Elementary Education Online, 2017; 16(4): 1917-1931
İlköğretim Online, 16(4), 1917-1931, 2017. [Online]: http://ilkogretim-online.org.tr
doi: 10.17051/ilkonline.2017.364505

The Effect of Latent Classes Formed According to the Affective Characteristics of Students on Differential Item Functioning Based on Sex

Öğrencilerin Duyuşsal Özelliklerine Göre Oluşan Gizil Sınıfların Cinsiyete Göre Farklılaşan Madde Fonksiyonu'na Etkisi

Seher Yalçın, *Ankara Üniversitesi, Eğitim Bilimleri Fakültesi, yalcins@ankara.edu.tr*

Abstract. This study is carried out in three stages. First, students' affective characteristics which may possibly explain the difference in the answering a science item correctly or wrong are determined according to the sex of the students who are at the same level of ability. The status of identified affective characteristics' forming latent classes is determined in the second stage. In the last stage, it is aimed to determine whether or not the same items display differential item functioning (DIF) in the emerged latent classes. The study group of this research, which is in a descriptive survey model consists of 875 students. In the first staged of data analysis, latent class analysis is used to determine the latent classes that are formed according to the students' affective characteristics. In the second phase of the analysis, Mantel-Haenszel method is used in order to determine the state of differential functioning of the items for the whole group and in the latent classes emerged according to the students' affective characteristics. According to the fact that the DIF analyses are carried out for the entire group, DIF is detected only in one of the items. However, it is found out that when the students are classified into latent classes according to some affective characteristics, this item did not display DIF in two latent groups and in one class, another item also displayed DIF.

Keywords: Differential item functioning, affective characteristics, latent class analysis, TIMSS, science

Öz. Bu çalışma üç aşamada gerçekleşmiştir. İlk olarak aynı yetenek düzeyindeki öğrencilerin cinsiyete göre, bir fen maddesini doğru ya da yanlış yanıtlamasında oluşan farklılığı olası açıklayabilecek öğrencilerin duyuşsal özellikleri belirlenmiştir. İkinci aşamada, belirlenen duyuşsal özelliklerin gizil sınıf oluşturma durumları tespit edilmiştir. Son aşamada, aynı maddelerin oluşan gizil sınıflarda farklılaşan madde fonksiyonu (FMF) gösterip göstermediğini belirlemek amaçlanmıştır. Betimsel tarama modelinde olan araştırmanın, çalışma grubunu 875 öğrenci oluşturmuştur. İlk aşamada, öğrencilerin duyuşsal özelliklerine göre oluşan gizil sınıfları belirlemek için gizil sınıf analizi kullanılmıştır. Analizlerin ikinci aşamasında, öğrencilerin duyuşsal özelliklerine göre oluşan gizil sınıflarda ve tüm grup için maddelerin farklı fonksiyonlaşma durumunu belirlemek için Mantel-Haenszel yöntemi kullanılmıştır. Tüm grup için yapılan FMF analizlerine göre ele alınan maddelerin sadece birinde FMF tespit edilmektedir. Ancak öğrenciler bazı duyuşsal özelliklerine göre gizil sınıflara ayrıldığında, bu maddenin iki gizil sınıfta FMF göstermediği, bir sınıfta ise farklı bir maddenin de FMF gösterdiği tespit edilmiştir.

Anahtar Sözcükler: Farklılaşan madde fonksiyonu, duyuşsal özelllikler, gizil sınıf analizi, TIMSS, fen bilimleri

ÖZET

Amaç ve Önem: Bu çalışmada, aynı yetenek düzeyindeki bireylerin bir maddeyi doğru ya da yanlış yanıtlamasında cinsiyete göre oluşan farklılığı olası açıklayabilecek, öğrencilerin duyuşsal özelliklerinin belirlenerek gizil sınıflara ayrılması ve ardından oluşan gizil sınıflardaki aynı maddelerin farklılaşan madde fonksiyonu (FMF) gösterip göstermediğini tespit etmek amaçlanmıştır. Bu çalışmada, olası FMF kaynaklarının FMF testinden önce belirlenerek FMF'ye etkisinin olup olmadığını tespit etmek üzere farklı bir yöntem önerilmektedir. Böylece, FMF'nin kaynağına yönelik istatistiksel bir bilgi sunmaktadır. Bu nedenle, FMF çalışmalarına farklı bir boyut kazandıracağı düşünülen bu yöntemin alan yazına katkı sağlayacağı düşünülmektedir.

Yöntem: Bu çalışmada, öğrencilerin duyuşsal özelliklerinin FMF üzerindeki etkisini tespit etmek için öğrenciler, ele alınan öğrenci özellikleri açısından gizil sınıflara ayrılarak hem her bir gizil sınıf için hem de tüm grup için FMF analizinin ayrı ayrı yapılması, var olan durumu ortaya

Geliş tarihi: 10/08/2017, Kabul tarihi: 26/10/2017, Yayımlanma tarihi: 11/12/2017

koyduğundan betimsel tarama modelindedir. Uluslararası Matematik ve Fen Eğilimleri Araştırmasında (TIMSS), öğrenciler 14 farklı kitapçıktan birindeki fen maddelerini yanıtlamaktadır. Bu çalışmada, çalışma grubunu geniş tutabilmek için 2015 yılı TIMSS Türkiye uygulamasında, kitapçıklar arası ortak madde sayısı en fazla olan ikinci ve üçüncü kitapçıklardaki ortak maddeler seçilmiş ve bu kitapçıkları alan toplam 875 öğrenci çalışma grubunu oluşturmuştur. Bu öğrencilerin 415'i kız (%47), 460'ı (%53) erkektir. Bu çalışmanın verilerini, öğrencilerin yanıtladığı 23 ortak fen maddesi ve öğrenci anketindeki öğrencilerin duyuşsal özelliklerine ilişkin seçilen indeks değişkenlere (fene yönelik tutum, fene karşı kendine güven, fen derslerine katılım ve fen öğrenmeye verilen değer) verilen tepkiler oluşturmaktadır. Analizler, iki aşamada gerçekleştirilmiştir. İlk aşamada, öğrencilerin fene yönelik tutumu, fene karşı kendine güveni, fen derslerine katılımı ve fen öğrenmeye verilen değerine göre oluşan gizil sınıfları belirlemek için gizil sınıf analizi kullanılmıştır. Analizlerin ikinci aşamasında, öğrencilerin duyuşsal özelliklerine göre oluşan gizil sınıflarda ve tüm grup için maddelerin farklı fonksiyonlaşma durumunu belirlemek için alan yazında en sık kullanılan yöntemlerden biri olan Mantel-Haenszel (MH) yöntemi kullanılmıştır. Analizlerin birinci aşamasında Latent Gold 5.1 ve ikinci aşamasında Xcalibre 4.2.2 paket programları kullanılmıştır.

Bulgular: Öğrencilerin duyuşsal özellikleri için yapılan gizil sınıf analizi sonucu, üç gizil sınıflı modelin veriye en iyi uyum sağladığı görülmüştür. Oluşan üç gizil sınıf için ayrı ayrı yapılan Mantel Haenzsel testi sonucu, ilk gizil sınıfta iki maddenin (1. ve 20. maddeler) kızlar lehine FMF gösterdiği diğer iki sınıfta bu maddelerin FMF göstermediği tespit edilmiştir. Ayrıca, FMF gösteren iki madde, konu alanı açısından değerlendirildiğinde, ilk maddenin Biyoloji öğrenme alanının "Organizmaların Yaşam Süreci ve Özellikleri" konusundan, ikinci maddenin Fizik öğrenme alanından "Kuvvet ve Hareket" konusundan olduğu, iki maddenin de bilişsel düzey açısından bilgi düzeyinde ve madde türü olarak çoktan seçmeli olduğu tespit edilmiştir. Gizil sınıf analizi yapılmadan tüm grup için FMF analizleri yapıldığında, birinci madde A düzeyinde yani ihmal edilebilir düzeyde FMF iken 20. madde C düzeyi (üst düzey) FMF göstermektedir. Gizil sınıf-1'deki bulgulardan farklı olarak birinci maddenin FMF gösterme durumunun, ihmal edilebilir düzeyde olduğu görülmüştür.

Tartışma ve Sonuç: FMF gösteren maddelerin olduğu gizil sınıf, genellikle fen öğrenmekten çok hoşlanan, fene karşı kendine çok güvenen, fen derslerine çok katılan ve fen öğrenmeye oldukça değer veren öğrencilerden oluşmaktadır. Eğer kızlar fende bu duyuşsal özelliklere yüksek düzeyde sahipse, bu maddeleri doğru yanıtlama oranlarının erkeklerden manidar bir şekilde daha yüksek olduğu ifade edilebilir. Ayrıca gizil sınıf-2 ve 3'te maddelerin FMF göstermemesi, aynı yetenek düzeyinde fene karşı düşük ya da orta düzey duyuşsal özelliklere sahip olan öğrencilerin cinsiyete göre maddeleri doğru yanıtlama davranışları arasında fark olmadığı şeklinde yorumlanabilir. Bu çalışmadan elde edilen bulgular, alan yazındaki TIMSS 2011 Türkiye fen bilimleri alanında FMF'nin incelendiği bir çalışmada, kızların fene karşı kendilerine güvenlerinin yüksek olmasının erkeklerle aralarındaki başarı farklılıklarının nedenlerinden biri olduğu yönünde elde edilen bulgular ile tutarlıdır (Yalçın ve Tavşancıl, 2015). Türkiye'de kızların fen alanındaki başarısının son yıllarda erkeklerden manidar bir şekilde yüksek olmasının, kızların fen bilimlerine ilişkin duyuşsal özelliklerinin daha olumlu olmasını sağlamış olabileceği düşünülmektedir. FMF analizlerine ilişkin sonuçlar genel olarak değerlendirildiğinde, alan yazında yaygın olarak yapılan FMF analizlerine göre ele alınan maddelerin sadece birinde (20. madde) FMF tespit edilmektedir. Ancak önerilen yöntem sayesinde, öğrenciler bazı duyuşsal özelliklerine göre gizil sınıflara ayrıldığında, yirminci maddenin iki grupta FMF göstermediği, bir grupta ise ayrıca birinci maddenin de FMF gösterdiği tespit edilmiştir. Bu bağlamda, ele alınan duyuşsal özelliklerin aynı yetenek düzeyindeki öğrencilerin cinsiyete göre maddeleri doğru yanıtlama durumlarını etkilediği ifade edilebilir. Bu durum, doğrudan madde yanlılığı olarak ifade edilememekle birlikte araştırmacılara FMF'nin kaynağına yönelik istatistiksel bir bilgi sunmaktadır. Ayrıca, öğrencilerin duyuşsal özelliklerinin fen başarısı üzerindeki etksinin cinsiyete göre değiştiği bulgusu göz önünde bulundurularak öğrencilerin duyuşsal özelliklerini geliştirici etkinlikler, okullarda fen bilimleri ders programlarına bütünleştirilerek uygulanabilir.

INTRODUCTION

Countries place importance on the field of science and education given in this field in order to follow technology-oriented developments, to understand the world they live in and to develop new systems and technologies. For this reason, the findings of studies conducted for determining the science achievement of students at international level are important. It is seen in the 2015 application of the Trends in International Mathematics and Science Study (TIMSS), which is one of the studies mentioned before, that the average science success rate of Turkish students is quite low with 493 points (Yıldırım, Özgürlük, Parlak, Gönen and Polat, 2016). Students' success in science is affected by plenty of factors. It is seen in numerous studies in the field literature that students' attitudes towards science (Anıl, 2009; Bayraktar, 2011; Ghagar, Othman ve Mohammadpour, 2011; Kahraman, 2014; Thomson ve Fleming, 2004; Thomson ve diğ., 2008; Tighezza, 2014), self-confidence in science (Atar ve Aktan, 2013; Atar ve Atar, 2012; Bayraktar, 2011; Ghagar et al., 2011; Kaya ve Rice, 2010; Kiamanesh, 2004; Thomson ve diğ., 2008; Thomson ve Fleming, 2004; Tighezza, 2014), engagement in science courses (Chang, Singh ve Mo, 2007; Kahraman, 2014; Mo, 2008; Mo, Singh ve Chang, 2013), and value attached to learning science (Chang, 2008; Ghagar et al., 2011; Mohammadpour, 2012; Thomson ve diğ., 2008) influence academic achievement in science course positively. It is also determined in the field literature that the science achievement of the students differs according to sex (Bursal, 2013; Bursal, Buldur ve Dede, 2015). This situation draws attention in the TIMSS findings as well. In the TIMSS 1999 application, the average success rate of male students (434) was three points higher than the rate of female students (431); whereas in TIMSS 2015, the averages of female students (503) are 19 points higher than of the average of male students (484). The change in the students' science achievement according to sex in years is shown in Figure 1.

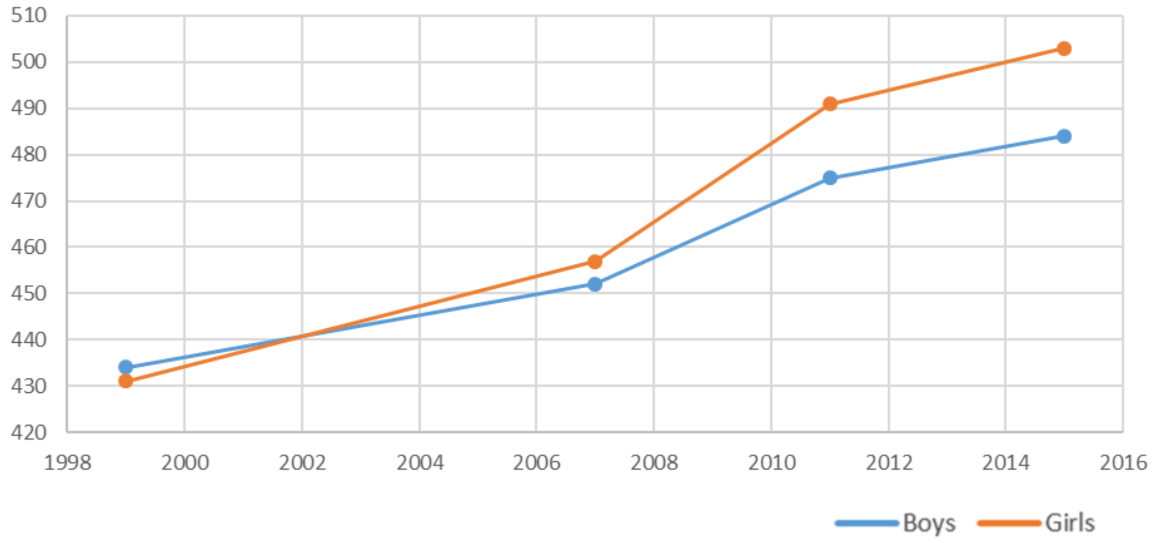

FIGURE 1. *The Students' Science Achievement According to Sex in Years*

As it is seen in the Figure 1, the difference between the science achievement scores of the male and female students' increase with the years and the girls' scores are observed to be higher than the boys' in the last two TIMSS applications. Differences in student achievement according to sex can be influenced by various factors (Wong, 2012). It is stated in the field literature that factors as social expectations (Kuzgun ve Sevim, 2004; Vatandaş, 2007), reading materials (Baker, 2002; Esen ve Bağlı, 2003; Eurydice, 2010; Kırbaşoğlu-Kılıç ve Eyüp, 2011), teachers (Eurydice, 2010; Kahle, Parker, Rennie ve Riley, 1993) and in-school factors (Eurydice, 2010) may create gender differences. In the field literature, differences were also determined in terms of the affective characteristics of students according to sex. Despite showing similar performances in many countries with boys, girls have lower self-concept in the field of science while men have higher self-sufficiency (Eurydice, 2010; Mo ve diğ., 2008; Thomson ve diğ., 2008; Thomson ve Fleming, 2004). On the other hand, Mohammadpour (2012) has seen in the study he conducted that girls' self-confidence in science is higher than boys. Thomson and Fleming (2004) indicate in their research that men's attitudes towards the field of science are higher than those of females. In this context, it is important that the determinations regarding students' achievement and affective characteristics are made according to sex.

Lower student success in TIMSS applications might be related to the increase in differences in success according to sex, and items' providing advantages or disadvantages to any subgroup. As a result of the measurement applications carried out, it is expected that the responses of the individuals in different groups which have equal abilities regarding the characteristic that is measured to be parallel, in other words, the measurements are expected to be invariable between different groups. Failure in having invariance invalidates the interpretation and comparison of the scores (Albano and Rodriguez, 2013). The accuracy of the decisions made based on the measurement results is closely related to the validity and reliability of the applications. One of the existing threats to the validity of the decisions is named as item bias (Clauser and Mazor, 1998). Bias is defined as systematic error in the measurement process (Osterlind, 1983). Test items' containing systematic errors causes the test to have less validity. In order to investigate whether the items which constitutes a test is biased or not, it is necessary to determine whether it is a differential item functioning (DIF). DIF can be defined as individuals' in different groups with the same ability level having different probabilities to answer an item correctly according to the subgroups (focus and reference) (Embretson ve Reise, 2000; Hambleton, Swaminathan ve Rogers, 1991; Mellenberg, 1989).

The existing statistical structure of the methods that are based on the Classical Test and Item Response Theory is rather limited in providing information on possible causes of the DIF and whether there is bias or not. Whether the items show differential item functions or not is generally investigated in the field literature in Turkey too. When the DIF is detected, the opinions of the experts are usually consulted to determine whether the item is biased or not (e.g. Çepni, 2011; Demirtaşlı ve Ulutaş, 2015; Kalaycıoğlu ve Kelecioğlu, 2011; Karakaya ve Kutlu, 2012; Kelecioğlu, Karabay ve Karabay, 2014; Özmen, 2014). It is observed that some recent studies have used explanatory and multilevel item response models in order to determine the sources and causes of the DIF. These studies are limited in Turkey (Yalçın and Tavşancıl, 2015) even though they are more common abroad (Albano ve Rodriguez, 2013; Balluerka, Gorostiaga, Gómez-Benito, and Hidalgo, 2010; Chaimongkol, 2005; Kamata, Chaimongkol, Genc ve Bilir, 2005; Kamata ve Binici, 2003; Zheng, 2009). The identification of the DIF sources also allows the test to abstain from the structure validity threat and to increase the accuracy of the estimations on ability parameter (Ong et al., 2011; Turhan, 2006). Since only a limited number of items are announced in large-scale applications such as TIMSS, the expert opinion cannot be obtained when DIF emerge in the unannounced items, and nothing can be stated as to the likely reason why the item is DIF and whether it is biased or not. For this reason, it is likewise important to determine the causes / multiple sources of the DIF as the determining of the DIF (Albano ve Rodriguez, 2013; Balluerka ve diğ., 2010; Beretvas, Cawthon, Lockhart ve Kaye, 2012; Kamata, 2001; Luppescu, 2002; Meulders ve Xie, 2004; Ong ve diğ., 2011; Turhan, 2006; Williams ve Beretvas, 2006). In the study Zumbo (2007) conducted and in which three generations of DIF studies are introduced, he named the studies on identification of the cause of the DIF's as the third generation DIF studies. Moreover, it is expressed that conceptual variables such as class size, socio-economic level (sel), teaching applications, and familial characteristics are not taken into consideration to a large extent in explaining DIF (and their causes) (Zumbo and Gelin, 2005). One of the DIF studies in the field literature (Yalçın and Tavşancıl, 2015) investigated the relationship between students' attitudes toward science, self-confidence in science, engagement in science courses, and the value attached to science learning and answering the items correctly according to sex. It is determined that the variable explaining the item that shows the maximum DIF is the "self-confidence in science".

Items' containing DIF in applications such as TIMSS, in which countries are compared at the international level, and important decisions concerning countries' education policies are made according to its results, damages the validity of the decisions made. Besides, in these kind of applications determining the causes/multiple sources of DIF is as important as the determining of the DIF. In this context, in order to determine the source of the DIF a new model is suggested in this study. In this research, a different method is employed in order to determine whether the potential DIF sources have influence on DIF by determining them before the DIF test. In the field literature, latent variables other than the observed variables play an important role in explaining structures that are dealt with such as student success. Creation of latent classes based on the responses of individuals to the observed variables gives opportunity to a better understanding of the structure that is addressed. In the latent class analysis (LCA), all observed variables are accepted as the cause of a latent variable that cannot be observed. It can be said that the relationship between the observed variables is conditional independence as a result of determining the latent variable as the control variable. Under this condition, determining of the latent variable which is the control variable, is carried out with LCA (Vermunt and Magidson, 2004). This study is carried out in three stages. First, students' affective characteristics which may possibly explain the difference in the answering a science item correctly or wrong are determined according to the sex of the students who are at the same level of ability. The status of identified affective characteristics' forming latent classes is determined in the second stage. In the last stage, it is aimed to determine whether or not the same items display DIF in the emerged latent classes. Thus, the effect of the DIF between the emerged latent classes can be seen. In addition, the effect of dividing into latent classes on the DIF will be determined by carrying out DIF analyses for the whole group. In this context, these are the questions to be answered in the study:

1. How are the latent classes that are emerged according to the students' attitude towards science, self-confidence in science, engagement in science classes, and the value attached to learning science?

2. How are the state of differential functioning of the items in the latent classes that are emerged according to the students' affective characteristics?

3. What is the status of the items' showing DIF for the whole group without making LCA?

METHOD

Model of the research

In this study, students are divided into latent classes with respect to the student characteristics that are approached in order to determine the effect of the students' affective characteristics on the DIF, and DIF analysis is conducted on an individual basis both for each latent class and for the whole group. In this context, this study is in the descriptive survey model since it puts forward the existing situation.

Population and Sampling

The population of the TIMSS 2015 application is composed of 1,187,893 students in the 8th grade in 2015. The sample is composed of 6079 students chosen with stratified multistage cluster sampling (Yıldırım et al., 2016). Since students response the science items in one booklet from 14 different booklets in TIMSS applications, in this study, in order to keep the working group far-reaching, common items in the second and third booklets, which have the highest number of common items among the booklets are chosen and a total of 875 students who took these booklets constitute the working group. 415 of these students are female (47%) and 460 (53%) are male.

Data and Collection

In the TIMSS applications, while the cognitive levels of students are determined by achievement tests, information regarding affective characteristics, home and family status, resources they have etc. are collected via student questionnaires. When the items in the achievement tests are placed in 14 different booklets, common items are used in both booklets in order to maintain equality between the booklets. In this study, in order to keep the working group far-reaching, the common items in the second and third booklets with the most common items, and are received by students at the most, and students who responded to these items are analysed. There are overall 23 of these items consisting nine from biology, seven from chemistry, four from physics and three from the field of earth sciences. When evaluated from the point of view of the type of item, it is seen that eight of them are multiple-choice, 11 are open-ended, and four are short-answers.

Selected index variables related to the affective characteristics of students in the student questionnaire are; attitude towards science (BSDGSLS), self-confidence in science (BSDGSCS), engagement in science courses (BSDGESL), and value attached to science learning (BSDGSVS). Detailed explanations regarding these index variables are presented below (Martin, Mullis, Foy & Hooper, 2016). The attitude towards science index consists of six items: a) I like to learn science, b) I wish I did not have to study science, c) Science is boring, d) I learn interesting things in science classes, e) I like science, and f) It is important for me to be good at science. The items have the Likert type of rating (1: Totally agree, 2: I agree, 3: I do not agree, 4: I totally disagree). When the index is being created, the ratings are converted to "0: Low (represents I more or less agree with the six items mentioned, in other words, it expresses that the attitude towards science is in a positive way), 2: High (represents I do not agree or agree a little to the six items mentioned, in other words, it expresses that the attitude towards science is in a negative way), 1: Medium (it points out other combinations, that is to say situations in which the attitude is neither positive nor negative) ".

The self-confidence in science variable is composed of four items: a) I am generally good at science b) I have more difficulty in science than my classmates c) Science is not one of the courses in which I am good at d) I learn science subjects rapidly. The items have Likert type of rating and they are converted to: "0: Low (represents I more or less agree with the four items mentioned, in other words, it expresses that the self-confidence in science is high), 2: High (represents I do not agree or agree a little to the four items mentioned, in other words, it means that the self-confidence in science is low), 1: Medium (it points out other combinations, in other words situations in which self-confidence in science is neither high nor low) ".

Student engagement in science courses variable consists of five items: a) I know what my teacher expects me to do b) I think about things that are not related to the lesson in science classes, c) It is easy to understand my teacher in science classes, d) I am interested in what my teacher says in science classes, e) My science teacher gives me interesting things to do. The items have Likert type rating. While creating the index, grades are converted to "0: Low (represents I more or less agree with the five items mentioned, in other words, it expresses that the student engagement in the science classes is high), 2: High (represents I do not agree or agree a little to the five items mentioned, in other words, indicating that the student participation in science classes is low), 1: Medium (refers to the other combinations, in other words situations in which student participation in science classes is neither high nor low)".

The value attached to learning science variable is based on the responses given to seven science-related situations; a) I would like to take more science courses at school b) I like to learn science c) I think learning science would make my daily life easier d) I need science to learn other school courses e) I need to be good at school to be able to go to the university I choose f) I want to work at a job that requires the use of science g) I need to do good at science in order to enter the job I want. The items have a Likert type rating. When the index is created, grades are converted to: "0: Low (represents I more or less agree with the seven items mentioned, in other words it points out that the value attached to the science learning is high), 2: High (represents I do not agree or agree a little to the seven items, in other words the value attached to the science learning is low), 1: Medium (points out other combination, in other words situations in which the value attached to learning science is neither high or low)".

Data Analysis

Analyses are conducted in two stages. In the first stage, latent class analysis (LCA) is used to determine the latent classes that are formed according to the students' attitudes towards science, self-confidence in science, engagement in science classes and value attached to science learning. All observed variables are accepted to be the cause of an unobservable latent variable in LCA (Vermunt and Magidson 2004). All possibilities, from the model with a latent class to a model that adapts the best are tried in the LCA. The simplest model which has the minimum latent class and least predictive parameter is preferred in model selection (Vermunt 2003; Vermunt and Magidson, 2004). In order to define the optimal number of classes, fit measures such as Log-Likelihood (LL) and Bayesian Information Criterion (BIC) are employed. In the simulation study Lukočienė, Varriale and Vermunt (2010) conducted, they observed that the BIC was the best criterion in model selection. For this reason, the BIC value is used in the model selection.

In the second phase of the analysis, Mantel-Haenszel (MH) method, which is one of the methods in field writing that is used oftentimes, is used in order to determine the state of differential functioning of the items for the whole group and in the latent classes emerged according to the students' affective characteristics. This method, which is developed by Mantel and Haenszel (1959), was first introduced by Holland and Thayer (1988) in order to determine the DIF. Being a non-parametric method, MH is based on comparison of groups which are matched according to the matching criterion with the help of 2x2 crosstabs in which the numbers of true and false responses that are separated by the focus and reference group indicator are shown (Holland & Thayer, 1988). In order to interpret the αMH value obtained as a result of the calculations easier, ΔMHi is obtained by applying a logarithmic transformation and the level of DIF is interpreted according to the ΔMHi value. If ΔMHi ≤ 1, then it is expressed as A level DIF (ignorable), if 1< ΔMHi <1.5, then it is expressed as B level DIF (medium level) and if ΔMHi≥ 1.5, then it is expressed as C level DIF (high level) (Dorans and Holland, 1992). It is used in many studies (Doğan and Öğretmen, 2008; Socha, DeMars, Zilberberg, & Phan, 2015; Zwick, 2012) for it is effective in determining items which include DIF content in different situations (DIF size, DIF type, sample size etc.). Being package programmes, Latent Gold 5.1 (Vermunt and Magidson 2013a, 2013b) is used in the first stage and Xcalibre 4.2.2 (Guyer & Thompson, 2014) is used in the second stage of the analyses.

RESULTS

Emerged Latent Classes

As a result of the LCA, which is carried out in order to determine the number of latent classes related to students' attitudes toward science, self-confidence in science, participation in science classes, and the value attached to learning science, it is observed that the model with three latent classes adapted to the data the best. The fit measures related to the models tested during the analyses are given in the Table 1.

Table 1. *Fit Measures of Formed Models Related to the Students' Affective Characteristics*

Model	LL	BIC (LL)	Npar
1-Class	-3264,7450	6583,4895	8
2-Class	-2909,2534	5906,2560	13
3-Class	*-2835,5467*	*5792,5922*	*18*
4-Class	-2822,9392	5801,1268	23
5-Class	-2810,7554	5810,5088	28
6-Class	-2807,2732	5837,2941	33

As it can be seen in the Table 1, the model with the lowest BIC value is the one that is composed of three classes. For this reason, this model is chosen and analyses are carried on afterwards. The results related to the state of the variables' being in classes according to the model with three classes are given in the Table 2.

Table 2. *Results on the State of Variables' Being in the Classes According to the Model with Three Classes*

	Class1	Class2	Class3	Wald	R^2
Class Size	0.45	0.41	0.14		

Students like learning science	Very Much Like Learning Science	0.97	0.23	0.00	74.5906*	0.71
	Like Learning Science	0.03	0.75	0.33		
	Do Not Like Learning Science	0.00	0.02	0.67		
Students confident in science	Very Confident in Science	0.63	0.08	0.01	103.8650*	0.49
	Confident in Science	0.34	0.44	0.16		
	Not Confident in Science	0.03	0.48	0.83		
Students' engaging teaching in science lessons	Very Engaging Teaching	0.92	0.60	0.16	91.4585*	0.36
	Engaging Teaching	0.08	0.33	0.43		
	Less than Engaging Teaching	0.00	0.07	0.41		
Students value science	Strongly Value Science	0.72	0.39	0.03	109.3228*	0.36
	Value Science	0.26	0.51	0.34		
	Do Not Value Science	0.02	0.10	0.63		

**p*<.05

As it can be seen in Table 2, the probability that all selected variables' to be included in the formed latent classes is meaningful according to the Wald statistics. The most effective one out of these variables is the attitude towards learning science (R^2: 0.71). Looking at the probability of students' to be included in the classes, 45% of the students are in the Class 1. 97% of students in this class very much like learning science, 63% of them are very self-confident in science, 92% of them are very engagement in science courses and 72% of them attach a great value to learning science. In the second class, which consists of 41% of the students, 75% of the students like to learn science, 48% of them are not self-confident in science, 60% are students who are very engagement in science courses and 51% of them are students who attach value to learning science. Class 3, which is the last class, consists of 14% of the students. It is seen that in this class, 67% of the students do not like learning science, 83% of them lack the self-confidence in science, 43% of them engage in science courses, 41% are less engagement in the science courses and 63% are students who do not attach value to learning science. The students in one of these three latent classes, which are formed according to their affective characteristics are divided into three groups.

DIF Analyses for Emerged Latent Classes

As a result of the Mantel Haenzsel test, which is carried out for the three emerged latent classes separately, it is determined that two items (S052261 and S052159) in the first latent class showed DIF favouring girls, yet these items did not show DIF in the other two classes. The first item shows B level (medium level) and the twentieth item shows C level (high level) DIF. The DIF results according to the emerged latent classes are given in the Table 3.

Tablo 3. *DIF Results According to Emerged Latent Classes*

	Latent Class 1 (N_{FE}: 205, N_{MA}:199)						**Latent Class 2** (N_{FE}: 165, N_{MA}:191)					**Latent Class 3** (N_{FE}: 45, N_{MA}:70				
	M-H	M-H D	M-H SE	z-test	p	Bias	M-H	M-H D	M-H SE	z-test	p	M-H	M-H D	M-H SE	z-test	p
1	**0.5309**	**1.4878**	**0.2745**	**2.3065**	**0.0211**	**FE**	0.9948	0.0123	0.3217	0.0163	0.9870	0.4898	1.6774	0.5390	1.3243	0.1854
2	1.1478	-0.3239	0.3108	-0.4436	0.6573		1.5512	-1.0318	0.3673	-1.1955	0.2319	0.9686	0.0751	0.6040	0.0529	0.9578
3	0.9349	0.1583	0.2810	0.2397	0.8106		1.0491	-0.1125	0.3273	-0.1463	0.8837	1.2793	-0.5789	0.6118	-0.4026	0.6872
4	0.8275	0.4451	0.2807	0.6748	0.4998		1.3072	-0.6296	0.3395	-0.7892	0.4300	1.2929	-0.6036	0.6219	-0.4130	0.6796
5	0.9427	0.1387	0.2857	0.2066	0.8363		0.9788	0.0503	0.3160	0.0678	0.9459	0.7901	0.5538	0.5823	0.4047	0.6857
6																
7	0.9983	0.0039	0.4085	0.0040	0.9968		2.3621	-2.0200	0.6225	-1.3808	0.1673	3.3208	-2.8205	1.2996	-0.9235	0.3558
8	0.6524	1.0035	0.3729	1.1452	0.2521		1.0540	-0.1237	0.5874	-0.0896	0.9286	1.5211	-0.9856	1.0403	-0.4031	0.6868
9	0.9434	0.1368	0.2837	0.2052	0.8374		1.6619	-1.1937	0.3414	-1.4880	0.1367	3.1316	-2.6826	0.6861	-1.6637	0.0962
10	0.9447	0.1336	0.2897	0.1963	0.8444		1.0746	-0.1691	0.3086	-0.2332	0.8156	1.0053	-0.0124	0.5512	-0.0096	0.9924
11	0.7473	0.6844	0.2825	1.0307	0.3027		0.8331	0.4290	0.3043	0.5999	0.5486	0.7194	0.7740	0.5283	0.6234	0.5330
12	1.1229	-0.2724	0.3847	-0.3013	0.7632		1.1062	-0.2371	0.3885	-0.2598	0.7950	0.9807	0.0458	0.6096	0.0320	0.9745
13	1.1067	-0.2382	0.3990	-0.2540	0.7995		1.0463	-0.1064	0.3938	-0.1149	0.9085	0.7845	0.5704	0.5919	0.4101	0.6818
14	0.9287	0.1739	0.3870	0.1912	0.8484		1.7505	-1.3158	0.4013	-1.3953	0.1629	1.9000	-1.5084	0.7085	-0.9060	0.3650
15	0.7325	0.7316	0.3726	0.8354	0.4035		1.4230	-0.8290	0.3486	-1.0121	0.3115	0.7122	0.7977	0.6107	0.5559	0.5783
16	1.1525	-0.3336	0.6474	-0.2193	0.8265		1.1190	-0.2642	0.4036	-0.2786	0.7805	1.0307	-0.0711	0.7077	-0.0427	0.9659
17	1.1345	-0.2965	0.2963	-0.4258	0.6702		1.4526	-0.8774	0.3317	-1.1257	0.2603	0.8190	0.4691	0.5373	0.3715	0.7103
18	0.8924	0.2675	0.2849	0.3995	0.6895		1.0458	-0.1053	0.3155	-0.1420	0.8871	1.1161	-0.2581	0.5768	-0.1904	0.8490
19	1.0154	-0.0360	0.3068	-0.0499	0.9602		1.2652	-0.5527	0.3962	-0.5937	0.5527	1.8744	-1.4765	0.6901	-0.9104	0.3626
20	**0.4023**	**2.1397**	**0.2889**	**3.1519**	**0.0016**	**FE**	0.5671	1.3328	0.2982	1.9016	0.0572	0.8989	0.2504	0.5601	0.1902	0.8491
21	0.5316	1.4850	0.3736	1.6916	0.0907		0.9672	0.0784	0.3860	0.0864	0.9312	0.6520	1.0050	0.6119	0.6989	0.4846
22	0.8065	0.5052	0.2763	0.7782	0.4365		1.3963	-0.7845	0.3244	-1.0292	0.3034	0.7421	0.7009	0.5741	0.5196	0.6034
23	0.8705	0.3260	0.2846	0.4873	0.6260		1.0690	-0.1569	0.3425	-0.1949	0.8455	1.4945	-0.9442	0.6171	-0.6511	0.5150

As it is seen in the Table 3, DIF analyses are not carried out by the program for the sixth item, because this item has two c parameters in three latent classes. Considering the fact that the latent class (LC1), which contains items that show DIF is most of the time composed of students who like to learn science, who are very high self-confidence in science, who are very engagement in science courses and who deeply attach value to learning science, if girls have these affective characteristics in a high level, it can be expressed that the ratio of girls' answering these items correctly is meaningfully higher than of boys'. In latent class-2, there is no difference with regards to the answering the items correctly behaviour between girls and boys who usually like learning science, who have self-confidence in science, who participate in science classes and who attach value to learning science and who are at the same level of ability. A similar situation is also valid for the Latent Class-3, which is composed of students who generally dislike learning science, does not have self-confidence in science, rarely participates in science classes and does not attach value to learning science.

DIF Analyses for the Whole Group

When the DIF analysis is carried out directly for a single class without conducting LCA, it is seen that the same two items just as in the Latent Class-1 showed the DIF. Detailed results regarding the DIF analysis are presented in the Table 4.

Tablo 4. *DIF Results for the Whole Group (N_{FE}: 415, N_{MA}:460)*

	M-H	M-H D	M-H SE	z-test	p	Bias
1	0.6791	0.9094	0.1931	2.0036	0.0451	FE
2	1.2898	-0.5981	0.2169	-1.1736	0.2406	
3	1.0643	-0.1464	0.1943	-0.3205	0.7486	
4	1.0579	-0.1323	0.1973	-0.2852	0.7755	
5	0.9619	0.0913	0.1954	0.1988	0.8424	
6						
7	1.5239	-0.9900	0.3291	-1.2799	0.2006	
8	0.8170	0.4749	0.2989	0.6762	0.4989	
9	1.3902	-0.7742	0.2076	-1.5870	0.1125	
10	1.0325	-0.0753	0.1966	-0.1629	0.8706	
11	0.7987	0.5283	0.1913	1.1754	0.2398	
12	1.1047	-0.2340	0.2623	-0.3795	0.7043	
13	0.9979	0.0049	0.2707	0.0077	0.9938	
14	1.3858	-0.7668	0.2705	-1.2061	0.2278	
15	0.9969	0.0073	0.2284	0.0136	0.9892	
16	1.0719	-0.1631	0.3261	-0.2129	0.8314	
17	1.1938	-0.4163	0.2020	-0.8769	0.3805	
18	0.9775	0.0536	0.1972	0.1156	0.9080	
19	1.2098	-0.4476	0.2281	-0.8351	0.4037	
20	0.5176	1.5474	0.1933	3.4063	0.0007	FE
21	0.7570	0.6541	0.2404	1.1580	0.2469	
22	1.0278	-0.0644	0.1959	-0.1399	0.8888	
23	1.0278	-0.0644	0.2063	-0.1328	0.8944	

As it is seen in the Table 4, while the first item is DIF at level A, which is a negligible level, the 20th item shows C level (high level) DIF. Different from the findings from the Latent class-1, the status of the first item's displaying DIF is at a negligible level.

DISCUSSION, CONCLUSION AND SUGGESTIONS

The state of students' dividing into latent classes is determined according to their attitudes towards science, self-confidence in science, engagement in science courses and the value attached to learning science by benefiting from the field literature. The first one of these latent classes is made up of students who quite like to learn science, who are very self-confident in science, who are very extremely engagement in science courses and who are attach a great value to learning science. Second latent class consists of students who like to learn science, who have a bit of self-confidence in science or none, who participate in science classes rarely or a lot and who attach value to science learning. As for the third latent class, it is made up of students who do not like learning science, who do not have self-confidence in science, who rarely participate in science courses and do not attach value to learning science.

Students are divided into three groups according to the state of their taking part in these latent classes which are formed according to affective characteristics. As a result of the Mantel Haenzsel test, which is carried out separately for the three emerged latent classes, it is determined that two items (S052261 and S052159) in the first latent class showed DIF in favour of girls, whereas these items in the other two classes did not show DIF. The first item shows B level (middle level) and the twentieth item shows C level (high level) DIF. Considering the fact that the latent class (LC1), which consists of items displaying DIF is comprise of students who usually quite enjoy learning science, who have high self-confidence in science, are very participative in science courses and who attach high value to learning science, if girls have a high level of these affective characteristics, it can be expressed that the ratio of girls' answering these items correctly is meaningfully higher than of boys'. Moreover, items' not displaying DIF in latent classes-2 and 3 can be interpreted as there is no difference according to sex in the answering correctly behaviours of the students who possess low or medium level affective characteristics towards science at the same ability level. Findings obtained in this study are consistent with the findings of a study analysing DIF in the field of science in Turkey in TIMSS 2011 (Yalçın and Tavşancıl, 2015), which found out that girls' having high self-confidence in science was one of the reasons for the difference in success between boys and girls. It is also determined in the study Mohammadpour (2012) conducted on Malaysia's 1999, 2003 and 2007 TIMSS application data that girls have more self-confidence in science than of boys. However, there are also findings in the field literature suggesting that boys have more positive attitudes towards science than females (Thomson and Fleming, 2004), that they are more self-confident (Eurydice, 2010; Mo ve diğ., 2008; Thomson ve diğ., 2008; Thomson ve Fleming, 2004) and that they attach more value to science than of girls (Chang, 2008). Although it is known that this situation varies from country to country, it is thought that girls' being meaningfully more successful than boys in the field of science in Turkey in recent years (Yıldırım et al., 2016) may made it possible for girls to have more positive affective characteristics regarding science.

When the two items that display DIF are evaluated from the point of view of the subject area, it is determined that the first item is from the topic of "The Course of Life and Characteristics of Organisms" in the field of Biology learning and the second item is from the "Force and Movement" in the field of Physics learning, both items are at the knowledge level with regards to cognitive level and they are multiple choice with regards to item type (IEA, 2016). Although in this study it is not meant to generalise with the results obtained solely from two items, the findings obtained from the study are consistent with the findings in the field suggesting that the biology items are in favour of girls (Berberoğlu, 1996; Calvert, 2002; Qian, 2011; Yenal, 1995; Yip ve diğ., 2004; Yung, 2006). However, is not consistent with the findings in the field literature suggesting that boys are more successful in physics items (Berberoğlu, 1996; Calvert, 2002; Qian, 2011; Yip ve diğ., 2004; Yung, 2006) and in multiple choice items (Le, 2009; Yip ve diğ., 2004; Yung, 2006) than girls. This situation is thought to be originated from students' learning styles and /or the increase in girls' success in science in the recent years.

When the DIF analyses are carried out for the entire group, the first item displays DIF at level A, which is negligible whereas the 20th item displays DIF at level C (high level). Different from the latent class-1 findings, the status of the first item's displaying DIF is at a negligible level. When the results regarding the DIF analyses are broadly evaluated, DIF is detected only in one of the items (20th item) that is dealt with according to the DIF analyses which are common in the field literature. However, thanks to the suggested method, it is found out that when the students are classified into latent classes according to some affective characteristics, the twentieth item did not display DIF in two groups and in one group, the first item also displayed DIF. In this context, it can be stated that the affective characteristics which are dealt with affect the status of answering the items correctly of the students with the same ability level according to sex. Even though this situation cannot be directly expressed as item bias, it provides statistical information regarding the source of the DIF to the researchers. For this reason, this method will contribute to the field literature for it is thought to give DIF studies a different dimension. In addition, considering the fact that the influence of the affective characteristics of the students on the science achievement varies according to sex, activities that can develop the affective characteristics of students can be executed in schools by integrating them into science curriculum.

There are also some limitations to the conducted study. In determining the source of the DIF, multilevel DIF models could have been used in this study for all booklets at the same time as well (Yalçın and Tavşancıl, 2015), however, because this study is a method suggestion and all individuals must have responded to all the items in order to be able to apply LCA, this study was limited to two booklets. In addition, the multilevel LCA analysis that includes the school level could not be carried out because of the low distribution frequency of the individuals who responded to the two booklets. Additionally, there are many methods employed in determining DIF in the field literature. In this study, more than one methods are used in determining DIF as whether items with DIF change in different situations is examined. One of the most often used DIF methods in the field literature is preferred. Interested researchers can make comparisons by using other DIF identification methods as well. Finally, if two items determined as the DIF can be accessed, why they provide advantage to girls and whether they are biased or not can be examined with expert opinions.

REFERENCES

Albano, D. A., & Rodriguez, M. C. (2013). Examining differential math performance by gender and opportunity to learn. *Educational and Psychological Measurement*, *73*(5), 836-856.

Anıl, D. (2009). Uluslararası öğrenci başarılarını değerlendirme programı (PISA)'nda Türkiye'deki öğrencilerin fen bilimleri başarılarını etkileyen faktörler. *Eğitim ve Bilim*, *34*, 87-100.

Atar, B. & Aktan, D. Ç. (2013). Örtük regresyon iki parametreli lojistik modeli. *Eğitim ve Bilim*, *38*(168), 59-68.

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

Baker, D. (2002). Where is gender and equity in science education? *Journal of Research in Science Teaching*, *39*, 659–663.

Balluerka, N., Gorostiaga, A., Gómez-Benito, J., & Hidalgo, D. (2010). Use of multilevel logistic regression to identify the causes of differential item functioning. *Psicothema*, 22(4), 1018-1025.

Bayraktar, Ş. (2011). Uluslararası fen ve matematik çalışması (TIMSS 2007) sonuçlarına göre Türkiye'de fen eğitiminin durumu: Fen başarısını etkileyen faktörler. *Selçuk Üniversitesi Ahmet Keleşoğlu Eğitim Fakültesi Dergisi*, *30*, 249-270.

Berberoğlu, G. (1996). The university entrance examinations in Turkey. *Studies in Educational Evaluation*, *22*(4), 363-373.

Beretvas, S. N., Cawthon, S. W., Lockhart, L. L., & Kaye, A. D. (2012). Assessing impact, DIF, and DFF in accommodated item scores: A comparison of multilevel measurement model parameterizations. *Educational and Psychological Measurement*, *72*, 754-773.

Bursal, M. (2013). İlköğretim öğrencilerinin 4-8. sınıf fen akademik başarılarının boylamsal incelenmesi: Sınıf düzeyi ve cinsiyet farklılıkları. *Kuram ve Uygulamada Eğitim Bilimleri*, *13*(2), 1141-1156.

Bursal, M., Buldur, S. & Dede, Y. (2015). Alt sosyo-ekonomik düzeyli ilköğretim öğrencilerinin 4-8. sınıflar fen ve matematik ders başarıları: Cinsiyet perspektifi. *Eğitim ve Bilim*, *40*(179), 133-145.

Calvert, T. (2002). *Exploring differential item functioning (DIF) with the rasch model: A comparison of gender differences on eighth grade science items in the United States and Spain*. Unpublished doctoral dissertation, University of Emery, Atlanta.

Chaimongkol, S. (2005). *Modeling differential item functionıng (DIF) using multilevel logistic regression models: A bayesian perspective*. Unpublished doctoral dissertation, University of Florida State, Tallahassee.

Chang, M., Singh, K., & Mo, Y. (2007). Science engagement and science achievement: Longitudinal models using NELS data. *Educational Research and Evaluation*, *13*(4), 349-371.

Chang, Y. (2008). *Gender differences in science achievement, science selfconcept, and science values*. Proceedings of the IRC, Chinese Taipei.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44.

Çepni, Z. (2011). *Değişen madde fonksiyonlarının SIBTEST, mantel haenszel, lojistik regresyon ve madde tepki kuramı yöntemleriyle incelenmesi.* Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Demirtaşlı, N. & Ulutaş, S. (2015). A study on detecting differential item functioning of PISA 2006 science literacy items in Turkish and American samples. *Eurasian Journal of Educational Research*, *58*, 41-60. http://dx.doi.org/10.14689/ejer.2015.58.3

Doğan, N. & Öğretmen, T. (2008). Değişen madde fonksiyonunu belirlemede mantel- haenszel, ki-kare ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, *33*(148), 100-112.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel Haenszel and standardization. In P. W. Holland, ve H. Wainer, (Eds.), *Differential item functioning* (pp. 35–66), New Jersey: USA.

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

Esen, Y. & Bağlı, M. T. (2003). İlköğretim ders kitaplarındaki kadın ve erkek resimlerine ilişkin bir inceleme. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *35*(1-2), 143-154.

Eurydice. (2010). *Eğitim çıktılarında cinsiyet farklılıkları: Avrupa'da alınan tedbirler ve mevcut durum.* Eurydice Raporu, Brüksel. 16 Eylül 2014 tarihinde http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/120TR.pdf adresinden erişilmiştir.

Ghagar, M. N., Othman, R., & Mohammadpour, E. (2011). Multilevel analysis of achievement in mathematics of Malaysian and Singaporean students. *Journal of Educational Psychology and Counseling*, 2, 285-304.

Guyer, R., & Thompson, N. A. (2014). *User's manual for Xcalibre item response theory calibration software, version 4.2.2 and later*. Woodbury MN: Assessment Systems Corporation.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the mantel-haenszel procedure. In H. Wainer, and H. I. Brown (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

IEA. (2016). *TIMSS 2015 international database*, item information tables. https://timssandpirls.bc.edu/timss2015/international-database/

Kahle, J. B., Parker, L. H., Rennie, L. J., & Riley, D. (1993). Gender differences in science education: Building a model. *Educational Psychologist*, *28*, 379-404.

Kahraman, N. (2014). Cross-grade comparison of relationship between students' engagement and TIMSS 2011 science achievement. *Education and Science*, *39*(172), 95-107.

Kalaycıoğlu, D. B. & Kelecioğlu, H. (2011). Öğrenci seçme sınavının madde yanlılığı açısından incelenmesi. *Eğitim ve Bilim*, *36*, 3-13.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79–93.

Kamata, A., & Binici, S. (2003, July). *Random effect DIF analysis via hierarchical generalized linear modeling*. Paper presented at the biannual International Meeting of the Psychometric Society, Sardinia, Italy.

Kamata, A., Chaimongkol, S., Genc, E., & Bilir, K. (2005, April). *Random-effect differential item functioning across group unites by the hierarchical generalized linear model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Karakaya, İ. & Kutlu, Ö. (2012). Seviye belirleme sınavındaki türkçe alt testlerinin madde yanlılığının incelenmesi. *Eğitim ve Bilim*, *37*, 165.

Kaya, S., & Rice, D. C. (2010). Multilevel effects of student and classroom factors on elementary science achievement in five countries. *International Journal of Science Education*, *32*(10), 1337-1363.

Kelecioğlu, H., Karabay, B. & Karabay, E. (2014). Investigation of placement test in terms of item biasness. *Elementary Education Online*, *13*(3), 934-953.

Kırbaşoğlu- Kılıç. L. & Bircan, E. (2011). İlköğretim türkçe ders kitaplarında ortaya çıkan toplumsal cinsiyet rolleri üzerine bir inceleme. *Sosyal Bilimler Araştırmaları Dergisi*, *6*(2), 129-148.

Kiamanesh, A. R. (2004, July). *Self-concept, home back-ground, motivation, attribution and their effects on Iranian students' science achievement*. Paper presented at the Third International Biennial SELF Research Conference, Berlin, Germany.

Kuzgun, Y. & Sevim, A. S. (2004). Kadınların çalışmasına karşı tutum ve dini yönelim arasındaki ilişki. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *37*(1), 14-27.

Le, L. T. (2009). Investigating gender differential item functioning across countries and test language of PISA science items. *International Journal of Testing*, 9, 122-133.

Lukočienė, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, *40*(1), 247-283. doi: 10.1111/j.1467-9531.2010.01231.x

Luppescu, S. (2002, April). *DIF detection in HLM item analysis*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Mantel, N. & Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.

Martin, M. O., Mullis, I. V. S., Foy, P. & Hooper, M. (2016). *TIMSS 2015 international results in science*. Retrieved from Boston College, TIMSS & PIRLS International Student Center website: http://timssandpirls.bc.edu/timss2015/international-results/wp-content/uploads/filebase/full%20pdfs/T15-International-Results-in-Science-Grade-8.pdf

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, 13, 127-143.

Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In P. De Boeck and M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 213-240). New York: Springer-Verlag.

Mo, Y. (2008). *Opportunity to learn, engagement, and science achievement: Evidence form TIMSS 2003 Data*. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Mo, Y., Singh, K., & Chang, M. (2013). Opportunity to learn and student engagement: A HLM study on eighth grade science achievement. *Educational Research for Policy & Practice*, 12(1), 3-19.

Mohammadpour, E. (2012). A multilevel study on trends in malaysian secondary school students' science achievement and associated school and student predictors. *Science Education*, *96*(6), 1013-1046.

Ong, Y. M., Williams, J., & Lamprianou, I. (2011). Exploring differential bundle functioning in mathematics by gender: The effect of hierarchical modelling. *International Journal of Testing*, 11, 271–293.

Osterlind, J. S. (1983). *Test item bias*. London: Sage Publications.

Özmen, D. T. (2014). PISA 2009 okuma testi maddelerinin yanlılığı üzerine bir çalışma. *Eğitim Bilimleri ve Uygulama*, 13(26), 147-165.

Qian, X. (2011). *A multi-level differential item functioning analysis of trends in international mathematics and science study: Potential sources of gender and minority difference among U.S. eighth graders' science achievement*. Unpublished doctoral dissertation, Faculty of the University of Delaware.

Socha, A., DeMars, C. E., Zilberberg, A. & Phan, H. (2015). Differential item functioning detection with the mantel-haenszel procedure: The effects of matching types and other factors. *International Journal of Testing*, *15*(3), 193-215, DOI: 10.1080/15305058.2014.984066

Thomson, S., & Fleming, N. (2004). *Examining the evidence: Science achievement in Australian schools in TIMSS 2002 (TIMSS Australia Monograph No 7).* Melbourne, VIC: ACER Press

Thomson, S., Wernert, N., Underwood, C., & Nicholas, M. (2008). *TIMSS 2007: Taking a closer look at mathematics and science in Australia*. Trends in International Mathematics and Science Study, Camberwell, Vic.

Tighezza, M. (2014). Modeling relationships among learning, attitude, self-perception, and science achievement for grade 8 Saudi students. *International Journal of Science and Mathematics Education*, *12*(4), 721-740.

Turhan, A. (2006). *Multilevel 2PL item response model vertical equating with the presence of differential item functioning*. Unpublished doctoral dissertation, University of Florida State, Tallahassee.

Vatandaş, C. (2007). Toplumsal cinsiyet ve cinsiyet rollerinin algılanışı. *Sosyoloji Konferansları*, 35, 29-56.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*(1), 213-239. doi: 10.1111/j.0081-1750.2003.t01- 1-00131.x

Vermunt, J. K., & Madigson, J. (2004). Local independence. In A. B. M. S. Lewis Beck (Ed.), *Encyclopedia of social sciences research methods* (pp. 732-733). Thousand Oaks: Sage Publications.

Vermunt, J. K., & Magidson, J. (2013a). *Latent GOLD 5.0 upgrade manual*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2013b). *LG-syntax user's guide: Manual for latent GOLD 5.0 syntax module*. Belmont, MA: Statistical Innovations Inc.

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement*, *30*, 22-42.

Wong, K. Y. (2012). *Gender differences in scientific literacy of HKPISA 2006: A multidimensional differential item functioning and multilevel mediation study*. Unpublished doctoral dissertation, The Chinese University of Hong Kong, Hong Kong.

Yalçın, S., & Tavşancıl, E. (2015). The factors explaining the differential item functioning in the administration of TIMSS 2011 science test according to gender. *Educational Sciences and Practice*, *14*(27), 1-21.

Yıldırım, A., Özgürlük, B., Parlak, B., Gönen, E., & Polat, M. (2016). *TIMSS 2015 ulusal matematik ve fen bilimleri ön raporu 4. ve 8. sınıflar*. MEB: Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü.

Yip, D. Y., Chiu, M. M., & Ho, S. C. (2004). Hong Kong student achievement in OECD-PISA study: Gender differences in science content, literacy skills and test item formats. *International Journal of Science and Mathematics Education*, *2*(1), 91-106.

Yung, B. H. W. (2006). *Learning from TIMSS implications for teaching and learning science at the junior secondary level*. Education and Manpower Bureau, Hong Kong (China): Government Logistics Department.

Zheng, X. (2009). *Multilevel item response modeling: Applications to large-scale assessment of academic achievement*. Unpublished doctoral dissertation, University of California, Berkeley.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233.

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies*, *5*, 1-2.

Zwick, R. (2012). *A review of ets differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement*. ETS, Research Report. https://www.ets.org/Media/Research/pdf/RR-12-08.pdf