Ocak 2015 Cilt:23 No:1 Kastamonu Eğitim Dergisi 227-240

PISA 2009 ÖĞRENCİ ANKETİ ALT ÖLÇEKLERİNDE (Q32-Q33) BULUNAN MADDELERİN DEĞİŞEN MADDE FONKSİYONU AÇISINDAN İNCELENMESİ

İbrahim Alper KÖSE
Abant İzzet Baysal Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü

İlk Kayıt Tarihi: ***11.02.2014*** **Yayına Kabul Tarihi: 09.05.2014**

Özet

Bu araştırmada PISA-2009 öğrenci anketinde bulunan okula ve öğretmenlere ilişki algı alt testlerinde yer alan maddelerin DMF açısından 4 farklı ülke (Türkiye, ABD, İrlanda ve İngiltere) ve cinsiyete göre ele alınarak karşılaştırmalı olarak araştırılması amaçlanmıştır. Araştırmada kullanılan maddeler çoklu puanlandığı için DMF analizinde aşamalı tepki modeli altında kullanılan olabilirlik oranı testi (likelihood ratio test) tercih edilmiştir. DMF analizlerine geçmeden önce öncelikle MTK'nın tek boyutluluk varsayımı test edilmiş ve veri gruplarına ayrı ayrı DFA uygulanmıştır. DFA analizleri sonucunda alt testler için test edilen tak faktörlü ölçme modelinin doğrulandığı ve MTK'nın tek boyutluluk varsayımının karşılandığı ortaya konmuştur. Ölçeklerdeki maddelerde DMF bulunup bulunmadığı MULTILOG programı yazılımı ile analiz edilmiş ve okul öğrenmelerine ilişkin algı alt ölçeğinde cinsiyete göre 2, IRL-ING örneklemlerinde 1, ABD-İngiltere'de 2 ve ABD-Türkiye'de 4 maddede, öğretmenlere yönelik algı alt ölçeğinde ise cinsiyete göre 1, İrlanda-İngiltere örneklemlerinde 1, ABD-İngiltere'de 2 ve ABD-Türkiye'de 4 maddede DMF bulunduğu istatistiksel olarak ortaya konmuştur.

Anahtar Kelimeler: *PISA, Değişen madde fonksiyonu, Algı, Aşamalı Tepki Modeli*

INVESTİGATİON OF ITEMS İN PISA 2009 STUDENT QUESTİONNAİRE SUBSCALES (Q32-Q33) İN TERMS OF DİFFERENTİAL ITEM FUNCTİONİNG

Abstract

This study was aimed to assess items of PISA-2009 student questionnaire's subscales of "sense of school learnings" and "sense of teachers" in terms of differential item functioning by using gender and 4 different country data comperatively. For the DIF analysis, likelihood ratio test was preferred in graded response theory based on item response theory (IRT) because of polytomously scored items. Confirmatory factor analysis (CFA) was used for data sets saperately to test the unidimensionality assumption of IRT. CFA analysis showed that unifactorial construct of subtests was confirmed and unidimensionality assumption of IRT was met. DIF analysis were carried out by MULTILOG software and 2 items for gender, 1 item for Ireland-GBR sample, 2 items for USA-GBR sample and 4 items for Turkey-USA sample

was flagged as DIF in the subscale of "sense of school learnings". In the subscale of "sense of teachers", 1 item for gender, 1 item for Ireland-GBR sample, 2 items for USA-GBR sample and4 items for USA-Turkey sample were flagged as DIF.

Keywords: *PISA, Differential Item Functioning, Attitude, Graded Response Model*

1. Giriş

Ekonomik İşbirliği ve Kalkınma Teşkilâtı - OECD (Organisation for Economic Co-Operation and Development) tarafından düzenlenen PISA, öğrencilerin, matematik, fen ve okuma becerileri alanlarındaki bilgi ve becerilerinin değerlendirildiği uluslararası en büyük eğitim araştırmalarından biridir. Üç yılda bir yapılan bu araştırmayla, OECD üyesi ülkeler ve diğer katılımcı ülkelerdeki (dünya ekonomisinin yaklaşık olarak %90'ı) 15 yaş grubu öğrencilerin modern toplumda yerlerini alabilmeleri için gereken temel bilgi ve becerilere ne ölçüde sahip oldukları değerlendirilmektedir.

Öğrencilerin üç temel konu alanındaki bilgi ve becerilerini değerlendirmenin yanında, PISA projesinde öğrencilerin öğrenme stratejileri, problem çözme becerileri ve değerlendirilen konu alanına yönelik ilgi ve tutumları da araştırılmaktadır. Okuma becerilerinin ağırlıklı alan olarak ele alındığı PISA 2009'da, öğrencilerin okuma etkinliklerine katılımına, kendi okuma ve öğrenme stratejileri hakkındaki düşüncelerine odaklanılmıştır. PISA 2009'a önce 33'ü OECD üyesi olmak üzere toplam 65 ülke, daha sonra 9 ülkenin katılımı ile 74 ülke katılmıştır.

TIMMS, PIRLS ve PISA sınavları gibi eğitim ve psikolojide çeşitli amaçlarla ölçme araçları kullanılmaktadır. Kullanılan bu araçlar, test alanları eşit şartlarda sınaması, herhangi bir gruba avantaj ya da dezavantaj sağlamaması gerekmektedir. Bu amacın gerçekleşmesi için ölçme araçlarının psikometrik özellikleri iyi incelenmesi gerekmektedir. Ölçme araçlarının psikometrik özelliklerinin yeterli düzeyde olmaması bu araçlara dayalı olarak verilen kararların doğruluğunu da tartışmalı hale getirecektir.

Ölçme araçlarında bulunması gereken en önemli özelliklerden birisi de geçerliktir. Geçerlik ölçme sonuçlarının gerçeği yansıtma derecesi olarak kısaca tanımlanabilir. Ölçme sonuçlarının değerlendirilmesinde geçerlik sağlanmadan, elde edilen sonuçlar veya bu sonuçlara dayalı alınan kararlar da anlamsız olacaktır. Bilgi ve teknolojideki ilerlemeler test sonuçlarının geçerliği ile ilgili düşüncelerin de değişmesine neden olmuştur. Geleneksel geçerlik çalışmalarında, çeşitli geçerlik türleri ön planda olmasına rağmen (kapsam ve ölçüt dayanaklı geçerlik), son çalışmalar (Messick, 1989) yapı geçerliğinin ön planda olması gerektiğini göstermektedir. Yapı geçerliği çalışmalarını madde analizleri izlemelidir. Geçerlik çalışmalarında kullanılan önemli çalışmalardan birisi de yanlılıktır (Zumbo, 1999). Yanlılık kavramı, test yanlılığı ve madde yanlılığı olmak üzere ikiye ayrılmaktadır.

Test yanlılığının tarihi gelişimi incelendiğinde, Alfred Binet'in 1910 yılında alt eko-

nomik statüdeki çocuklar üzerinde yapmış olduğu çalışmalar göze çarpmaktadır. Bu çalışmada Binet test maddelerini incelemiş ve bazı maddelerin zihinsel kapasiteden daha çok, kültürel yetişme tarzlarına odaklandığını ortaya koymuştur (Camilli ve Shepard, 1994). Binet'i daha sonra Cleary'nin (1968) çalışmaları izlemektedir. Bilindiği üzere test performansı ve ölçüt performanslar arası ilişkiler, regresyon eşitlikleri yardımı ile ortaya konmaktadır. Cleary (1968), yapmış olduğu çalışmada regresyon doğrusu üzerinden yordanan ölçüt puanların alt örneklemlerde çok yüksek veya çok düşük olduğunu bularak, ilk olarak test yanlılığı kavramını ortaya atmıştır (Aktaran; Lee, 2003).

Test yanlılığı, testle ölçülen özelliklerin dışında belli bir guruba ait ölçme sonuçlarının geçersizliği veya sistematik hata karışması olarak tanımlanabilir. Bir başka tanımlama ile de, test dışı faktörlerden birisi olan, testin kapsamının alt gruplarda farklılık göstermesinin ortaya çıkardığı varyans, belirli alt gruplarda farklılık yaratıyorsa bu durum testin yanlı olduğu anlamına gelmektedir (Angoff, 1982). Test yanlılığı grup için geçerli bir kavram olmakla beraber grubun içindeki bir birey için anlamlı bir kavram değildir. Test yanlılığı genellikle farklı etnik gruplar veya farklı cinsiyet grupları için söz konusudur. Test yanlılığının sebepleri test dışı faktörlerden ve testin içindeki maddelerden kaynaklanabilir. Test dışı faktörler yordama geçerliği modellerinden, test içi faktörler; testte bulunan maddelerin yapısından kaynaklanabilir (Camilli ve Shepard, 1994). Bu maddelerdeki farklı yapılar madde yanlılığı kavramını ortaya çıkarmaktadır.

Madde yanlılığı, aynı yetenek düzeyindeki bireylerin oluşturduğu alt gruplarda bir maddenin doğru cevaplanma olasılığının farklı olmasıdır (Ackerman, 1992; Camilli ve Shepard, 1994; Bolt, 2002). Madde yanlılığı, 1970'li yıllardaki test yanlılığının ortaya konması için gerekli ölçüt kriterlerinin karşılanması zorluğundan yola çıkılarak ortaya atılmış bir kavramdır. Madde güçlüğünün alt gruplarda farklılaşmasından (bağıl güçlük-relative difficulty), madde yanlılığının kaynaklandığı ifade edilse de, bu farklı güçlüğün testin ölçtüğü yapıdan kaynaklanmamış olması gerekmektedir (Camilli ve Shepard, 1994).

Yanlılığın madde düzeyinde ortaya konması, değişen madde fonksiyonu (DMF) analizleri ile mümkündür. Bir maddede yanlılık olması, maddenin DMF içerdiğinin göstergesidir. Ancak DMF gösteren maddelerin yanlı olduğu kesin değildir (Kamata ve Vaughn, 2004). Angoff (1982) ve Zumbo (1999) madde yanlılığının sadece istatistiksel yöntemlerle belirlemenin yeterli olmayacağını, eğitim ve psikolojiye dayanan bulgularla desteklenmesi gerektiğini vurgulamışlardır. Madde yanlılığı belirleme yöntemleri, maddenin yanlılığının ortaya konmasında sadece bir adım olarak düşünülmelidir. Bir maddede değişen madde fonksiyonu (DMF) belirlenmişse, madde yanlılığı için bir işaret olarak algılanmalı, içerik analizi, olgusal değerlendirme…vb yöntemlerle madde yanlılığının varlığı ortaya konmalıdır.

İki tip DMF bulunmaktadır. Bunlar tek biçimli (uniform) ve tek biçimli olmayan (non-uniform) DMF olarak tanımlanır. Madde, belirli bir grubun bütün yetenek

düzeylerinde DMF gösteriyorsa tekbiçimli DMF, örneğin sadece bir grubun yüksek puan alan bireylerinde DMF gösteriyorsa tek biçimli olmayan DMF olarak tanımlanır (Kamata ve Vaughn, 2004; Van Dam, Earleywine ve Forstyh, 2009).

Yanlılık iki yöntemle ortaya konabilir. Bunlar, yargılara ve istatistiğe dayalı yöntemlerdir. Yargılara dayalı yöntem basit olarak uzman görüşünün alınmasıdır. Uzmanın testte bulunan maddeleri inceleyerek olası yanlı maddeleri belirlemesidir. İstatistiksel yöntemler ise potansiyel DMF analizlerinin yapılmasıdır (Zumbo, 1999).

DMF alanyazınında maddenin puanlanma biçimi önemli bir yer tutmaktadır. Test maddeleri iki ve çoklu puanlanan maddeler olmak üzere ikiye ayrılmaktadır. DMF yöntemlerinin çoğu iki kategorili (doğru-yanlış) cevaplanan maddeler için geliştirilmiştir. Ancak çoklu puanlanan maddeler için de DMF yöntemleri bulunmaktadır. Bu yöntemler, iki kategorili puanlanan maddeler için geliştirilmiş DMF yöntemlerinin bir uzantısı olarak kabul edilmektedir (Zumbo, 1999).

DMF analizlerinde klasik test kuramına ve madde tepki kuramına dayalı yöntemler geliştirilmiştir. DMF çalışmalarında madde tepki kuramının (MTK), klasik test kuramı üzerinde önemli avantajları vardır. İlk olarak MTK'da madde parametreleri (madde güçlüğü ve madde ayırt ediciliği), kestirildikleri örneklemden bağımsızdır. İkinci olarak, maddenin farklı alt gruplardaki değişen fonksiyonları, MTK'da daha kesin olarak belirlenebilmektedir. Son olarak da, MTK'da madde parametrelerinin grafiksel gösterimi hazır olarak verilmektedir, bu özellik DMF gösteren maddelerin anlaşılmasını kolaylaştırmaktadır (Camilli ve Shepard, 1994).

MTK'ya dayalı DMF belirleme yöntemlerinin çoğu 1-0 şeklinde ikili puanlanan maddeler için geliştirilmiştir. Yetenek ve başarı testlerinde kullanılan maddeler bu maddelere örnek olarak verilebilir. Çoklu puanlanan maddeler için MTK altında geliştirilmiş modeller ikili maddeler için geliştirilmiş modellere göre daha sınırlıdır (Cohen, Kimve Baker, 1993). MTK'da DMF analizleri temel olarak, madde parametrelerinin referans ve odak gruplarda farklılaşıp farklılaşmadığının incelenmesidir. Bu farklılaşma, maddenin doğru cevaplanma olasılığının referans ve odak gruplarda farklılaşması veya DMF olduğunu anlamına gelebilmektedir (Embretson ve Reise, 2000). Bu farklılaşma, madde karakteristik eğrilerinin her iki grupta incelenmesi ile ortaya konmaktadır. En temel anlamda MTK'da DMF, her iki grup için oluşturulan madde karakteristik eğrileri arasında kalan alanların hesaplanması ile belirlenmektedir (Hambleton ve Swaminathan, 1985; Lautenschalger, Flaherty ve Park, 1994; Meij-de Meij, Kelderman ve van der Flier, 2010).

Aşamalı Tepki Modelinde DMF

Çoklu puanlanan maddelerde MTK'ya dayalı olarak DMF analizleri için, Samejima'nın (1969) aşamalı tepki modeli (ATM) tercih edilmektedir. Çoklu puanlanan bir maddede, aşamalı tepki modeline göre eğer referans ve odak grupları için Madde Gerçek Puan Fonksiyonları-MGPF (Item True Score Fucti-

ons) farklı ise bu maddede DMF bulunduğu iddia edilir. Bu durum matematiksel olarak;

$T_{jR}(\theta) \neq T_{jO}(\theta)$ ile ifade edilir.

R……..Referans grup
O…….Odak grup.
MGPF ölçekteki herhangi bir madde için aşamalı tepki modelinde;

$$T_j(\theta) = \sum_{k=1}^{m_j} u_{jk} P_{jk}(\theta)$$

eşitliği ile ifade edilir. Eşitlikte;

u_{jk} ……..kategori numarasını,
j……..testteki herhangi bir maddeyi,
$P_{jk}(\theta)$….j maddesinin k kategorisinde tepki verme olasılığını ifade eder.

Aşamalı tepki modelinde çoklu puanlanan bir madde için bir tane madde ayırt edicilik parametresi ve kategori sayısının bir eksiği kadar madde eşik parametresi tanımlanır. DMF oluşabilmesi için üç farklı durumdan bir tanesinin gerçekleşmesi gerekir.

1. $a_{jR} \neq a_{jO} ve b_{jR} = b_{jO}$ madde ayırt edicilik parametreleri eşit değil, kategori eşik parametreleri eşittir.

2. $a_{jR} = a_{jO} ve b_{jR} \neq b_{jO}$ madde ayırt edicilik parametreleri eşit, kategori eşik parametreleri eşit değildir.

3. $a_{jR} \neq a_{jO} ve b_{jR} \neq b_{jO}$ madde ayırt edicilik parametreleri ve kategori eşik parametreleri eşit değildir (Cohen, Kim ve Baker, 1993).

Çoklu puanlanan maddelerde DMF analizleri için, poly-SIBTEST, DFIT ve the likelihood ratio (LR) yöntemleri sıklıkla kullanılmaktadır. Poly-SIBTEST ve DFIT yöntemleri, aynı yetenek düzeyindeki bireylerin maddeyi doğru cevaplama olasılıklarının farklılaşıp farklılaşmadığı esasına göre DMF analizlerini yapmaktadır.

Thissen, Steinberg ve Gerard (1986) tarafından ortaya atılan LR testi, DMF içerin maddeleri belirlemek için madde parametreleri arasındaki farklara odaklanan bir testtir. Bu yöntemde DMF analizi yapılan madde *çalışılan madde* (studied item) olarak tanımlanır. DMF analizinde her bir madde için temel (compact) ve sınırlandırılmış (augmented) modellerin uyum değerleri karşılaştırılır. Temel modelde, testte bulunan her maddenin parametreleri referans ve odak gruplarda eşit, sınırlandırılmış modelde ise sadece çalışılan maddenin parametrelerinin farklı olduğu varsayımına göre kestirilir. Her iki model arasındaki -2xloglikelihood değeri LR yönteminde G^2 değeri olarak hesaplanır. Bu G^2 değeri belirlenen anlamlılık düzeyindeki χ^2 değerini aşarsa, maddede DMF bulunduğu yönünde yorum yapılır (Bolt, 2002). Bu çalışmada MULTILOG programı altında LR yöntemi ile DMF analizleri yapılmıştır.

Bireyler hakkında bilgi toplamak amacıyla ölçme araçları kullanılmaktadır. Bu ölçme araçlarından elde edilen sonuçlar kullanılarak bireyler hakkında çeşitli kararlar verilmektedir. Bu kararların yerindeliği için ölçme araçlarından elde edilen sonuçların geçerliği ve güvenilir olması gerekmektedir. Yanlılık, ölçme araçlardan elde edilen puanların geçerliği ve güvenirliğini tehdit eden önemli unsurlardan birisidir. TIMMS, PIRLS ve PISA testleri de uluslararası düzeyde öğrenci edimlerini karşılaştıran güncel testlerdir. Bu testlerde bulunan maddelerin uygulandığı her ülkedeki öğrencilere aynı şekilde hitap edip etmediği veya belirli bir gruba yanlı olup olmadığı alanyazında süregelen bir tartışmadır. Bu araştırmanın genel amacı PISA 2009 öğrenci anketinde yer alan alt ölçeklerde (Q32-Q33) bulunan maddelerinin değişen madde fonksiyonu açısından incelenmesidir.

2. Yöntem

Araştırmanın Amacı

Bu araştırmanın amacı PISA 2009'da yer alan öğrenci anketinde yer alan alt ölçeklerdeki (Q32-Q33) maddelerin aşamalı tepki modeli altında değişen madde fonksiyonu açısından incelenmesidir. Araştırmada cinsiyet ve farklı dil ve kültürler bağımsız değişkenler olarak seçilmiştir.

Araştırma Grubu

Bu çalışmada araştırma grubu olarak PISA 2009 uygulamasına katılan Türkiye, Amerika Birleşik Devletleri (ABD), İngiltere ve İrlanda örneklemlerindeki öğrenciler seçilmiştir. Bu ülkelerin seçilmesinin nedeni dilsel ve kültürel olarak benzerlik ve farklılıkları karşılaştırabilmektir. İngiltere-İrlanda ile aynı dil-benzer kültür, ABD-İngiltere ile aynı dil-farklı kültür, ABD-Türkiye ile farklı dil-farklı kültür karşılaştırmaları yapılmıştır. Bunun yanında tüm ülkelerden elde edilen bir veri grubu ile cinsiyete göre maddelerin DMF gösterip göstermediği incelenmiştir. Araştırmada seçilen ülkelere ve cinsiyete göre öğrenci sayıları Tablo 1.'de verilmiştir.

Tablo 1. Çalışma Grubu

	Ülkeler	Frekans	%			Frekans	%
ÜLKE	Türkiye	4522,00	18,97	CİNSİYET	Kız	11994	50,3
	ABD	4846,00	23,33				
	İngiltere	11107,00	46,59		Erkek	11844	49,7
	İrlanda	3364	14,11				
	Toplam:	23838,00	100,00		Toplam	23838	100,0

Veri Toplama Aracı

Bu çalışma OECD 'nin 2009 yılında düzenlemiş olduğu Uluslar arası Öğrenci Değerlendirme Programı (PISA)'nın öğrenci anketinde bulunan "okul öğrenmelerine ilişkin algı (Q32 kodlu-4 madde)" ve "öğretmenlere yönelik algı (Q33 kodlu-5

madde)" alt ölçeklerinde bulunan 9 madde üzerinden yürütülmüştür. Ölçekte bulunan maddeler dörtlü likert tipinde olup, maddelere tamamen katılıyorum, katılıyorum, katılmıyorum ve tamamen katılmıyorum düzeyinde tepki verilebilmektedir.

Verilerin Analizi

Bu çalışmada PISA-2009 öğrenci anketinde bulunan Q32 ve Q33 kodlu alt ölçeklerde bulunan maddelerin DMF içerip içermediğinin ortaya konması amaçlanmıştır. Alt ölçeklerde bulunan maddeler çoklu puanlandığı için, DMF analizleri MTK altında ATM kullanılarak MULTILOG programı yardımıyla yapılmıştır. Araştırmada anlamlılık seviyesi .05 olarak seçilmiştir. DMF analizlerine geçilmeden önce MTK varsayımlarının test edilmesi gerekmektedir. Bu varsayımlar tek boyutluluk ve yerel bağımsızlıktır. Tek boyutluluk varsayımı, test performansını etkileyen tek bir başat bileşen veya faktörün olması şeklinde tanımlanabilir (Hambleton, Swaminathan ve Rogers, 1991). Bu varsayımın ihlal edilmesi, veri grubunu açıklayan modelin yetersizliğine ve elde edilen sonuçların tartışılmasına yol açacaktır. Bu da ölçeğin yapı geçerliği üzerinde önemli bir tehdittir (Sheng, 2005). İkinci varsayım olan yerel bağımsızlık, cevaplayıcının testteki bir maddeye vermiş olduğu tepkinin, testteki diğer maddelerden etkilenmemesi ya da bağımsız olması anlamına gelmektedir (Sijtsma & Hemker, 2000). McDonald (1982) yerel bağımsızlık varsayımının, tek boyutluluk varsayımına dayandığını belirtmektedir. Bu nedenle bu araştırmada bu varsayım test edilmeyecektir.

Tek boyutluluk varsayımının test edilmesi için veri grubuna doğrulayıcı faktör analizi (DFA) uygulanmıştır. DFA, yapısal eşitlik modellemesi çerçevesinde kullanılan ve faktör yapılarının geçerliğini test eden bir analizdir (Byrne, 2001). DFA veri grubunun seçilen modele uyumunu gösteren birçok uyum istatistiği üretmektedir. Bu istatistikler "uyumun iyiliği istatistikleri" olarak adlandırılmaktadır (Gillapsy, 1996). Model ile veri grubu arasındaki uyum bu istatistikler sayesinde niceleştirilmektedir (Hu & Bentler, 1995).

Model ile verilerin uyumunu test etmek amacıyla günümüzde X^2 (Kay-Kare Uyum İyiliği; Chi-Square Goodness of fit), X^2/sd (kay-kare/serbestlik derecesi), uyum indeksleri olarak bilinen uyum iyiliği (Goodness of fit, GFI), Bentler'in karşılaştırmalı uyum indeksi (comparative fit index-CFI), ortalama karekök değeri yaklaşımı (Root Mean Squre of Approximation-RMSEA) ve yaklaştırmanın standart ortalama karekök değeri (SRMR) yaygın olarak kullanılmaktadır (Stapleton, 1997).

Araştırmada kullanılan PISA-2009 öğrenci anketinde yer alan ölçeklerin tek boyutluluk varsayımının sınanması için veri grubuna her bir ölçek için ayrı ayrı doğrulayıcı faktör analizi yapılmış ve bulgular Tablo 2.'de özetlenmiştir. Tablo 2. İncelendiğinde 4 ve 5'er maddeden oluşan ölçme modellerinin uyum değerlerinin RMSEA haricinde güçlü uyuma işaret ettiği görülmektedir. Veri grubu ile model arasındaki uyumu test etmek için hangi uyum hesaplama indeks ve kurallarını uygulanacağı konusunda eğitim araştırmacıları birden fazla uyum indeksi kullanmanın daha geçerli

sonuçlar vereceklerini ifade etmişlerdir (Byrne, Shavelson ve Muthen, 1989; Taub, 2001). Bu noktadan hareketle PISA-2009 öğrenci anketinde yer alan Q32 ve Q33 kodlu alt ölçeklerin tek boyutluluk varsayımını karşıladığı iddia edilebilir.

Tablo 2. Okula ve Öğretmenlere Algı Ölçeklerinin DFA Sonuçları

		X^2	sd	RMSEA	SRMR	GFI	CFI
BİRLEŞİK	Q32 (4 MAD)	645,19	2	0,081	0,044	0,97	0,95
	Q33 (5MAD)	879,35	5	0,086	0,023	0,99	0,99
TÜRKİYE	Q32 (4 MAD)	209,64	2	0,152	0,051	0,98	0,94
	Q33 (5MAD)	129,35	5	0,074	0,019	0,99	0,99
ABD	Q32 (4 MAD)	356,23	2	0,191	0,050	0,96	0,94
	Q33 (5MAD)	241,04	5	0,099	0,024	0,98	0,99
İRLANDA	Q32 (4 MAD)	92,08	2	0,116	0,030	0,99	0,98
	Q33 (5MAD)	185,22	5	0,104	0,029	0,98	0,98
İNGİLTERE	Q32 (4 MAD)	579,33	2	0,164	0,042	0,97	0,95
	Q33 (5MAD)	406,55	5	0,085	0,023	0,99	0,99

3. Bulgular ve Yorum

Bu çalışmada PISA-2009 öğrenci anketinde bulunan iki alt ölçekteki 9 maddenin DMF içerip içermediği amaçlanmıştır. Bu araştırmada verilerin analizine geçmeden önce veri dosyaları MULTILOG programına uygun hale getirilmiştir. Veri dosyaları referans ve odak gruplarına ait maddelerin yan yana eklenmesi ile oluşturulmuş olup veri dosyasının bir örneği EK-1'de sunulmuştur. İkinci aşamada analiz için temel model ve sınırlandırılmış model için uygun kodlar (syntax) ayrı ayrı yazılmış ve bu kodlar EK-2 ve EK-3'te verilmiştir. Son aşamada bu araştırmada test edilecek değişkenler (cinsiyet, dil ve kültür) için oluşturulan veri grupları MULTILOG yazılımında analiz edilmiş ve analiz sonuçları Tablo 3. ve Tablo 4.'de özetlenmiştir.

Tablo 3.'de PISA-2009 öğrenci anketinde yer alan alt ölçeklerdeki 9 maddenin DMF analizi sonuçları ve Tablo 4.'de DMF içeren maddelerin oranı yer almaktadır. Sonuçlar incelendiğinde Okul öğrenmelerine ilişkin algı alt boyuttaki maddelerde cinsiyete göre 2 maddede DMF içeren madde bulunmuştur. Dil ve kültür değişkenlerine göre ise, İrlanda-İngiltere örneklemlerinde 1, ABD-İngiltere örneklemlerinde 2 maddede, Türkiye-ABD örneklemlerinde ise dört maddenin hepsinde DMF tespit edilmiştir.

Öğretmenlere ilişkin algı alt ölçeğinde ulunan 5 maddeye ilişkin ise, cinsiyet değişkeni açısından 1 maddede DMF bulunmuştur. Dil ve kültür değişkenlerine göre ise, İrlanda-İngiltere örnekleminde 1, ABD-İngiltere örnekleminde 2 maddede, Türkiye-ABD örnekleminde ise 4 maddede DMF tespit edilmiştir. Tablo 4. incelendiğinde DMF tespit edilen maddelerin ölçekteki ağırlıkları yer almaktadır. Bulgulara göre, kültür ve dil değişkeni açısından dil ve kültür benzerliğinde DMF'li madde oranı az iken, kültür ve dil farklılığı arttıkça, ki bu durum Türkiye-ABD karşılaştırmasında açıkça

Tablo 3. Çalışma Stratejileri Alt Ölçeği DMF Analizi Sonuçları

		Aynı Dil-Benzer Kültür								Aynı Dil-Farklı Kültür				Farklı Dil-Farklı Kültür			
		Cinsiyet				IRL-ING				ABD-ING				ABD-TUR			
		X^2	G^2	sd	DMF	X^2	G^2	sd	DMF	X^2	G^2	sd	DMF	X^2	G^2	sd	DMF
	Temel Model	-60998,3	8,7		YOK	-34485,5				-39318,1				-1424,3			
Q32	M1	-60990,0	-10,5	4	YOK	-34482,2	-3,3	4	YOK	-39329,1	9	4	YOK	-1512,9	88,6	4	**VAR**
	M2	-61165,5	167,2	4	**VAR**	-35035,6	550,1	4	**VAR**	-39740,0	421,9	4	**VAR**	-2115,7	691,4	4	**VAR**
	M3	-61001,5	3,2	4	YOK	-34497,5	8,1	4	YOK	-39322,8	4,7	4	YOK	-2111,6	687,3	4	**VAR**
	M4	-61094,6	96,3	4	**VAR**	-34498,3	12,8	4	YOK	-39513,4	195,3	4	**VAR**	-1441,6	17,3	4	**VAR**
	M1	-61000,8	2,5	4	YOK	-34477,8	-7,7	4	YOK	-39313,8	-4,3	4	YOK	-1458,2	33,9	4	**VAR**
	M2	-61006	7,7	4	YOK	-34484,4	-1,1	4	YOK	-39315,8	-2,3	4	YOK	-2079,4	655,1	4	**VAR**
Q33	M3	-61007,5	9,2	4	YOK	-34488,1	2,6	4	YOK	-39320,7	2,6	4	YOK	-1425,8	1,5	4	YOK
	M4	-60992,6	-5,7	4	YOK	-34845,4	359,9	4	**VAR**	-39396,5	78,4	4	**VAR**	-1365,2	-59,1	4	**VAR**
	M5	-61051,4	53,1	4	**VAR**	-34480	-5,5	4	YOK	-39357,3	39,2	4	**VAR**	-2680,7	1256,4	4	**VAR**

Temel Model: Madde parametrelerinin referans ve odak gruplarda eşit olduğunu varsayan model

*<9,488(sd=4)

Tablo 4. Cinsiyet, Kültür ve Dil Farklılıklarına Göre DMF Gösteren Maddelerin Dağılımı

		Madde Sayısı	DMF'Lİ madde sayısı	%DMF
	Cinsiyet	9	3	33
Dil ve Kültür	Aynı dil-Benzer Kültür (IRL-GBR)	9	2	22
	Aynı dil-Farklı Kültür (IRL-GBR)	9	4	44
	Farklı dil-Farklı Kültür (IRL-GBR)	9	8	89

görülmektedir, DMF içeren madde oranında önemli artışlar gözlenmektedir. Türkiye-ABD karşılaştırmasında istatistiksel olarak elde edilen X^2 farklılıklarının belirgin bir biçimde fazla olması bu maddelerde DMF olduğunu desteklemektedir.

4. Sonuç ve Öneriler

Bu araştırmada PISA-2009 öğrenci anketinde yer alan "okul öğrenmelerine ilişkin algı (Q32-4 madde)" ve "öğretmenlere yönelik algı (Q33-5 madde)" alt ölçeklerinde bulunan 9 maddenin DMF içerip içermediği MTK'da aşamalı tepki modeli altında incelenmiştir. Analizler öncesi MTK'nın en öncelikli varsayımlarından biri olan tek boyutluluk varsayımı DFA ile test edilmiş ve RMSEA uyum indeksi hariç bütün uyum değerlerinin mükemmel uyuma işaret ettiği belirlenmiştir. Uyum indekslerinin değerlendirilmesinde çoklu kriterlerin olması gerektiği göz önüne alınırsa, ölçeklerin MTK'nın tek boyutluluk varsayımı desteklediği ifade edilebilir.

İkinci aşamada yapılan DMF analizlerinde bütün veri gruplarında bazı maddelerde DMF bulunduğu belirlenmiştir. DMF bulunan veri grupları incelendiğinde kültürel ve dilsel benzerlik artıkça DMF oranın azaldığı belirlenmiştir. Bu durum maddelerdeki ifadelerin kültürel ve dilsel benzerliği olan ülkelerde aynı anlamı ifade etmesi ile açıklanabilir. Kültürel ve dilsel farklılıklar arttıkça ölçeklerdeki DMF oranı da artmaktadır. Maddelerdeki dilsel karşılıkların anlamı diğer dilde aynı anlamı taşımaması ve ölçeklerdeki ölçülen özellikler olan okul öğrenmeleri ve öğretmenlere yönelik algıların farklı kültürlerde farklı olması DMF'nin *olası* sebepleri arasında gösterilebilir. Bu çalışmadan elde edilen bulgular Asil ve Gelbal 'ın (2012) çalışması ile paralellik göstermektedir. Asil ve Gelbal (2012) yapmış oldukları çalışmada fen ilimlerine yönelik tutumlarından "Bilimsel sorgulamaya verilen destek" alt boyutundaki maddelerin farklı dil ve kültürlerde DMF içerip içermediğini incelemişler, kültürler arası farklılıklar arttıkça DMF içeren madde sayısında artış olduğunu ortaya koymuşlardır. Atalay (2006) PISA 2006 öğrenci anketinde yer alan tutum maddelerini DMF açısından farklı yöntemler altında incelemiş ve farklı yöntemler altında DMF'li made sayısında benzerlikler tespit etmiştir. Alanyazın araştırmaları ve bu araştırmanın sonuçları incelendiğinde, kültürler arası farklılıklar arttıkça ölçeklerdeki DMF içeren madde sayısının arttığını desteklemektedir.

Bir maddede DMF bulunması, o maddenin yanlı olduğu anlamına gelmemekle beraber, yanlılığa da önemli bir işrettir. DMF içeren maddelerin yanlı olup olmadığı konusunda bir uzman grubuna incelme yaptırılması bu konuda verilecek olan kararlar için önemli bir adımıdır. Bunun yanında tek bir yöntemle yapılan DMF çalışmaları yerine çoklu yöntemlerle yapılan karşılaştırmalar daha sağlıklı sonuçlar verecektir.

Messick (1989) geçerlik çalışmalarında yapı geçerliğinin ön planda olması gerektiğini belirtmiş ve DMF içeren maddelerin yapı geçerliği üzerinde önemli bir tehdit olduğunu vurgulamışlardır. Elde edilen araştırma sonuçları bu kuramsal bilgi ile paralellik göstermektedir. Türkiye-ABD karşılaştırmasında DMF'li madde oranının en yüksek olduğu gözlenmiş ve bu ölçeklerden elde edilen DFA sonuçları ile karşılaştırıldığında,

en kötü uyum değerlerinin bu iki ülkede elde edildiği kısmen de olsa ifade edilebilir. Bu araştırmanın bulguları Messick (1989)'in tespitleri ile paralellik göstermektedir.

Bu noktadan hareketle araştırmadan elde edilen sonuçlar karşısında yeni araştırmalara yol açacak öneriler sunulabilir. Öncelikle MTK altında farklı yöntemlerle DMF analizleri aynı ve farklı gruplarda tekrarlanabilir. DMF analizleri aynı kuram altındaki farklı yöntemler ve farklı kuram altındaki yöntemlerle karşılaştırılabilir. Bunun yanında maddelerde tespit edilen değişen madde fonksiyonunun hangi grup lehine olduğu da incelenebilir.

5. Kaynakça

Asil, M. ve Gelbal, S. (2012). PISA öğrenci anketinin kültürler arası eşdeğerliği. Eğitim ve Bilim, 37(166), 236-249.

Atalay, K. (2010). PISA 2006 öğrenci anketinde yer alan tutum maddelerinin değişen madde fonksiyonu açısından incelenmesi. Yayınlanmamış Yüksek Lisans Tezi. Hacettepe Üniversitesi, Ankara.

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact and item validity from a multidimensional perspective. Journal of Educational Measurement, 29(1), 67-91.

Bolt, D.M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*(2), 113-141.

Byrne, B.M. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validty of a measuring instrument. *International Journal of Testing, 1*(1), 55-86.

Byrne, B.M., Shavelson, R.J. ve Muthen, B. (1989). Testingfor the equivalence of factor covariance and mean structures: the issueof partial measurement invariance. **Psychological Bulletin, 105** (3), 456-466.

Camilli, G. veShepard, L.A. (1994). Methods for Identifying Biased Test Items. ThousandOaks. Sage Publications.

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

Embretson, S.E. and Reise, S.P. (2000). Item Response Theory for Psychologists. Lawrence Erlbaum Associates, Inc., New Jersey, USA.

Gillapsy, J.A. (1996). A Primer on Confirmatory Factor Analysis. Paper presented at the Annual Meeting of the Southwest Educational Research Association. New Orleans, LA. (Eric document reproduction service no: ED 395 040).

Hableton, R.K. and Swaminathan, H. (1985). Item Response Theory. Kluwer-Nijhoff Publishing, MA, USA.

Hambleton, R. K., Swaminathan, H. and Rogers, H. (1991). *Fundamentals of Item Response Teory***.** Newbury Park CA: Sage.

Hu, L. & Bentler, P.M. (1995). Evaluating model fit. In R.H. Hoyle (Ed.), *Structural Equation Modeling: Concepts, Issues and Applications* (pp. 76-99). Thousand Oaks, CA: Sage.

Lautenschlager, G.J., Flaherty, V.L. and Park, D.G. (1994). IRT Differential item functioning: An examination of ability scale purifications. *Educational and Psychological MeasureMent, 54,* 21.

Lee, K. (2003). Parametric and Nonparametric IRT Models for Assessing Differential Item Functioning. Unpublished Doctoral Dissertation. Wayne State University, USA.

Kamata, A. and Vaughn, B. (2004). An introduction to Differential item functioning analysis. *Learning Disabilities: A Contemporary Journal, 2*(7), 49-69.

Kim, S.H. ve Cohen, A.S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22,* 345-355.

Maij-de Meij, A.M., Kelderman, H. and van der Flier, H. (2010). Improvement in detection ofdifferential item functioning using a mixture item response theory model. *Multivaria-te Behavioral Research, 45*, 975-999.

McDonald, R.P. (1982). Linear versus models in item response theory *Applied Psychological Measurement, 6*, 379-396.

MEB (2009). PISA 2009 Uluslararası Öğrenci Değerlendirme Programı Ulusal Ön Rapor. MEB, Ankara.

Messick, (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18(2), 5-11.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph*, No: 17.

Sheng, Y. (2005). *Bayesian Analysis of Hierarchical IRT models: Comparing and Combining the Unidimensional and Multidimensional IRT models.* Unpublished Doctoral Dissertation. University of Missouri-Columbia.

Sijstma, K. & Hemker, B.T. (2000). A taxonomy of IRT models for ordering persons anditems using simple sum scores. *Journal of Educational and Behavioral Statistics, 25* (4), 391-415.

Stapleton, C. D. (1997). Basic concepts in exploratory factor analysis as a tool to evaluate score validity: A right-brained approach. 25Kasım 2006 tarihinde http://ericea.net/ft/tamu/Efa.htmadreinden erişildi.

Taub, G. E. (2001). A confirmatory analysis of the wechsleradult intelligence scale-third edition: is the verbal/ performance discrepancy justified?**Practical Assessment, Research and Evaluation, 7**(22).

Thissen, D., Steinberg, L. and Gerard, M. (1986). Beyond man group differences: The concept of item bias. Psychological Bulletin, 99, 118-128.

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.

EK-1: dat. Uzantılı veri dosyası örneği

```
  1 12414233344444
  2 13344444444444
  3 12344333444444
  4 13444242444444
  5 13434323343243
  .
  .
  .
996                23444444444444
997                22424244444444
998                23142442344444
999                23424233244444
1000               21434443443444
```

EK-2: Temel modele ait kodlar (syntax)

```
From Ibrahim Alper Kose <i.alper.kose@GMAIL.com> on 21 JUL 2013
        DIF model
>PROBLEM  RANDOM, INDIVIDUAL, DATA='ABDGBR.dat',
        NITEMS=8, NGROUPS=2, NEXAMIEES=15952, NCHARS=5;
>TEST     ALL, GRADED, NC=(4(0)8), HIGH=(4(0)8);
>EQUAL    AJ,
        ITEMS=(1,2,3,4,5,6,7,8),
        WITH= (9,10,11,12,13,14,15,16);
>EQUAL    BK= (1,2,3),
        ITEMS=(1,2,3,4,5,6,7,8),
        WITH= (9,10,11,12,13,14,15,16);
>ESTIMATE NCYCLES=100;
>SAVE;
>END;
4
1234
11111111
22222222
33333333
44444444
(5A1,1X,I1,8A1)
```

EK-3: Sınırlandırılmış modele ait (1. Madde serbest bırakılmış) kodlar (syntax)

```
From Ibrahim Alper Kose <i.alper.kose@GMAIL.com> on 21 JUL 2013
        DIF model
>PROBLEM  RANDOM, INDIVIDUAL, DATA='ABDGBR.dat',
        NITEMS=8, NGROUPS=2, NEXAMIEES=15952, NCHARS=5;
>TEST     ALL, GRADED, NC=(4(0)8), HIGH=(4(0)8);
>EQUAL    AJ,
        ITEMS=(2,3,4,5,6,7,8),
        WITH= (10,11,12,13,14,15,16);
>EQUAL    BK= (1,2,3),
        ITEMS=(2,3,4,5,6,7,8),
        WITH= (10,11,12,13,14,15,16);
>ESTIMATE NCYCLES=100;
>SAVE;
>END;
4
1234
11111111
22222222
33333333
44444444
(5A1,1X,I1,8A1)
```

EXTENDED ABSTRACT

Purpose

Measurement tools are used to get informations about individuals. Various decisions are made for these individuals with the help of the outcomes of these measurement tools. Measurement tools should valid and reliable for the relevancy of decisions. Bias is one of the important threats on relability and validity of these instruments. TIMMS, PIRLS and PISA are tests which compares student achievements in the international area. There are everlasting discussions for these tests whether biased to any cultures or gender or not. DIF analysis are the first step for the item bias. For these reasons, this study was aimed to examine Differential Item Functioning (DIF) analysis in terms of gender, language and culture with the items of student questionnaire subscales of "sense of school learnings" and "sense of teachers" in Programme for International Student Assessment (PISA) 2009. For this aim DIF analysis were carried out with the help of the likelihood ratio test in graded respnse model based on item response theory.

Method

This research was aimed to perform DIF analysis on the items available in the PISA 2009 student questionnaire subtests of "sense of school learnings (Q32)" and "sense of teachers (Q33)". 23838 students, who were participated to PISA 2009 from Turkey, USA, Ireland and England, selected as research group. Also, all groups were combined to form a unique data set to exaimine DIF for gender. The main point for selecting these countries is to compare lingual and cultural similarities and distinctions. Confirmatory factor analysis was used for the data sets separately to test the unidimensionality assumption of IRT. CFA analysis showed that unifactorial construct of subtests was confirmed and unidimensionality assumpotion of IRT was met. For the DIF analysis, the likelihood ratio test was preferred under graded response theory based on item response theory (IRT) because of items which are polytomously scored. The likelihood ratio test, compares X^2 with the degrees of freedom that is equal to estimated item parameters, between compact model and augmented model. Before peforming DIF analysis, data sets were adapted to MULTILOG and syntaxes for the analyses were made ready.

Results

DIF analysis were carried out by MULTILOG software and 2 items for gender, 1 item for Ireland and England sample, 2 items for USA-England sample and 4 items for Turkey and USA sample was flagged as DIF in the subscale of "sense of school learnings". In the subscale of "sense of teachers", 1 item for gender, 1 item for Ireland-England sample, 2 items for USA-England sample and 4 items for USA-Turkey sample were flagged as DIF.

Discussion and Conclusion

This research was aimed to perform DIF analysis on the items available in the PISA 2009 student questionnaire subtests of "sense of school learnings (Q32-4 items)" and "sense of teachers (Q33-5 items)". At the first stage of the study unidimensionality assumption of IRT was investigated and findings showed that this assumption was met. At the second stage DIF analysis were performed and DIF was flagged for certain items in all data sets. Results showed that DIF ratio increases, while cultural and lingual differencs increases. This finding can be explained by the different meaning of the translated items. Furthermore, students sensation of school learnnigs and teachers may not be equal between these cultures and gender. This can be another reason for DIF. DIF is an important sign for the item bias but not exact. Also, with the DIF analysis source of DIF cannot be identified. For these reasons experts should examine the potential sources of DIF and bias as judgemental reviews. This study used only one DIF analysis method but multi methodological comparisons would be more reliable. For the future research, different analysis techniques under IRT may be recommended. Also, DIF analysis can be compared with the fingings of classic test theory and iem response theory.