Sırgancı, G. & Çakan, M. (2020). Sıralı lojistik regresyon ve poly-sıbtest yöntemleri ile değişen madde fonksiyonunun belirlenmesi. *Bolu Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 20* (1), 705-717.

Makalenin Türü / Article Type : Araştırma Makalesi / Researh Article
Geliş Tarihi / Date Received : 26.12.2019
Kabul Tarihi / Date Accepted : 28.02.2020
Yayın Tarihi / Date Published : 06.03.2020

https://dx.doi.org/10.17240/aibuefd.2020.20.52925-665084

SIRALI LOJİSTİK REGRESYON VE POLY-SIBTEST YÖNTEMLERİ İLE DEĞİŞEN MADDE FONKSİYONUNUN BELİRLENMESİ*

Gözde SIRGANCI[1], Mehtap ÇAKAN[2]

ÖZ

Bu çalışmanın amacı psikometride önemli bir geçerlik sorunu teşkil eden madde yanlılığının belirlenmesinde kullanılan farklı değişen madde fonksiyonu yöntemlerini kullanarak yöntemlerin uygulamadaki benzerlik ve farklılıklarının ortaya konmasıdır. Bu amaçla alanyazında kullanılan iki yöntem olan sıralı lojistik regresyon ve poly-SIBTEST yöntemleri ile değişen madde fonksiyonu analizleri yapılmıştır. Yöntemlere ilişkin değişen madde fonksiyonu analizleri, PISA 2006'da uygulanan bir öğrenci anketi maddelerine uygulanmıştır. Çalışma grubu ise kültürel ve dilsel farklılıkları yansıtan Avustralya, Yeni Zelanda, Amerika Birleşik Devletleri ve Türkiye olmak üzere dört ülke verisinden oluşmaktadır. Verilerin analizi aşamasında öncelikle öğrenci anketinin her bir ülkedeki faktör yapısı doğrulayıcı faktör analizi ile incelenmiştir. Devamında, sıralı lojistik regresyon ve poly-SIBTEST yöntemleri ile ölçme aracının benzer-farklı kültür ve dillerde değişen madde fonksiyonu analizi yapılmış olup yöntemlerin değişen madde fonksiyonunu belirleme benzerlikleri incelenmiştir. Doğrulayıcı Faktör Analizi bulguları ölçme aracının her bir ülkede aynı faktör yapısına sahip olduğunu göstermiştir. Sıralı lojistik regresyon ve poly-SIBTEST yöntemleri ile yapılan değişen madde fonksiyonu analiz bulguları ise ülkeler arasında kültürden ve dilden kaynaklı farklılıklar arttıkça değişen madde fonksiyonlu madde sayısında artış olduğunu göstermiştir. Yöntemlerin uyumları incelendiğinde ise her iki yöntemin değişen madde fonksiyonu belirlemede uyumlu olduğu; ancak poly-SIBTEST yöntemi ile daha hassas DMF analizinin yapıldığı belirlenmiştir.

Anahtar Kelimeler: PISA, değişen madde fonksiyonu, doğrulayıcı faktör analizi, sıralı lojistik regresyon, poly-SIBTEST

DETECTING DIFFERENTIAL ITEM FUNCTIONING BY USING SEQUENTIAL LOGISTIC REGRESSION AND POLY-SIBTEST METHODS

ABSTRACT

The aim of this study is to compare the differential item functioning methods used in the determination of item bias, which is an important validity problem in psychometry, and to reveal the similarities and differences in the application. For this purpose, the two preferred methods, sequential logistic regression and poly-SIBTEST, were used to analyze the DIF. PISA 2006 student questionnaire was examined to determine whether there was DIF items according to cultural and linguistic differences. Australia, New Zealand, United States and Turkey were comparatively investigated samples in this study. First, the factor structure of the student questionnaire was examined by confirmatory factor analysis. Subsequently, the sequential logistic regression and poly-SIBTEST analyzes were used to determine whether the questionnaires showed DIF between different cultures and languages. Then, the similarities of DIF determination methods were examined. According to results of the confirmatory factor analysis showed that the measurement model had the same factor structure in all cultures. Sequential logistic regression and poly-SIBTEST findings show that there are items that have different item functioning among countries, and if the linguistic and cultural differences between countries increase, the number of items that have differential item functions increases. When the compatibility of the methods was examined, it was found that both methods were compatible in determining differential item functioning, but more sensitive DIF analysis was performed with poly-SIBTEST method.

Keywords: PISA, differential item functioning, confirmatory factor analysis, sequential logistic regression, poly-SIBTEST

* *Bu çalışma, 2012 yılında Doç. Dr. Mehtap ÇAKAN danışmanlığında Abant İzzet Baysal Üniversitesi Eğitim Bilimleri Enstitüsü Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda sunulan yüksek lisans tezinden üretilmiştir.*

[1] *Bozok Üniversitesi, Eğitim Fakültesi, gozdesirganci@gmail.com,* https://orcid.org/0000-0003-4824-5413
[2] *Gazi Üniversitesi, Eğitim Fakültesi, cakanmehtap@hotmail.com,* https://orcid.org/0000-0001-6602-6180

1.GİRİŞ

Test puanları sadece öğrencilerin akademik başarılarının bir ölçüsü değildir; aynı zamanda yüksek riskli kararların alındığı birer değerlendirme kriteridir. Bu kararların bireysel, sosyal ve politik açıdan önemli sonuçları olabilir. Dolayısıyla ölçme araçları bireyin var olan özelliğini belirleme konusunda adil olmalıdır. Bu durum test adaleti kavramının ortaya çıkmasına neden olmuştur. APA (Amerikan Psikologlar Birliği), AERA (Amerikan Eğitim Araştırmaları Birliği) ve NCME (Eğitimde Ölçme Ulusal Konseyi) gibi kuruluşlar tarafından 1990'lı yılların başından bu yana test adaleti ile ilgili çeşitli tanımlamalar yapılmış ve bir dizi standart ortaya konmuştur. Bu standartlarda ölçme aracından alınan puanların farklı gruplarda bulunan bireyler için karşılaştırılabilir olmaması durumunda adilliğin sağlanamayacağı vurgulanmıştır (APA, AERA, NCME 1999). Bireylerin ait oldukları alt gruplar için tek bir test maddesinin bile adil olmadığına ilişkin bir kanıtın elde edilmesi, test puanlarının geçerliğini düşürmektedir. Dolayısıyla ölçme aracının her bir maddesinin ölçülen özellik açısından bireyler için eşit olduğunu gösteren kanıtların sunulması gerekmektedir.

TIMSS (Uluslararası Matematik ve Fen Eğilimleri Araştırması), PISA (Uluslararası Öğrenci Değerlendirme Programı) ve PIRLS (Uluslararası Okuma Becerilerinde Gelişim Çalışması) gibi uluslararası geniş ölçekli uygulamalarda, özellikle kültürel ve dilsel farklılıklardan kaynaklanan adaletsizliği en aza indirmek adına testin hazırlaması ve uygulanması sürecinde büyük kaynaklar sarf edilmektedir. Çünkü farklı kültür ve dildeki bireyleri karşılaştırma çalışmaları ancak ölçme araçlarının farklı gruplardan gelen bireyler için değişmez olduğu durumda anlamlıdır (Gierl, 2000). Dolayısıyla test puanlarının testi alan tüm bireyler için karşılaştırılabilir olduğunu söyleyebilmek için geçerlik çalışmalarının yapılması gerekmektedir.

Test puanlarının yapı geçerliğinin sağlanmasında en büyük tehditlerden birisi yanlılıktır (Kristanjansonn, Aylesworth, McDowell ve Zumbo, 2005). Yanlılık, bireylerin yer aldığı gruplara özgü farklılıklarından dolayı test puanlarının gruba bağlı olarak değişmesidir (Angoff, 1993). Yanlılık söz konusu olduğunda alt gruplardaki hetorojenlik, ölçme aracı ile ölçülmek istenen yapıdan farklı bir varyans kaynağı şeklinde test puanlarına etki eder. Yanlılık analizleri, test ve madde düzeyinde yapılabilir olsa da genellikle araştırmacılar madde yanlılığı ile ilgilenmektedir.

Madde yanlılığı belirleme çalışmaları, değişen madde fonksiyonları (DMF) analizleri ile yapılmaktadır. DMF, benzer yetenek düzeyine sahip olup farklı gruplarda bulunan bireylerin, bir maddeyi doğru yanıtlama olasılıklarının değişmesidir (Clouser ve Mazor, 1998; Zumbo, 1999). İstatistiki olarak bir maddenin DMF göstermesi onun olası yanlı madde olduğu anlamına gelmektedir ve bu maddenin belli bir grup için yanlılık gösterip göstermediğine uzman görüşleri doğrultusunda karar verilir (Kamata ve Vaughn, 2004). DMF yöntemleri araştırmacılar tarafından farklı sınıflar altında toplanmıştır. Örneğin, Benito ve Navas-Ara (2000) DMF belirleme yöntemlerini KTK, faktör analizi, ki-kare ve MTK'ye dayalı yöntemler olarak sınıflamışlardır. Panfield ve Camilli (2007) DMF yöntemlerini oran farkları, genel odds oranı, lojistik regresyon ve ortalama farklara dayalı yöntemler biçiminde sınıflamışlardır. Gelin (2001) ise DMF belirleme yöntemlerini gözlenen puana (observed score) ve gizil değişkene (latent variable) dayalı olmak üzere iki temel grup altında sınıflandırmıştır. Gözlenen puana dayalı DMF belirleme yöntemlerinde, ölçülen psikolojik özellik bakımından benzer bireylerin yer aldığı en az iki grup karşılaştırılır. Bu gruplar odak (focal) ve referans (reference) grup olarak adlandırılır. Odak grup, referans gruba göre ait olduğu gruptan dolayı adalet ve eşitlik açısından dezavantajlı olduğu düşünülen gruptur (Bilir, 2009). Gözlenen gruplarla DMF belirleme yöntemlerinde, odak ve referans gruplarından kestirilen madde parametreleri karşılaştırılarak DMF analizi yapılır. Bu parametrelerin arasında anlamlı fark olması, DMF'nin olası kaynağının incelenen grup değişkeni (cinsiyet, ırk, etnik köken vb.) olabileceğini gösterir (Clauser ve Mazor, 1998).

Gözlenen puana dayalı DMF belirleme yöntemlerinin işlevselliği ağırlıklı olarak bilişsel testlere yöneliktir. TIMSS, PISA gibi geniş ölçekli uygulamalarda öğrenci, öğretmen, okul ya da aile anketleri uygulanmakta olup bu tür çok kategorili puanlanan maddelerin yer aldığı araçlar için DMF belirleme yöntemlerinin işlevselliğinin test edildiği araştırmalar yetersizdir (Kristanjansonn, ve diğerleri, 2005; Schulz, 2003, 2005, 2008). Dolayısıyla bu çalışmanın problemiçok kategorili puanlanan maddeler için önerilen iki farklı yöntemin kültürel ve dilsel farklılıklara göre DMF belirlemedeki performansının araştırılmasıdır.

1.1. Araştırmanın Amacı

Bu çalışmanın amacı, PISA 2006 öğrenci anketinde yer alan "bilimsel sorgulamaya verilen destek" ölçme modelinde yer alan maddelerin değişen madde fonksiyonunun Sıralı Lojistik Regresyon (SLR) ve poly-SIBTEST teknikleri ile belirlenerek yöntemlerin uygulamadaki avantaj ve dezavantajlarınının belirlenmesidir. DMF analizleri kültürel ve dilsel benzerlik ve farklılıkları yansıtması açısından Avustralya, Yeni Zelanda, ABD ve Türkiye örneklemleri üzerinden yürütülmüştür. Araştırmanın amacı doğrultusunda aşağıdaki sorulara yanıt aranmıştır.

PISA 2006 öğrenci anketinde ölçülen "bilimsel sorgulamaya verilen destek" ölçme modelinin maddeleri;

1- Avustralya ve Yeni Zelanda
2- Avustralya ve ABD
3- Avustralya ve Türkiye

örneklemleri arasında sıralı lojistik regresyon ve Poly-SIBTEST yöntemlerine dayalı yapılan analizlere göre değişen madde fonksiyonu göstermekte midir?

1.2. Araştırmanın Önemi

Ölçme araçlarında aranan en önemli özelliklerden biri geçerliktir ve DMF, geçerliği düşüren bir faktördür. Yapılan sınavlarda kullanılan ölçme araçlarında DMF içeren madde olması bu sınavların sonuçlarının geçerliğini olumsuz bir şekilde etkilemektedir ve sınav sonuçlarına bağlı olarak yanlış yorumlar ve düzeltmeler yapılmasına neden olmaktadır. DMF belirleme yöntemleri özellikle iki kategorili puanlanan maddeler için çok iyi yapılandırılmıştır. Çok kategorili puanlanan maddeler için ise DMF belirleme ile ilgili birçok yöntem geliştirilmiş olmasına rağmen bu yöntemlerin uygulamadaki işlevselliğinin belirlenmesine de ihtiyaç vardır (Zumbo ve McDowell, 2005).

Örneklem büyüklüğü, verilerin yapısı, maddelerin puanlanma biçimi, DMF belirlemede kullanılan teknikler gibi etkenler maddelerin DMF düzeylerinin farklı araştırmalarda farklı şekilde belirlenmesine neden olabilmektedir. Bu nedenle de hangi tekniği kullanmanın uygulamada daha iyi olacağına karar verebilmek için DMF belirleme tekniklerinin karşılaştırılmasına ihtiyaç duyulmaktadır.

Çalışma çok kategorili puanlanabilen maddelerin değişen madde fonksiyonunu belirlemek için kullanılabilecek istatistiksel yöntemleri tanıtması, bu yöntemlerin benzerlik ve farklılıklarını ortaya koyması ve gelecek araştırmalara ışık tutması yönünden önemlidir. Ayrıca farklı kültürel ve dilsel özelliklere sahip örneklemler üzerinde çalışılarak kültürden ve dilden kaynaklanan yanlılık sorunlarının olup olmadığı da araştırılmıştır. Bu da ölçme eşdeğerliği olmayan uygulamalardan elde edilen verilere dayanarak ortaya konan sonuçların ne derece geçerli ve güvenilir olduğunun sorgulanması açısından önem arz etmektedir.

2. YÖNTEM

2.1. Araştırmanın modeli

Araştırma PISA 2006 öğrenci anketinin farklı kültürel ve dilsel özelliklere göre değişen madde fonksiyonlarının incelenmesi yönüyle betimsel; DMF belirleme yöntemlerinden sıralı lojistik regresyon ve Poly-SIBTEST tekniklerini karşılaştırılması yönüyle de kuramsal özellik taşımaktadır.

2.2. Araştırmanın çalışma grubu

Çalışma grubunu 2006 PISA uygulamasına Avustralya, Yeni Zelanda, ABD ve Türkiye'den katılan toplam 24203 öğrenci oluşturmaktadır. Karşılaştırmalarda kültürel ve dilsel benzerlik ve farklılıkları yansıtması açısından Avustralya, Yeni Zelanda, Amerika Birleşik Devletleri ve Türkiye örneklemleri ile çalışılmıştır. Öğrencilerin ülkelere dağılımı 11644'i (%48.1) Avustralya, 3503'i (%14.5) Yeni Zelanda, 4519'i (%18.7) ABD ve 4537'si (% 18.7) Türkiye biçimindedir.

2.3. Veri toplama araçları ve süreci

Veri toplama aracı olarak PISA 2006 uygulamasında öğrenci anketinde yer alan "bilimsel sorgulamaya verilen destek" ölçme modeli kullanılmıştır. Ölçme aracı, her birinde beşer maddenin yer aldığı "genel değer" ve "kişisel değer" boyutlarından oluşmaktadır. Modeli oluşturan tüm maddeler dörtlü likert türünde olup madde kategorileri tümüyle katılıyorum (1)'den, hiç katılmıyorum (4)'e kadardır. Ölçme aracının verilerine PISA'nın resmi internet sitesinden erişilmiştir (www.pisa.oecd.org).

2.4. Verilerin analizi

DMF analizleri öncesinde ölçme aracının her bir kültürde ve dilde aynı yapıya sahip olduğunun belirlenmesi gerekmektedir. Bunun için her bir ülke veri setine doğrulayıcı faktör analizi (DFA) uygulanmıştır. DFA uygulanmadan önce her bir ülke verisinin kayıp değer, aşırı uç değer, normallik, tekli bağlantılılık, çoklu doğrusal bağlantı ve doğrusallık varsayımlarına uygunluğu test edilmiştir ve varsayımların sağlandığı sonucuna ulaşılmıştır. Ardından her bir ülkenin veri setine DFA uygulanmış, sonrasında ise maddelerin kültürel ve dilsel farklılıklara göre DMF gösterip göstermedikleri sıralı lojistik regresyon ve poly-SIBTEST yöntemleri ile incelenmiştir.

Sıralı lojistik regresyon (SLR)

Lojistik regresyon yöntemi kullanılarak DMF belirlemede, maddeye doğru cevap verme olasılığı toplam puan ve toplam puan ile grup ilişkisinden yararlanılarak tahmin edilir. French ve Miller (1996) lojistik regresyonun bir uzantısının çok kategorili maddelerde DMF belirlemede uygun olabileceğini önermişlerdir. Zumbo (1999) tarafından da çok kategorili veriler için DMF etki büyüklüğü geliştirilmiştir. SLR ile DMF analizi üç aşamada yapılır ve her aşamada ki-kare değeri hesaplanır. Üçüncü aşama ile birinci aşamadan elde edilen ki-kare değerleri serbestlik derecesinin iki olduğu ki-kare dağılımı ile karşılaştırılarak DMF testinin anlamlılığı belirlenir (Zumbo, 1999). DMF etki büyüklüğü ise üçüncü model ile birinci modelden elde edilen R^2 değerlerinin farkından hesaplanır.

Poly SIBTEST

SIBTEST parametrik olmayan örtük değişken modelidir (Potenza ve Dorans, 1995, aktaran: Atalay, 2010: 12). Chang, Mazzeo ve Roussos (1996), SIBTEST yönteminin çok kategorili veriler için farklı bir formu olan Poly-SIBTEST yöntemini geliştirmişlerdir. Poly- SIBTEST örtük değişkenler için tanımlanmıştır ancak MTK yetenek ve madde parametrelerini kullanmaz. Poly-SIBTEST yöntemi ile incelenecek olan maddeler çalışan test ve eşleştirme testi olmak üzere iki alt teste ayrılır. Çalışan test olası DMF içeren maddelerden; eşleştirme testi ise testtin geri kalan maddelerinden oluşur. Bireylerin yetenek düzeyini temsil eden toplam puan SIBTEST ve Poly-SIBTEST bilgisayar programlarında elde edilir (Henderson, 1999). Referans ve odak grubu oluşturan bireyler, toplam puana ve kategori sayısına göre eşleştirilir. Her bir gruptaki bireylerin, testle ölçülen özellik bakımından aynı yetenekte olduğu varsayılarak referans ve odak gruptaki bireylerin performansları her bir "çalışan test" için K grupta karşılaştırılır (Henderson, 1999).

Betimsel istatistikler, DFA'nın varsayımlarının testi ve sıralı lojistik regresyon analizi için SPSS 15.0 programı kullanılmıştır. DFA, Mplus7 (Muthen ve Muthen, 2012); poly-SIBTEST testi ise Poly-SIBTEST (Sheally ve Stout, 1993) programları ile yapılmıştır.

3. BULGULAR

Bu bölümde öncelikle ölçme aracının farklı kültür ve dillerde test edilen yapı geçerliği sonuçları daha sonra ise araştırma sorularına ilişkin bulgular sunulmuştur.

3.1. Yapı geçerliğine ilişkin bulgular

Ölçme aracının farklı kültür ve dillerdeki yapı geçerliği DFA ile incelenmiştir. DFA sonucu elde edilen faktör yükleri Avustralya verisinde 0.54 - 0.88; Yeni Zelanda verisinde 0.47 - 0.85; ABD verisinde 0.58 - 0.81 ve Türkiye verisinde 0.58 - 0.76 arasında yer almaktadır. Faktör yüklerinin 0.40 ve üzerinde olduğu görülmektedir. Model veri uyumu RMSEA, CFI, TLI ve SRMR değerleri dikkate alınarak incelenmiştir. Ki kare testi örneklem büyüklüğüne duyarlıdır ve aşırı kestirim yaptığı için bu çalışmada raporlanmamıştır (Cheung ve Rensvold 2002, Yuan ve Chan 2016). Hu ve Bentler (1999) CFI ve TLI değerlerinin ≤0.95 olduğunda model veri uyumunun mükemmel; Brown (2006) SRMR değerinin ≤ 0.08, Hooper, Coughlan ve Mullen (2008) RMSEA değerinin ≤ 0.08 olduğu durumda model veri uyumunun kabul edilebilir olduğunu belirtmişlerdir. Ölçme modeline ait uyum indeksleri Tablo 1'de sunulmuştur. Tablo 1'deki model veri uyumu indeks değerleri incelendiğinde bilimsel sorgulamaya verilen destek ölçme modelinin her bir ülke ve birleşmiş veri seti için iyi uyum gösterdiği görülmektedir.

Tablo 1.
Model Veri Uyumu

	RMSEA	CFI	TLI	SRMR
Avusturalya	0.080	0.97	0.96	0.052
Yeni Zelanda	0.082	0.97	0.95	0.058
ABD	0.078	0.97	0.97	0.046
Türkiye	0.076	0.97	0.96	0.045

Verilerin her bir ülke için güvenirliğine ilişkin Cronbach alfa değerleri incelendiğinde güvenirlik değerlerinin Avustralya örneklemi için 0.89, 0.81 ve 0.86; Yeni Zelanda örneklemi için 0.88, 0.78, 0.84; ABD örneklemi için 0.89, 0.82, 0.85; ve Türkiye örneklemi için 0.86, 0.78, 0.81 bulunmuştur.

Sonuç olarak, DFA sonuçlarına göre bilimsel sorgulamaya verilen destek ölçme modelinin her bir kültür ve dil için iki faktörlü bir yapıya sahip; ölçme modelinin her bir kültür ve dil için güvenilir olduğu ortaya konmuştur. Bu bulgu doğrultusunda benzer-farklı kültür ve diller için değişen madde fonksiyonu analizleri yapılmıştır.

3.2. Avustralya ve Yeni Zelanda örneklemlerine ilişkin SLR ve Poly-SIBTEST bulguları

Benzer kültür ve dil yapısını yansıtan Avustralya ve Yeni Zelanda'nın ölçme modeli maddelerine ilişkin SLR yöntemi ile yapılan DMF analizine ilişkin bulgular Tablo 2'de sunulmuştur. DMF değeri, SLR çıktısındaki üçüncü aşamada elde edilen X2değerinden birinci aşamada elde edilen X2değerinin çıkartılması ile bulunmuştur. Her bir madde için bu değerin anlamlılığının değerlendirilmesinde Bonferroni düzeltmesi kullanılmıştır. Buna göre 10 maddeden elde edilen düzeltilmiş anlamlılık seviyesi 0.005'tir. Tablo 2 incelendiğinde ST18Q04, ST18Q06, ST18Q03, ST18Q07 ve ST18Q010 maddelerinin DMF değerinin anlamlı olduğu görülmüştür. Dolayısıyla bu maddelerde DMF'nin varlığından söz edilebilir. DMF'nin etki düzeyi incelendiğinde ST18Q04 kodlu maddenin B düzeyinde yani orta derecede DMF gösterdiği görülmektedir. ST18Q04 dışındaki maddelerin DMF düzeyleri ise A düzeyinde olduğu için bu maddelerde DMF yoktur veya ihmal edilebilir düzeydedir.

Tablo 2.
Avustralya ve Yeni Zelanda Örneklemlerinde SLR Yöntemi ile Yapılan DMF Sonuçları

Boyutlar	Maddeler	DMF X^2	p	DMF ΔR^2	Düzey
	ST18Q01	0.6	0.741		
	ST18Q02	1.37	0.504		
GD	ST18Q04	35.69*	0.000	0.15	B
	ST18Q06	20.47*	0.000	0.08	A
	ST18Q09	1.04	0.595		
	ST18Q03	20.377*	0.000	0.08	A
	ST18Q05	1.48	0.471		
KD	ST18Q07	12.53*	0.002	0.03	A
	ST18Q08	9.92	0.007		
	ST18Q10	18.22*	0.000	0.03	A

Avustralya ve Yeni Zelanda örneklemlerine uygulanan SLR analizinde tek bir maddenin B düzeyinde; diğer dört maddenin A düzeyinde yani ihmal edilebilir düzeyde DMF içerdiği görülmektedir. Bu ülkelerde aynı dilin konuşulması ve benzer kültürel özelliklere sahip olunması, maddelerin her iki ülkedeki öğrenciler tarafından benzer şekilde algılandığı ve bundan ötürü yüksek düzeyde DMF çıkmaması sonucunu desteklemektedir.

Tablo 3'te Avustralya ve Yeni Zelanda örneklemlerinin ölçme modeline ilişkin poly-SIBTEST yöntemi ile yapılan DMF analizi bulguları sunulmuştur.

Tablo 3.
Avustralya ve Yeni Zelanda Örneklemlerinde Poly-SIBTEST Yöntemi ile Yapılan DMF Sonuçları

Boyutlar	Maddeler	Beta kestirimi	SH	p	Düzey
	ST18Q01	0.018	0.014	0.212	
	ST18Q02	0.008	0.014	0.540	
GD	ST18Q04	-0.083	0.015	0.000	B
	ST18Q06	0.032	0.015	0.033	A
	ST18Q09	0.003	0.018	0.879	
	ST18Q03	0.020	0.018	0.276	
	ST18Q05	0.006	0.016	0.720	
KD	ST18Q07	0.037	0.015	0.016	A
	ST18Q08	-0.023	0.016	0.150	
	ST18Q10	-0.046	0.018	0.010	A

Tablo 3 incelendiğinde, maddelerden dört tanesinin beta değerinin 0.05 düzeyinde anlamlı olması bu maddelerde DMF'nin var olduğunu göstermektedir. Ancak bu maddeler incelendiğinde fen bilimlerine verilen genel değer faktöründeki ST18Q04 kodlu maddenin B, yani orta düzeyde, yine aynı faktördeki ST18Q06 ve fen bilimlerine verilen kişisel değer faktöründe yer alan ST18Q07 ile ST18Q10 maddelerinde ise A düzeyi yani ihmal edilebilir düzeyde DMF olduğu görülmektedir.

Avustralya ve Yeni Zelanda örneklemlerine uygulanan poly-SIBTEST analizinde tek bir maddenin B düzeyinde; diğer üç maddenin A düzeyinde yani ihmal edilebilir düzeyde DMF içerdiği bulunmuştur. Bu ülkelerde aynı dilin konuşulması ve benzer kültürel özelliklere sahip olunmasından ötürü maddelerin her iki ülkedeki öğrenciler tarafından benzer şekilde algılandığı ve bundan ötürü maddelerin yüksek düzeyde DMF göstermediği söylenebilir.

3.3. Avustralya ve ABD örneklemlerine ilişkin SLR ve Poly-SIBTEST bulguları

Aynı dil farklı kültür yapısını yansıtan Avustralya ve ABD örneklemleri üzerinde ölçme modeli maddelerine ilişkin SLR yöntemi ile yapılan DMF analizine ilişkin bulgular Tablo 4'de sunulmuştur.

Tablo 4.
Avustralya ve ABD Örneklemlerinde SLR Yöntemi ile Yapılan DMF Sonuçları

Boyutlar	Maddeler	DMF X^2	p	DMF ΔR^2	Düzey
	ST18Q01	7.82	0.020		
	ST18Q02	5.1	0.078		
GD	ST18Q04	11.15*	0.004	0.06	A
	ST18Q06	17.62*	0.000	0.07	A
	ST18Q09	17.77*	0.000	0.03	A
	ST18Q03	78.217*	0.000	0.32	C
	ST18Q05	27.79*	0.000	0.07	A
KD	ST18Q07	15.66*	0.000	0.03	A
	ST18Q08	17.19*	0.000	0.07	A
	ST18Q10	21.97*	0.000	0.04	A

Tablo 4 incelendiğinde SLR yöntemi ile yapılan DMF analizi sonucunda iki madde dışındaki tüm maddelerin DMF içerdiği görülmektedir. Öte yandan DMF gösteren maddelerden sadece biri C düzeyinde yani yüksek düzeyde DMF gösterirken diğerleri A düzeyinde yani göz ardı edilebilir düzeyde DMF göstermiştir. Bulgular incelendiğinde bu ülkelerde aynı dil konuşulmasına rağmen kültürel özelliklerin farklı olmasından dolayı DMF gösteren madde sayısının artmış olabileceği düşünülebilir. Ayrıca DMF gösteren maddelerin Avustralya ve ABD örneklemlerinde aynı biçimde anlaşılmamış olabileceği de söylenebilir.

Avustralya ve ABD örneklemleri dikkate alınarak poly-SIBTEST yöntemi ile yapılan DMF analiz sonuçları Tablo 5'de verilmiştir.

Tablo 5.
Avustralya ve ABD Örneklemlerinde Poly-SIBTEST Yöntemi ile Yapılan DMF Sonuçları

Boyutlar	Maddeler	Beta kestirimi	SH	p	Düzey
	ST18Q01	0.003	0.015	0.837	
	ST18Q02	0.045	0.010	0.001	A
GD	ST18Q04	0.020	0.015	0.182	
	ST18Q06	0.059	0.014	0.000	B
	ST18Q09	-0.055	0.018	0.002	A
	ST18Q03	-0.093	0.018	0.000	C
	ST18Q05	0.071	0.016	0.000	B
KD	ST18Q07	0.045	0.015	0.003	A
	ST18Q08	0.044	0.016	0.006	A
	ST18Q10	0.015	0.018	0.394	

Tablo 5 incelendiğinde maddelerden yedi tanesinin beta değerinin 0.05 düzeyinde anlamlı olması bu maddelerde DMF'nin var olduğunu göstermektedir. DMF'li maddeler incelendiğinde fen bilimlerine yönelik kişisel değer faktöründeki ST18Q03 kodlu maddenin yüksek, ST18Q05 kodlu maddenin orta, ST18Q07 ve ST18Q08 kodlu maddelerin ise ihmal edilebilir düzeyde DMF'li olduğu; fen bilimlerine yönelik genel değer faktöründeki ST18Q06 kodlu maddenin orta, ST18Q02 ve ST18Q09 kodlu maddelerin ise hafif veya ihmal edilebilir düzeyde DMF'li olduğu görülmektedir.

Genel olarak poly-SIBTEST sonucuna bakıldığında yedi maddeden bir tanesi C düzeyinde, iki tanesi B düzeyinde ve geri kalan dört tanesi ise A düzeyinde DMF göstermişlerdir. DMF'li madde sayısındaki artışın aynı dilin konuşulduğu Avustralya ve ABD ülkelerindeki kültürel farklılıklardan kaynaklandığı düşünülebilir.

3.4. Avustralya ve Türkiye örneklemlerine ilişkin SLR ve Poly-SIBTEST bulguları

Avustralya ve Türkiye örneklemleri için SLR yöntemi ile yapılan DMF analiz sonuçları Tablo 6'de verilmiştir.

Tablo 6 incelendiğinde tüm maddelerin DMF içerdiği sonucuna varılmıştır. Aynı zamanda DMF içeren bu maddelerden sadece ST18Q06 kodlu madde A düzeyinde yani hafif ya da göz ardı edilebilir düzeyde DMF gösterirken diğer maddeler C düzeyinde yani yüksek düzeyde DMF göstermişlerdir.

Avustralya ve Türkiye örneklemleri arasında tüm maddeler SLR yöntemi ile DMF göstermiştir. Bu durumun hem ülkelerdeki kullanılan dilin faklılığından hem de kültürel farklılıklardan ortaya çıkmış olma olasılığı yüksektir ve maddelerin Avustralya ve Türkiye'deki öğrenciler için aynı anlama gelmediği sonucuna varılabilir.

Tablo 6.
Avustralya ve Türkiye Örneklemlerinde SLR Yöntemi ile Yapılan DMF Sonuçları

Boyutlar	Maddeler	DMF X^2	p	DMF ΔR^2	Düzey
GD	ST18Q01	125.25	0.000	0.31	C
	ST18Q02	329.2	0.000	0.65	C
	ST18Q04	380.02	0.000	0.52	C
	ST18Q06	47.33	0.000	0.01	A
	ST18Q09	1268.24	0.000	1.96	C
KD	ST18Q03	453.61	0.000	1.73	C
	ST18Q05	239.61	0.000	0.43	C
	ST18Q07	1513.53	0.000	2.93	C
	ST18Q08	1200.64	0.000	2.7	C
	ST18Q10	1568.82	0.000	3.97	C

Avustralya ve Türkiye örneklemleri için poly-SIBTEST yöntemi ile yapılan DMF analiz sonuçları Tablo 7'de sunulmuştur.

Tablo 7 incelendiğinde Avustralya ve Türkiye örneklemleri arasında yapılan poly-SIBTEST analizine göre, ST18Q06 kodlu madde dışındaki tüm maddelerde beta değeri anlamlıdır ve hepsi yüksek düzeyde DMF göstermektedir. Bu durumun hem ülkelerdeki kullanılan dilin farklılığından hem de kültürel farklılıklardan ortaya çıkmış olabileceği düşünülmektedir. Bu bağlamda, maddelerin Avustralya ve Türkiye'deki öğrenciler için aynı anlama gelmediği sonucuna varılabilir.

Tablo 7.
Avustralya ve Türkiye Örneklemlerinde Poly-SIBTEST Yöntemi ile Yapılan DMF Sonuçları

Boyutlar	Maddeler	Beta kestirimi	SH	p	Düzey
GD	ST18Q01	-0.229	0.069	0.001	C
	ST18Q02	-0.389	0.059	0.000	C
	ST18Q04	-0.431	0.081	0.000	C
	ST18Q06	-0.105	0.070	0.131	
	ST18Q09	-0.635	0.071	0.000	C
KD	ST18Q03	0.273	0.025	0.000	C
	ST18Q05	0.111	0.022	0.000	C
	ST18Q07	-0.597	0.020	0.000	C
	ST18Q08	0.345	0.021	0.000	C
	ST18Q10	-0.536	0.023	0.000	C

4.TARTIŞMA ve SONUÇ

Bilimsel sorgulamaya verilen destek ölçme modelini oluşturan faktörlerin yapısına ilişkin doğrulayıcı faktör analizi sonuçları ve uyum indeksleri, iki faktörlü ölçme modelinin üç ülke örnekleminde de var olduğunu göstermiştir.

Avustralya referans, Yeni Zelanda, ABD ve Türkiye odak grup alınarak yapılan SLR ve poly-SIBTEST analizleriyle bilimsel sorgulamaya verilen destek ölçme modelini oluşturan maddelerin DMF gösterip göstermediğine ilişkin ortaya çıkan genel sonuçlar Tablo 8'de özetlenmiştir.

Maddelere ilişkin Avustralya ve Yeni Zelanda örneklemleri arasında yapılan DMF analizleri sonucunda SLR yöntemiyle beş, poly-SIBTEST yöntemi ile dört maddenin farklı fonksiyonlaştığı ortaya çıkmıştır. Poly-SIBTEST yöntemi ile DMF gösteren her madde SLR yöntemi ile de farklı fonksiyonlaşmış ve her iki yöntemle de aynı düzeyde DMF göstermişlerdir. Poly-SIBTEST'ten farklı tek bir madde SLR yöntemi ile farklı fonksiyonlaşsa da bu maddede ihmal edilebilir düzeyde DMF bulunmuştur. Yöntemler DMF belirlemede %80'lik bir uyum göstermiş olup önemli düzeyde DMF tespitinde (B ve üzeri) tam uyum göstermişlerdir. Ayrıca bu iki ülke arasında önemli düzeyde DMF gösteren madde sayısının bir olması ülkelerin benzer dil ve kültürel özellikler göstermeleri yani öğrencilerin maddelere aynı anlamı yükledikleri şeklinde yorumlanmıştır.

Avustralya ve ABD'ye ilişkin yapılan DMF analizleri sonucunda SLR yöntemiyle sekiz, poly-SIBTEST yöntemi ile yedi maddenin farklı fonksiyonlaştığı ortaya çıkmıştır. Her iki analizde de DMF gösteren madde sayısı altıdır ve bunların dördü aynı düzeyde DMF göstermişlerdir. Bu dört maddeden biri her iki yöntemde de yüksek düzeyde farklı fonksiyonlaşmış ve ABD lehine, tek boyutlu olmayan DMF göstermiştir. Diğer üç madde her iki yöntemde ihmal edilebilir düzeyde DMF göstermişlerdir. Her iki yöntemde DMF gösteren diğer iki madde ise SLR yöntemiyle ihmal edilebilir düzeyinde farklı fonksiyonlaşırken poly-SIBTEST yöntemiyle orta düzeyde farklı fonksiyonlaşmıştır. Dolayısıyla her iki yöntem de DMF'li maddeleri tespit etmiş ancak iki madde de DMF düzeylerini farklı ortaya koymuşlardır. Dolayısıyla poly- SIBTEST'in daha hassas olduğu söylenebilir. Öte yandan genel anlamda yöntemlerin uyum gösterdikleri de görülmektedir. Avustralya ve ABD örneklemleri arasında önemli düzeyde DMF gösteren madde sayısı üçtür ve bu durumun aynı dili konuşan iki ülkenin kültürel yapısındaki farklılıklardan kaynaklanmakta olduğu düşünülmektedir.

Tablo 8.
Kültür ve Dil Farklılıklarına Göre SLR ve Poly-SIBTEST Sonuçları

		SLR		Poly-SIBTETS	
	Madde sayısı	**DMF'li madde sayısı ve düzeyi**	**%DMF**	**DMF'li madde sayısı ve düzeyi**	**%DMF**
Aynı dil- Benzer kültür	10	1 tane B	%10	1 tane B	%10
(AUS- Yeni.Z)		4 tane A	%40	3 tane A	%30
Aynı dil- Farklı kültür	10	1 tane C	%10	1 tane C	%10
(AUS-ABD)		7 tane A	%70	2 tane B	%20
				4 tane A	%40
Farklı dil-Farklı kültür	10	1 tane A	%10	-	-
(AUS-TUR)		9 tane C	%90	9 tane C	%90

Avustralya ve Türkiye örneklemleri arasında yapılan DMF analizleri sonucunda ise SLR yöntemiyle tüm maddelerin, poly-SIBTEST yöntemi ile dokuz maddenin farklı fonksiyonlaştığı ortaya çıkmıştır. Her iki yöntemde de maddelere ait ki-kare farklılık değerlerinin diğer karşılaştırmalara oranla oldukça yüksek olduğu tespit edilmiştir. Bir madde dışındaki tüm maddelerde her iki yöntemde de yüksek düzeyde DMF olduğu sonucuna varılmıştır. Farklı olarak sadece bir madde SLR analizi ile ihmal edilebilir düzeyde farklı fonksiyonlaşmıştır. Bu durumda yöntemlerin oldukça yüksek uyum gösterdikleri sonucuna ulaşılmıştır. Avustralya ve Türkiye arasında dokuz madde önemli düzeyde DMF göstermiştir. Bu durumun bu iki ülkenin farklı dil ve kültürel yapıya sahip olmalarından kaynaklanmış olabileceği düşünülmektedir. Ancak maddelerin çoğu odak grup olan Türkiye örnekleminin lehine farklı fonksiyonlaşmıştır.

Genel olarak doğrulayıcı faktör analizi sonucunda ölçme modelinin faktör yapısının tüm kültürlerde benzer yapıda olduğu ortaya çıkmış olmasına rağmen bu sonuç bilimsel sorgulamaya verilen destek ölçme modelinin çalışmaya dâhil edilen dört ülke için eşdeğer olduğu sonucunu beraberinde getirmemiştir. Tablo 11'de görüldüğü üzere ülkeler arasındaki dil ve kültür farklılıkları arttıkça DMF gösteren madde sayısı artmış ve DMF'nin düzeyi yükselmiştir. Elde edilen bulgular benzer çalışmalar tarafından da desteklenmektedir (Asil ve Gelbal, 2012; Atalay Kabasakal ve Kelecioğlu, 2012; Ercikan, Gierl, McCreith, Puhan ve Koh, 2002; Ercikan ve Koh, 2005; Karakoç Alatlı, Ayan, Polat Demir ve Uzun, 2012; Köse, 2015; Tiryaki, 2019). Yine benzer şekilde sonuçlar Atalay'ın (2010) ve Asil'in (2010) çalışmasında bulduğu kültürler arası farkın maddelerin farklı fonksiyonlaşmasına neden olduğu sonucunu da destekler niteliktedir. Sonuç olarak kültür ve dil farklılıkları bilimsel sorgulamaya verilen destek ölçme modeli maddelerinin farklı fonksiyonlaşmasına neden olmuştur. Bu sonuca göre ülkeler arası yapılan karşılaştırmaların sadece Avustralya ve Yeni Zelanda arasında daha anlamlı olacağı düşünülmektedir.

Diğer taraftan, yöntemler açısından bakıldığında, poly-SIBTEST'in SLR'ye göre daha hassas olduğu ancak genel olarak her iki yöntemin de uyumlu çalıştığı ve karşılaştırılabilir olduğu sonucuna varılmıştır. Diğer bir ifade ile yöntemlerin farklı dil ve kültürlere göre DMF belirlemede avantajlı veya dezavantajlı oldukları bir durum yoktur. Bu bulgular alanyazınla tutarlık göstermiştir (Gierl, Khaliq ve Boughton, 1999; Lyons-Thomas, Sandilands ve Ercikan, 2014; Atalay, 2010; Demir, 2013).

Benzer analizlerin farklı yöntemler kullanılarak ve farklı ülkeleri referans grubu alarak yapılmasının, ayrıca, DMF'nin olası sebeplerinin incelenmesinin alana katkı getireceği düşünülmektedir. Bu çalışmada çok kategorili maddelerde DMF belirleme yöntemlerinin uygulamadaki benzerlik ve farklılıkları incelenmiştir. Bir sonraki aşamada bu yöntemlerin DMF belirlemedeki özgüllük ve duyarlılık incelemeleri yapılarak yöntemler karşılaştırılabilir. Ayrıca bu çalışmada kültürel ve dilsel benzerlik ve farklılığı yansıtan gerçek veri kullanılmıştır. Yöntemsel karşılaştırma çalışmaları gerçek verilerden simule edilen koşullar altında gerçekleştirilerek hangi yöntemin hangi durumda avantajlı ve dezavantajlı olduğu incelenebilir.

KAYNAKÇA

Asil, M. (2010). *Uluslararası öğrenci değerlendirme programı (PISA) 2006 öğrenci anketinin kültürler arası eşdeğerliğinin incelenmesi.* (Yayımlanmamış Doktora Tezi). Hacettepe Üniversitesi, Ankara, Türkiye.

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

Angoff, W. H. (1993). *Perspectives on differential item functioning methodology*. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 3-23). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

American Psycholocigal Assosiation, American Educational Reserach Association & National Council on Measurement in Education. (1999). Standarts for educational and psychological testing. Washington: American National Research Education.

Atalay Kabasakal, K. ve Kelecioğlu, H (2012). PISA 2006 öğrenci anketinde yer alan maddelerin değişen madde fonksiyonu açısından incelenmesi. *Ankara Üniversitesi Eğitim Bilimleri Dergisi, 45(2)*, 77-96.

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

Bilir, M. K. (2009). *Mixture İtem Response Theory-MIMIC model: simultaneous estimation of differential item functioning for manifest groups and latent classes* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No: 3399179)

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. (First Edition). NY: Guilford Publications, Inc.

Chang, H., Mazzeo, J. ve Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33(3),* 333-353.

Cheung, G. W. ve Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling, 9(2),* 233-255.

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

Demir, S. (2013). *PISA 2009 matematik okuryazarlığı alt testinde bulunan maddelerinin Mantel-Haenszel, Sibtest ve Lojistik Regresyon yöntemleri ile değişen madde fonksiyonunun incelenmesi,* (Yayımlanmamış Yüksek Lisans Tezi). Abant İzzet Baysal Üniversitesi, Bolu, Türkiye.

Ercikan, K., Gierl, M., McCreith, T., Puhan, G., ve Koh, K. (June, 2002). *Comparability of English and French versions of SAIP for reading, mathematics and science items.* Paper presented at the annual meeting of the Canadian Society for Studies in Education, Toronto, ON.

Ercikan, K. ve Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5(1),* 23-35.

Gelin, M. N. (2001). *Type I error rates of the DIF MIMIC approach using Joreskog's covariance matrix with ML and WLS estimation* (Doctoral dissertation). Retrieved from https://open.library.ubc.ca/cIRcle/collections/ubctheses/831/items/1.0054355

Gierl, M. J. (2000). Construct Equivalence on Translated Achievement Tests. *Canadian Journal of Education, 25(4),* 280-296

Gierl, M., Khaliq, S. N. ve Boughton, K. (1999). *Gender Differential Item Functioning in Mathematics and Science: Prevalence and Policy Implications,* Paper presented at the Canadian Society for the Study of Education.

Gómez-Benito, J. ve Navas-Ara, M. J. (2000). A Comparison of $\chi 2$, RFA and IRT Based Procedures in the Detection of DIF. *Quality and Quantity, 34(1*), 17-31.

Henderson, D. L. (1999). *Investigation of differential item functioning in exit Examinations across item format and subject area*. Unpublished doctor dissertation, University of Alberta, Edmonton, Alberta, Canada.

Hooper, D., Coughlan, J. ve Mullen, M. (2008). Structural equation modeling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods. 6(1),* 53-60.

Kamata, A. ve Vaughn, B. K. (2004). An introduction to differential ıtem functioning analysis. *Learning Disabilities: A Contemporary Journal, 2(2),* 49-69.

Karakoc Alatlı, B., Ayan, C., Polat Demir, B. Ve Uzun, G. (2016). Examination of the TIMSS 2011 Fourth Grade Mathematics Test in terms of cross-cultural measurement invariance. *Eurasian Journal of Educational Research, 66,* 389- 406.

Kristanjansonn E., R. Aylesworth, McDowell, I. ve B. D. Zumbo (2005). A comparison of four methods for detecting differeantial item functioning in ordered response model. *Educational and Psychological Measurement. 6,* 935-953.

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

Lyons-Thomas, J., Sandilands, D. ve Ercikan, K. (2014). Gender differential item functioning in mathematics in four international jurisdictions, *Eğitim ve Bilim, 39 (172),* 20-32.

Muthén, L. K. ve Muthén, B. O. (2012). *MPlus:Statistical analysis with latent variables*--User's guide.

Penfield, R. D. ve Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay, and C. R. Rao (Eds.), *Handbook of statistics, 26*, (pp. 125-167). North Holland.

Schulz, W. (2003). *Validating questionnaire constructs in ınternational studies. Two examples from PISA 2000.* Paper Presented at the Annual Meetings of the American Educational Research Association (AERA) in Chicago, 21-25

Schulz, W. (2005). *Testing parameter ınvariance for questionnaire ındices using confirmatory factor analysis and ıtem response theory.* Paper Presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11

Schulz, W. (2008). *Questionnaire construct validation in the ınternational civic and citizenship education study.* Paper presented to the 3rd IEA International Research Conference in Taipei, 23-27

Shealy, R. T. ve Stout, W. F. (1993). A model based standardization approach that separates true bias/DIF. *Psychometrika, 58*, 197–239.

Swaminathan, H. ve Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27(4),* 361- 370.

Tabachnick, B. G. ve Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Boston MA: Allyn & Bacon.

Tiryaki, F. (2019). *PISA 2015 öğrenci tutum anketlerinin değişen madde fonksiyonu ve ölçme değişmezliğinin incelenmesi.* (Yayınlanmamış Yüksek Lisans Tezi). Ankara Üniversitesi, Ankara, Türkiye.

Yuan, K. H. ve Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological methods, 21(3),* 405.

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) ıtem scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

EXTENDED ABSTRACT

1. Introduction

Test scores are not merely a measurement of students' academic achievement, but also evaluation criteria on which high risk decision are made. These decisions may be important individual, social and political consequences. Therefore, the applied assessment tool should be fair to every individual in expressing their psychological properties. A justly designed test helps achieve educational, social and economic goals more promptly. For instance, the controller of large-scale international applications such as TIMSS (Trends in International Mathematics and Science Study), PISA (Programme for International Student Assessment) and PIRLS (Progress in International Reading Literacy Study) spend enormous resources in the test preparation phase just to minimize bias, which stems from especially cultural and linguistic differences between countries. Comparative studies, conducted on subjects who come from different cultures, are only relevant so long as assessment tools are uniform for all subjects, even for subjects in different groups (Gierl, 2000). Consequently, in order to be able to say that the test score is comparable for all subjects that took the test, validity studies should be conducted.

The main threat to a test's construct validity is bias (Kristanjansonn, Aylesworth, McDowell & Zumbo, 2005). Bias is the deviation of test scores with respect to groups, which is related to different properties of a group that a subject resides in (Angoff, 1993). Bias analyses are conducted at test and item levels, and researchers typically focus on item biases.

Item bias identification studies are conducted by utilizing Differential Item Functioning (DIF) analyses. DIF is the change in the probability of a correct answer to an item given by participants in different groups but have similar ability (Clouser & Mazor, 1998; Zumbo, 1999). Statistically, an item exhibiting a DIF indicates the probability of bias for that item and summation is made whether the item is biased towards a group by consulting expert view (Kamata & Vaughn, 2004). DIF methods are gathered into different classes by researchers. For example, Gelin (2001) categorized DIF determination methods under two main groups, which were based on observed score and latent variable. In observed score DIF methods, at least two groups with participants, who have similar psychological properties that are being assessed, are compared. These groups are named focal and reference. Focal group is thought to have a disadvantage with respect to fairness and equality when compared to the reference group (Bilir, 2009). In DIF determination methods with observed groups, a DIF analysis is conducted by comparing the predicted item parameters from the focal and reference groups. The presence of a significant difference between these parameters indicates the probable source of DIF stems from the examined group variable (gender, race, ethnic origin, etc.) (Clauser & Mazor, 1998). The functionality of observed score based DIF determination methods are mainly oriented during cognitive tests. In large-scale applications such as TIMSS and PISA, student, instructor, school or family questionnaires are applied and for these type of test, in which there are polychotomous scoring items, the studies that examine the functionality of DIF detection methods are not adequate (Kristanjansonn, et al., 2005; Schulz, 2003, 2005, 2008). Thereby, the problem of this study is to compare the techniques of Sequential Logistic Regression (SLR) and Poly-SIBTEST, which are recommended for DIF determination in polychotomous items, and to examine their pros and cons in application. In methodology comparison, items from the scale model of "support for scientific inquiry", which is featured in PISA 2006 student questionnaire, were used. During comparison, sample sets from Australia, New Zealand, USA and Turkey were included in order to reflect the cultural and linguistic differences.

2. Method

The study group consisted of 24,203 students who attended the 2006 PISA from Australia, New Zealand, the United States and Turkey. Study group consist of, 11,644 of the students (48.1%) from Australia, 3,503 of the students (14.5%) from New Zealand, 4,519 of the students (18.7%) from the United States and finally 4,537 of the students (18.7%) from Turkey.

The scale model of "support for scientific inquiry", which is featured in PISA 2006 student questionnaire, were used as the data collection tool. The questionnaire comprises of two dimensions such as "general value" and "personal value". All items that constitutes the model are 4-point Likert type scale ranging from Strongly Agree (1) to Strongly Disagree (4). The data of the questionnaire was obtained from PISA' s official website (www.pisa.oecd.org).

It is necessary to verify that the items measure the same structure in each culture and language before performing DIF analysis. Therefore, in the process of data analysis, Confirmatory Factor Analysis (CFA) assumptions were tested for each countries data. Subsequently, SLR and poly-SIBTEST methods were used to determine whether the items exhibited differential item functioning according to cultural and linguistic differences, and to what extent DIF exhibited and which type of DIF worked in favor of the group.

SPSS was used to determine descriptive statistics, CFA assumptions and DIF detection with SLR method. Mplus7 (Muthen ve Muthen, 2012) was used for CFA. A Computer program developed by Shealy and Stout (1993) was used for Poly-SIBTEST.

3. Findings, Discussion and Results

The confirmatory factor analysis results and compatibility indices about the structure of factors, which constitute the "support for scientific inquiry" measurement model, showed that the two-factor model exists in all three sample spaces of the countries.

As a result of DIF analysis concerning Australia and New Zealand sample sets, 5 items were found to be differential functioning in SLR method, while 4 items were functioning differentially in poly-SIBTEST method. Every item that exhibits DIF in poly-SIBTEST method, was also shows DIF in SLR method and all items have same level of DIF in both methods. Only a single item in SLR method was found to have DIF compared to poly-SIBTEST however, DIF level was insignificant. The methods were discovered to be 80% compatible in DIF detection, and the fact that there was only 1 item with significant DIF was interpreted as both countries speak the same language and have similar cultures, hence the students assigned same meaning to same questions.

As a result of DIF analysis on Australia and USA sample sets, it was revealed that while 8 items appeared to function differently in SLR method, 7 items did function differently in poly-SIBTEST method. In both methods, 6 items exhibited DIF, 4 of which had same levels. One of these 4 items functioned highly different in both methods and exhibit a non-uniform DIF that was biased toward USA. Other 3 items had negligible DIF levels in both methods. The remaining 2 items showed insignificant levels of DIF in SLR method, whereas they exhibited modest levels of DIF in poly-SIBTEST method. Therefore, poly-SIBTEST can be considered the more sensitive method. On the other hand, it was seen that two methods enjoyed compatibility in general. There were 3 items, which showed significant levels of DIF across Australian and American sample sets and this can be explained by the fact that while two countries speak the same language, there are cultural differences to consider.

As a result of DIF analysis on Australia and Turkey sample sets, it was shown that while all items appeared to function differently in SLR method, 9 items did function differently in poly-SIBTEST method. In both methods, it was determined that chi-square differences were quite high considering other comparisons. Aside from 1 item, it was seen that all items exhibited high levels of DIF. Distinctly, just 1 item showed insignificant level of DIF using SLR method. This situation indicates a high compatibility between methods. Nine items across Australian and Turkish sample sets had significant levels of DIF. This is thought to stem from the fact that not only two countries speak separate languages, but also, they have cultural differences. However, most of the items functioned with a bias towards focal group Turkey.

In general, even though confirmatory factor analysis results indicate a similar pattern for factor structure of the scale model, this result did not imply that "support for scientific inquiry" scale model was equivalent for all 4 countries included in the study. As it can be seen in Table 11, when the cultural and linguistic differences between countries increase, the number of items exhibiting DIF also increases and DIF levels rose accordingly. The findings are supported by similar studies (Asil and Gelbal, 2012; Atalay Kabasakal and Kelecioğlu, 2012; Ercikan, Gierl, McCreith, Puhan and Koh, 2002; Ercikan and Koh, 2005; Karakoç Alatlı, Ayan, Polat Demir and Uzun, 2012; Köse, 2015; Tiryaki 2019). Furthermore, the results are in alignment with studies conducted by Atalay (2010) and Asil (2010), which stated that cultural differences cause items to function in a differential manner. In conclusion, cultural and linguistic differences caused items of "support for scientific inquiry" to function differentially. According to this finding, international comparisons will only be meaningful if done between Australia and New Zealand.

Nonetheless, from a method-oriented perspective, it is concluded that while poly-SIBTEST is more sensitive than SLR, two methods were found to be compatible and comparable. These results are also in alignment with the literature (Gierl, Khaliq and Boughton, 1999; Lyons-Thomas, Sandilands and Ercikan, 2014; Atalay, 2010; Demir, 2013). It is concluded that conducting similar analyses by using different methods and taking different countries as reference groups and investigating the possible sources of DIF will advance the literature.

ETİK BEYANNAME

Yapılan bu araştırmanın yazım sürecinde bilimsel ve etik kurallara tüm araştırmacılar tarafından uyulmuş, farklı eserlerden yararlanması durumunda atıfta bulunulmuş, kullanılan verilerde herhangi bir tahrifat yapılmamış, araştırmanın tamamı veya bir kısmı farklı bir akademik yayın platformunda yayınlatılmak üzere gönderilmemiştir. Tüm bu durumlardan araştırmada ismi bulunan yazarların bilgisi olduğunu ve gerekli kurallara uyulduğunu beyan ederim. 02/03/2020

İmza
Gözde Sırgancı
Araştırmanın Sorumlu Yazarı