Elementary Education Online, 2018; 17(2): pp. 909-925
İlköğretim Online, 2018; 17(2): s.909-925. [Online]: http://ilkogretim-online.org.tr
doi 10.17051/ilkonline.2018.419339

Mantel-Haenszel, Lojistik Regresyon ve Olabilirlik Oranı Değişen Madde Fonksiyonu İnceleme Yöntemlerinin Farklı Yazılımlar Kullanılarak Karşılaştırılması*

Comparison of Mantel-Haenszel, Logistic Regression and Likelihood Ratio Methods to Evaluate Differential Item Functioning by Using Different Computer Software†

İlhan Koyuncu, Adıyaman Üniversitesi, *ilhankync@gmail.com*
Gökhan Aksu, Adnan Menderes Üniversitesi, *gokhanaksu1983@hotmail.com*
Hülya Kelecioğlu, Hacettepe Üniversitesi, *hulyakelecioglu@gmail.com*

Öz. Bu çalışmanın amacı, 2012 yılında yapılan Uluslararası Öğrenci Değerlendirme Programı (PISA) uygulaması öğrenci anketinden alınan bir veri ile klasik test ve madde tepki kuramı yöntemleri kullanılarak yapılan değişen madde fonksiyonu (DMF) analizlerinin sonuçlarını karşılaştırmaktır. Araştırmanın çalışma grubu matematik çalışma disipliniyle ilgili 9 maddeye ait kayıp verisi olmayan Türk öğrenciler arasından rastgele örnekleme yöntemiyle seçilen 990 öğrencidir. Matematik başarısı bakımından, öğrencilerin %45'i alt grupta, %55'i ise üst grupta yer almıştır. Veri analiz yöntemleri olarak Mantel-Haenszel, lojistik regresyon ve olabilirlik oranı yöntemleri kullanılmıştır. Analizlerde JMETRIK, DIFAS, Zumbo SPSS Syntax, IRTLRDIF ve IRTPRO yazılımları kullanılmıştır. Araştırma sonuçlarına göre, tüm yöntemlere ve yazılımlara göre 'ödevi zamanında tamamlama' maddesinde DMF olduğu belirlenmiştir. Tüm yöntemlerin ve yazılımların sonuçları kısmen benzerlikler göstermekle birlikte farklılıklar da içermektedir. Araştırmadan elde edilen bulgular ve ilgili alan yazın ışığında teori ve pratiğe dönük bazı önerilerde bulunulmuştur.

Anahtar Sözcükler: Değişen madde fonksiyonu, Lojistik regresyon, Madde yanlılığı, Mantel-Haenszel, Olabilirlik oranı

Abstract. The aim of this study is to compare the results of differential item functioning (DIF) analysis in terms of classical test and item response theory with different computer software by using the Programme for International Student Assessment (PISA) 2012 student questionnaire data. Participants are 990 students who were randomly selected from Turkish students answering all 9 mathematics work ethics items. In terms of mathematics performance factor variable, 45% of the participants were from lower and 55% of them were from upper group. DIF was evaluated with Mantel-Haenszel, logistic regression and likelihood ratio methods. Data was analyzed by using JMETRIK, DIFAS, Zumbo SPSS Syntax, IRTLRDIF and IRTPRO software. The results showed that there was DIF in the item "homework completed in time" in terms of all analysis software and methods. Moreover, it was observed that the results of all analysis software and methods contain similarities and differences. In the light of the findings of the study and related literature, some suggestions were made for theory and practice.

Keywords: Differential item functioning, Logistic regression, Mantel-Haenszel, Item bias, Likelihood ratio

* Bu çalışma, 4. Uluslararası Avrasya Eğitim Araştırmaları Kongresinde sözlü bildiri olarak sunulmuştur.
†This study was presented as an oral presentation at the IVth International Eurasian Educational Research Congress.

Geliş tarihi: 29.11.2017 Kabul tarihi: 17.03.2018 Yayımlanma tarihi: 27.04.2018

SUMMARY

Introduction

Item bias are systematical errors emerging from difference between estimated and predicted values of items parameters. These errors affect construct and predictive validity of measurement instrument (Osterlind, 1983). In differential item functioning (DIF), for the participants at the same proficiency level and from different subgroups (e.g. gender, socio-economic status, etc.), the probability of answering an item correctly differs (Zumbo, 1999). The reason for this discrepancy might be item bias or difference in real knowledge, ability, etc. (Özdemir, 2003). After detecting item bias with DIF, professionals and content analysts in the field are consulted to determine the source/s of this bias (Doğan & Öğretmen, 2008).

There are different classifications for DIF detecting algorithms in the literature. The most basic classification was made based on the theory that the algorithm relies on. Chi-square, logistic regression (LR), Mantel-Haenszel (MH), transformed item index, factor analysis, standardizing method and analysis of variance were categorized as classical test theory (CTT) techniques. Some of item response theory (IRT) algorithms are signed and unsigned area indices, item parameters, likelihood ratio (IRT-LR) and Lord's Chi-square (Camili ve Stephard, 1994; Hambleton, Swaminathan, & Rogers, 1991; Mellenberg, 1989; Raju, 1990; Zumbo, 1999). Potenza and Dorans (1995) classified logistic regression, Mantel-Haenszel and standardizing methods as observed score techniques; SIBTEST, Lord's Chi-square and likelihood ratio methods as latent ones.

MH is one the most widely used algorithm which is based on Chi-square statistics. This method, in which the probability ratios were evaluated, cannot discriminate between uniform and non-uniform DIF. LR compares statistically probability of responding an item correctly for individuals who are at the same proficiency level but from different groups (Zumbo, 1997). The analyses which are performed with likelihood ratio (IRT-LR) base item response theory models. In this method, the hypothesis which claims no difference between focal and reference groups' item parameters is tested (Atalay, Gök, Kelecioğlu, & Arsan, 2012).

The computer software used for detecting items with DIF differs in terms of its theoretical bases (IRT or CTT-based) and DIF's structure (uniform or non-uniform). Past studies revealed that different DIF detection algorithms have different results (Gao & Wang, 2005). For this reason, in related literature, it is advised to use more than one DIF detection methods together (Ferreres & Muniz, 2005). The aim of this study is to compare the results of MH, LR and IRT-LR methods for evaluating differential item functioning with different computer software by using the Programme for International Student Assessment (PISA) 2012 student questionnaire data. In this study, DIFAS and JMETRIK software were used for MH, EZDIF was selected for both MH and LR, Zumbo SPSS Syntax was used for LR, and IRT-LR was evaluated by using IRTPRO and IRTLRDIF software.

Method

This study is a basic research as it is aimed to compare the results of analysis performed with different computer software for different DIF detection algorithms. In basic research, it is intended to contribute new knowledge to the existing literature (Karasar, 2005). It also shows the properties of descriptive (survey) research because it is aimed to reveal and describe an existing situation. Participants are 990 students who were randomly selected from students answering all nine mathematics work ethics items. In terms of mathematics performance factor variable, 45% of the participants were from lower and 55% of them were from upper group. DIF was evaluated with Mantel-Haenszel, logistic regression and likelihood ratio methods. Data was analyzed by using JMETRIK, DIFAS, Zumbo SPSS Syntax, IRTLRDIF and IRTPRO software.

Results

The results showed that there was DIF in the item "homework completed in time" in terms of all analysis software and methods. Each of first six items shows DIF for at least two algorithms. Moreover, it was observed that the results of all analysis software and methods contain similarities and differences. According to the results of IRTLRDIF and JMETRIK, DIF detected for items 1 and 3 is not negligible. Except this software, for all other software and algorithms, DIF was detected for item 2. According to the results of IRTPRO and Zumbo SPSS Syntax, DIF detected for item 6 is not negligible. LR algorithm results revealed that except item 4, first 6 items show DIF. For other software and algorithms, DIF was detected for at least two items. Although MH, LR and IRT-LR algorithms have different statistical bases, the results of all methods revealed that first three items show DIF. This result provides significant evidence that these items shows bias for certain sub-groups.

Discussion and Conclusion

It was observed that the results of all software and methods contain similarities and differences. Past studies revealed that different DIF detection algorithms have different results (Gao & Wang, 2005). For this reason, in related literature, it is advised to use more than one DIF detection methods together (Ferreres & Muniz, 2005).

The number of DIF detected items for LR is more than the one for MH. This result is not consistent with the finding of Gomez-Benito and Navas-Ara (2000), and Atalay, Gök, Kelecioğlu and Arslan (2012). However, the findings of those studies were obtained from simulated data and that might be the reason for this situation. Besides that, the number of DIF detected items with LR and IRT-LR is consistent. This result shows similarity with the findings of Thissen, Steinberg and Wainer (1988), and Yıldırım (2008).

The analyses performed with different computer software for the algorithms having the same statistical bases revealed different results. However, the same analyses for the algorithms having different statistical bases revealed similar results. These findings were consistent with the idea that using more than one algorithm and computer software in order to determine items showing DIF will provide stronger evidence to the professionals and content analysts to decide on item biases.

Determining items showing DIF with the same computer software by using simulated data to examine the efficacy of the algorithms will contribute the accuracy of inferences done from the analysis performed with real data. In future studies, algorithms' Type I error rates and powers could be evaluated under certain conditions by using as much datasets as possible.

GİRİŞ

Yanlılık, madde parametrelerinin gerçek değerleri ile kestirilen değerleri arasındaki farklılıktan dolayı ortaya çıkan sistematik hatalardır. Ölçme aracının yapı ve yordama geçerliliğine etki eder (Osterlind, 1983). Aynı yetenek düzeyinde olmasına rağmen bir testin alt gruplarındaki bireylerin maddeyi doğru cevaplama olasılıklarının değişiklik gösterebilmektedir (Zumbo, 1999). Bu değişmenin sebebi madde yanlılığı ya da gerçek bilgi, beceri vb. farklılığından kaynaklanmaktadır (Doğan ve Öğretmen, 2008). Yanlılık, genel bir kavram olmasına karşın değişen madde fonksiyonu (DMF) yanlılığın işlevsel, objektif bir göstergesidir (Özdemir, 2003). DMF birçok kaynakta benzer şekilde tanımlanmaktadır. En genel çerçevede DMF, aynı yetenek düzeyinde olan ancak farklı alt gruplarda yer alan bireylerin test maddelerine cevap verme olasılıklarının farklılık göstermesi olarak tanımlanmaktadır (Osterlind, 1983; Raju, 1990; Mellenberg, 1989; Zumbo, 1999). DMF ile madde yanlılığı belirlendikten sonra yanlılığın kaynağını saptamak amacıyla uzman ve içerik analizlerinin görüşlerine başvurulabilmektedir (Doğan ve Öğretmen, 2008).

Değişen madde fonksiyonu tek biçimli (uniform) ve tek biçimli olmayan (nonuniform) olmak üzere ikiye ayrılmaktadır. Eğer bir maddenin doğru cevaplanma olasılığı tüm yetenek düzeyleri için belirli bir grup lehine DMF içeriyorsa tek biçimli; farklı yetenek düzeylerinde farklı gruplar lehine DMF içeriyorsa tek biçimli olmayan DMF olarak ele alınır (Zumbo, 1999). Örneğin, cinsiyet değişkeni açısından tüm yetenek düzeylerinde doğru cevaplandırma olasılıkları kızlar lehine işliyorsa tek biçimli DMF'den söz edilebilir. Ancak, bu olasılık düşük yetenek düzeylerinde kızlar lehine, yüksek yetenek düzeylerinde erkekler lehine işliyorsa tek biçimli olmayan DMF vardır (Özdemir, 2003).

Araştırmalarda DMF'yi değerlendirmek amacıyla kullanılan yöntemler için farklı sınıflandırmalar yapılmıştır. Yapılan en temel sınıflandırma klasik test kuramına ve madde tepki kuramına göre olmak üzere ikiye ayrılmaktadır. Klasik test kuramına dayanan teknikler Ki-Kare, lojistik regresyon, Mantel-Haenszel, dönüştürülmüş madde indeksi, faktör analizi, madde ayırıcılık gücü, standartlaştırma yöntemi Ki-Kare ve varyans analizidir. Madde tepki kuramına dayanan teknikler ise işaretli ve işaretsiz alan indeksleri, madde parametreleri, Lord'un Ki-Kare'si ve madde tepki kuramı olabilirlik oranıdır(Camili ve Stephard, 1994; Hambleton, Swaminathan, & Rogers, 1991; Mellenberg, 1989; Raju, 1990; Zumbo, 1999). Potenza ve Dorans (1995) ise lojistik regresyon, Mantel-Haenszel ve standartlaştırma yöntemlerini gözlenen puan teknikleri; madde tepki kuramı olabilirlik oranı testleri, Lord'un Ki-Kare'si, ve SIBTEST tekniklerini gizil puan teknikleri olarak ele almışlardır.

Mantel-Haenszel (MH) tekniği, Ki-Kare istatistiğine dayanan en yaygın kullanılan yöntemlerdendir. Olasılık oranlarına dayanan bu teknik tek biçimli ve tek biçimli olmayan DMF'yi ayırt edememektedir. Bu teknik maddelerin olasılık oranlarına dayanır. Eşleştirme kriteri olarak toplam puanlar süreksiz veri olarak kullanılır. Bunun yanında, MH test istatistiği olumsallık çizelgelerine dayanır ve 2x2'lik bir olumsallık çizelgesinde aşağıdaki gibi hesaplanır (Osterlind, 1983).

$$MH = \Omega \frac{N_{11}N_{22}}{N_{21}N_{12}}$$

Bu denkleme göre elde edilen MH istatistiği eğer 1'den büyükse referans grup lehine; eğer 1'den küçükse odak grup lehine DMF var; eğer 1'e eşit veya yakın bir değer ise DMF olmadığı anlamına gelir. Bu denkleme logaritmik dönüşüm yapıldığında,

$$\Delta_{MH} = -2{,}35.\ln(MH)$$

denklemi elde edilir. Bu denkleme göre elde edilen Δ_{MH} istatistiği eğer 0'dan büyükse odak grup lehine; eğer 0'dan küçükse referans grup lehine DMF var; eğer 0'a eşit veya yakın bir değer ise DMF olmadığı anlamına gelir. DMF'nin düzeyini için, eğer $|\Delta_{MH}|$ değeri 1'den küçük ise A düzeyinde (ihmal edilebilir), eğer 1 ile 1,5 arasında ise B düzeyinde (orta), eğer 1,5'e eşit ya da büyükse C düzeyinde (önemli) yorumu yapılır (Dorans ve Holland, 1993).

Lojistik Regresyon, aynı yetenek düzeyindeki bireylerin farklı grup üyeliklerinde doğru cevap verme olasılıklarının istatistiksel olarak karşılaştırılmasını temel alır. Oluşturulacak modele öncelikle toplam puanlar katılır. Daha sonra grup değişkeni modele atılır. Son olarak da etkileşim

terimi modele eklenir (Zumbo, 1999). Eşleştirme kriteri olarak toplam puanlar sürekli veri olarak kullanılır. Grupların modele eklenmesiyle elde edilen Ki-Kare değerinin manidarlığı tek biçimli DMF'yi; etkileşimin olduğu modelin Ki-Kare değerlerinin manidarlığı ise tek biçimli olmayan DMF'ye ilişkin kanıt oluşturur (Doğan ve Öğretmen, 2008). Lojistik regresyonla elde edilen DMF'nin önemliliğini belirlemek için ikinci ve üçüncü adımlarda elde edilen R^2 ve ΔR^2 değerlerinin büyüklüklerine bakılır. Zumbo ve Thomas (1997) ile Gierl ve Khaliq (2001) belirlediği ölçütler Tablo 1'de verilmiştir.

Tablo 1. *Lojistik regresyon için R^2 ölçütleri*

Gierl ve Khaliq	Zumbo ve Thomas	DMF düzeyi	Yorum
$R^2<0,035$	$R^2<0,13$	A	Yok
$0,035<R^2<0,070$	$0,13<R^2<0,26$	B	Orta düzeyde
$R^2\geq0,070$	$R^2\geq0,26$	C	Yüksek düzeyde

Olabilirlik oranı (MTK-OO) yönteminde madde tepki kuramına dayanan yöntemlere göre analizler yapılır. Bu yöntemde odak ve referans grup madde parametreleri arasında fark olup olmadığı hipotezi test edilir. Buna göre sınırlandırılmış ve genelleştirilmiş modeller oluşturularak birbirine oranları test edilir (Atalay, Gök, Kelecioğlu ve Arslan, 2012). Bu olabilirlik oranın logaritması alınarak G^2 değeri elde edilir ve serbestlik derecesi (MTK parametre sayısı) ile Ki-Kare tablosundan kontrol edilir. Bu değerin anlamlı olması DMF olduğu anlamına gelir (Thissen, 2001). G^2 değerleri DMF'nin büyüklüğü hakkında bilgi vermektedir. Thissen (2001) tarafından önerilen Tablo 2'deki ölçütlere bakılarak bu konuda yorum yapılabilmektedir.

Tablo 2. *G^2değerlerinin yorumlanması*

Düzey	Kriterler	Yorum
A	$3,84<\mathbf{G^2}< 9,4$	Yok veya ihmal edilebilir düzeyde
B	$9,4\leq \mathbf{G^2}< 41,9$	Orta düzeyde
C	$\mathbf{G^2}\geq 41,9$	Yüksek düzeyde

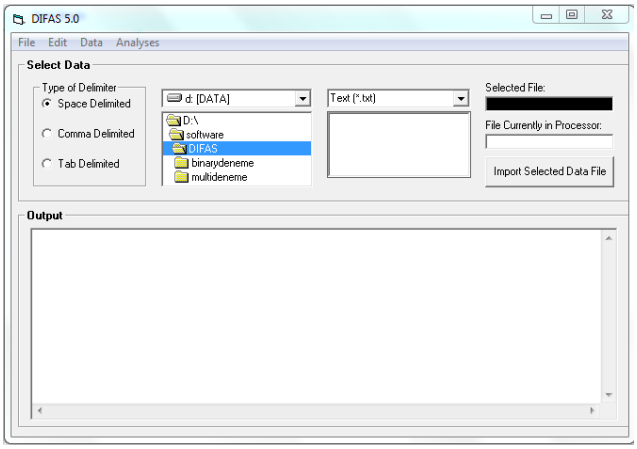
SIBTEST ise DMF analizi yapılacak maddeler ile diğer maddeler karşılaştırılır. SIBTEST'te β değeri, etki büyüklüğü olarak yorumlanır ve bu değer DMF'nin düzeyini ifade eder. $\beta<|0,059|$ ise, maddede ihmal edilebilir düzeyde (A düzeyi), $|0,059|\leq\beta<|0,088|$ ise orta düzeyde (B düzeyi), $\beta\geq|0,088|$ ise önemli düzeyde (C düzeyi) DMF olduğu kabul edilmektedir (Rousses ve Stout, 1996). Değişen madde fonksiyonu belirlemede kullanılan programlar DMF'nin KTK ve MTK'ya dayalı olmasına ve DMF'nin tek biçimli ya da tek biçimli olmamasına göre farklılık göstermektedir. Tablo 3'te analizlerde kullanılan belli başlı programlar verilmiştir.

Tablo 3. DMF yazılımları

Yazılımlar	Kullanılan Yöntemler
DIFAS	MH yöntemi
EZDIF	LR ve MH yöntemi
JMETRIK	MH yöntemi
IRTLRDIF	MTK Olabilirlik yöntemi
GMHDIF	MH yöntemi
ConQuest	MTK Olabilirlik yöntemi
difR Paketi	MH, LR ve MTK olabilirlik yöntemleri
Zumbo SPSS Syntax	LR yöntemi
BILOG-MG	MTK yöntemi
SIBTEST	Uniform ve Nonuniform DMF
lordif paketi	MTK'ya dayalı LR yöntemi
EASY-DIF	MH yöntemi
IRTPRO	MTK yöntemi
MULTILOG	MTK yöntemi

Bu yazılımlardan DIFAS, JMETRIK, EZDIF, Zumbo SPSS Syntax, IRTPRO ve IRTLRDIF bu çalışmada kullanılmış ve yazılımlar ile ilgili özet bilgiler yöntem bölümünde verilmiştir. MH yöntemi ile DIFAS ve JMETRIK, MH ve LR yöntemleri ile EZDIF, LR yöntemi ile Zumbo SPSS Syntax, MTK olabilirlik yöntemi ile IRTPRO ve IRTLRDIF yazılımları DMF analizleri yapmaktadır.

DIFAS programı (Penfield, 2005) ile DMF, değişen test fonksiyonu (DTF) ve çok kategorili maddeler için değişen adım fonksiyonu (DAF) incelenebilir. Yazılımın arayüzü Şekil 1'de verilmiştir.

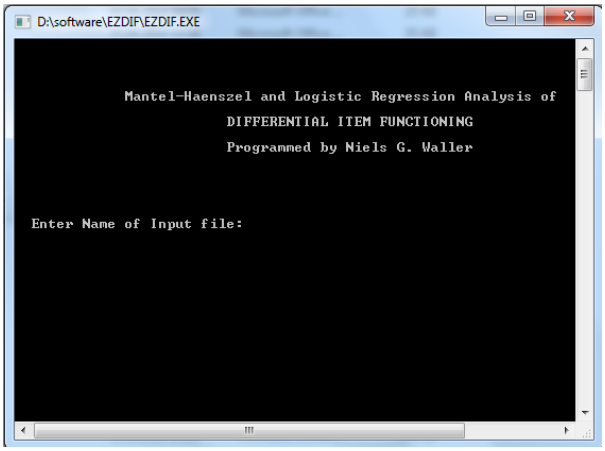


ŞEKİL 1. *DIFAS arayüzü*

DIFAS yazılımı, parametrik olmayan analizler yaparak test ve madde istatistikleri verir. Analiz sonucu elde edilen çıktıda her bir madde için MH Ki-Kare değerleri, Log-odds oranları ve bunlara ilişkin hata değerleri, log-odds oranlarının standart değerleri, Breslow ve Day tarafından elde edilen Ki-Kare değerleri ve DMF nin düzeyine ilişkin bilgiler yer almaktadır.Programın kullanma kılavuzu ve indirmesine http://erm.uncg.edu/measurement-software/ internet adresinden ulaşılabilmektedir.

JMETRIK (Meyer, 2014)**,** Java tabanlı, açık kaynak kodlu ücretsiz bir programdır. Programda analiz yapabilmek için öncelikle bir veri tabanı oluşturulur. Programda madde analizleri, DMF (MH), Rasch modelleri, MTK madde kalibrasyonu ve yetenek kestirimleri yapılabilmektedir. Programın avantajı ikili ve çok kategorileri maddeleri kolaylıkla analiz yapabilmesidir. Program, odak ve referans gruplarının madde karakteristik eğrilerini de vermektedir.

EZDIF (Waller, 1998)**,** MS DOS komutları ile çalışmaktadır. Programın arayüzü Şekil 2'de verilmiştir.



ŞEKİL 2. *EZDIF arayüzü*

EZDIF yazılımı, http://www.psych.umn.edu/faculty/waller/downloads.htm adresinden ücretsiz indirilebilir. EZDIF ile Lord'un Ki-Karesi ve anlamlılık düzeyi, iki farklı yanıt fonksiyonu için elde edilen işaretlenmiş ve işaretlenmemiş alan ölçüleri,ESA ve EUA ölçütleri için z değerleri, Raju'nun dengeleyici ve dengeleyici olmayan DMF indeksleri, Raju'nun DTF indeksi ve bu değerin Ki-Kare ve anlamlılık düzeyielde edilebilmektedir.

Zumbo SPSS Syntax, Bruno Zumbo tarafından yazılmıştır. Eşleştirme kriteri olarak toplam test puanı sürekli değişken olarak ele alınır. Tek biçimli ve tek biçimli olmayan DMF lojistik regresyon yöntemi ile incelenir. Çok kategorili veri için SPSS Syntax Şekil 3'te verilmiştir.

```
include file='ologit2.inc'.
execute.


GET
FILE='multicategory.sav'.
 EXECUTE .

compute item= item2.
compute total= total.
compute grp= group.

* Regression model with the conditioning variable, total score, in alone.
* Step #1 .
ologit var = item total
    /output=all.
execute.

* Regression model adding uniform DIF to model.
* Step #2.
ologit var = item total grp
    /contrast grp=indicator
    /output=all.
execute.


* Regression model adding non-uniform DIF to the model.
* Step #3.
ologit var = item total grp total*grp
    /contrast grp=indicator
    /output=all.
execute.
```

ŞEKİL 3. *Çok kategorili veri için SPSS Syntax*

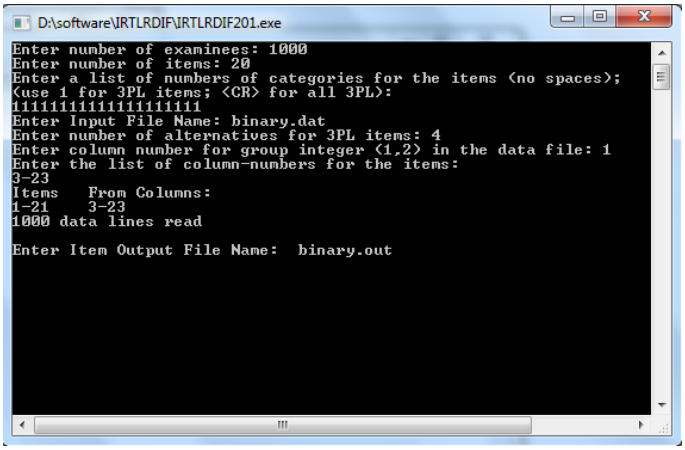
Zumbo SPSS Syntax ile elde edilen sonuçların yorumlanmasında, iki kategorili veriler için Nagelkarke; çok kategorili veriler için McKelvey & Zavoina R^2 değerleri kullanılır. Analizlerde üç aşamalı bir yol izlenir:

1. İlk aşamada ölçüt değişken olarak toplam puanlarla model kurulur.
2. İkinci aşamada grup değişkeni modele eklenir.
3. Üçüncü aşamada regresyon denklemine etkileşim terimi eklenir.

Model uyumu için grup değişkenin olduğu ve olmadığı modellerin R^2 değerleri karşılaştırılır. Bu değerlerin farkı ile elde edilen ΔR^2 değeri ile DMF'nin miktarı ve düzeyi hakkında bilgi verir. ΔR^2değerleri ile ilgili tablo giriş bölümünde verilmiştir.

IRTPRO (Cai, Du Toit ve Thissen, 2011), *t*ek boyutlu ve çoklu grup analizleri, çok boyutlu analiz ve MTK yetenek kestirimleri yapabilmektedir. IRTPRO programı, 1 PL, 2 PL, 3 PL, Graded Response, Genelleştirilmiş Kısmi Kredi, Nominal madde tepki modelleri ile analiz yapar. Madde parametrelerinin kestiriminde Maximum Likelihood (ML) veya Maximum a Posteriori (MAP) kullanır. Ayrıca maddelere verilen yanıtlara ilişkin grup bazında veya birlikte madde karakteristik eğrileri alınabilmektedir.

IRTLRDIF (Thissen, 2001), madde tepki kuramının parametrelerine ilişkin hipotezlerin doğrudan incelenmesi için testler içerdiğinden farklılaşan güçlükler, eğimler ve şans faktörlerinden kaynaklanan DMF'yi belirleyebilir. Programın arayüzü Şekil 4'te verilmiştir.



ŞEKİL 4. *IRTLRDIF arayüzü*

Bu program, model uyumu sağlandığında büyük örneklemler için en iyi sonucu verir. CAT uygulamalarından elde edilen verilerde DIF bulmada kolay bir geçiş sağlamaktadır. MULTILOG gibi programlarda birçok analizin yapılması ve sonuçların karşılaştırmasını gerektirmeyecek kullanımı kolay bir programdır. Çıktıda elde edilen G^2 değeri Ki-Kare istatistiğine karşılık gelir ve Tablo 2'de verilen G^2 tablosundaki aralıklara göre DMF'nin düzeyi belirlenir. IRTLRDIF, komutlarla çalışan bir yazılımdır. Bu yazılım, http://www.unc.edu/~dthissen/dl.html adresinden indirilebilir.

Bu çalışmada, alan yazında değişen madde fonksiyonuyla ilgili yapılan çalışmalar, madde yanlılıklarının incelendiği ile yöntemlerin karşılaştırıldığı çalışmalar şeklinde iki kategoride ele alınmıştır. Madde yanlılıklarının incelendiği çalışmalarda, DMF yöntemlerinin ya da kullanılan programların etkililiğine veya karşılaştırılmasına dönük bulgular elde etmek ya da çıkarımlarda bulunmak yerine incelenen değişkenlerin farklı gruplara yanlılık gösterip göstermediğine ilişkin kanıtlar elde etmek amaçlanmıştır. Alan yazında, madde yanlılıklarının incelendiği bazı araştırmalarda;

- Çoktan seçmeli testlerde iki kategorili ve önsel ağırlıklı puanlama yöntemleri kullanmanın DMF'ye etkisinin incelenmesi (Özdemir**,** 2003),
- Değişen madde fonksiyonu ve kaynaklarını örtük sınıf modeli yaklaşımı ile incelenmesi (Oliveri, Ercikan ve Zumbo, 2013),
- Madde tepki kuramına dayalı bir DMF belirleme tekniği ile boyut sayısı belli olmayan çok boyutlu iki kategorili maddeler için değişen madde fonksiyonun belirlenmesi (Snow ve Oshima, 2009),
- 2005 yılı ÖSS sınavında yer alan maddelerin cinsiyete göre DMF içerip içermediğinin incelenmesi (Bakan-Kalaycıoğlu ve Kelecioğlu, 2011),
- 2009 yılı SBS sınavında yer alan maddelerin cinsiyete göre yanlılık gösterip göstermediğinin belirlenmesi (Karakaya, 2012),
- SBS sınavının Türkçe alt testinde yer alan maddelerin cinsiyet ve okul türü değişkenine göre yanlılık gösterip göstermediği MH ve LR teknikleri ile incelenmesi (Karakaya ve Kutlu, 2012),
- Kanada, Çin (Şangay), Finlandiya ve Türkiye olmak üzere dört farklı ülkede matematik alanında cinsiyete göre DMF gösteren maddelerin belirlenmesi (Lyons Thomas, Sandilands ve Ercikan, 2014),
- 2009 yılında yapılan Uluslararası Öğrenci Değerlendirme Programı (PISA) öğrenci anketi tutum maddelerinin farklı kültürlere göre DMF gösterip göstermediğinin Poly-SIBTEST, ordinal lojistik regresyon ve MTK olabilirlik oranı ile incelenmesi (Gök, Atalay Kabasakal ve Kelecioğlu, 2014),
- PISA 2006 öğrenci anketinde yer alan bazı maddelerin cinsiyete ve kültüre göre Değişen Madde Fonksiyonu açısından Ordinal Lojistik Regresyon (OLR) ve Poly-SIBTEST yöntemleri ile incelenmesi (Atalay Kabasakal ve Kelecioğlu, 2012),

- PISA 2009 anketinde okula ve öğretmenlere ilişkin algıyı ölçen toplam 9 maddenin DMF göstermesi bakımından cinsiyete ve Türkiye, Amerika, İrlanda ve İngiltere olmak üzere ülkelere göre incelenmesi (Köse, 2015) amaçlanmıştır.

Yöntemsel karşılaştırmışların yapıldığı çalışmalarda ise, DMF yöntemlerinin ya da kullanılan programların farklı koşullardaki etkililiğine veya karşılaştırılmasına dönük bulgular elde etmek ya da çıkarımlarda bulunmak amaçlanmıştır. Örneğin, Doğan ve Öğretmen (2008) tarafından yapılan çalışmada değişen madde fonksiyonu belirlemede Mantel-Haenszel, Ki-Kare ve Lojistik Regresyon tekniklerinin karşılaştırılması amaçlanmıştır. Her üç teknik ile hesaplanan Ki-Kare değerleri arasında sıra farkları korelasyon katsayıları incelendiğinde istatistiksel olarak anlamlı ve 0,79 ile 0,93 arasında yüksek düzeyde ilişki olduğu belirlenmiştir.

Yıldırım (2008) tarafından 2003 yılı PISA verisi ve bu verinin kontrol edilmesine ve açıklanmasına yönelik simülasyon verisi ile yapılan çalışmada Sınırlandırılmış Faktör Çözümlemesi (SFÇ) yönteminin, Mantel-Haenszel (MH) ve Olabilirlik Oranı Analizi (OOA) ile karşılaştırılması amaçlanmıştır. Gerçek veriyle yapılan analizlerde, MH ile OOA yöntemleri arasındaki uyumun %82, MH ile SFÇ arasındaki uyumun %72, OOA ile SFÇ arasındaki uyumun %64 olduğu belirlenmiştir. Simülasyon çalışmasında ise SFÇ yönteminin karşılaştırılan grup ortalamaları eşit ya da farklı olduğu durumlarda MH ve OOA yöntemlerine göre DMF'li maddeleri daha doğru olarak tespit etmiştir.

Elosua ve Wells (2013) tarafından yapılan simülasyon çalışmasında çok kategorili maddelerde DMF belirleme yöntemlerinden model tabanlı yöntemler olan ortalama ve kovaryans yapısı modeli ile madde tepki kuramı ve gözlenen puan modellerinden lojistik regresyon yönteminin 1. Tip hata ve güçlerini karşılaştırmak amaçlanmıştır. 1. Tip hata oranı bazı maddeler DIF içerdiğinde MTK'ya dayalı testler ve sıralı lojistik regresyon için yüksek çıkmıştır. Küçük örneklem büyüklüğünde, ortalama ve kovaryans yapısı modeli ile madde tepki kuramı benzer güç gösterirken, sıralı lojistik regresyon için bu güç biraz daha yüksek çıkmıştır. Tek biçimli olmayan DMF için MTK modeli, diğer iki modele göre çok daha fazla güç düzeyi göstermiştir.

Padila, Hidalgo, Benitez ve Benito (2012), Maentel Haenszel yöntemine göre EASY-DIF, DIFAS ve EZDIF olmak üzere üç ayrı yazılımda simulasyon ve gerçek veri ile elde edilen sonuçları aynı veri üzerinde karşılaştırmışlardır. DIFAS ve EASY-DIF zayıf karşılaştırma stratejisi durumlarında daima denk sonuçlar verirken EZDIF daha az doğru sonuçlar vermektedir. DIFAS ve EASY-DIF, özellikle yeni uygulama yapanlar için, çalıştırılması en kolay ve DMF belirlemek için anahtar özellikler için daha geniş bir sonuç ranjı sunmaktadır.

Gök, Kelecioğlu ve Doğan (2010) tarafından gerçek bir veri ile yapılan çalışmada Değişen Madde Fonksiyonu belirlemede Mantel-Haenszel ve Lojistik Regresyon tekniklerinin karşılaştırılması amaçlanmıştır. MH ve LR teknikleriyle elde edilen χ^2 değerleri arasında ilişki olup olmadığı belirlemek için Spearman'ın sıra farkları korelasyon katsayısı kullanılmış ve teknikler arasında istatistiksel olarak anlamlı bir ilişki olmadığı belirlenmiştir.

Atar ve Kamata (2011) tarafından yapılan Monte Carlo simülasyon çalışmasında çok kategorili puanlanan maddeler için MTK olabilirlik oranı testi ve kümülatif lojit ordinal lojistik regresyon yöntemlerinin değişen madde fonksiyonunu belirlemede ortaya çıkan I. Tip hata oranları ve güçleri incelenmiştir. Elde edilen bulgulara göre hem MTK olabilirlik oranı testi hem de kümülatif lojit ordinal regresyon yöntemlerinde 54 farklı simülasyon durumunda I. Tip hataların 0,05 değerinin altında olduğu belirlenmiştir. Çalışmada MTK olabilirlik oranı testinde hem örneklem büyüklüğü hem de DMF büyüklüğü arttıkça testin gücünün arttığı belirlenmiştir. Bunun yanında ordinal lojistik regresyon yönteminin gücü büyük örneklem ve büyük DMF koşulunun dışındaki diğer tüm durumlarda kabul edilemez düzeyde düşük çıkmıştır.

Acar (2011) tarafından yapılan çalışmada farklı örneklem büyüklüklerinde DMF belirleme yöntemlerinden Genelleştirilmiş Aşamalı Doğrusal Modelleme (GADM) ile belirlenen DMF'li madde sayısının belirlenmesidir. Araştırma bulgularına göre örneklem büyüklüğü arttıkça DMF'li maddelerin sayısının da arttığı belirlenmiştir.

Atalay vd. (2012) tarafından yapılan simülasyon çalışmasında gözlenen puan yöntemlerinden MH ve LR ile örtük puan yöntemlerinden MTK-OO ve SIBTEST yöntemlerinin karşılaştırılması amaçlanmıştır. Elde edilen bulgulara göre örtük özelliğe dayalı MTK-OO ve SIBTEST yöntemlerinin gözlenen puana dayalı MH ve LR yöntemlerinden daha duyarlı ve etkili

olduğu belirlenmiştir. Gözlenen puanlara dayalı yöntemlerden DMF gösteren maddeleri belirleme açısından MH tekniğinin tek biçimli DMF'yi belirlemede, LR tekniğinin ise tek biçimli olmayan DMF'yi belirlemede daha etkili olduğu belirlenmiştir. Örtük puana dayalı yöntemlerden ise her iki yöntemin tek biçimli olan ve tek biçimli olmayan DMF'yi belirleme açısından eşit güçte oldukları belirlenmiştir.

Sonuç olarak, yapılan araştırmalar farklı DMF belirleme yöntemlerinin farklı sonuçlar ortaya çıkardığını göstermektedir (Gao ve Wang, 2005). Bu durum, yöntemsel araştırmalarda görüldüğü gibi, madde yanlılıklarının incelendiği bazı araştırmalarda (Atalay Kabasakal ve Kelecioğlu, 2012; Gök, Atalay Kabasakal ve Kelecioğlu, 2014; Karakaya ve Kutlu, 2012) da görülebilmektedir. Bu sebeple ilgili alan yazında birden fazla DMF belirleme yönteminin bir arada kullanılması önerilmektedir (Fidalgo, Ferreres ve Muniz, 2004). Bu çalışmanın amacı, gerçek bir veri üzerinde KTK ve MTK yöntemleri ile yapılan DMF analizlerinin sonuçlarını karşılaştırmaktır. Çalışmada kullanılan veri PISA (2012) çalışmasından alınmıştır. PISA uygulaması, üç yılda bir dünya genelindeki eğitim sistemlerini değerlendirmek amacıyla 15 yaş grubunda öğrencilerin matematik, fen ve okuma alanlarındaki bilgi ve becerilerinin test edildiği bir araştırmadır (OECD, 2014). Her uygulamada üç alandan da testler yer almasına rağmen, uygulamanın odağında belirli bir alan yer alır. 2012 yılında yapılan uygulamada matematik alanına ağırlık verilmiş ve bu alanla ilgili kapsamlı veri elde edilmiştir. Bu çalışmada, 2012 yılındaki uygulamada matematikle ilgili öğrenci anketinde yer alan matematik çalışma disiplini ile ilgili maddelerin matematik performansı açısından alt ve üst gruplarda yer alan bireylere DMF gösterip göstermediği incelenmiştir. Bu doğrultuda, DIFAS, JMETRIK, EZDIF, Zumbo SPSS Syntax, IRTPRO ve IRTLRDIF yazılımları kullanılarak MH, LR ve MTK olabilirlik yöntemleri ile elde edilen sonuçlar karşılaştırılmıştır.

YÖNTEM

Bu araştırmada, gerçek bir veri üzerinde değişen madde fonksiyonu yöntemleri için farklı yazılımlardan elde edilen sonuçlar karşılaştırıldığından temel bir araştırmadır. Temel araştırmalarda var olan bilgilere yeni bilgiler katmak amaçlanır (Karasar, 2005). Araştırmada, var olan bir durum ortaya çıkarılarak betimlediğinden tarama araştırmalarının özelliklerine sahiptir. Tarama araştırmalarında, bireylerin ya da durumların özelliklerinin var olduğu gibi betimlenmesini amaçlanır (Büyüköztürk, Kılıç Çakmak, Akgün, Karadeniz ve Demirel, 2013; Fraenkel ve Wallen, 2011). Araştırmanın evreni, PISA uygulayıcıları tarafından Türkiye'den tabakalı örnekleme yöntemiyle seçilip PISA (2012) uygulamasına katılan 15 yaş grubundaki öğrencilerdir. Araştırmanın çalışma gurubu ise çalışma disiplini değişkeninin gösterge maddelerine ait kayıp verisi olmayan öğrenciler arasından rastgele örnekleme yöntemiyle seçilen 990 öğrencidir. PISA (2012) öğrenci anketinde çalışma disiplini değişkenine ait 9 madde yer almaktadır. Araştırmanın faktör değişkeni olan matematik performansının ortalaması alınarak alt ve üst gruplar oluşturulmuştur. Buna göre, öğrencilerin %45'i alt grupta, %55'i ise üst grupta yer almıştır. Bireylerin maddelere verdiği yanıtlar 4'lü Likert tipinde puanlanmıştır. Matematik okuryazarlığı puanları bakımından başarılı ve başarısız gruplar belirlendikten sonra matematik çalışma disiplini maddelerinin başarı değişkenine göre DMF içerip içermediği belirlenmiştir. Ölçeğin Cronbach Alpha güvenirlik katsayısı 0,90 olarak bulunmuştur. Ölçek maddelerine ilişkin betimsel istatistikler Tablo 4'te verilmiştir.

Tablo 4. Maddelere yönelik betimsel istatistikler

	N	Min,	Maks,	Ort,	Ss	Çarpıklık		Basıklık	
							Std, Hata		Std, Hata
1,Ödevi zamanında tamamlama	990	1	4	2,05	0,842	0,505	0,078	-0,297	0,155
2, Ödeve sıkı çalışma	990	1	4	2,36	0,851	-0,079	0,078	-0,736	0,155
3, Sınavlara çalışma	990	1	4	1,72	0,742	0,915	0,078	0,708	0,155
4, Quizlere sıkı çalışma	990	1	4	2,18	0,834	0,228	0,078	-0,596	0,155
5, Konuya anlayana kadar çalışma	990	1	4	2,07	0,835	0,341	0,078	-0,568	0,155
6, Derslere önem verme	990	1	4	2,05	0,810	0,390	0,078	-0,395	0,155
7, Dersleri dinleme	990	1	4	1,75	0,702	0,804	0,078	0,865	0,155
8, Çalışırken rahat edici şeylerden kaçınma	990	1	4	2,02	0,800	0,421	0,078	-0,341	0,155
9, Sistemli çalışma	990	1	4	2,22	0,848	0,213	0,078	-0,619	0,155
Başarı Durumu	990	0,00	1,00	0,55	0,498	-0,207	0,078	-1,961	0,155

Tablo 4 incelendiğinde, değişkenlere ait kayıp veri bulunmamaktadır. Değişkenlerin çarpıklık ve basıklık değerleri, dağılımların normalden önemli sapmalar göstermediğine işaret etmektedir. Araştırmada, veri analiz yöntemleri olarak Mantel-Haenszel, lojistik regresyon ve madde tepki kuramı olabilirlik oranı yöntemleri kullanılmıştır. Analizlerde JMETRIK, DIFAS, Zumbo SPSS Syntax, IRTLRDIF ve IRTPRO yazılımlarından yararlanılmıştır. Yöntemlerin çalışma prensiplerinin neler olduğuna, analiz sonucunda elde edilen çıktıların nasıl yorumlanacağı, kullanılan yazılımlarla ilgili bilgilere bu çalışmanın giriş kısmında yer verilmiştir.

BULGULAR

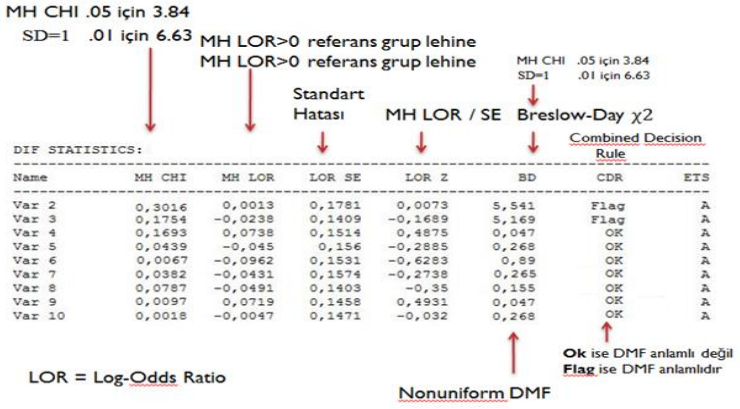
Çalışmada, Mantel-Haenszel yöntemi için DIFAS, JMETRIK ve EZDIF yazılımları; Lojistik regresyon yöntemi için Zumbo SPSS Syntax; Olabilirlik oranı yöntemi için IRTPRO ve IRTLRDIF yazılımları kullanılmıştır. Her bir yöntem için farklı yazılımlardan elde edilen bulgular aşağıda sırasıyla açıklanmıştır.

Mantel-Haenszel Yöntemi ile Elde Edilen Sonuçlar

Bu yöntem ile DIFAS ve JMETRIK yazılımlarında analizler yapılmıştır.

DIFAS sonuçları

Mantel-Haenszel yöntemi için DIFAS 5.0 programı kullanılmıştır. Analiz sonucuna ilişkin çıktılar Şekil 5'te verilmiştir.



ŞEKİL 5. *DIFAS analiz sonuçları*

Ki-Kare (MH CHI) değeri için 0,05 anlamlılık düzeyinde Ki-Karenin beklenen değeri 3,84 olarak kabul edilmekte ve bu değeri aşan değerler anlamlı bulunmaktadır. Analiz sonucunda 'CDR'

sütununda yer alan 'OK' ifadesi DMF'nin anlamlı olmadığını, 'Flag' ifadesi ise DMF'nin anlamlı olduğunu (χ^2>3,84) göstermektedir. Son sütunda yer alan ETS ise DMF'nin ne düzeyde olduğunu göstermektedir. Buna göre DMF'li olarak belirlenen 1. ve 2. maddelerin ihmal edilebilir düzeyde DMF'ye sahip olduğu görülmektedir.

JMETRIK sonuçları

JMETRIK programı yardımıyla gerçekleşen analiz sonucunda elde edilen bulgular Şekil 6'da gösterilmiştir.

DIF ANALYSIS
disidisi.DSDS
Nisan 26, 2015 17:17:34

Item	Chi-square	p-value	Valid N	E.S. (95% C.I.)	Class
var2	23,04	0,00	925	-0,18 (-0,25, -0,11)	BB-
var3	11,01	0,00	923	-0,11 (-0,18, -0,05)	AA
var4	25,10	0,00	860	0,15 (0,09, 0,21)	BB+
var5	5,30	0,02	927	-0,08 (-0,15, -0,01)	AA
var6	6,67	0,01	927	0,10 (0,03, 0,17)	AA
var7	9,67	0,00	923	0,10 (0,03, 0,16)	AA
var8	2,06	0,15	927	0,04 (-0,02, 0,10)	AA
var9	2,09	0,15	929	0,05 (-0,02, 0,12)	AA
var10	2,95	0,09	927	-0,06 (-0,14, 0,02)	AA

ŞEKİL 6. *JMETRIK analiz sonuçları*

Şekil 6'da görüldüğü üzere 9 maddenin tamamı için Ki-Kare değerleri, anlamlılık değerleri, güven aralıkları ve DMF düzeyine ilişkin bilgiler yer almaktadır. Elde edilen sonuçlara göre sadece 1. (Var2) ve 3. (Var4) maddede orta düzeyde DMF olduğu diğer maddelerde ise ihmal edilebilir düzeyde DMF olduğu belirlenmiştir.

Lojistik Regresyon Yöntemi ile Elde Edilen Sonuçlar

Lojistik regresyon analizi için Zumbo tarafından geliştirilen SPSS Syntax kullanılmıştır. Buna göre, çalışma disiplini maddelerinin başarı durumuna göre DMF durumları Tablo 5'te verilmiştir.

Tablo 5. Lojistik regresyon DMF değerleri

Maddeler	ΔX^2	p	p (etkileşim)	ΔR^2	DMF durumu	DMF çeşidi	DMF düzeyi
m1	20,744	*0,000*	*0,000*	-0,02	Var	TBO	A
m2	10,3	*0,001*	*0,001*	0,03	Var	TBO	A
m3	28,545	*0,000*	*0,000*	0,03	Var	TBO	A
m4	5,229	0,022	0,017	0,07	Yok	-	-
m5	7,968	*0,005*	*0,004*	0	Var	TBO	A
m6	10,46	*0,001*	*0,001*	0,1	Var	TBO	C
m7	2,057	0,156	0,133	0,01	Yok	-	-
m8	2,353	0,125	0,075	0,04	Yok	-	-
m9	3,012	0,083	0,048	0,14	Yok	-	-

Not: ΔX^2: Lojistik regresyon Ki-Kare değerleri, p: ΔX^2 değerleri olasılıkları, p (etkileşim): Toplam puan x grup etkileşim değerleri olasılıkları, ΔR^2: Etki büyüklüğü değerleri, TBO: Tek biçimli olmayan DMF.

Değerlendirme ölçütü olarak, p değerlerinin 0,01'den küçük olması Ki-Kare değerlerinin anlamlı olduğu ve bu maddelerin DMF gösterdiği anlamına gelir. Buna göre, 1, 2, 3, 5 ve 6. maddelerde DMF olduğu belirlenmiştir. Tek biçimli olduğu belirlenen bu maddelerden 6. maddenin DMF düzeyi yüksek iken diğer maddelerin DMF düzeyi ihmal edilebilecek düzeydedir. Buna göre, lojistik regresyon analizi sonucunda sadece 6. maddede önemli düzeyde DMF vardır.

Olabilirlik Oranı Yöntemi ile Elde Edilen Sonuçlar

Olabilirlik oranı yöntemi için IRTPRO ve IRTLRDIF yazılımları kullanılmıştır.

IRTLRDIF sonuçları

Bu programın çıktısındaki tablonun ilk sütunundaki G^2 değerleri DMF'nin büyüklüğü hakkında bilgi vermektedir. Thissen (2001) tarafından önerilen Tablo 2'deki kriterlere bakılarak bu konuda yorum yapılabilmektedir. Değerlendirme kriteri olarak, p değerlerinin 0,01'den küçük olması Ki-Kare değerlerinin anlamlı olduğu ve bu maddelerin DMF gösterdiği kabul edilmiştir. DMF'nin düzeyi için Tablo 2'deki kriterler göz önünde bulundurulmuştur. Tablo 6'da IRTLRDIF ile elde edilen analiz sonuçları verilmiştir.

Tablo 6. *IRTLRDIF analiz sonuçları*

Maddeler	ΔG^2	p	DMF durumu	DMF düzeyi
m1	20,7	*0,000*	Var	B
m2	9,7	0,046	Yok	-
m3	18,6	*0,000*	Var	B
m4	5,1	0,277	Yok	-
m5	5,1	0,277	Yok	-
m6	10,9	0,028	Yok	-
m7	1,9	0,754	Yok	-
m8	2,1	0,717	Yok	-
m9	5,0	0.287	Yok	-

Tablo 6'ya göre, 1 ve 3. Maddelerde B düzeyinde (orta) DMF olduğu belirlenmiştir.

IRTPRO sonuçları

Çok kategorili puanlanan 9 maddelik matematik çalışma disiplinine ilişkin verilerin analizi için öncelikle veri dosyası IRTPRO programı için .txt formatına dönüştürülmüş ve sonrasında hangi modelin kullanılması gerektiğini belirlemek amacıyla -2LL değerleri karşılaştırılmıştır. Analiz sonuçları Şekil 7'de verilmiştir.

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on the loglikelihood	
-2loglikelihood:	23321.64
Akaike Information Criterion (AIC):	23481.64
Bayesian Information Criterion (BIC):	23874.18

Likelihood-based Values and Goodness of Fit Statistics

Statistics based on the loglikelihood	
-2loglikelihood:	23190.08
Akaike Information Criterion (AIC):	23430.08
Bayesian Information Criterion (BIC):	24018.89

ŞEKİL 7. *IRTPRO analizi model sonuçları*

Analiz sonucunda -2LL değerleri arasında farklılık bulunmaktadır. Ancak ne aşamalı tepki modeli ne de sınıflamalı tepki modeli için elde edilen değerlerin 1-2-3PL modellerde olduğu gibi karşılaştırılması mümkün olmadığı için ilgili alan yazında en çok önerilen yöntemlerden biri olan sınıflamalı tepki modelinin kullanılmasına karar verilmiştir (Embretson ve Reise, 2000). Daha sonra analiz için gerekli işaretlemeleri tamamladıktan sonra belirlenen serbestlik derecesinde her bir maddenin odak ve referans grubu lehine DMF içerip içermediği belirlemek amacıyla p değerleri incelenmiştir. Çalışmada ilk olarak aşamalı tepki modeline ilişkin analiz sonuçları Şekil 8'de verilmiştir.

DIF Statistics for Graded Items (Back to TOC)

Item numbers in:

Group 1	Group 2	Total X^2	d.f.	p	X^2_a	d.f.	p	X^2_{cla}	d.f.	p
1	1	10.0	4	0.0401	0.0	1	0.9459	10.0	3	0.0184
2	2	4.3	4	0.3639	0.0	1	0.9940	4.3	3	0.2289
3	3	8.7	4	0.0686	0.4	1	0.5531	8.4	3	0.0391
4	4	3.1	4	0.5461	0.7	1	0.3972	2.4	3	0.5025
5	5	3.3	4	0.5128	0.4	1	0.5219	2.9	3	0.4130
6	6	3.4	4	0.4956	0.8	1	0.3635	2.6	3	0.4649
7	7	0.8	4	0.9357	0.0	1	0.9103	0.8	3	0.8477
8	8	2.4	4	0.6623	1.2	1	0.2675	1.2	3	0.7603
9	9	5.6	4	0.2344	3.5	1	0.0620	2.1	3	0.5532

ŞEKİL 8. *IRTPRO aşamalı tepki modeli analiz sonuçları*

Şekil 8 incelendiğinde sadece 1. maddenin DMF gösterdiği diğer maddelerde ise DMF olmadığı belirlenmiştir. Bilindiği üzere IRTPRO ile sadece DMF olup olmadığı belirlenebildiğinden DMF'nin düzeyi hakkında bilgi almak amacıyla ilerleyen bölümlerde IRTLRDIF programından yararlanılacaktır. Bunun yanında çalışma disiplini sorularının başarılı ve başarısız olarak belirlenen öğrencilerin verdiği yanıtlar bakımından DMF gösterip göstermediğini belirlemek amacıyla sınıflamalı tepki modeli ile analizler yapılmış ve sonuçlar Şekil 9'da verilmiştir.

DIF Statistics for Nominal Items (Back to TOC)

Item numbers in:

Group 1	Group 2	Total X^2	d.f.	p	X^2_s	d.f.	p	$X^2_{a\|s}$	d.f.	p
1	1	18.6	6	0.0048	7.7	2	0.0212	0.0	1	0.8547
2	2	16.2	6	0.0126	8.4	2	0.0153	0.1	1	0.7546
3	3	5.7	6	0.4574	3.8	2	0.1483	1.0	1	0.3161
4	4	10.4	6	0.1099	4.4	2	0.1119	0.5	1	0.4679
5	5	8.7	6	0.1879	4.0	2	0.1338	1.1	1	0.2901
6	6	15.5	6	0.0167	8.7	2	0.0126	4.2	1	0.0412
7	7	6.7	6	0.3465	4.1	2	0.1260	0.7	1	0.4140
8	8	5.8	6	0.4441	2.1	2	0.3525	2.6	1	0.1108
9	9	7.2	6	0.3004	2.0	2	0.3669	3.5	1	0.0625

Şekil 9. *IRTPRO sınıflamalı tepki modeli analiz sonuçları*

Şekil 9 incelendiğinde 1, 2 ve 6. maddelerin DMF gösterdiği diğer maddelerde ise DMF olmadığı belirlenmiştir.

Sonuçların Karşılaştırılması

Tüm analizler sonucunda elde edilen bulgular Tablo 7'de özetlenmiştir.

Tablo 7. *Farklı yazılımlar ve farklı yöntemlerle elde edilen tüm sonuçlar*

DIF Yöntemi	Kullanılan Yazılım	DMF İçeren Maddeler
Mantel-Haenszel	JMETRIK	**1 ve 3**
	DIFAS	1 ve 2
Lojistik regresyon	Zumbo SPSS Syntax	1, 2, 3, 5 ve **6**
Olabilirlik oranı	IRTLRDIF	**1 ve 3**
	IRTPRO	**1, 2 ve 6**

Not: Koyu renkli olmayan maddelerde DMF düşük veya ihmal edilebilir düzeydedir.

Tablo 7'ye göre, tüm yöntemlere ve yazılımlara göre 1. maddede DMF vardır. Tabloya göre her bir madde, en az iki yöntemde DMF'li çıkmıştır. Genel olarak, yöntemlerin ve yazılımların benzer sonuçlar gösterdiği görülmektedir. IRTLRDIF ve JMETRIK sonuçlarına göre 1. ve 3. maddelerde bulunan DMF ihmal edilebilir değildir. Bu iki yazılımının dışında diğer yöntem ve yazılımlarda 2. Maddenin DMF gösterdiği saptanmıştır. IRTPRO ve Zumbo SPSS Syntax yazılımlarında 6. maddede bulunan DMF oranı ihmal edilebilir değildir. Lojistik Regresyon analizi sonuçlarına göre 4. madde dışında ilk 6 maddede DMF bulunmuştur. Diğer yöntem ve yazılımlarda ise en az ikişer maddede DMF olduğu görülmüştür. Yöntemsel olarak farklı istatistiksel alt yapıya sahip olan MH, MTK-OO ve LR yöntemlerinin aynı sonucu vermesi, 1., 2. ve 3. maddelerde DMF olabileceğini doğrular niteliktedir.

TARTIŞMA ve SONUÇ

Bu araştırmada, MH, MTK-OO ve LR yöntemleri ile farklı programlarda yapılan analizlerden elde edilen sonuçlar karşılaştırılmıştır. Buna göre, bazı farklılıkların olmasının yanında önemli benzerlikler de tespit edilmiştir. Aynı teorik alt yapıya (KTK ya da MTK) dayanan yöntemler için farklı yazılımlarda yapılan analizlerde farklı sonuçlara ulaşılmıştır. Bununla birlikte farklı teorik alt yapıya sahip yöntemlerle farklı yazılımlarda aynı sonuçlar elde edilmiştir. Bu sonuçlar, ilgili alan yazında var olan çalışmalardan elde edilen sonuçlarla paralel göstermektedir. Nitekim yapılan araştırmalar farklı DMF belirleme yöntemlerinin farklı sonuçlar ortaya çıkardığını göstermektedir (Gao ve Wang, 2005). Bu sebeple, alan yazında önerildiği gibi (Fidalgo, Ferreres

ve Muniz, 2004), yanlılık belirleme çalışmalarında birden fazla DMF belirleme yönteminin bir arada kullanılması madde yanlılıklarının belirlenmesinde uzmanlara daha etkili kanıtlar sağlayacaktır.

Çalışmada farklı yöntemlere göre DMF gösteren madde sayısı 1 ile 5 arasında değişmektedir. Çalışmada MH yöntemiyle 2, LR yöntemiyle 5, MTK-OO yöntemi ile 2, IRTPRO ile 3 maddede DMF olduğu belirlenmiştir. Söz konusu yöntemlere dayalı farklı programlar yardımıyla elde edilen DMF'li madde sayıları arasında özellikle MH ve MTK-OO arasında benzerlik görülmektedir. Bunun yanında LR yöntemiyle daha fazla maddede DMF olduğu belirlenmiştir.

Bu araştırmada kullanılan Mantel-Haensel yöntemine dayalı JMETRIK ve DIFAS programları, lojistik Regresyon yöntemine dayalı Zumbo SPSS programı ve MTK-OO yöntemine dayalı IRTLRDIF ve IRTPRO programlarından elde edilen sonuçlar karşılaştırılmıştır. Çalışmada LR ile tespit edilen DMF'li madde sayısının MH ile tespit edilenden fazla olması Gomez-Benito ve Navas-Ara (2000) ile Atalay, Gök, Kelecioğlu ve Arslan (2012) tarafından yapılan çalışmanın bulgularıyla farklılık göstermektedir. Ancak ilgili çalışmalarda simülasyon datasıyla çalışmanın ortaya çıkan bu farklılığın sebebi olabileceği düşünülmektedir. Bunun yanında LR ve MTK-OO yöntemleriyle belirlenen DMF'li madde sayılarının uyumlu olması Thissen, Steinberg ve Wainer (1988) ile Yıldırım (2008) tarafından yapılan çalışmaların bulgularıyla benzerlik göstermektedir.

Çalışmada ilgili alan yazında en çok kullanılan yöntemlerden biri olan IRTLRDIF ile hem ücretsiz hem de açık kaynak kodlu bir yazılıma sahip olan JMETRIK programında DMF gösteren maddeler aynıdır. Elde edilen bu sonuca göre JMETRIK programının IRTLRDIF programına alternatif bir program olacağı düşünülmektedir. Çalışmada DMF gösteren maddelerin yanlı olup olmadığının belirlenmesi amacıyla alana uzmanlarından görüş alınması gerektiği önerilmektedir.

Elde edilen bulgulara dayalı olarak hangi yöntemin daha avantajlı olduğu sorusuna cevap verebilmek amacıyla aynı programlarla bir de benzetim verisi üreterek gerçekten DMF gösteren madde sayısı belirlenmesi ve böylece programlardan hangisinin daha iyi kestirim yaptığının belirlenmesi önerilmektedir. Bundan sonraki çalışmalarda birden fazla veri seti kullanılarak yöntemlerin I. Tip hata oranları ve güçlerinin hesaplanması önerilmektedir.

Bu araştırmada DMF belirleme yöntemlerinin ve bilgisayar programlarının karşılaştırılması amaçlandığından sınırlı sayıda kişi ile PISA 2012 öğrenci anketinde yer alan matematik çalışma disiplini maddelerinin matematik başarı durumlarına göre yanlılığına bakılmıştır. Başka bir çalışmada, Türkiye'den çalışmaya katılan tüm öğrencilere ait veriden yararlanılarak, başarılı veya başarısız öğrencilere DMF gösteren maddelerin yanlılık kaynaklarının neler olduğu incelenebilir. Ayrıca, çalışma disiplini dışında diğer duyuşsal değişkenler açısından da maddelerin yanlılık durumları araştırılabilir.

KAYNAKÇA

Acar, T. (2011). Maddenin farklı fonksiyonlaşmasında örneklem büyüklüğü: Genelleştirilmiş aşamalı doğrusal modelleme uygulaması. *Kuram ve Uygulamada Eğitim Bilimleri, 11*(1), 279-288.

Atar, B. & Kamata, A. (2011) MTK olabilirlik oranı testi ve lojistik regresyon değişen madde fonksiyonu belirleme yöntemlerinin karşılaştırılması. *Hacettepe Eğitim Fakültesi Dergisi, 41*, 36-47.

Atalay, K., Gök, B., Kelecioğlu, H. & Arsan, N. (2012). Değişen madde fonksiyonunun belirlenmesinde kullanılan farklı yöntemlerin karşılaştırılması: Bir simülasyon çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 43*, 270-281.

Atalay Kabasakal, K., & Kelecioğlu, H. (2012). Evaluation of attitude items in PISA 2006 student questionnaire in terms of differential item functioning. *Ankara University Journal of Faculty of Educational Sciences, 45*(2), 77-96.

Bakan-Kalaycıoğlu, D. & Kelecioğlu, H. (2011). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi. *Eğitim ve Bilim, 36*, 161, 3-13.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2013). *Bilimsel araştırma yöntemleri.* Ankara: Pegem Akademi.

Cai, L., Du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling* [Computer software]. Chicago, IL: Scientific Software International.

Camili, G. & Stephard, L. A. (1994). *Methods for identifying biased test items*. London: Sage Publications.

Doğan, N. & Öğretmen, T. (2008). Değişen madde fonksiyonunu belirlemede Mantel-Haenszel, Ki-Kare ve lojistik regresyon yöntemlerinin 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.

Elosua, P. & Wells, C. S. (2013). Detecting DIF in polytomous items using MACS, IRT and Ordinal Logistic Regression. *Psicológica, 34*, 327-342.

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

Fidalgo, Á. M., Ferreres, D., & Muñiz, J. (2004). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II error rates. *The Journal of Experimental Education*, *73*(1), 23-39.

Fraenkel R.J. & Wallen E.N. (2011). *How to design and evaluate research in education*. New York: McGraw-Hill.

Gao, L. & Wang C. (2005). *Using five procedures to detect DIF with passage-based testlets.* A paper prepared for the poster presentation at the Graduate Student Poster Session at the annual meeting of the National Council of Measurement inEducation, Montreal, Quebec.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement, 38*, 164–187.

Gómez-Benito, J., & Navas-Ara, M. J. (2000). A Comparison of χ2, RFA and IRT based procedures in the detection of DIF. *Quality & Quantity*, *34*(1), 17-31.

Gök, B., Kabasakal, K. A., & Kelecioğlu, H. (2014). PISA2009 öğrenci anketi tutum maddelerinin kültüre göre değişen madde fonksiyonu açısından incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *5*(1), 72-87.

Gök, B., Kelecioğlu, H. & Doğan, N. (2010) Değişen madde fonksiyonunu belirlemede Mantel–Haenszel ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim,* 35 (156), 3-16.

Hambleton, R K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage Publication.

Karakaya, İ. (2012). Seviye belirleme sınavındaki fen ve teknoloji ile matematik alt testlerinin madde yanlılığı açısından incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 12*(1*),* 215-229.

Karakaya, I. & Kutlu, Ö. (2012). Seviye belirleme sınavındaki Türkçe alt testlerinin madde yanlılığının incelenmesi. *Eğitim ve Bilim, 37* (165), 348-362.

Karasar, N. (2005). *Bilimsel araştırma yöntemi*. Ankara: Nobel Yayın Dağıtım.

Köse, I. A. (2015). PISA 2009 öğrenci anketi alt ölçeklerinde (q32-q33) bulunan maddelerin değişen madde fonksiyonu açısından incelenmesi. *Kastamonu Eğitim Dergisi, 23*(1), 227-240.

Lyons-Thomas, J., Sandilands, D. & Ercikan, K. (2014). Gender differential item functioning in mathematics in four international jurisdictions. *Education and Science, 39*(172), 20-32.

Mellenberg, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research: Applications of Item Response Theory, 13*, 123–144.

Meyer, J. P. (2014). *Applied measurement with jMetrik*. Routledge.

Oliveri, M. E., Ercikan, K. & Zumbo, B. (2013). Analysis of sources of latent class differential item functioning in international assessments. *International Journal of Testing*, *13*(3), 272-293.

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

Özdemir, D. (2003). Çoktan seçmeli testlerde iki kategorili ve önsel ağırlıklı puanlamanın değişen madde fonksiyonuna etkisi ile ilgili bir araştırma. *Eğitim ve Bilim, 28*(129), 37-43.

Padilla, J. L., Hidalgo, M. D., Benítez, I. & Benito, J. G. (2012). Comparison of three software programs for evaluating DIF by means of the Mantel_Haenszel procedure: EASY_DIF, DIFAS and EZDIF. *Psicológica, 33,* 135-156.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system, computer program exchange. *Applied Psychological Measurement*, *29*(2), 150-151.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*(1), 23–37.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197–207.

Snow, T. K. & Oshima, T. C. (2009). A comparison of unidimensional and three-dimensional differential item functioning analysis using two-dimensional data. *Educational and Psychological Measurement, 69*(5), 732-747.

The Organisation for Economic Co-operation and Development [OECD] (2014). *PISA 2012 technical report*. Paris: OECD.

Thissen, D. (2001). *IRTLRDIF v. 2.0 b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning.* Chapel Hill, NC: LL Thurstone Psychometric Laboratory.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-172). Hillsdale, NJ: Lawrence Erlbaum Associates.

Waller, N. G. (1998). EZDIF: Detection of uniform and nonuniform differential item functioning with the Mantel-Haenszel and logistic regression procedures. *Applied Psychological Measurement*, *22*(4), 391-391.

Yıldırım, S. (2008). Farklı isleyen maddelerin belirlenmesinde sınırlandırılmış faktör çözümlemesinin Olabilirlik-Oranı ve Mantel-Haenszel yöntemleriyle karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education), 34,* 297-307.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as sa unitary framework for binary and likert-type (ordinal) item scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF* (Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science). Prince George, Canada: University of Northern British Columbia.