments have been administered in paper-and-pencil form and the achievement tests have 14 different booklets which are linked to each other by common items, i.e. anchor items. In the booklets, there are both multiple-choice and open-ended items. Also, there have been anchor items between any consecutive TIMSS assessments so that test equating becomes feasible across assessments.

Purpose of the Study

The purpose of this study is to investigate the optimum CAT algorithm alternative to the paper and pencil based TIMSS eighth grade mathematics assessments. The data of two participating countries in the TIMSS eighth grade mathematics assessments in 2007, 2011 and 2015, Turkey and USA, were used for item calibration and the item pool. Then, a series of simulations covering different scenarios were tested to compare different starting rules, ability estimation methods and test termination rules. Additionally, item exposure rates were calculated in order to determine the effect of item exposure control strategy. The research questions of the study are given below.

1. What is the optimum CAT algorithm of TIMSS eighth grade mathematics assessments regarding different starting rules, ability estimation methods and test termination rules?

2. How does the item exposure control strategy affect the optimum CAT algorithm which is developed as an alternative to TIMSS eighth grade mathematics assessments?

METHOD

This part contains information related with the participants, data collection instruments and data analysis.

Participants

Table 1 gives information about the TIMSS sample sizes of Turkey and United States of America in 2007, 2011 and 2015 eighth grade mathematics administrations.

Table 1. Participants of Turkey and USA in TIMSS eighth grade mathematics assessments

Year	Turkey	USA	Total
2007	4498	7377	11875
2011	6928	10477	17405
2015	6079	10221	16300
Total	17505	28075	45580

Data Collection Instruments

As mentioned before, 14 different booklets were used in TIMSS eighth grade mathematics assessments and these booklets were linked to eachother with anchor items. Table 2 shows the number of items in these booklets.

Table 2. Test length of TIMSS eighth grade mathematics booklets

Booklet	TIMSS 2007	TIMSS 2011	TIMSS 2015
1	29	26	35
2	31	32	33
3	32	32	28
4	29	29	32
5	29	32	34
6	32	33	32
7	33	30	32
8	32	34	30
9	31	34	28
10	32	31	28
11	32	32	29
12	28	32	30
13	28	33	29
14	30	27	30
Mean	30.6	31.2	30.7

Table 2 gives information about the average test length of TIMSS eighth grade mathematics achievement tests, which is about 30 items. The response patterns of Turkey and the USA participants were merged by using anchor items to obtain incomplete data matrix. Data collection design of these assessments is shown in Figure 1.

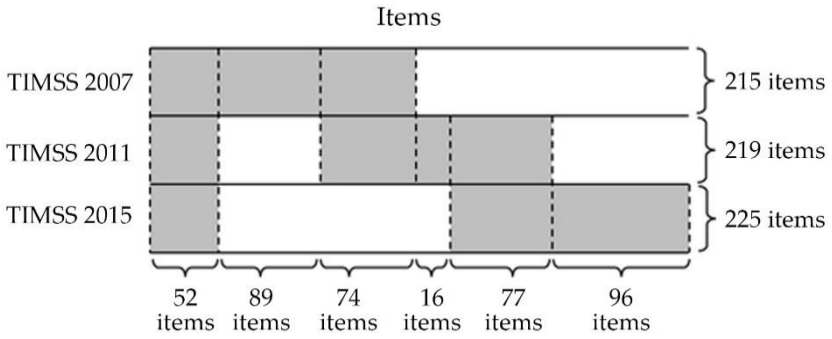


Figure 1. Data collection design of TIMSS eighth grade mathematics assessments

The data matrix contained 45,580 rows (participants) and 404 columns (items). However, 11 of the items (M042273, M062345BA, M062345BB, M062345BC, M062345BD, M062345B, M062342, M062048A, M062048B, M062048C and M062048) were taken out of the analysis since they had all missing responses. Out of the 393 items, dichotomously scored 360 items were calibrated by using 2 Parameters Logistic (2PL) model and polytomously scored 33 items were calibrated by using Partial Credit Model (PCM). In the item pool, all the multiple-choice items were dichotomously scored. However, some of the open-ended items were dichotomously scored and the remaining were polytomously scored. PCM is a unidimensional model for the responses scored in two or more ordered categories (Masters, 2016). MIRT (Glas, 2010) program was used for item analysis and calibrating both dichotomously and polytomously scored items. The item parameter distribution of dichotomously scored items (item difficulty versus item discrimination) is given in Figure 2.

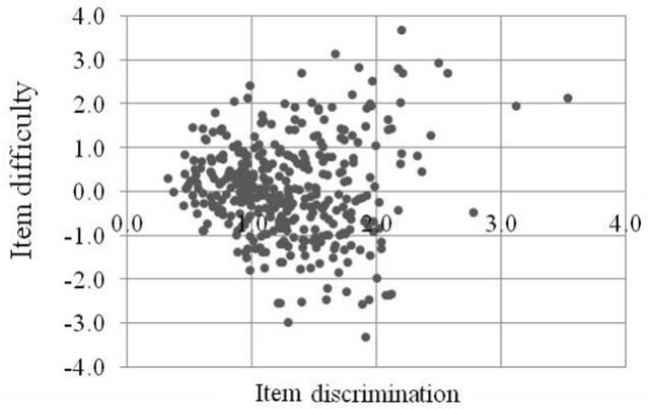


Figure 2. Distribution of item parameters calibrated by 2PL model

In addition to the item parameters, MIRT program also reported ability estimations and standard error values of these estimations based on WML and EAP methods. Statistical information about the ability estimations are given in Table 3.

Table 3. Mean values of ability estimations and standard error values in item calibration

Statistic	Ability estimation method	
	WML	EAP
Ability estimation	-.054	-.063
Mean SE	.371	.328

As shown in Table 3, mean value of ability estimations were -.054 and -.063 for WML and EAP methods, respectively. Also, the mean values of standard errors were .371 in WML and .328 in EAP.

Data Analysis

Test equating and scaling of TIMSS assessments were conducted based on IRT framework (Martin, Mullis & Hooper, 2016) so the assumptions were supposed to be satisfied. The item calibration were conducted based on the unidimensional IRT model by using MIRT software package (Glas, 2010). In this analysis, 360 items were calibrated by 2PL model and 33 items were calibrated by PCM. Afterwards, these item parameters were used in simulation studies. A sample of 1000 simulated test takers were drawn from normal distribution N(0,1) and three sets of simulations were designed. Afterwards, based on the item parameters and drawn ability values, a response matrix having 1000 rows and 393 columns was formed.

In the first set of simulations, variable-length tests were used and .20, .30 and .40 reference values were set for standard error. Next, (a) correlation between true theta and estimated theta, (b) average test length and (c) distribution of item exposure rates, (d) root mean square error (RMSE) and (e) bias were compared for each standard error value. Here, item exposure rate stands for the ratio of the participants facing the item to the total number of participants. For example, if 130 out of 1000 participants saw an item during a test administration, then the item exposure rate for this item would be .13. The RMSE and bias are the values representing the differentiation between predicted (true theta values) and observed (estimated theta values) ability estimations.

Second set of simulations was focused on the comparison of fixed-length tests with 10, 20 and 30 items based on (a) correlation between true theta and estimated theta, (b) mean standard errors and (c) distribution of exposure rates, (d) root mean square error (RMSE) and (e) bias.

Third set of simulations was conducted to indicate the effect of using item exposure control in CAT algorithm whereas the fourth set of simulations were implemented to analyze the efficiency of ability estimation methods.

In these simulations, different number of random items were administered at the beginning of the test as test starting rules, Fisher's information was used as item selection, WML and EAP methods were compared as ability estimation method, variable-length test and fixed-length test were used as test termination rule. Also, the effect of Randomesque method (Kingsbury & Zara, 1989) on the CAT algorithm was examined.

RESULTS

First set of simulations were conducted and 36 conditions were compared to determine the optimum CAT algorithm by comparing three types of starting rules (ability estimations without any constraint i.e. standard version, after three random items or six random items), two different ability estimation methods (EAP or WML) and six different termination rules (fixed-length tests with 10, 20 or 30 items; variable-length tests terminated after reaching .20, .30 or .40 standard error values).

a) Simulations based on variable-length tests

Here, simulations were conducted to compare the effects of the determined situations on variable-length tests so that the average test length and correlation coefficient between true and estimated theta values were calculated. The results are shown in Table 4.

Table 4. Test lengths and correlation coefficients between true and estimated theta in variable-length tests

Method	initial ability estimation	Test termination rule					
		SE < .20		SE < .30		SE < .40	
		test length	r	test length	r	test length	r
EAP	after first item	33	.978	12	.961	6	.931
	after 3 random items	35	.980	13	.955	7	.926
	after 6 random items	36	.978	15	.960	9	.922
WML	after first item	36	.979	13	.953	7	.929
	after 3 random items	36	.980	14	.957	9	.934
	after 6 random items	38	.980	16	.958	10	.933

Table 4 shows that a better measurement precision was obtained with higher correlations but this cost more items as expected. This can be explained by the relationship between the standard errors and the reliability of the test scores. Also, average test length was directly related with the same context. In other words, the algorithm gave more items to the participant so as to reach a standard error less than .20. Decreasing the standard error reference from .40 to .30 almost doubled the test length and tripled when the standard error reference changed from .30 to .20. Using more random items before initial ability estimations increased the test length in variable-length tests. More specifically, variable-length tests needed more items since random items were used in the algorithm rather than selecting the most informative item. Finally, when the effect of EAP and WML ability estimationmethods were analyzed in variable-length tests, there was no prominent differentiation occurs among test lengths and correlation coefficients.

Table 5 indicates the item exposure rate distributions, RMSE and bias of different ability estimations in variable-length tests.

Table 5. Item exposure rate distributions, RMSE and bias of different ability estimations in variable-length tests

Method	Initial ability estimation	Test termination rule	Item exposure rate				RMSE	Bias
			<.01	.01-.20	.21-.40	>.40		
EAP	after first item	SE < .20	0	334	38	21	.210	-.008
		SE < .30	168	205	14	6	.286	-.012
		SE < .40	235	148	6	4	.359	.001
	after 3 random items	SE < .20	0	338	35	20	.209	.000
		SE < .30	163	211	13	6	.296	-.006
		SE < .40	203	181	7	2	.375	.014
	after 6 random items	SE < .20	0	339	34	20	.210	-.004
		SE < .30	150	226	14	3	.280	.017
		SE < .40	214	173	4	2	.382	-.021
WML	after first item	SE < .20	0	333	35	25	.202	-.013
		SE < .30	116	257	15	5	.303	-.017
		SE < .40	198	184	8	3	.383	-.039
	after 3 random items	SE < .20	1	334	35	23	.198	-.019
		SE < .30	138	236	15	4	.284	-.016
		SE < .40	168	215	7	3	.376	-.026
	after 6 random items	SE < .20	1	333	35	24	.202	-.007
		SE < .30	131	245	14	3	.292	-.022
		SE < .40	189	197	5	2	.365	-.015

The effect of variable-length tests and different termination criteria on item exposure rates, RMSE and bias were analyzed and as shown in Table 5, the decrease in the standard error reference value ended up with the decrease in the number of items with underexposure (exposure rates less than .01) This seems to be a positive outcome but at the same time it increased the number of items with overexposure (exposure rates greater than .40). When the effect of variable-length tests on RMSE and bias was examined, stricter test termination rules (smaller standard error reference values) ended up with smaller RMSE and bias.

Although there was no obvious differentiation of EAP and WML methods when RMSE and bias were compared, EAP had less bias. Moreover, WML provided negative bias values in all conditions interpreting that this method had higher observed values (estimated theta) than the predicted values (true theta).

When comparing the test starting rules, it was found that using more random items at the beginning of the test had a positive impact on decreasing the number of items with overexposure (exposure rates greater than .40).

b) Simulations based on fixed-length tests

Second set of simulations were conducted to observe the effect ability estimation methods and starting rules on fixed-length tests containing 10, 20 and 30 items. Table 6 shows the mean standard errors and correlation coefficients between true and estimated theta in fixed-length tests.

Table 6. Mean standard errors and correlation coefficients between true and estimated theta in fixed-length tests

Method	Initial ability estimation	Test termination rule					
		10 items		20 items		30 items	
		mean SE	R	mean SE	r	mean SE	r
EAP	after first item	.310	.946	.231	.973	.194	.980
	after 3 random items	.328	.947	.237	.969	.198	.979
	after 6 random items	.375	.918	.249	.969	.205	.977
WML	after first item	.329	.943	.244	.971	.206	.978
	after 3 random items	.348	.935	.245	.969	.210	.980
	after 6 random items	.430	.912	.260	.966	.212	.976

When Table 6 is examined, the increase in the test length decreased the mean standard error values and increased the correlation coefficients between true and estimated theta. In almost all conditions, an increase in the number of random items at the beginning of a test decreased the correlation coefficients and increased the mean standard errors. In a general perspective, intervening the item selection algorithm has a cost of an increase in test length in order to preserve the reliability. Therefore, using 6 random items at the beginning of the test had smaller correlation coefficients and higher standard error values compared with other two cases (after first item and after 3 random items). Finally, when the ability estimation methods were compared EAP method provided comperatively better results than WML method.

Table 7 shares the item exposure rate distributions, RMSE and bias of different ability estimations in fixed-length tests.

Table 7. Item exposure rate distributions, RMSE and bias of different ability estimations in fixed-length tests

Method	Initial ability estimation	Test termination rule	Item exposure rate				RMSE	Bias
			<.01	.01-.20	.21-.40	>.40		
EAP	after first item	10 items	247	129	11	6	.324	-.020
		20 items	209	142	27	15	.229	.001
		30 items	182	148	36	27	.202	.000
	after 3 random items	10 items	224	154	11	4	.330	-.008
		20 items	193	162	26	12	.246	.006
		30 items	163	171	35	24	.207	.002
	after 6 random items	10 items	233	152	6	2	.398	-.008
		20 items	201	160	24	8	.256	-.012
		30 items	170	170	31	22	.213	-.013
WML	after first item	10 items	236	141	10	6	.334	.005
		20 items	205	146	29	13	.252	.006
		30 items	174	158	34	27	.211	.007
	after 3 random items	10 items	219	160	10	4	.360	-.006
		20 items	189	167	26	11	.246	-.004
		30 items	154	176	37	26	.209	-.001
	after 6 random items	10 items	229	157	4	3	.453	-.007
		20 items	196	164	23	10	.272	-.006
		30 items	165	175	31	22	.223	-.013

In fixed-length tests, longer tests had a positive impact on increasing the number of items having underexposure (exposure rates less than .01) but at the same time had a negative impact on increasing the number of items having overexposure (exposure rates greater than .40). In all cases, RMSE values decreased as the test length increased. Also, administering 6 random items before the initial ability estimation had a positive effect on the item exposure rates.

Although the item exposure rates were different across test lengths with 10, 20 and 30 items, the results were not sufficient to determine the superiority of the ability estimation methods. In other words, EAP and WML methods seemed to have similar item exposure rate distributions. However, when the RMSE values were on focus, EAP provided more comparable results than WML.

Up to this point, fixed-length tests provided better results than variable-length tests. In variable-length tests, especially low and high achievers were given more than 100 items (or even all the 393 items) in order to satisfy termination rule. Even the termination rule could not achieve to decrease the standard error to the set value after implementing all the items in the pool to a participant. On the other hand, the length of the test was 4 items for some of the participants. If an additional minimum and maximum values for test length are not defined in CAT algorithm, using variable-length tests do not seem to be convenient in TIMSS's adaptive testing practices. So, what could be the optimum test length: 10, 20 or 30? In general, an increase in the test length provided evidence to the content validity of the test. Additionally, a test having 10 items did not give stable correlation coefficients across starting rules and ability estimation methods. It seems more reasonable to use tests containing either 20 items or 30 items in CAT algorithm. The scores from a test with 20 items had correlation coefficient of .97 and a test with 30 items test had a correlation coefficient of .98 with true theta. Fixed-length tests containing 20 items had standard error values between .23 and .26 but tests containing 30 items had standard error values between .19 and .21. Hence, tests with 20 items had .93 reliability and tests with 30 items had .96 reliability. The RMSE took values between .229 and .272 in fixed-length tests with 20 items and took values between .202 and .223 in fixed-length tests with 30 items.

When all these results are interpreted, fixed-length tests with 20 items seem to be the optimum condition for CAT algorithm since these tests provide high correlation coefficients and reliability values.

c) Simulations based on item exposure rates

Third set of simulations focus on the item exposure control and the effect of Randomesque method on fixed-length tests having 20 items was analyzed. Based on the results of previous simulations, item exposure rates were defined for each item. These rates were used to decrease the number of overexposed and to increase the number of underexposured items. The results are given in Table 8.

Table 8. Item exposure rate distributions, RMSE and bias of different ability estimations in using item exposure control

Method	Initial ability estimation	Exposure control	Item exposure rate				RMSE	Bias
			< .01	.01 - .20	.21 - .40	> .40		
EAP	after first item	no	209	142	27	15	.229	.001
		yes	197	156	27	13	.245	-.003
	after 3 random items	no	193	162	26	12	.246	.006
		yes	178	175	32	9	.242	.005
	after 6 random items	no	201	160	24	8	.256	-.012
		yes	189	172	25	9	.266	-.008
WML	after first item	no	205	146	29	13	.252	.006
		yes	179	174	31	12	.244	-.006
	after 3 random items	no	189	167	26	11	.246	-.004
		yes	170	189	25	13	.267	-.010
	after 6 random items	no	196	164	23	10	.272	-.006
		yes	183	179	25	11	.274	-.008

According to Table 8, using item exposure control decreased the number of underexposure items (exposure rates less than .01). When RMSE and bias was concerned, there were some differentiation in the values but it did not seem to have a pattern.

Analysis were conducted to determine whether it was more convenient to estimate abilities with either EAP or WML methods and after first item, after 3 random items or after 6 random items. For all of the cases, item exposure rates were calculated for the items in the pool. In order to observe the changes more clearly, these rates were sorted from high to low. The graphs showing the efficiency of item exposure control for different ability estimation methods and starting rules are shared in Figure 3. In the figure, vertical axis stands for the item exposure rates. The horizontal axis indicates the items on which the items with high exposure rates locate to the left and the items with low exposure rates locate to the right.

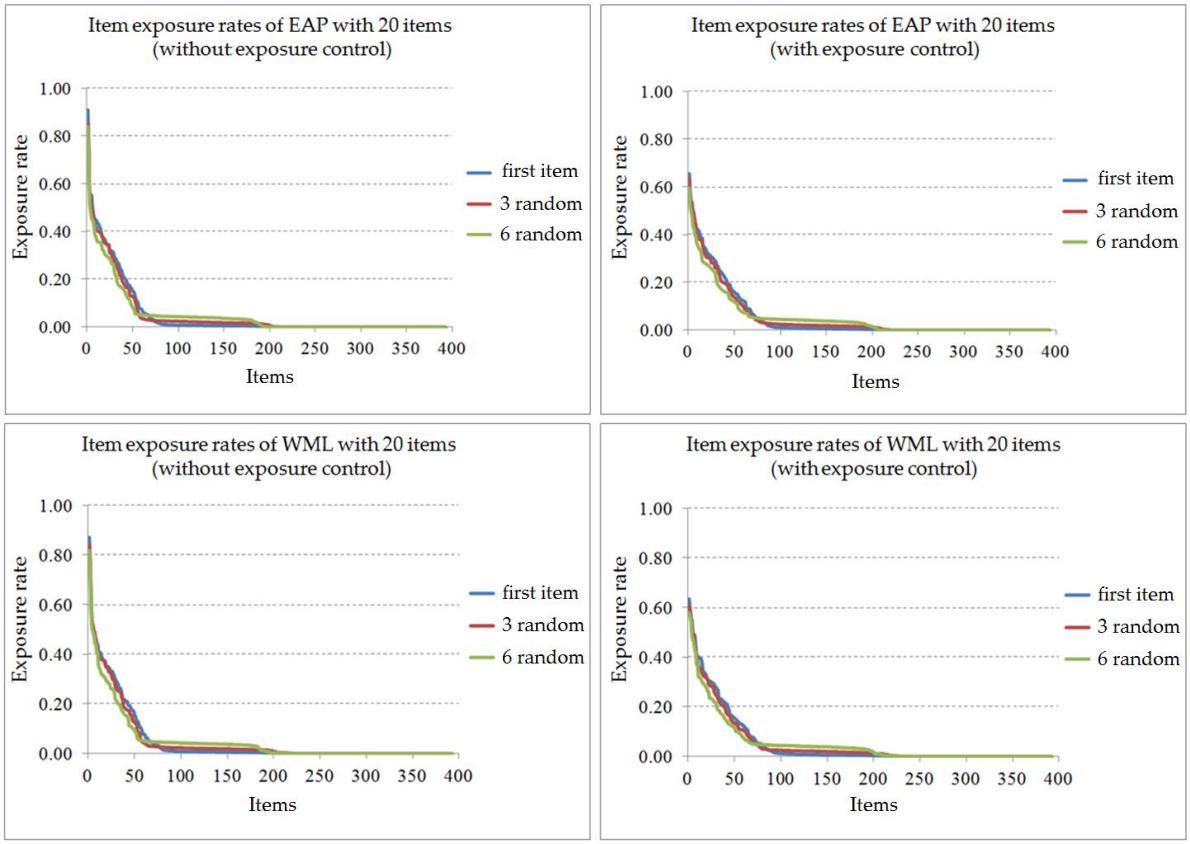


Figure 3. Distribution of item exposure rates by different ability estimation methods either with or without exposure control

In Figure 3, although item exposure control had a positive impact on the item exposure rates of the items in the pool, there was a major problem that almost half of the items were not used in any of the test administrations. When test starting rules were compared, item exposure rates of overexposure items decreased evidently. The main reason behind this is directly related with providing a way to present not used items. Hence, it is believed that using 6 random items at the beginning of the test ensures the effective usage of the item pool so it could be a good starting rule for the optimum algorithm of TIMSS eighth grade mathematics assessments.

Up to this point, the simulation results provide similar results for both EAP and WML.

d) Simulations based on ability estimation methods

Fourth group of simulations were conducted to determine the effectiveness of EAP and WML methods. In the simulations, test starting rule was set to administer 6 random items before initial ability estimation and test termination rule was set to fixed-length tests with 20 items. Moreover, item exposure conrol was used in the comparisons and the relationship between true and estimated theta is given in Figure 4.

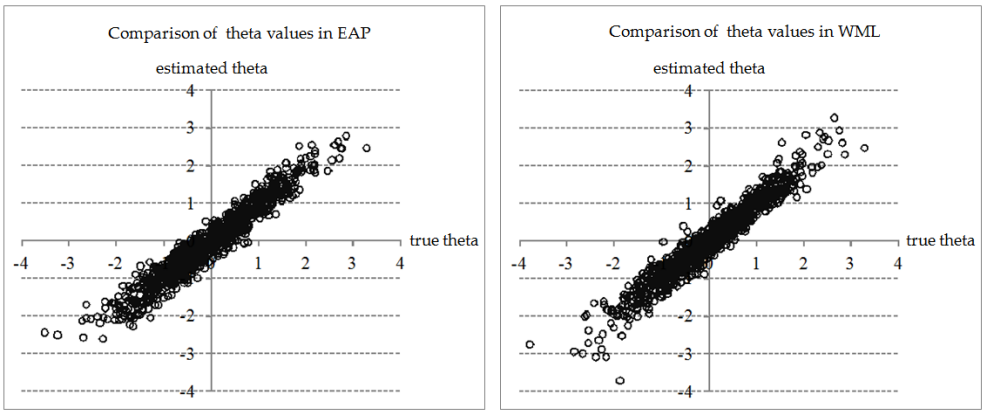


Figure 4. Comparison of estimated theta and true thata in EAP and WML

According to Figure 4, theta values located at very low and very high values in WML were scattered more than as they are in EAP. So, EAP seems to provide a better estimation for the participants from especially quite low and high theta values compared to WML.

To summarize the results of this study, starting ability estimations after six random items as the starting rule, using EAP as the ability estimation method, terminating the test after 20 items and using item exposure control indicated the optimum condition for TIMSS eight grade mathematics assessments. In this case, the mean SE estimation was .253 (.135 as minimum and .468 as maximum) and the correlation between true and estimated theta was .964.

DISCUSSION and CONCLUSION

The aim of this study is to determine the optimum CAT algorithm which is an alternative to the paper and pencil based TIMSS eight grade mathematics assessments. In the simulations, different starting rules, ability estimation methods and termination rules were compared and the effectiveness of item exposure control was analyzed.

As a starting rule, initial ability estimations after first item, after 3 random items and after 6 random items were compared. Although, using more random items at the beginning of the test had a negative effect on RMSE values, its positive impact on the item exposure rates made it indispensible for optimum algorithm. However, it was more convenient to use 6 random items in longer tests. In other words, it was not convenient to use 6 random items in fixed-length tests with 10 items or in variable-length tests with .40 standard error reference because 6 items probably consituted the major part of the test in such cases.

When the ability estimation methods were compared, EAP and WML gave similar results but EAP provided better estimations for especially low and high achievers, which is very similar to the findings of Gu and Reckase (2007).

In order to determine the optimum test termination criteria, variable-length and fixed-length tests were compared. When the standard error was set to .20 in variable-length tests, the correlation coefficients were calculated to be higher but in some of the cases the algorithm presents all the items in the bank but it was not successful to diminish the standard error value below .20. Therefore, it was not practical to use variable-length tests for low achievers and high achievers. To be more specific, then the algorithm could not succeed in decreasing the standard error to .20 even after using all 393 items in the pool. Similar results were interpreted in the study by Gökçe and Berberoğlu (2015). Hence, using a fixed-length test becomes more reasonable in TIMSS eighth grade mathematics assessments. When fixed-length tests with 10, 20 and 30 items were compared, test with 20 items provided more comparable results for TIMSS eight grade mathematics assessments. In the study, Randomesque exposure control was used and the results indicated that this method balanced the item usage by

increasing the exposure rates of underexposure items and decreasing the exposure rates of overexposure items. However, in any case, almost half of the items in the pool were not used for any of the participants. In CAT administrations, one of the major problems related with the items is underexposure and overexposure of items (Eggen, 2001; Eggen & Straetmans, 2000). For further studies, it would be better to compare different exposure control methods in TIMSS assesments.

In TIMSS eight grade mathematics assessments, the number of items contained in eighth grade mathematics booklets is about 30. These tests estimated ability with a mean SE value of .328 by EAP method. On the other hand, the optimum CAT algorithm estimated theta values with a mean SE value of .253 with 20 items (with a 35.5% shorter test) by the same method. This results is one of the main advantages of CAT applications. There are many studies indicating that computerized adaptive tests provide more reliable estimations with shorter tests and decreases the testing time (Eggen, 2007; Hambleton et al., 1991; Meijer & Nering, 1999; Mills & Stocking, 1996; Verschoor & Straetmans, 2010).

In all of the cases, there were high correlation coefficients between true that and estimated theta. There are studies reporting that there would be similar ability estimations when different starting rules, ability estimation methods and test termination rules are used in the algorithm (Kalender, 2011; Kezer & Koç, 2014).

This study investigated the applicability of TIMSS eighth grade mathematics assessments as computerized adaptive test and has some limitations. In the literature, starting rules are related with the difficulty of the items at the beginning of the test but instead the effect of starting the test with a group of random items was investigated in this study. Moreover, there are two types of items in the pool either dichotomous or polytomous. In paper and pencil based TIMSS assessments, it is easy to control the number of dichotomous and polytomous items but this study did not focus on balancing item type. In TIMSS eighth grade mathematics assessments, there are 4 learning areas (numbers, geometry, algebra and data-probability) and tests developers can control the number of items for each learning area. However, this study did not consider any control based on content. Finally, open-ended items existed in the item pool of the CAT simulations. Although the ability estimations were carried out by using these items, it would be difficult to use such items in real CAT practices because of their scoring. This is another limitation of the study.

In the study, the data sets of Turkey and United States of America were used. For further studies, the data of other participating countries from TIMSS 1995, 1999, 2003, 2007, 2011 and 2015 mathematics assessments could be analyzed and compared with the results of this study. Also, since this study used eighth grade mathematics data set, further studies could focus on the TIMSS fourth grade mathematics data and check whether to obtain comparable results across grade levels.

REFERENCES

Davey, T., & Pitoniak, M. J. (2006). Designing computerized adaptive tests. *Handbook of Test Development*, 543-574. Routledge.

Eggen, T. J. H. M. (2001). Overexposure and underexposure of items in computerized adaptive testing. *Measurement and Research Department Reports, 1*.

Eggen, T. J. H. M. (2004). Contributions to the Theory and Practice of Computerized Adaptive Testing. Dissertation. Print Partners Ipskamp B.V., Enschede.

Eggen, T. J. H. M. (2007). Choices in CAT models in the context of educational testing. In D. J. Weiss (Ed.), Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, 60(5), 713-734.

Glas, C. A. W. (2010) MIRT: Multidimensional Item Response Theory. (Computer Software). University of Twente. Retrieved from https://www.utwente.nl/nl/bms/omd/Medewerkers/medewerkers/glas/#soft-ware

Glas, C. A. W., & Geerlings, H. (2009). Psychometric aspects of pupil monitoring systems. *Studies in Educational Evaluation*, 35, 83-88.

Gu, L., Reckase M. D. (2007). Designing optimal item pools for computerized adaptive tests with Sympson-Hetter exposure control. In D. J. Weiss (Ed.), Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory* (Vol. 2). Sage.

Kalender, I. (2011). Effects of different computerized adaptive testing strategies on recovery of ability. Yayınlanmamış Doktora Tezi. Middle East Technical University, Ankara.

Kezer, F. & Koç, N. (2014). Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması [A comparison of computerized adaptive testing strategies]. *Eğitim Bilimleri Araştırmaları Dergisi - Journal of Educational Sciences Research*, 4 (1), 145-174.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2(4), 359-375.

Luecht, R. M. & Sireci, S. G. (2012). A review of models for computer-based testing. *Research Report RR-2011-12*. New York: The College Board.

Masters, G. N. (2016). Partial credit model. In *Handbook of Item Response Theory, Volume One* (pp. 137-154). Chapman and Hall/CRC.

Meijer, R. R. & Nering M. L. (1999). Computerized adaptive testing: overview and introduction. *Applied Psychological Measurement*, 23, 187-194.

Mills, C. N. & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, 9 (4), 287-304.

Mullis, I., Martin, V. & Loveless, T. (2016). 20 years of TIMSS, international trends in mathematics and science achievement, curriculum, and instruction. IEA, TIMSS&PIRLS International Study Center Lynch School of Education, Boston College.

Sireci, S. G., Baldwin, P., Martone, A., Zenisky, A. L., Kaira, L., Lam, W., & Hambleton, R. K. (2008). Massachusetts Adult Proficiency Tests Technical Manual, Version 2. *Center for Educational Assessment Research Report No, 677*.

Smits, N., van Straten, A., & Cuijpers, P. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, 188(1), 147-155.

van der Linden, W. J. (1995). Advances in computer applications. In T. Oakland & R. K. Hambleton (Eds.), *International Perspectives on Academic Assessment*, (pp. 105-124). Kluwer Academic Publishers.

van der Linden, W. J. (2001). Computerized test construction. *Research Report*. Twente University, Enschede (Netherlands).

van der Linden, W. J. (2010). Item selection and ability estimation in adaptive testing. *Elements of Adaptive Testing*, 3-30. Springer.

Verschoor, A. J., & Straetmans, G. J. J. (2010). MATHCAT: A flexible testing system in mathematics education for adults. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing*, (pp. 137-149). Statistics for Social and Behavioral Sciences. Springer.

Wainer, H. (2000). Computerized Adaptive Testing: A Primer. Mahvah, NJ: Erlbaum.

Zenisky A. L., & Sireci, S. G. (2002) Technological innovations in large-scale assessment, *Applied Measurement in Education*, 15:4, 337-362.