Araştırma Makalesi

Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 52, 281-300, [2021]
doi: 10.9779.pauefd.825631

Öğretmen Yapımı Bir Testteki Maddelerin Değişen Madde Fonksiyonu Bağlamında İncelenmesi

Eren Can AYBEK* **Metin YAŞAR**** **Seval KULA KARTAL*****

• ***Geliş Tarihi:*** *15.11.2020* • ***Kabul Tarihi:*** *09.02.2021* • ***Çevrimiçi Yayın Tarihi:*** *09.02.2021*

Öz

Bu araştırmanın amacı öğretmen yapımı bir testte yer alan maddelerin klasik test kuramı ve madde tepki kuramı temelli değişen madde fonksiyonu (DMF) gösterip göstermeme durumunu farklı DMF belirleme yöntemlerine göre incelemektir. Bu amaç doğrultusunda 435 üniversite öğrencisinin ölçme ve değerlendirme dersi dönem sonu sınavı puanları kullanılmıştır. Dönem sonu sınavı 50 çoktan seçmeli maddeden oluşmaktadır. Veriler uzaktan eğitim sistemi üzerinden çevrimiçi olarak toplanmıştır. Veri analizi R üzerinde ShinyItemAnalysis paketi kullanılarak gerçekleştirilmiştir. Araştırma sonucunda, bir maddenin cinsiyete göre dört farklı yönteme göre DMF gösterdiği bulunmuştur. Üniversiteye göre yapılan inceleme sonucunda ise delta plot yöntemi hiçbir maddede DMF bulunamazken, Mantel-Haenszel, Lojistik Regresyon ve Lord $\chi 2$ yöntemlerine göre beş maddede DMF'ye rastlanmıştır. Hem cinsiyete hem de üniversiteye göre DMF bulunan maddeler incelendiğinde DMF'ye yol açabilecek ifadelere rastlanmamıştır.

Anahtar sözcükler: öğretmen yapımı test, yanlılık, değişen madde fonksiyonu

Atıf:

Aybek, E.C., Yaşar, M. ve Kula Kartal, S. (2021). Öğretmen yapımı bir testteki maddelerin değişen madde fonksiyonu bağlamında incelenmesi. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 52,* 281-300.doi: 10.9779.pauefd.825631

* Dr., Pamukkale Üniversitesi, https://orcid.org/0000-0003-3040-2337, erencan@aybek.net
** Dr., Pamukkale Üniversitesi, https://orcid.org/0000-0002-7854-1494, myasar@pau.edu.tr
*** Dr. Pamukkale Üniversitesi, https://orcid.org/0000-0002-3018-6972, seval.kula@hotmail.com

Giriş

Ölçme araçlarından elde edilen puanların taşıması gereken en temel iki özellik güvenirlik ve geçerliktir (Crocker ve Algina, 2008). Puanların güvenirliğinin yüksek olması, o araçla yapılan tekrarlı ölçümlerde benzer puanların alınabileceğini, başka bir deyişle ölçme sonuçlarına karışan tesadüfi hataların az miktarda olduğunu göstermektedir (Popham, 2014). Geçerlik ise ölçme aracı ile ölçülmek istenilen özelliğin ne derece ölçülebildiği ile ilgilidir. Örneğin, bir trafik işaretinin İngilizce karşılığını soran bir test maddesini doğru yanıtlamak, yalnızca İngilizce bilgisine sahip olmak ile mümkün değildir. Aynı zamanda trafik işaretinin anlamının da bilinmesini gerektirir. Dolayısıyla böyle bir maddenin İngilizce testinde yer alması, ölçülmek istenilen özellik olan İngilizce bilgisinin, trafik bilgisini de dahil ederek ölçülmesi anlamına gelmektedir. Bu örnek, ölçme ve değerlendirme alan yazınında geçerlik ile ilgili yapılan sık tanımlardan biri olan ölçülmek istenilen özelliğin araya başka bir değişken sokulmadan ölçülebilmesi tanımını da açıklamaktadır.

Geçerlik sorunu oluşturan durumlar bir maddenin birden fazla özelliği ölçmesi ile sınırlı değildir. Aynı zamanda maddenin belirli bir grup için yanlı olması, yani aynı yetenek düzeyindeki bir grubun, o maddeyi doğru yanıtlama olasılığının, başka bir gruba göre daha fazla olması da bir geçerlik sorunu oluşturmaktadır. Bu durum, yanlılık olarak anılmaktadır (Osterlind, 1983). Yanlılık çalışmaları incelendiğinde ölçme değişmezliği (Millsap, 2011), değişen madde fonksiyonu (Reise ve Revicki, 2015), değişen çeldirici fonksiyonu (Osterlind, 1983) gibi yöntemlerden yararlanıldığı görülmektedir. Ölçme değişmezliği çalışmalarında, faktör analitik tekniklerden yararlanılmakta ve faktör yapısının, faktör yüklerinin ve hata varyanslarının gruba göre farklılaşıp farklılaşmadığı incelenmektedir. Değişen çeldirici fonksiyonu çalışmalarında ise aynı yetenek düzeyinde bulunan iki grubun, belli bir seçeneği farklı olasılıklarda seçip seçmediği araştırılmaktadır.

Bu araştırmanın da konusunu oluşturan değişen madde fonksiyonu çalışmalarında, aynı yetenek düzeyindeki iki grubun, bir maddeye doğru yanıt verme olasılığının farklılaşıp farklılaşmadığı incelenmektedir (Gamerman, Gonçalves ve Soares, 2018). Bir maddede değişen madde fonksiyonu bulunması durumunda, bu maddenin ilgili gruba göre yanlılık gösterebilecek ifadeler içerip içermediği incelenmektedir. Maddelerde, tek biçimli ya da tek biçimli olmayan şekilde iki tür DMF olabilimektedir. Tek biçimli değişen madde fonksiyonu, tüm yetenek düzeylerinde maddenin bir grubun, diğer gruba göre daha yüksek olasılıkla maddeye doğru yanıt vermesi şeklinde ortaya çıkmaktadır. Tek biçimli olmayan

değişen madde fonksiyonu ise düşük yetenek düzeylerinde bir grup, yüksek yetenek düzeylerinde ise diğer grup üyelerinin maddeyi daha yüksek olasılıkla doğru yanıt vermesi durumunda ortaya çıkmaktadır (de Ayala, 2009). Bir maddenin DMF içermesi, doğrudan o maddenin yanlı olduğu anlamına gelmemektedir. DMF'ye neden olabilecek değişkenin ölçülmek istenen özellikle alakalı olmaması halinde, uzman görüşlerine de başvurarak maddenin yanlı olduğuna karar verilebilir (Karami, 2012).

DMF belirleme yöntemleri McNamara ve Roever (2006, s.93) tarafından madde güçlüğüne dayalı yöntemler, paremetrik olmayan yaklaşımlar, madde tepki kuramı (MTK) temelli yaklaşımlar ve diğer yaklaşımlar olmak üzere dört kategoride sınıflandırılmıştır. Madde güçlüğüne dayalı yöntemlerden ilki dönüştürülmüş madde güçlüğünü kullanan delta plot yöntemidir. Bir grup için kolay olan bir madde, diğer grup için güç ise bu durumda maddede DMF olduğu şeklinde yorumlanmaktadır (Angoff, 1982). Bu yöntemin en büyük sınırlılığı, grupların yetenek düzeylerinin eşleştirilmemiş olmasıdır. Mantel-Haenszel (MH) yöntemi kontenjans tablosu temelli parametrik olmayan bir tekniktir. Bu teknikte referans ve odak grubun bir maddeyi doğru ve yanlış yanıtlama olasılıkları dikkate alınmakta ve buradan elde edilen MHΔ mutlak değeri ve p-düzeyine göre sınıflandırma yapılmaktadır. Bu değerin 1.5'tan büyük olması ya da .05 düzeyinde 1.00'dan anlamlı düzeyde farklılaşması halinde geniş DMF gösterdiği şeklinde yorumlanmaktadır (McNamara ve Roever, 2006). MTK modellerinde bir maddeye doğru yanıt verme olasılığı, farklı yetenek düzeylerine göre kalibre edilir ve her bir madde için bu olasılıkları gösteren bir madde karakteristik eğrisi (MKE) oluşturulur (Hambleton ve Swaminathan, 1985). Bu eğrinin gruplara göre farklılaşması halinde maddede DMF olduğu şeklinde yorumlanır (Lord, 1980). Burada önemli olan nokta, seçilecek olan MTK modelidir. Yalnızca madde güçlüğünü dikkate alan modeller 1 parametreli lojistik model (1PLM) veya Rasch modeliyken, aynı zamanda madde ayırt ediciliğini dikkate alan 2PLM ve bunların yanında şansla doğru yanıtlama olasılığını (3PLM) ve yüksek yetenekli bireylerin dikkatsizlikle maddeyi yanlış yanıtlamasını dikkate alan 4PLM modeller tercih edilebilir (de Ayala, 2009). Diğer yaklaşımlar ise aslında DMF belirleme amacıyla geliştirilmemiş, ancak DMF belirlemede de kullanışlı olan teknikleri kapsamaktadır. Bunlardan en yaygın olarak kullanılanı Lojistik Regresyon (LR) tekniğidir. Bu teknikte madde puanları, yalnızca toplam puan ve toplam puanla birlikte grup üyeliği değişkenleri ile yordanmaya çalışılmaktadır (Zumbo, 1999).

Rogers ve Swaminathan (1993), simülatif veriler kullanarak MH ve LR tekniklerini karşılaştırmış, MH tekniğinin kolay ve zor testlerde DMF belirlemede iyi bir performans

sergilediğini, orta güçlükteki testlerde ise yeterince başarılı olmadığını ifade etmiştir. LR'nun ise tek biçimli DMF belirlemede en az MH kadar iyi bir teknik olduğu; tek biçimli olmayan DMF belirlemede ise MH'den daha iyi sonuçlar verdiğini bulmuşlardır. Narayanan ve Swaminathan (1994) ise simülatif verilerle çalıştıkları araştırmada, MH ve SIBTEST'in tek biçimli DMF belirlemede benzer şekilde etkili olduğunu, bu iki istatistiğin de örneklem büyüklüğünden etkilendiğini ifade etmiştir.

Doğan ve Öğretmen (2008)'in, Ortaöğretim Kurumları Seçme ve Yerleştirme sınavı fen başarı testi verileri ile cinsiyete göre DMF tespit etmede MH, LR ve ki-kare tekniklerini karşılaştırdığı çalışmada, en fazla yanlılık tespit eden tekniğin MH olduğu, LR tekniği ile hiçbir maddede DMF bulunamadığı ifade edilmiştir. Gök, Kelecioğlu ve Doğan (2010), ortaöğretim kurumları sınavının (OKS) matematik ve fen bilgisi alt testlerinin cinsiyet ve okul türüne göre DMF gösterip göstermediğini incelemiştir. Bu amaçla Mantel-Haenszel (MH) ve Lojistik Regresyon (LR) teknikleri karşılaştırılmış; matematik alt testinde cinsiyete göre MH yöntemi ile dokuz, LR yöntemiyle ise üç maddede göz ardı edilebilir düzeyde DMF bulunmuştur. Okul türüne göreyse, 25 maddeden 15'i A, 3'ü B ve 1'i C düzeyinde olmak üzere toplam 19 maddede DMF bulunmuştur. Fen bilgisi alt testinde ise cinsiyete göre MH yöntemi ile 10, LR yöntemiyle 9 maddede göz ardı edilebilir düzeyde DMF'ye rastlanmıştır. Okul türüne göre yapılan karşılaştırmalarda ise MH yöntemi 14 maddede A, 1 maddede ise B düzeyinde DMF tespit etmiştir. LR yöntemine göre ise 18 maddede DMF olduğu sonucuna ulaşılmıştır. Çıkrıkçı Demirtaşlı ve Ulutaş (2015), PISA 2006 fen okuryazarlığı testini çok gruplu doğrulayıcı faktör analizi (DFA) kullanarak Türk ve Amerikalı öğrenciler için ölçme değişmezliğini incelemiş, ardından MH, Simultaneous Item Bias Test (SIBTEST) ve MTK temelli olabilirlik oranı kullanarak DMF gösteren maddeleri araştırmışlardır. Her üç yöntemle 38 maddenin DMF gösterdiğinin bulunduğu çalışmada, uzman görüşlerine de başvurulmuş ve bu görüşler sonrasında dokuz maddede DMF olduğuna karar verilmiştir. Koyuncu, Aksu ve Kelecioğlu (2018) ise PISA 2012 verilerini MH, LR ve olabilirlik oranı tekniklerini farklı yazılımlar kullanarak karşılaştırmıştır. Farklı DMF belirleme yöntemlerine göre, DMF bulunan madde sayısı bir ile beş arasında değişim göstermiştir. Ayva Yörü ve Atar (2019) ise 2012 yılının ortaöğretime giriş sınavını cinsiyet ve okul değişkenine göre SIBTEST, Breslow-Day, Lord's $\chi 2$ ve Raju'nun ölçme alanı tekniklerini kullanarak incelemişler ve farklı yöntemlerin farklı sayıda DMF maddesi gösterdiğini bulmuşlardır. Bu bağlamda DMF belirleme çalışmalarında en az iki farklı yöntemin kullanılması gerektiğini önermişlerdir.

Alan yazındaki çalışmaların genellikle merkezi ve geniş ölçekli testlerde yoğunlaştığı görülmektedir. Ancak ölçmenin geçerli ve yansız olması, yalnızca merkezi ve geniş ölçekli testlerde değil sınıf içi ölçmelerde de aranan bir özelliktir. Dolayısıyla öğretmen yapımı bir testin değişen madde fonksiyonu bağlamında incelendiği daha çok çalışmaya ihtiyaç duyulmaktadır. Bu bağlamda çalışmanın amacı, öğretmen yapımı bir testteki maddelerin cinsiyet ve üniversiteye göre değişen madde fonksiyonu gösterip göstermediğini, klasik test kuramı temelli delta uzaklığı, Mantel-Haenszel, lojistik regresyon ve madde tepki kuramı temelli Lord χ^2 teknikleriyle incelemek ve karşılaştırmaktır.

Yöntem

Bu araştırmada, öğretmen yapımı bir testteki maddelerin çeşitli yöntemlere göre DMF gösterip göstermediğinin araştırılması amaçlanmaktadır. Bu bağlamda, madde özelliklerinin farklı gruplara göre değişip değişmediği incelenmiş, dolayısıyla ilişkisel tarama türünde bir araştırma yürütülmüştür.

Çalışma Grubu

Araştırmaya konu olan veriler 2019 – 2020 öğretim yılı yaz döneminde ölçme ve değerlendirme dersini alan 435 öğrenciden toplanmıştır. Bu öğrencilerden 126'sı erkek ve 309'u kadındır. Benzer şekilde 126 öğrenci Pamukkale Üniversitesi (PAÜ) Eğitim Fakültesi'ne devam etmekte iken, 309 öğrenci ise diğer üniversitelerin eğitim fakültelerine devam etmektedir. Narayanan ve Swaminathan (1994) DMF çalışmalarında odak grup büyüklüğünün, referans gruba ait örneklem büyüklüğüne göre DMF belirlemede daha etkili olduğunu vurgulamıştır. Bu nedenle erkek öğrenciler ve PAÜ'ye devam etmekte olan öğrenciler referans grup olarak kabul edilmiş; daha büyük bir grubu oluşturan kadın öğrenciler ile diğer üniversitelerden gelen öğrenciler odak grubunu oluşturmuştur. MTK temelli DMF belirleme yöntemleri ise maddelerin bir MTK modeline göre kalibrasyonuna ihtiyaç duymaktadır. Şahin ve Anıl (2016) yaptıkları simülasyon çalışmalarında 1 parametreli lojistik model (1PLM) için 10, 20 ve 30 maddelik testlerde 150 kişilik bir örneklem büyüklüğünün yeterli olabileceğini ifade etmiştir. Simülasyon çalışmaları ışığında, ele alınan testin öğretmen yapımı bir test olma özelliği de göz önüne alındığında araştırma grubu büyüklüğünün böyle bir çalışma için yeterli olduğuna karar verilmiştir.

Çalışmada Kullanılan Ölçme Aracı

Başarı testi

Veriler 50 maddelik çoktan seçmeli bir başarı testi ile toplanmıştır. Başarı testi, eğitimde ölçme ve değerlendirme konularını kapsamaktadır. Testin kapsamında yer alan konular ve konulara ilişkin hazırlanan madde sayıları şu şekildedir: Temel kavramlar (15); ölçmede hata (4), ölçme araçlarında aranılan özellikler (18) ve geleneksel ölçme araçları (13). Testten elde edilen puanlara ait iç tutarlılık katsayısı KR-20 .835 [.812 - .856] olarak bulunmuştur. KR-20 için güven aralığı değerleri F dağılımına bağlı olarak Feldt, Woodruff ve Salih (1987) tarafından önerilen şekilde hesaplanmıştır. Bu yöntemde, KR-20 katsayısının üst sınırı, $1 - [(1 - \text{KR-20}).F(\alpha/2)]$ formülüyle, alt sınırı ise $1 - [(1 - \text{KR-20}).F(1 - \alpha/2)]$ formülüyle elde edilmektedir. Bu bağlamda testten elde edilen verilerin oldukça güvenilir olduğu görülmektedir. Test maddeleri, doğru yanıt için 1 ve yanlış-boş yanıtlar için 0 şeklinde puanlanmaktadır.

Verilerin Toplanması

Veriler bir dönem sonu sınavında dersi alan öğrencilerden uzaktan eğitimde kullanılan öğrenme yönetim sistemi aracılığıyla toplanmıştır. Uygulama için öğrencilere 60 dakika süre tanınmıştır. Sistemsel sorunları azaltmak amacıyla, maddeler beşerli gruplar halinde öğrencilere sunulmuş ve ilk beş madde yanıtlandıktan sonra, bir sonraki beş maddenin görünmesi sağlanmıştır. Öğrenciler kendilerine tanınan süre içerisinde diledikleri maddeye geri dönebilmiş ve yanıtlarını gözden geçirebilmişlerdir. Veri toplama süreci E.93803232-622.02-13963 numaralı etik kurul onayı ile yürütülmüştür.

Veri Analizi

Veriler R 4.0.2 (R Core Team, 2020) üzerinde ShinyItemAnalysis (Martinková ve Drabinová, 2016) paketi kullanılarak analiz edilmiştir. ShinyItemAnalysis, test ve madde istatistiklerini hesaplayabilen, MTK kalibrasyonlarını ve değişen madde fonksiyonu, değişen çeldirici fonksiyonu analizlerini gerçekleştirebilen bir pakettir. Araştırmanın veri analizi ihtiyaçlarının tümünü karşılayabilen bir paket olduğu için bu çalışmada kullanılmasına karar verilmiştir.

Testteki maddelerin DMF gösterip göstermediğinin belirlenmesinde Delta plot, Mantel-Haenszel, Lojistik Regresyon ve MTK temelli Lord $\chi 2$ tekniklerinden yararlanılmıştır. Giriş kısmında açıklanan bu tekniklerden, Lord $\chi 2$, MTK'ya dayandığından, analizlerin bir MTK modeline bağlı olarak gerçekleştirilmesi gerekmektedir. Daha basit bir

model olması nedeniyle, analizler Rasch modeli temelinde yürütülmüştür. Rasch modelinde, bir maddenin doğru yanıtlanma olasılığının yalnızca maddenin güçlük düzeyine bağlı olduğu; madde ayırt ediciliğinin ise tüm maddeler için aynı ve 1.00 olduğu kabul edilmektedir (de Ayala, 2009).

Bulgular

Madde Analizi Bulguları

Başarı testine ait madde güçlüğü (p) ve madde ayırt ediciliği (rjx) indeksleri hesaplanmış ve Tablo 1'de verilmiştir:

Tablo 1. *Başarı testine ait madde istatistikleri*

Madde no	p	r_{jx}	Madde no	p	r_{jx}
1	0,48	*0,10*	26	0,72	0,50
2	0,94	0,23	27	0,92	0,37
3	0,73	0,25	28	0,60	*0,15*
4	0,77	*0,03*	29	0,64	0,33
5	0,87	0,33	30	0,83	0,36
6	0,61	0,23	31	0,94	0,32
7	0,79	0,54	32	0,17	*0,14*
8	0,93	*0,19*	33	0,96	0,27
9	0,86	0,38	34	0,86	0,46
10	0,77	0,51	35	0,82	0,47
11	0,70	0,37	36	0,69	0,30
12	0,79	0,44	37	0,78	0,62
13	0,86	0,56	38	0,46	0,24
14	0,33	*0,18*	39	0,72	0,45
15	0,21	*-0,08*	40	0,85	0,54
16	0,60	0,25	41	0,97	0,25
17	0,81	0,49	42	0,23	*0,03*
18	0,60	0,40	43	0,29	*0,00*
19	0,04	*-0,19*	44	0,87	0,33
20	0,79	0,54	45	0,20	*0,06*
21	0,45	*0,16*	46	0,72	0,30
22	0,08	*-0,09*	47	0,84	0,55

23	0,64	0,30	48	0,58	*0,12*
24	0,07	*0,05*	49	0,49	0,26
25	0,89	0,46	50	0,16	*0,09*

Çok kolay madde — Çok zor madde

Madde istatistikleri incelendiğinde; 1, 4, 8, 15, 19, 21, 22, 24, 28, 32, 42, 43, 45, 48 ve 50. maddeler olmak üzere toplam 15 maddenin ayırt edicilik indekslerinin .20'den daha küçük olduğu bulunmuştur. Bu maddeler başarılı ve başarısız öğrenciyi ayırt etmede yetersiz maddeler olmuştur. Ancak bu maddelerden, 8, 22, 24, 32 ve 50. maddelerin çok kolay (grubun çoğunluğu tarafından doğru yanıtlanmış) ya da çok zor (grubun çoğunluğu tarafından yanlış yanıtlanmış) maddeler olduğu görülmektedir. Grubun çoğunluğu tarafından doğru ya da yanlış yanıtlanan maddelerin, başarılı ve başarısız öğrencileri ayırt etmede yetersiz kalmış olabileceği göz önüne alınmalıdır. Testin ortalama güçlüğü ise .64 olarak bulunmuştur. Buna göre, testin orta güçlükte olduğu ifade edilebilir.

Cinsiyete ve Kuruma Göre Değişen Madde Fonksiyonu Bulguları

Cinsiyete göre DMF analizleri sonucunda elde edilen bulgular Tablo 2'de verilmiştir.

Tablo 2. *Cinsiyete Göre DMF İstatistikleri*

Madde no	Delta Plot	MH	LR χ^2	p_{LR}	Lord χ^2	p_L	Avantajlı Grup
1	-0,21	0,99	0,75	0,69	0,94	0,33	
2	0,27	0,33	5,72	0,06	1,16	0,28	
3	0,23	0,90	1,47	0,48	0,42	0,51	
4	-0,53	0,99	0,56	0,76	1,11	0,29	
5	-0,32	0,94	0,88	0,64	0,05	0,82	
6	-0,08	0,88	0,04	0,98	0,14	0,71	
7	0,24	0,93	0,55	0,76	0,67	0,41	
8	0,57	0,43	3,44	0,18	2,74	0,10	
9	0,08	0,97	0,51	0,77	0,35	0,55	
10	-0,09	0,46	0,32	0,85	0,00	0,96	
11	0,49	0,62	3,68	0,16	1,50	0,22	

Tablo 2. *Cinsiyete Göre DMF İstatistikleri (Devamı)*

Madde no	Delta Plot	MH	LR χ^2	p_{LR}	Lord χ^2	p_L	Avantajlı Grup
12	0,06	0,54	1,46	0,48	0,15	0,70	
13	-0,18	0,97	2,69	0,26	0,00	0,99	
14	0,45	0,49	0,66	0,72	0,20	0,65	
15	-0,92	0,03	5,60	0,06	8,97	0,00	E (MH, Lord)
16	-0,59	0,09	2,64	0,27	2,82	0,09	
17	0,48	0,35	2,10	0,35	1,99	0,16	
18	-0,03	0,15	1,02	0,60	0,07	0,79	
19	-0,29	0,88	1,47	0,48	1,53	0,22	
20	-0,36	0,34	1,38	0,50	0,36	0,55	
21	0,15	0,68	2,86	0,24	0,01	0,92	
22	-0,18	0,99	2,26	0,32	1,32	0,25	
23	-0,13	0,48	0,82	0,66	0,21	0,65	
24	0,20	0,81	0,04	0,98	0,20	0,65	
25	-0,18	0,93	1,92	0,38	0,02	0,89	
26	0,46	0,35	1,19	0,55	1,44	0,23	
27	-0,87	0,08	1,49	0,48	0,91	0,34	
28	0,82	0,03	5,80	0,06	3,54	0,06	K (MH)
29	-0,34	0,11	2,03	0,36	0,96	0,33	
30	0,51	0,14	3,16	0,21	2,31	0,13	
31	0,01	0,82	0,42	0,81	0,40	0,53	
32	0,03	0,93	0,31	0,86	0,58	0,45	
33	-0,08	0,77	3,70	0,16	0,25	0,62	

Tablo 2. *Cinsiyete Göre DMF İstatistikleri (Devamı)*

Madde no	Delta Plot	MH	LR χ^2	p_{LR}	Lord χ^2	p_L	Avantajlı Grup
34	-0,67	0,16	1,93	0,38	0,98	0,32	
35	-0,43	0,62	8,35	0,02	0,41	0,52	E (LR)
36	0,90	0,02	8,58	0,01	5,08	0,02	K (MH, LR, Lord)
37	0,56	0,10	2,46	0,29	2,44	0,12	
38	0,52	0,86	5,02	0,08	0,72	0,40	
39	0,10	0,93	0,10	0,95	0,08	0,77	
40	0,72	0,21	5,75	0,06	4,07	0,04	
41	0,41	0,42	1,70	0,43	1,58	0,21	
42	-0,61	0,28	4,94	0,09	4,99	0,03	
43	-0,44	0,77	1,57	0,46	3,22	0,07	
44	-1,74	0,01	10,83	0,00	7,45	0,01	E (Delta, LR, Lord), K (MH)
45	0,20	0,99	0,10	0,95	0,13	0,72	
46	-0,18	0,65	0,16	0,92	0,16	0,69	
47	0,11	0,91	1,40	0,50	0,37	0,54	
48	0,56	0,30	3,20	0,20	1,40	0,24	
49	0,08	0,89	1,66	0,44	0,05	0,82	
50	0,24	0,92	3,32	0,19	0,09	0,77	

Tablo incelendiğinde, 44. madde tüm yöntemlere göre DMF göstermektedir. 36. madde ise delta plot yöntemi dışındaki tüm yöntemlere göre DMF göstermektedir. 15. madde MH ve Lord yöntemlerine göre, 28. Madde yalnızca MH ve 35. Madde yalnızca LR yöntemine göre DMF göstermektedir. Tüm yöntemlerin, cinsiyete göre DMF'ye işaret ettiği 44. madde incelendiğinde;

> *Öğretmen adaylarının, öğretmenlik uygulamaları dersindeki ortaya koydukları performans puanları ile ilerideki öğretmenlik performanslarına dayalı olarak alacakları puanlar arasında bir korelasyondan söz edildiğinde aşağıdaki geçerlilik türlerinden hangisine işaret edilmiş olur?*

A. Kapsam

B. Yordama

C. Uygunluk

D. Yapı

E. Görünüş

Cinsiyete göre DMF oluşturabilecek herhangi bir etmene rastlanamamıştır. Benzer şekilde üç yöntemin DMF olduğuna işaret ettiği 36. madde de aşağıda sunulmuştur:

Diğerlerine göre puanlama hatası en az olan ölçme aracı aşağıdaki seçeneklerin hangisinde verilmiştir?

A. Yazılı yoklamalar

B. Kısa cevap gerektiren testler

C. Çoktan seçmeli testler

D. Sözlü yoklamalar

E. Ödev ve projeler

Maddenin cinsiyete göre DMF göstermesine neden olabilecek ifadeleri içermediği görülmektedir. İki yönteme göre DMF gösteren 15. madde ise aşağıdadır:

Aşağıdaki ölçme sonuçlarından hangisi oranlı ölçekle elde edilebilir?

A. Anne-babanın eğitim düzeyi

B. Öğrencinin Türkçe dersindeki performansı

C. Oyun parkındaki çocuk sayısı

D. Ateşi ölçülen hastanın sıcaklık değeri

E. Öğrencilerin öz değerlendirmeye yönelik davranışları

Bu maddenin de kök ve seçeneklerindeki ifadelerin cinsiyete göre DMF'ye neden olabilecek ifadeler bulundurmadığı anlaşılmaktadır. Kuruma göre DMF gösteren maddeler incelenmiş ve elde edilen bulgular Tablo 3'te verilmiştir.

Tablo 3. *Kuruma Göre DMF İstatistikleri*

Madde no	Delta Plot	MH	LR χ^2	p_{LR}	Lord χ^2	p_L	Avantajlı Grup
1	0,66	0,11	3,92	0,14	1,48	0,22	
2	-0,22	0,82	0,38	0,83	0,14	0,71	
3	-0,70	0,08	9,61	0,01	2,23	0,14	PAÜ (LR)
4	-0,63	0,60	4,08	0,13	1,37	0,24	
5	-0,39	0,84	0,88	0,64	0,06	0,80	
6	0,36	0,11	0,68	0,71	0,57	0,45	
7	0,26	0,65	5,29	0,07	0,90	0,34	
8	-0,02	0,88	1,18	0,56	0,49	0,48	
9	0,03	0,90	0,72	0,70	0,37	0,54	
10	-0,60	0,09	3,65	0,16	1,29	0,26	
11	-0,45	0,37	2,72	0,26	1,05	0,31	
12	0,71	0,15	4,26	0,12	4,18	0,04	Diğer (Lord)
13	-0,37	0,48	0,86	0,65	0,07	0,79	
14	0,18	0,97	0,75	0,69	0,12	0,73	
15	-0,20	0,40	1,84	0,40	2,24	0,13	
16	-0,49	0,07	3,70	0,16	1,97	0,16	
17	0,70	0,05	4,24	0,12	4,20	0,04	Diğer (Lord)
18	0,33	0,82	1,01	0,60	0,42	0,51	
19	0,45	0,82	0,64	0,73	0,19	0,66	

Tablo 3. *Kuruma Göre DMF İstatistikleri (Devamı)*

Madde no	Delta Plot	MH	LR χ^2	p_{LR}	Lord χ^2	p_L	Avantajlı Grup
20	0,41	0,47	1,34	0,51	1,67	0,20	
21	-0,27	0,35	1,59	0,45	1,49	0,22	
22	-0,59	0,04	9,73	0,01	5,49	0,02	PAÜ (MH, LR, Lord)
23	-0,31	0,55	1,82	0,40	0,71	0,40	
24	-0,87	0,02	7,99	0,02	8,09	0,00	PAÜ (MH, LR, Lord)
25	-0,64	0,48	1,93	0,38	0,38	0,54	
26	0,61	0,20	4,25	0,12	2,71	0,10	
27	-0,83	0,42	1,90	0,39	0,57	0,45	
28	-0,15	0,35	2,24	0,33	0,31	0,58	
29	0,43	0,47	1,12	0,57	1,03	0,31	
30	0,50	0,14	2,39	0,30	2,56	0,11	
31	-0,50	0,96	0,97	0,62	0,01	0,93	
32	-0,13	0,17	2,37	0,31	1,94	0,16	
33	-0,39	0,81	0,57	0,75	0,03	0,87	
34	-0,29	0,75	2,73	0,26	0,01	0,92	
35	0,04	0,77	0,10	0,95	0,27	0,61	
36	0,88	0,02	6,34	0,04	5,21	0,02	Diğer (MH, LR, Lord)
37	0,89	0,00	13,91	0,00	6,04	0,01	Diğer (MH, LR, Lord)
38	0,10	0,60	0,10	0,95	0,07	0,79	
39	-0,32	0,44	1,59	0,45	0,44	0,51	
40	0,45	0,17	1,99	0,37	2,26	0,13	

Tablo 3. *Kuruma Göre DMF İstatistikleri (Devamı)*

Madde no	Delta Plot	MH	LR χ^2	p_{LR}	Lord χ^2	p_L	Avantajlı Grup
41	-0,33	0,95	2,35	0,31	0,09	0,76	
42	-0,80	0,00	10,80	0,01	8,56	0,00	PAÜ (Lord)
43	0,41	0,51	0,23	0,89	0,01	0,92	
44	0,57	0,06	3,18	0,20	3,28	0,07	
45	0,05	0,61	2,94	0,23	0,87	0,35	
46	0,52	0,29	5,30	0,07	2,01	0,16	
47	0,45	0,20	2,65	0,27	2,26	0,13	
48	0,35	0,70	1,04	0,59	0,43	0,51	
49	0,54	0,36	2,38	0,30	0,85	0,36	
50	-0,43	0,07	4,92	0,09	4,37	0,04	PAÜ (Lord)

PAÜ ve diğer üniversiteler arasında DMF gösteren maddeler incelendiğinde, 22, 24, 36, 37 ve 42. maddelerin üç yönteme göre DMF gösterdiği anlaşılmaktadır. Bu maddeler sırasıyla aşağıda sunulmuştur:

Madde 22:

Aşağıdakilerden hangisi bir testte tesadüfi hata *bulunduğunun göstergesidir?*

A. Öğrenci kâğıtlarındaki puanlar toplanırken farkında olmadan hata yapılması

B. Ön sıralarda oturup ders dinleyen öğrencilere 5'er puan eklenmesi

C. Testin uygulandığı tüm öğrencilere 10 puan eklenmesi

D. Her öğrenciye puanının %10'unun eklenmesi

E. Sınıf kurallara uymayan öğrencilerden 5 puan silinmesi

Madde 24:

Test puanları için KR-21 katsayısını 0.80 bulan bir öğretmenin, aşağıdaki yorumlardan hangisini yapması doğru olur?

A. Öğrenciler %80 oranında başarılı olmuştur

B. Ölçmek istenilen özellik ölçülebilmiştir

C. Sınav puanlarına %20 hata karışmıştır

D. Sınavın iç tutarlılığı yüksektir

E. Sorular, konulara dengeli dağılmıştır

Madde 36:

Diğerlerine göre puanlama hatası en az olan ölçme aracı aşağıdaki seçeneklerin hangisinde verilmiştir?

A. Yazılı yoklamalar

B. Kısa cevap gerektiren testler

C. Çoktan seçmeli testler

D. Sözlü yoklamalar

E. Ödev ve Projeler

Madde 37:

Bir öğretmenin uygulamış olduğu bir ölçme işleminde aşağıdakilerden hangisi ölçme kuralı olarak belirlemesi doğru olmaz?

A. Boş bırakılan cevaplara puan verilmez

B. İlk sorunun doğru cevaplandırılması durumunda 10 puan verilecektir

C. Her imla kuralı hatasında 2 puan kırılacaktır

D. Dönem sonu ortalama puanı 60 olan dersten geçer

E. Paragraf girintisi mutlaka 8 karakter boş bırakılarak yapılacaktır

Madde 42:

I. Standart sapma

II. Ortalama

III. KR-21 katsayısı.

Bir sınavdan alınan puanlar için ölçmenin standart hatasını hesaplamak isteyen bir öğretmen, verilen istatistiklerden hangisi ya da hangilerini kullanmalıdır?

A. Yalnız I

B. I ve II

C. I ve III

D. II ve III

E. I, II ve III

Madde kökü ve seçeneklerde yer alan ifadeler incelendiğinde, üniversiteye göre yanlılık oluşturabilecek ifadelere rastlanmadığı görülmektedir. Öte yandan, aracın çevrimiçi ortamda uygulanması nedeniyle, üniversite öğrencilerinin sınavın uygulandığı ortama olan aşinalığı göz önünde bulundurulmalıdır. Fakat, yine de bu maddelerin diğer maddelerle aynı ortamda, benzer şekilde sunulmuş olması nedeniyle, bu değişkenin de DMF'ye neden olmayacağına karar verilmiştir.

Tartışma

Öğretmen yapımı bir başarı testinde yer alan maddelerin, cinsiyet ve üniversiteye göre DMF gösterip göstermediğini incelemeyi amaçlayan bu araştırmada, cinsiyete göre bir madde (44) tüm DMF belirleme yöntemlerine göre yanlı bulunmuştur. Üniversiteye göre DMF incelemesinde ise Delta yöntemine göre hiçbir madde DMF göstermezken, beş maddenin (22, 24, 36, 37 ve 42) üç yönteme göre DMF gösterdiği bulunmuştur. Ancak maddeler incelendiğinde cinsiyet ya da üniversiteye göre yanlılık oluşturabilecek ifadelere rastlanmadığı görülmüştür. DMF gösteren maddelerin farklı yöntemlerle düşük uyum göstermesi Rogers ve Swaminathan (1993), Gierl, Khaliq ve Boughton (1999), Doğan ve Öğretmen (2008), Gök ve diğerlerinin (2010) bulgularıyla örtüşmektedir. Ayrıca Gierl, Khaliq ve Boughton (1999), MH yönteminin DMF bulmada SIBTEST ve LR'a göre daha tutucu olduğunu ve daha az sayıda maddenin DMF gösterdiği şekilde işaretlediğini ifade etmiştir. Bu araştırma kapsamına SIBTEST yöntemi girmese de; MH ve LR yöntemleri karşılaştırıldığında özellikle cinsiyete göre MH yönteminin daha çok maddeyi DMF'li olarak işaretlediği bulunmuştur.

Öğretmen yapımı testleri alan öğrenci sayısının az olması, DMF çalışmalarını daha geniş örneklem büyüklükleri ile çalışmaya olanak sağlayan PISA, TIMMS, PIRLS gibi testlere ve ulusal ölçekte uygulanan kurumlar arası geçiş testlerine ait verilere odaklamıştır (Çıkrıkçı Demirtaşlı ve Ulutaş, 2015; Doğan ve Öğretmen, 2008; Gök ve diğ. 2010; Koyuncu, Aksu ve Kelecioğlu, 2018; Ayve Yörü ve Atar, 2019).

Sonuç

Öğretmen yapımı testlerden alınan puanlarla, öğrenciler hakkında başarılı – başarısız kararı verilmekte ve öğrencilerin dersi tekrar etmesi gerekebilmektedir. Her ne kadar bu testler, geniş ölçekli testlerin eğitim politikalarına yön verme amacını taşımasa da öğrenciler hakkında geçti – kaldı kararı verilirken, belirli bir gruba yanlı davranan bir testten elde edilen puanları temel almak, geçerli kararlar verilmesini engelleyecektir. Bu bağlamda,

öğretmen yapımı testler üzerinde daha çok DMF çalışmasının yürütülmesi ve bu testlerde DMF'ye neden olabilecek faktörlerin belirlenmesi öğretmen eğitimine de katkı sağlayacaktır.

Araştırmada DMF belirleme amacıyla delta plot, MH, LR ve Lord $\chi 2$ yöntemleri kullanılmıştır. Diğer MTK modelleri ve DMF belirleme yöntemlerinin kullanılacağı gelecek araştırmalara ihtiyaç bulunmaktadır. Ayrıca bu araştırmanın en büyük sınırlılığı örneklem büyüklüğünün geniş ölçekli testlere göre daha küçük olmasıdır. Daha büyük örneklemlere sahip öğretmen yapımı testlerle benzer çalışmalar yürütülebilir. Koyuncu, Aksu ve Kelecioğlu (2018) farklı DMF yöntemlerini farklı yazılımlar kullanarak incelemiştir. Bu çalışma ise yalnızca ShinyItemAnalysis yazılımı ile sınırlıdır. Öğretmen yapımı testlerde diğer yazılımların da test edileceği yeni araştırmalar yapılabilir. Bu araştırmada bazı maddelerin DMF göstermesine rağmen, maddeler incelendiğinde DMF oluşturabilecek ifadelere rastlanmamıştır. DMF gösteren maddelerin, uzman görüşlerine başvurularak incelenmesinin, eğitim fakültelerinde okutulmakta olan ölçme ve değerlendirme dersi ve test geliştirmeye yönelik yürütülen hizmet içi eğitimlerin içeriğine de katkı sağlayacağı düşünülmektedir.

Etik Kurul İzin Bilgisi: *Bu araştırma, Pamukkale Üniversitesi Sosyal ve Beşeri Bilimler Araştırma ve Yayın Etiği Kurulunun 03/02/2021 tarihli E-93803232-622.02-13963 sayılı kararı ile alınan izinle yürütülmüştür.*

Çıkar Çatışması: *Yazarların beyan edeceği bir çıkar çatışması yoktur.*

Yazar Katkısı: *Birinci yazar, problem durumunun belirlenmesi ve verilerin analizi aşamalarında; ikinci yazar, çalışma grubunun belirlenmesi, verilerin toplanması ve problem durumunun belirlenmesi aşamasında, üçüncü yazar ise verilerin analizi aşamasında çalışmaya katkı sağlamıştır. Öte yandan tüm yazarlar, literatür taraması, tartışma ve raporlama aşamasında çalışmaya katkı vermiştir.*

Kaynakça

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp.96-116). Baltimore: Johns Hopkins University.

Ayva Yörü, F.G. ve Atar, H.Y. (2019). Determination of differential item functioning (DIF) according to SIBTEST, Lord's $\chi 2$, Raju's area measurement and Breslow-Day methods. *Journal of Pedagogical Research, 3*(3), 139-150. doi: 0.33902/jpr.v3i3.137

Crocker, L. ve Algina, J. (2008). *Introduction to classical and modern test theory*. Ohio: Cengage Learning.

Çıkrıkçı Demirtaşlı, N. ve Ulutaş, S. (2015). A study on detecting of differential item functioning of PISA 2006 science literacy items in Turkish and American samples. *Eurasian Journal of Educational Research, 58*. 41-60.

de Ayala, R.J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Doğan, N. ve Öğretmen, T. (2008). Değişen madde fonksiyonu belirlemede Mantel-Haenszel, ki-kare ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim, 33*(148). 100-112.

Feldt, L.S., Woodruff, D.J. ve Salih, F.A. (1987). Statistical inference for coefficient alpha. *Applied psychological measurement, 11*(1), 93-103.

Gamerman, D., Goncalves, F. B. ve Soares, T. M. (2018). Differential item functioning. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 3, pp. 67–86). Boca Raton, FL: CRC Press.

Gierl, M., Khaliq, S. N., & Boughton, K. (1999, June). Gender differential item functioning in mathematics and science: Prevalence and policy implications. In *Annual Meeting of the Canadian Society for the Study of Education*, Canada.

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

Hambleton, R.K. ve Swaminathan, H. (1985). *Item response theory: Principles and applications*. New York: Springer Science+Bussiness Media.

Karami, H. (2012). An introduction to differential item functioning. The International *Journal of Educational and Psychological Assessment, 11*(2), 59–76.

Koyuncu, İ., Aksu, G. ve Kelecioğlu, H. (2018). Comparison of Mantel-Haenszel, logistic regression and likelihood ratio methods to evaluate differential item functioning by using different computer software. *Elementary Education Online*, *17*(2). 909-925. doi: 10.17051/ilkonline.2018.41933

Lord, M.D. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrance Erlbaum Associates.

Martinková, P. ve Drabinová, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, 10(2), 503-515. doi: 10.32614/RJ-2018-074

McNamara, T. ve Roever, C. (2006). *Language testing: The social dimension*. Massachusetts: Blackwell Publishing.

Millsap, R.E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Narayanan, P. ve Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*(4). 315-328.

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

Popham, W.J. (2014). *Classroom assessment: What teachers need to know*. Boston: Pearson.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reise, S.P. ve Revicki, D.A. (2015). *Handbook of item response theory modeling*. New York: Routledge.

Rogers, H. J. ve Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

Şahin, A. ve Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice, 17*(1). 321-335. doi: 10.12738/estp.2017.1.0270

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

Research Article

Pamukkale University Journal of the Faculty of Education, 52, 281-300, [2021]
doi: 10.9779.pauefd.825631

Examination of Items in a Teacher-Made Test in the context of Differential Item Functioning

Eren Can AYBEK* **Metin YAŞAR**** **Seval KULA KARTAL*****

• ***Received:*** *15.11.2020* • ***Accepted:*** *09.02.2021* • ***Online First:*** *09.02.2021*

Abstract

The present study aims to examine whether the items in a teacher-made test show a classical test theory and item response theory based differential item functioning (DIF) with different DIF determination methods. For this purpose, final exam scores of 435 students regarding the course of measurement and evaluation were used. Final exam consisted of 50 multiple-choice items. Data were collected online through the learning management system. Data analysis was performed using the ShinyItemAnalysis package on R. As a result of the study, it was found that one item showed DIF according to four different methods by gender. As a result of analysis made by university, on the other hand, none of the items showed DIF according to delta plot method, while five items showed DIF according to Mantel-Haenszel, Logistic Regression and Lord $\chi 2$ techniques. When the statements in the items showing DIF by both gender and university are examined, on the other hand, it was observed that there was no statement that can create DIF.

Key words: teacher-made test, bias, differential item functioning

Cited:

Aybek, E.C., Yaşar, M., & Kula Kartal, S. (2021). Examination of items in a teacher-made test in the context of differential item functioning. *Pamukkale University Journal of Education, 52,* 281-300 doi:10.9779.pauefd.825631

* PhD., Pamukkale University, https://orcid.org/0000-0003-3040-2337, erencan@aybek.net
** PhD, Pamukkale University, https://orcid.org/0000-0002-7854-1494, myasar@pau.edu.tr
*** PhD, Pamukkale University, https://orcid.org/0000-0002-3018-6972, seval.kula@hotmail.com

Introduction

Two most important features that scores obtained from measurement tools need to have are reliability and validity (Crocker & Algina, 2008). The high rates of reliability of scores indicate that similar scores can be obtained in the repeated measures with that tool, in other words, this shows that there is a small amount of random errors involved in the measurement results (Popham, 2014). The validity, on the other hand, is the degree to which the tool measures what it claims to measure. For example, answering a test item questioning the English equivalent of a traffic sign correctly is not possible only knowing English. One also should know what this traffic sign means. Hence, the inclusion of such an item in an English test means that English knowledge, the feature to be measured, is measured by including traffic information, too. This example also explains the definition that the desirable characteristic, one of the common definitions of validity in the educational measurement literature, can be measured without including another variable.

Situations that create validity problems are not limited to an item measuring more than one trait. Besides, when the item is biased for a particular group, that is, if a group of the same trait level is more likely to respond that item correctly than another group, this also poses a validity problem. This is referred to as bias (Osterlind, 1983). When we examine the studies on bias, we see that various methods have been benefitted from, such as measurement invariance (Millsap, 2011), differential item functioning (Reise & Revicki, 2015), and differential distractor functioning (Osterlind, 1983), etc. Factor analytical techniques are used in measurement invariance studies and it is examined whether the factor structure, factor loadings and error variances differ according to the group. In differential distractor functioning studies, on the other hand, it is investigated whether two groups of the same trait level chose a certain option with different possibilities.

The studies on differential item functioning, which is the subject of this research, examines whether the probability of correctly responding to an item between two groups who have equal overall ability differs (Gamerman, Gonçalves, & Soares, 2018). It is investigated whether an item with differential item functioning has statements that may show bias according to the group. In items, two types of DIF can occur: uniform and non-uniform. Uniform differential item functioning occurs when one group performs better at all ability levels than the other group. The non-uniform differential item functioning, on the other hand, arises when one group is more likely to answer the item correctly at lower ability levels while the other group to answer correctly at most ability levels (de Ayala, 2009). The

presence of DIF does not necessarily indicate the presence of bias. If the variance likely to cause DIF is irrelevant to the construct being measured by the test, then, this item can be considered to be biased only after consulting a panel of experts (Karami, 2012).

McNamara and Roever (2006, p.93) have classified methods of DIF detection into four categories: analyses based on item difficulty, nonparametric approaches, item-response-theory-based (IRT) approaches, and other approaches. The first one of the analysis based on item difficulty is the delta plot method which uses transformed item difficulty index. The fact that the item is easy for one group but hard for another group gives the signal of the presence of DIF in the item (Angoff, 1982). The major limitation of this method is that the ability levels of the groups are not matched. The Mantel-Haenszel (MH) method is a nonparametric technique based on contingency table. In this method, focal and reference groups' probability of success on the item is compared and a classification is made according to the MHΔ absolute value and p-level obtained here. Items are identified as showing DIF in large scale if absolute value is greater than 1.5 or it is significantly different from 1.0 (McNamara & Roever, 2006). In the IRT models, the probability of giving the correct answers to an item is calibrated according to different ability levels and a particular item characteristic curve (ICC) showing these probabilities is created for each item (Hambleton & Swaminathan, 1985). The differentiation of this curve according to the groups is interpreted as the presence of DIF in the item (Lord, 1980). The important point here is the IRT model to be chosen. The One-Parameter Logistic Model (1PLM) or the Rasch model, which involves only item difficulty, or 2PLM, which considers item discrimination, and 3PLM that concentrates on the probability of the correct response, and 4PLM, which assumes that even high ability examinees can make mistakes (e.g. due to careless), can be preferred (de Ayala, 2009). Other approaches include techniques that were not originally developed for the purpose of DIF detection but are also useful in DIF detection. The most widely used of these is the Logistic Regression (LR) method. In this method, item scores are tried to be predicted with total score and the interaction of the grouping membership and the test score (Zumbo, 1999).

Rogers and Swaminathan (1993) compared MH and LR methods using simulated data sets and reported that MH method showed a good performance in detecting DIF in easy and hard tests, but it was not powerful enough in medium difficulty tests. They found that LR is a good method as MH in detecting uniform DIF and even it gives better results than MH in detecting non-uniform DIF. In their study with simulated data, Narayanan and

Swaminathan (1994) stated that MH and SIBTEST were similarly effective in determining uniform DMF and that both statistics were affected by the sample size.

Doğan and Öğretmen (2008) compared Mantel-Haenszel (MH), logistic regression (LR), and Chi-square techniques in determining DIF according to gender using data of science subtests of Secondary Schools Exam (SSE), and they found that the technique that determined the most bias was MH, and that no DIF was found in any item with the LR technique. Gök, Kelecioğlu, and Doğan (2010) examined whether maths and science subtests of Secondary Schools Exam (SSE) showed DIF according to gender and school type. They compared Mantel-Haenszel (MH) and logistic regression (LR) techniques for this purpose, and in maths subtest according to gender, negligible DIF was found in nine items with MH method and in three items with LR method. According to school type, on the other hand, they found DIF in a total of 19 items out of 25 items, including type A DIF in 15 items, type B DIF in 3 items, and type C DIF in 1 item. In terms of science subtest, on the other hand, negligible DIF was found according to gender in 10 items with MH method and 9 items with LR method. In the comparisons according to school type, type A DIF was found in 14 items with MH method and type B DIF was found in 1 item. It was concluded that there was DIF in 18 items with LR method. Çıkrıkçı Demirtaşlı and Ulutaş (2015) examined measurement invariance of PISA 2006 science literacy items in Turkish and American students using multi-group confirmatory factor analysis (MCFA) at first, and then, they investigated items showing DIF using MH, Simultaneous Item Bias Test (SIBTEST) and IRT-based likelihood ratio. In the study where 38 items that showed DIF by each of the three methods were accepted as having DIF, expert opinions were consulted and after these opinions, it was decided that nine items had DIF. Koyuncu, Aksu and Kelecioğlu (2018) compared MH, LR, and likelihood ratio methods by using PISA 2012 data on different computer software. According to different DIF determination methods, the number of items having DIF was found to vary between 1 and 5. Ayva Yörü and Atar (2019) used SIBTEST, Breslow-Day, Lord's $\chi 2$ and Raju's area measurement methods on the data of Centralized High School Entrance Placement Test (HSEPT) in 2012 according to gender and school type, and they found that the number and DIF levels of the items with DIF differed depending on the different methods. In line with the findings, they suggested that at least two methods should be used to determine the DIF.

It is seen that the studies in the literature are generally focused on nation-wide and large-scale tests. But a valid and unbiased measure is required not only in nation-wide and

large-scale tests, but also in classroom measurements. Therefore, more studies are needed to examine a teacher-made test in the context of differential item functioning. In this regard, the aim of this study is to examine and compare whether the items in a teacher-made test have DIF according to gender and university variables, using classical test theory-based delta plot, Mantel-Haenszel, Logistic Regression, and item response theory based Lord χ^2 techniques.

Method

This study aims to investigate whether items in a teacher-made test show DIF, according to different methods. In this context, it was examined whether item features varied according to different groups, therefore, a correlational research was conducted.

Study Group

The data of research were collected from 435 students who took the lesson of measurement and evaluation in summer semester of 2019-2020 school year. Of these students, 126 were male and 309 were female. Of the participants, similarly, 129 students are studying at the Faculty of Education of Pamukkale University (PAU), and 309 students at the faculties of education of other universities. In their studies examining DIF, Narayanan and Swaminathan (1994) underlined that while determining DIF, focal group size is more effective than the sample size of the reference group. Therefore, male students and the students who study in PAU were regarded as reference group, and female students who formed a larger group and students who study in other universities were regarded as focal group. IRT-based DIF determination methods, on the other hand, require the items to be calibrated according to an IRT model. In their simulation study, Şahin and Anıl (2016) suggested that a sample size of 150 was deemed acceptable in one-parameter logistic model (1PLM) for all test lengths (10, 20, and 30 items). In the light of simulation studies and considering that the test in question is a teacher-made test, it was decided that the size of the research group was deemed acceptable for such a study.

Measurement Tool Used in the Study

Achievement test

Data were collected using a 50-item multiple-choice achievement test. Achievement test contains the subjects of measurement and evaluation in education. The subjects covered in the test and the number of items prepared for the subjects were as follows: Basic concepts (15), error in measurement (4), required qualifications of measurement tools (18), and

classical measurement tools (13). Internal consistency coefficient of the scores obtained from the test was KR-20 .835 [.812 - .856]. Confidence interval values of KR-20 were calculated according to F distribution, as suggested by Feldt, Woodruff and Salih (1987). In this method, the upper bound of the KR-20 coefficient was calculated using the formula 1 – [(1 – KR-20).F(α/2)] while the lower bound with 1 – [(1 – KR-20).F(1 – α/2)]. In this regard, it is seen that the data obtained from the test were fair reliable. Test items are scored as 1 for correct responses and 0 for incorrect-blank responses.

Data Collection

Data were collected from a final exam using learning management system from students taking the course. Students were given 60 minutes to practice. To reduce systemic problems, the items were given to the students in groups of five and after the first five items were responded, the next five items were enabled to appear on the screen. Within the time given, the students were able to return to the item they wished and reviewed their answers. The present study also has ethics committee approval with decision no: E.93803232-622.02-13963.

Data Analysis

The data was analyzed using the ShinyItemAnalysis (Martinková & Drabinová, 2016) package on R 4.0.2 (R Core Team, 2020). ShinyItemAnalysis is an R package that can calculate tests and their item statistics, perform IRT calibrations and differential item functioning, and differential distractor functioning analyses. Since it is a package that can meet all the data analysis needs of the research, this package was decided to be used in this study.

While determining whether test items show DIF, several techniques were used, such as delta plot, Mantel-Haenszel, Logistic Regression, and IRT-based Lord $\chi 2$ methods. Of these techniques described in the introduction, as Lord $\chi 2$ was based on IRT, analyses need to be performed based on an IRT model. Because it is a simpler model, the analyzes were conducted on the basis of the Rasch model. The Rasch model asserts that the probability of an item being responded correctly depends only on the difficulty level of the item and that item discrimination is same (1.00) for all items (de Ayala, 2009).

Results

Findings of Item Analysis

Item difficulty (p) and item discrimination (r_{jx}) indexes of the achievement test were calculated and given in Table 1:

Table 1. *Item statistics for achievement test*

Item no	p	r_{jx}	Item no	p	r_{jx}
1	.48	*.10*	26	.72	.50
2	.94	.23	27	.92	.37
3	.73	.25	28	.60	*.15*
4	.77	*.03*	29	.64	.33
5	.87	.33	30	.83	.36
6	.61	.23	31	.94	.32
7	.79	.54	32	.17	*.14*
8	.93	*.19*	33	.96	.27
9	.86	.38	34	.86	.46
10	.77	.51	35	.82	.47
11	.70	.37	36	.69	.30
12	.79	.44	37	.78	.62
13	.86	.56	38	.46	.24
14	.33	*.18*	39	.72	.45
15	.21	*-.08*	40	.85	.54
16	.60	.25	41	.97	.25
17	.81	.49	42	.23	*.03*
18	.60	.40	43	.29	*.00*
19	.04	*-.19*	44	.87	.33
20	.79	.54	45	.20	*.06*
21	.45	*.16*	46	.72	.30
22	.08	*-.09*	47	.84	.55
23	.64	.30	48	.58	*.12*
24	.07	*.05*	49	.49	.26
25	.89	.46	50	.16	*.09*

Very easy item — Very difficult item

Analyzing item statistics, it was found that the discrimination indexes of 15 items in total (Items 1, 4, 8, 15, 19, 21, 22, 24, 28, 32, 42, 43, 45, 48, and 50) were less than .20. These items were insufficient in the discrimination of successful and unsuccessful students. However, it was seen that items 8, 22, 24, 32, and 50 were very easy (correctly responded by most of the group) or very difficult (incorrectly responded by most of the group). It should be taken into consideration that the items responded correctly or incorrectly by most of the group may have been insufficient in discriminating successful and unsuccessful students. The average difficulty of the test was found to be .64. Accordingly, it can be stated that the test is of medium difficulty.

Item Functioning Results That Vary According to Gender and School

Table 2 shows the findings obtained as a result of DIF analyses by gender.

Table 2. *DIF Statistics by Gender*

Item no	Delta Plot	MH	LR χ^2	p_{LR}	Lord χ^2	p_L	Favored Group
1	-.21	.99	.75	.69	.94	.33	
2	.27	.33	5.72	.06	1.16	.28	
3	.23	.90	1.47	.48	.42	.51	
4	-.53	.99	.56	.76	1.11	.29	
5	-.32	.94	.88	.64	.05	.82	
6	-.08	.88	.04	.98	.14	.71	
7	.24	.93	.55	.76	.67	.41	
8	.57	.43	3.44	.18	2.74	.10	
9	.08	.97	.51	.77	.35	.55	
10	-.09	.46	.32	.85	.00	.96	
11	.49	.62	3.68	.16	1.50	.22	

Table 2. *DIF Statistics by Gender (cont.)*

Item no	Delta Plot	MH	LR χ^2	p_{LR}	Lord χ^2	p_L	Favored Group
12	.06	.54	1.46	.48	.15	.70	
13	-.18	.97	2.69	.26	.00	.99	
14	.45	.49	.66	.72	.20	.65	
15	-.92	.03	5.60	.06	8.97	.00	M (MH, Lord)
16	-.59	.09	2.64	.27	2.82	.09	
17	.48	.35	2.10	.35	1.99	.16	
18	-.03	.15	1.02	.60	.07	.79	
19	-.29	.88	1.47	.48	1.53	.22	
20	-.36	.34	1.38	.50	.36	.55	
21	.15	.68	2.86	.24	.01	.92	
22	-.18	.99	2.26	.32	1.32	.25	
23	-.13	.48	.82	.66	.21	.65	
24	.20	.81	.04	.98	.20	.65	
25	-.18	.93	1.92	.38	.02	.89	
26	.46	.35	1.19	.55	1.44	.23	
27	-.87	.08	1.49	.48	.91	.34	
28	.82	.03	5.80	.06	3.54	.06	F (MH)
29	-.34	.11	2.03	.36	.96	.33	
30	.51	.14	3.16	.21	2.31	.13	
31	.01	.82	.42	.81	.40	.53	
32	.03	.93	.31	.86	.58	.45	
33	-.08	.77	3.70	.16	.25	.62	

Table 2. *DIF Statistics by Gender (cont.)*

Item no	Delta Plot	MH	LR χ^2	p_{LR}	Lord χ^2	p_L	Favored Group
34	-.67	.16	1.93	.38	.98	.32	
35	-.43	.62	8.35	.02	.41	.52	M (LR)
36	.90	.02	8.58	.01	5.08	.02	F (MH, LR, Lord)
37	.56	.10	2.46	.29	2.44	.12	
38	.52	.86	5.02	.08	.72	.40	
39	.10	.93	.10	.95	.08	.77	
40	.72	.21	5.75	.06	4.07	.04	
41	.41	.42	1.70	.43	1.58	.21	
42	-.61	.28	4.94	.09	4.99	.03	
43	-.44	.77	1.57	.46	3.22	.07	
44	-1.74	.01	1.83	.00	7.45	.01	M (Delta, LR, Lord), F (MH)
45	.20	.99	.10	.95	.13	.72	
46	-.18	.65	.16	.92	.16	.69	
47	.11	.91	1.40	.50	.37	.54	
48	.56	.30	3.20	.20	1.40	.24	
49	.08	.89	1.66	.44	.05	.82	
50	.24	.92	3.32	.19	.09	.77	

Examining the table, item 44 showed DIF according to all methods. Item 36 showed DIF according to all methods other than delta plot. Item 15 showed DIF according to MH and LR methods, while item 28 and 35 showed DIF according to MH method only and LR method only, respectively. Examining item 44, where all methods indicate a DIF by gender, the item is given below:

> When it is mentioned about a correlation between preservice teachers' performance scores in the lecture of teaching practices and the scores that they will get based on their future teaching performances, which of the following types of validity is pointed out?
>
> A. Content validity
> B. Predictive validity
> C. Convergent validity
> D. Construct validity

E. Face validity

When the statements in the item were examined, it was seen that there was no factor that might cause DIF by gender. Similarly, item 36, where three methods indicated the presence of DIF, is given below:

> Which of the following options shows the measurement tool with the least scoring error compared to the others?
>
> A. Written exams
> B. Short-answer type of tests
> C. Multiple-choice type of tests
> D. Oral exams
> E. Homework and projects

This item was observed not to involve any statement that could cause DIF by gender. Item 15 showing DIF according to two methods is given below:

> Which of the following measurement results can be obtained with ratio scale?
>
> A. Parents' education level
> B. Student's performance in Turkish lesson
> C. The number of children in the playground
> D. Temperature value of the patient whose fever was measured
> E. Students' behaviors towards self-assessment

It is understood that the statements in the root and options of this item do not contain expressions that may cause DIF by gender. Items showing DIF by school were examined and the findings obtained were given in Table 3.

Table 3. *DIF Statistics by School*

Item no	Delta Plot	MH	LR χ^2	p_{LR}	Lord χ^2	p_L	Favored Group
1	.66	.11	3.92	.14	1.48	.22	
2	-.22	.82	.38	.83	.14	.71	
3	-.70	.08	9.61	.01	2.23	.14	PAU (LR)
4	-.63	.60	4.08	.13	1.37	.24	
5	-.39	.84	.88	.64	.06	.80	
6	.36	.11	.68	.71	.57	.45	
7	.26	.65	5.29	.07	.90	.34	
8	-.02	.88	1.18	.56	.49	.48	
9	.03	.90	.72	.70	.37	.54	
10	-.60	.09	3.65	.16	1.29	.26	
11	-.45	.37	2.72	.26	1.05	.31	
12	.71	.15	4.26	.12	4.18	.04	Other (Lord)
13	-.37	.48	.86	.65	.07	.79	
14	.18	.97	.75	.69	.12	.73	
15	-.20	.40	1.84	.40	2.24	.13	
16	-.49	.07	3.70	.16	1.97	.16	
17	.70	.05	4.24	.12	4.20	.04	Other (Lord)
18	.33	.82	1.01	.60	.42	.51	
19	.45	.82	.64	.73	.19	.66	

Table 3. *DIF Statistics by School (cont.)*

Item no	Delta Plot	MH	LR χ^2	p_{LR}	Lord χ^2	p_L	Favored Group
20	.41	.47	1.34	.51	1.67	.20	
21	-.27	.35	1.59	.45	1.49	.22	
22	-.59	.04	9.73	.01	5.49	.02	PAU (MH, LR, Lord)
23	-.31	.55	1.82	.40	.71	.40	
24	-.87	.02	7.99	.02	8.09	.00	PAU (MH, LR, Lord)
25	-.64	.48	1.93	.38	.38	.54	
26	.61	.20	4.25	.12	2.71	.10	
27	-.83	.42	1.90	.39	.57	.45	
28	-.15	.35	2.24	.33	.31	.58	
29	.43	.47	1.12	.57	1.03	.31	
30	.50	.14	2.39	.30	2.56	.11	
31	-.50	.96	.97	.62	.01	.93	
32	-.13	.17	2.37	.31	1.94	.16	
33	-.39	.81	.57	.75	.03	.87	
34	-.29	.75	2.73	.26	.01	.92	
35	.04	.77	.10	.95	.27	.61	
36	.88	.02	6.34	.04	5.21	.02	Other (MH, LR, Lord)
37	.89	.00	13.91	.00	6.04	.01	Other (MH, LR, Lord)
38	.10	.60	.10	.95	.07	.79	
39	-.32	.44	1.59	.45	.44	.51	
40	.45	.17	1.99	.37	2.26	.13	

Table 3. *DIF Statistics by School (cont.)*

Item no	Delta Plot	MH	LR χ^2	p_{LR}	Lord χ^2	p_L	Favored Group
41	-.33	.95	2.35	.31	.09	.76	
42	-.80	.00	1.80	.01	8.56	.00	PAU (Lord)
43	.41	.51	.23	.89	.01	.92	
44	.57	.06	3.18	.20	3.28	.07	
45	.05	.61	2.94	.23	.87	.35	
46	.52	.29	5.30	.07	2.01	.16	
47	.45	.20	2.65	.27	2.26	.13	
48	.35	.70	1.04	.59	.43	.51	
49	.54	.36	2.38	.30	.85	.36	
50	-.43	.07	4.92	.09	4.37	.04	PAU (Lord)

Examining the items showing DIF between PAU and other universities, it was observed that items 22, 24, 36, 37, and 42 showed DIF according to three methods. These items are given below:

Item 22:

Which of the following is an example of a random error in a test?

A. Making unintentional mistakes while summing up the scores in the student sheets
B. Adding 5 points to each students sitting at front desks and listening to lecture
C. Adding 10 points to all students tested
D. Adding 10% of their score to each student
E. Breaking 5 points from each Student who break the rules of classroom

Item 24:

Which of the following comments would be correct for a teacher, who found KR-21 coefficient for test scores as 0.8 to make?

A. Students have achieved 80% success
B. The desirable characteristic was measured
C. There was a 20% error in exam scores
D. The internal consistency of the exam is high
E. The distribution of questions to course subject is balanced

Item 36:

Which of the following options shows the measurement tool with the least scoring error, compared to the others?

A. Written exams
B. Short-answer type of tests
C. Multiple-choice type of tests
D. Oral exams
E. Homework and projects

Item 37:

Which of the following would not be correct for a teacher to determine as a measurement rule in a measurement procedure he/she applies?

A. Answers left blank are not given points
B. 10 points will be given if the first question is responded correctly.
C. 2 points will be broken for each spelling rule error
D. Students with end-term mean score of 60 pass the course
E. The indents will definitely be typed by leaving 8-character blank.

Item 42:
I. Standard deviation
II. Mean
III. KR-21 coefficient.
Which one(s) of the above-mentioned statistics should a teacher, who wants to calculate the standard error of measurement for scores of an exam, use?

A. Only I
B. I and II
C. I and III
D. II and III
E. I, II and III

When the statements in item root and options are examined, it is seen that there were no statements that can create bias by universities. However, since the measurement tool was applied online, the familiarity of university students with the environment in which the exam was administered should be considered. But still, since these items were presented in the same environment and similarly with other items, it was decided that this variable would not cause DIF either.

Discussion

In this study, which aims to examine whether the items of a teacher-made achievement test show DIF by gender and university, one item (item 44) was found to be biased according to all DIF determination methods by gender. When examining DIF by university, on the other

hand, it was seen that none of the items showed DIF according to Delta method, while five items (22, 24, 36, 37, and 42) showed DIF according to three methods. But when the items were analyzed, it was observed that there were no statements that can create bias by gender or university. Low compatibility of DMF-showing items with different methods is consistent with the findings of Rogers and Swaminathan (1993), Gierl, Khaliq and Boughton (1999), Doğan and Öğretmen (2008), Gök et.al. (2010). Besides, Gierl et.al. (1999) stated that the MH method is more conservative in determining DIF compared to SIBTEST and LR, and that less items showed DIF. Although SIBTEST method was not included in the scope of this research; comparing MH and LR methods, it was found that MH method marked more items with DIF, especially by gender.

The lack of the number of students taking the teacher-made tests has focused the DIF studies on data of tests that allow studying with larger sample sizes, such as PISA, TIMMS, PIRLS, and inter-institutional transfer tests applied at a national scale (Ayva Yörü and Atar, 2019; Çıkrıkçı Demirtaşlı and Ulutaş, 2015; Doğan and Öğretmen, 2008; Gök et.al., 2010; Koyuncu et.al., 2018).

Conclusion

Students are being deemed successful-unsuccessful through the scores they took from the teacher-made tests and they may need to repeat the grade level, as occasion requires. Although these tests are not intended to guide educational policies of large-scale tests, while the "passed/failed" decision is made about the students, using the scores taken from a test biased towards a particular group as a base would prevent us giving a valid decision. In this context, conducting more DIF studies on teacher-made tests and determining the factors that may cause DIF in these tests would also contribute to teacher education.

In our study, we used delta plot, MH, LR and Lord $\chi 2$ techniques to determine DIF. Further studies which would use other DIF determination methods and IRT calibration methods is needed. Additionally, the greatest limitation of this study was that the sample size was smaller than large-scale tests. Similar studies can be conducted with teacher-made tests with larger samples. Koyuncu et.al. (2018) examined different DMF methods using different software. This study is limited to ShinyItemAnalysis software only. Further studies can be carried out, where other software is tested in the teacher-made tests. Although some items showed DIF in this study, when the items were examined, no statements that can create DIF were found. It is thought that analyzing items showing DIF by consulting expert

opinions would also contribute to the content of the measurement and evaluation courses, taught in the faculties of education, and in-service trainings made for test development.

Ethical Approval: *This research was conducted with the permission of the Pamukkale University ethics committee with the decision no E.93803232-622.02-13963 dated 03.02.2021*

Conflict of Interest: *Authors have no conflict of interest to declare.*

Author Contributions: *The contribution of the first author to the study was at the determining the problem statement and data analysis phases. The second author was contributed at the determining the problem statement and study group, and data collection, and the third author was contributed at the data analysis of the study. Besides, all the authors contributed to the literature review, discussion, and reporting.*

References

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp.96-116). Baltimore: Johns Hopkins University.

Ayva Yörü, F.G. & Atar, H.Y. (2019). Determination of differential item functioning (DIF) according to SIBTEST, Lord's χ2, Raju's area measurement and Breslow-Day methods. *Journal of Pedagogical Research, 3*(3), 139-150. doi: 0.33902/jpr.v3i3.137

Crocker, L. & Algina, J. (2008). *Introduction to classical and modern test theory*. Ohio: Cengage Learning.

Çıkrıkçı Demirtaşlı, N. & Ulutaş, S. (2015). A study on detecting of differential item functioning of PISA 2006 science literacy items in Turkish and American samples. *Eurasian Journal of Educational Research, 58*. 41-60.

de Ayala, R.J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

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

Feldt, L.S., Woodruff, D.J. & Salih, F.A. (1987). Statistical inference for coefficient alpha. *Applied psychological measurement, 11*(1), 93-103.

Gamerman, D., Goncalves, F. B. & Soares, T. M. (2018). Differential item functioning. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 3, pp. 67–86). Boca Raton, FL: CRC Press.

Gierl, M., Khaliq, S. N., & Boughton, K. (1999, June). Gender differential item functioning in mathematics and science: Prevalence and policy implications. In *Annual Meeting of the Canadian Society for the Study of Education*, Canada.

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. (1985). *Item response theory: Principles and applications*. New York: Springer Science+Bussiness Media.

Karami, H. (2012). An introduction to differential item functioning. The International *Journal of Educational and Psychological Assessment, 11*(2), 59–76.

Koyuncu, İ., Aksu, G. & Kelecioğlu, H. (2018). Comparison of Mantel-Haenszel, logistic regression and likelihood ratio methods to evaluate differential item functioning by using different computer software. *Elementary Education Online*, *17*(2). 909-925. doi: 10.17051/ilkonline.2018.41933

Lord, M.D. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrance Erlbaum Associates.

Martinková, P. & Drabinová, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, 10(2), 503-515. doi: 10.32614/RJ-2018-074

McNamara, T. & Roever, C. (2006). *Language testing: The social dimension*. Massachusetts: Blackwell Publishing.

Millsap, R.E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Narayanan, P. & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*(4). 315-328.

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

Popham, W.J. (2014). *Classroom assessment: What teachers need to know*. Boston: Pearson.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reise, S.P. & Revicki, D.A. (2015). *Handbook of item response theory modeling*. New York: Routledge.

Rogers, H. J. & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

Şahin, A. & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice, 17*(1). 321-335. doi: 10.12738/estp.2017.1.0270

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