ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);112-127

WÇZÖ-IV Maddelerinin Cinsiyet ve Sosyo-Ekonomik Düzey Açısından İşlev Farklılığının Belirlenmesinde Kullanılan Yöntemlerin Karşılaştırılması*

Gender and Socioeconomic Status DIF on The WISC-IV Turkish Form Items: A Comparison of DIF Detection Tecniques

Elif Bengi ÜNSAL ÖZBERK ** Nizamettin KOÇ ***

Öz

Bu araştırmanın amacı Wechsler Çocuklar için Zekâ Ölçeği (WÇZÖ) IV Türkçe formundaki maddelerin sosyo-ekonomik düzey ve cinsiyete göre işlev farklılığı gösterip göstermediğinin, birden fazla yöntemle incelenerek birbiri ile uyumlu sonuçlar verip vermediğinin ortaya konulmasıdır. Araştırmada WÇZÖ-IV Türkiye Uyarlama ve Standardizasyon Çalısmasına ait 819 kişilik ön uygulama verileri kullanılmıştır. Çalışma WÇZÖ-IV'ün çoklu puanlanan maddelerden oluşan Küplerle Desen, Benzerlikler, Sayı Dizisi, Sözcük Dağarcığı, Harf-Rakam Dizisi, Kavrama alttestleri ve ikili puanlanan maddelerden oluşan Resim Kavramları, Mantık Yürütme Kareleri, Resim Tamamlama, Genel Bilgi, Aritmetik, Sözcük Bulma olmak üzere toplam 12 alttest ve bu alt testlerde yer alan 315 madde üzerinden yürütülmüştür. Maddelerin işlev farklılığı içerip içermediğine karar vermek için ikili puanlanan maddeler açısından Rasch Modeli, Mantel-Haenszel, SIBTEST yöntemleri, çoklu puanlanan maddeler açısından da Kısmi Puan Modeli, Mantel Test ve Poly-SIBTEST yöntemleri kullanılmış ve 12 alt test açısından madde işlev fonksiyonu içeren maddeler belirlenmiştir. Araştırmada kullanılan MİF belirleme yöntemleri incelendiğinde, genel olarak yöntemlerin birbiriyle tutarlı sonuçlar verdiği ancak cinsiyet açısında MİF içeren maddelerin tespitinde Mantel-Haenszel, SIBTEST ve Mantel Test, Poly-SIBTEST istatistikleri daha tutarlı sonuçlar verirken sosyoekonomik düzeye göre MİF içeren maddelerin tespitinde Mantel-Haenszel, Rasch Modeli ve Mantel Test, Kısmi Puan Modelinin daha tutarlı sonuçlar verdiği saptanmıştır.

Anahtar Kelimeler: Wechsler çocuklar için zekâ ölçeği IV, Rasch modeli, Mantel-Haenszel, SIBTEST, kısmi puan modeli, Mantel test, Poly-SIBTEST.

Abstract

The purpose of this study is to investigate potential gender and socio-economic status bias in theWechler Intelligence Scale for Children: Fourth Edition (WISC-4) by using several differential item functioning detection techniques. In this study, WISC-4 Turkish standardization test pilot data including 817 children were used. In accordance with the purpose of the study, 315 items were used both in polytomously scored subtests such as Block Design, Similarities, Digit Span, Vocabulary, Letter-Number Sequencing, Comprehension, and dichotomously scored subtests such as Picture Concepts, Matrix Reasoning, Picture Completion, Information, Arithmetic, and Word Reasoning. While Rasch Model, Mantel-Haenszel, and SIBTEST DIF detection techniques were used for dichotomously scored items, Partila Credit Model, Mantel, and Poly-SIBTEST techniques were used for polytomously scored items. In terms of DIF techniques, Mantel-Haenszel, SIBTEST and Mantel Test, Poly-SIBTEST analyses provided similar results when DIF based on gender was investigated. In addition Mantel-Haenszel, Rasch estimations and Partial Credit Model, Mantel Test results were similar while investigating DIF according to socioeconomic status.

Keywords: Wechsler intelligence scale for children IV, rasch model, Mantel-Haenszel, SIBTEST, partial credit model, Mantel test, Poly-SIBTEST.

* Bu çalışma, birinci yazarın Prof. Dr. Nizamettin KOÇ danışmanlığında tamamlanan doktora tezinden türetilmiştir.

** Dr., Adalet Bakanlığı Ceza ve Tevkifevleri Genel Müdürlüğü, Personel Eğitim Bürosu, Şube Müdürü, Ankara-Türkiye, elifbengiunsal@gmail.com
*** Prof. Dr., Ankara Üniversitesi, Eğitimde Ölçme ve Değerlendirme ABD, Ankara-Türkiye, Emekli, nkoc@ankara.edu.tr

DOI: *10.21031/epod.287577*
Geliş Tarihi: 24.01.2017
Kabul Tarihi: 02.03.2017

GİRİŞ

Psikolojik testlerin genel amacı; kişiler arasındaki farkları belirleyip karşılaştırabilmek (bireylerarası-interindividual) ya da aynı bireyin zaman içerisinde değişen (bireyiçi-intraindividual) yeteneğini ölçebilmektir (Cronbach, 1990). Bu yüzden psikolojik testlerdeki değerlendirmenin amacı bireylerarası ve bireyiçi farklılıkları belirlemektir. Bu noktada eğitsel ve psikolojik ölçmeler ve buna bağlı olarak ölçme araçlarının geliştirilmesi, bireyler arası ve birey içi fark kavramından doğmuştur (Thorndike, 1982). Test sonuçlarının testi alan alt gruplara göre değişiklik göstermesi, gruplar arasındaki gerçek yetenek veya başarı farklılığından kaynaklanmış olabileceği gibi testin sebep olduğu sistematik bir adaletsizliğin sonucu da olabilir.

Cronbach (1990)'a göre testlerin kullanım amaçları, bireylerin belli bir özelliğe göre sıralanması, seçilmesi, bir üst sınıf ya da kademeye geçirilmesi kararının verilmesi olarak sıralanmıştır. Murphy ve Davidshofer (1994)'in belirttiği gibi testler bireyler hakkında önemli kararlar almak için kullanılmaktadırlar. Benzer bir şekilde, Türkiye'de özel eğitime ihtiyacı olan çocukların belirlenmesi ve özel eğitime devam kararlarının verilmesinde ayrıca üstün yetenekli çocukların belirlenmesi ve Bilim Sanat Merkezleri'nde eğitim almaları gibi önemli kararlarının verilmesinde Wechsler Çocuklar için Zeka Ölçeği kullanılmaktadır.

Böylesine önemli kararların verildiği, testlerin sonuçlarına bağlı olarak adil kararlar verebilmek için testlerin psikometrik özelliklerinin kabul edilebilir seviyelerde olması gerekmektedir (Horst 1966; Reynold, Livingston ve Willson, 2006). Bu amaçla, testlerden beklenen de hatalardan olabildiğince arınık olması ve ölçmek istediği özelliği başka değişkenler karışmadan ölçebilmesidir. Bir testin ölçmek istediği özellik dışında başka değişkenlerin karışmasına örnek oluşturan durumlardan biri de testin yanlılık içermesidir. Yanlılık, farklı alt gruplardaki bireylerin test puanlarının bulundukları gruba bağlı olarak sistematik hata içermesidir (Camilli ve Shepard, 1994; Tittle, 1988; Zumbo, 1999). Test yanlılığı, bir testin sonuçlarına dayanılarak alınan bir karar, tüm gruplar için adil değilse ya da bir gruba, diğerine göre eşit olmayan bir etki yapıyorsa ortaya çıkar (Osterlind ve Everson, 2009). Eğer bir test yanlı ise, testi alan farklı gruplardaki bireylere adil davranılmamasına neden olmaktadır (Tittle, 1988). Madde yanlılığı ise birçok araştırmacı tarafından, aynı yetenek düzeyinde olan, fakat cinsiyet, sosyoekonomik düzey, etnik köken, vb. gibi farklı gruplardan gelen bireylerin, test koşullarından ya da maddenin bazı özelliklerinden, test maddelerine doğru cevap verme olasılıklarının diğer gruba göre az ya da çok olması biçiminde tanımlanmıştır (Adams ve Rowe, 1988; Mellenberg, 1983; Osterlind, 1983; Raju, 1990; Rodney ve Drasgow, 1990; Shepard, Camilli ve Williams, 1984; Tittle, 1988; Zumbo, 1999).

Madde yanlılığı çalışmaları özellikle test geliştirme ve uyarlama sürecine önemli katkı sağlar. Özellikle testlerin denkliğini sağlama açısından gerekli bir süreçtir. Madde yanlılığı çalışması ile her bir maddeler gözden geçirilerek, her bir madde için ayrı ayrı geçerlik kanıtı toplanmış olur. Madde yanlılığı çalışmaları, hem istatistiksel hem de uzman kanılarına dayalı yargısal süreçleri gerektirir. Test maddelerinin yanlı olup olmadığının belirlenmesinin ilk adımı madde işlevinin farklılığını (MİF) belirlemeye yönelik istatistiksel bir süreçtir. Belirli istatistiksel işlemler sonrası MİF gösteren bir madde olası yanlı madde olarak değerlendirilir (Kamata ve Vaughn, 2004). Bu açıdan madde yanlılığı çalışmalarında, madde etkisi ve MİF kavramları arasındaki farka değinmek önemlidir. Madde etkisi, farklı alt gruplardan gelen bireylerin bir maddeyi doğru yanıtlama olasılıklarının, ilgili madde ile ölçülmek istenen psikolojik özellik bakımından farklılaşmasını ifade etmektedir (Zumbo, 1999). Bu farklılık, madde etkisi söz konusu olduğunda, madde ile ölçülen psikolojik özellik bakımından gruplar arasındaki gerçek farklardan kaynaklanırken, MİF söz konusu olduğunda grupların yetenek düzeylerindeki farklılaşmadan değil, MİF'ten kaynaklanmaktadır (Camilli ve Shepard, 1994). Bir maddenin yanlılık içermesi, maddenin MİF içerdiğinin göstergesidir. Ancak MİF gösteren maddelerin yanlı olduğu kesin değildir (Kamata ve Vaughn, 2004). Bir maddenin yanlı olduğu kararını vermek için o maddenin yanlızca MİF göstermesi yeterli olmamakla birlikte maddenin MİF göstermesi, maddenin yanlılığının ortaya konması sürecinde bir ilk adımdır.

MİF belirleme çalışmalarının başladığı günden bugüne birçok MİF belirleme yöntemi kullanılmıştır. Örneğin Zumbo (2007) sınıflandırmayı olasılık tablosuna dayalı yöntemler, madde tepki kuramına dayalı yöntemler, çok boyutlu yöntemler biçiminde yaparken Camilli, Shepard (1994) varyans analizine dayalı yöntemler, olasılık tablosuna dayalı yöntemler olarak yapmıştır. MİF çalışmalarında birden çok yönteme dayalı inceleme yapılması önerilmektedir (Holland ve Wainer, 1993; Osterlind ve Everson, 2009). Bu araştırmada da, cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddeler, Wechsler Çocuklar için Zeka Ölçeği IV Türkçe formundaki ikili puanlanan maddeler açısından Rasch MODELİ, SIBTEST ve Mantel-Haenszel, çoklu puanlanan maddeler açısından da Kısmi Puan Modeli, Poly-SIBTEST ve Mantel Test ile yapılan analizlerle incelenmiştir.

İkili puanlanan maddelerde cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddelerin tespiti açısından kullanılan yöntemlerden biri olan Rasch modeli, Rasch tarafından geliştirilen MTK kapsamında tek parametreli bir modeldir (Rasch, 1960). Rasch modelinde sadece güçlük parametresi (bi) kullanılır ve model ayırıcılık parametresini tüm maddeler için sabit tutar, şans parametresi ise 0'dır. Rasch modelinde madde karakteristik eğrisi aşağıdaki fonksiyonla bulunmaktadır (Hambleton, Swaminathan ve Rogers, 1991).

$$P(X_{is} = 1|\theta_s, \beta_i) = \frac{\exp(\theta_s - \beta_i)}{1 + \exp(\theta_s - \beta_i)} \quad (1)$$

θ_S bireyin yetenek düzeyini β_i, i maddesinin güçlük indeksini ifade etmektedir. $P(X_{IS}=1|\ \theta_S, \beta_i)$de, β_i güçlük düzeyindeki maddeyi, θ_S yetenek düzeyindeki bireyin doğru cevaplama olasılığını tanımlamaktadır (Embretson ve Reise, 2000). İkili puanlanan maddelerde cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddelerin tespiti açısından kullanılan bir diğer yöntem olan Mantel Haenszel yöntemi, Mantel ve Haenszel (1959) tarafından eşleştirilmiş gruplarda uygulanacak olan ki-kare tekniği olarak geliştirilmiştir. Bu yöntem, 1998 yılında Holland ve Thayer tarafından güncellenerek MİF belirleme yöntemi olarak ölçme değerlendirme literatürüne kazandırılmıştır. MH yöntemi χ^2 istatistiğine dayanır, MH yöntemi için odak (focal) ve referans gruplarındaki bireyler gösterdikleri performansa göre eşleştirilir. Bu eşleştirmenin yapılabilmesi için referans ve odak grupta yer alan cevaplayıcıların testten aldıkları toplam puanlara göre 4 ya da 5 yetenek grubu oluşturulur. Heterojen puan dağılımlarında yetenek grubu sayısı artırılabilir. Odak ve referans grupların toplam puanları eşleştirildikten sonra her madde için 2 (gruplar) x 2 (madde puanları) x M (puan düzeyi) olasılık çizelgesi olarak isimlendirien üç boyutlu bir matris oluşturulur. MH_{χ}^2, 1 serbestlik derecesinde, yaklaşık normal dağılıma sahip bir χ^2 istatistiğidir. MH_{χ}^2 istatistiği için kritik değer 0.05 manidarlık düzeyinde 3.84, 0.01 manidarlık düzeyinde ise 6.63'tür (Penfield, 2013). Son olarak bu çalışmada ikili puanlanan maddelerde cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddelerin tespiti açısından kullanılan yöntem olan SIBTEST, bir veya birde çok maddede MİF değerlendirmede kullanılan ve hipotez testi yöntemini kullanan nonparametrik bir yöntemdir. Bu yöntem Shealy ve Stout (1993)'ün çok boyutlu madde tepki kuramına göre MİF belirleme yöntemlerine dayanmaktadır. Model daha sonra Roussos ve Stout (1996) tarafından geliştirilmiştir. SIBTEST, testin ölçmek istediği örtük özellik bakımında bireyleri birbirleri ile eşleştirir ve bir veya birden çok madde üzerindeki performanslarda bir farklılık olup olmadığını inceler. SIB istatistiği, β_U, hesaplanırken ilk başta N maddelik test n maddelik alt testlere ayrılır ve N-n kadar şüpheli madde içerir. SIBTEST istatistiği yanlı madde olmadığı durumda N(1,0) olacak biçimde normal dağılıma sahiptir ($\beta_U = 0$) Bu yüzden odak gruba karşı MİF hipotezi H_0: $\beta_U = 0$ ve H_A: $\beta_U > 0$ ile test edilebilir.

Çoklu puanlanan maddelerde cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddelerin tespiti açısından kullanılan yöntemlerden biri olan Kısmi puan modeli, Rasch modelinin, çoklu puanlanan maddeler için uyarlanmış bir uzantısı olarak düşünülebilir. Kısmi puan modelinde, Rasch modelindeki maddenin güçlük parametresinin yerini,δ_{ij} , madde adım güçlüğü (item step difficulty) parametresi almaktadır. δ_{ij} , ardışık kategori yanıt eğrilerinin kesiştiği noktanın yetenek düzeyi

olarak yorumlanabilir. Bu yüzden madde parametreleri kategori kesişim parametresi (category intersection parameter) olarak da adlandırılmaktadır (Embretson ve Reise, 2000). δ_{ij} değeri ne kadar büyürse, bir maddeyi yanıtlamak için gerekli adımlardan ilgili olanının o kadar güç olduğu şeklinde yorumlanır. Kısmi puan modeli için herhangi bir kategoride yanıt verme olasılığı aşağıdaki formülle gösterilir:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{j=0}^{x}(\theta-\delta_{ij})\right]}{\sum_{r=0}^{mi}\left[\exp\sum_{j=0}^{r}(\theta-\delta_{ij})\right]} \quad (2)$$

Formulde mi; kategori eşik paremetre sayısını, δ_{ij}; j ile puanlanan kategoriye ait adım güçlüğü parametresini, x tepki kategorilerini temsil etmektedir. Çoklu puanlanan maddelerde cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddelerin tespiti açısından kullanılan yöntemlerden bir diğeri olan Mantel test, 1993 yılında ise Zwick ve diğerleri tarafından Mantel-Haenszel Testinin bir uzantısı olarak çok kategorili puanlanan maddelerde MİF belirlemede kullanılması önerilmiştir. MH χ^2 yöntemi çoklu puanlanan maddelere uyarlanmak istenildiğinde madde yanıt kategorilerinin sıralı ve gruplar arasında karşılaştırılabilir olduğu varsayımı öne çıkar. Mantel testi, hedef gruplardaki puanların eşleştirilmesine bağlı olarak madde ile grup üyelikleri arasındaki ilişkiyi inceler. Burada bahsedilen puan maddelerin toplamından elde edilen gözlenen puandır. Mantel test, MİF'e ilişkin yokluk hipotezini test etmek üzere bir serbestlik derecesinde ki-kare dağılımına ilişkin istatistik vermektedir (Zwick, Donoghue ve Grima, 1993). Son olarak bu çalışmada çoklu puanlanan maddelerde cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddelerin tespiti açısından kullanılan yöntem olan Poly-SIBTEST'te ise, Shealy ve Stout (1993) tarafından geliştirilen SIBTEST ikili puanlanan maddelerde MİF analizi için düzenlenen formunun, Chang ve arkadaşları (1996) tarafından yapılan araştırmada ikili puanlanan maddelerin dışında benzer prosedür sıralama yanıtlarına dayanan çok kategorili maddeler için uygulanmış halidir.

Geçmişten bugüne madde işlev farklılığını belirlemek için kullanılan araştırmalarda yöntemlerin belirli koşullar atında üstün yanları ve sınırlılıkları tartışılmaktadır. Örneğin Ackerman ve Evans'ın 1992 yılında yaptığı çalışmada çok sayıda maddenin MİF gösterdiği durumlarda SIBTEST'in MH yönteminden daha iyi performans gösterdiği sonucuna ulaşmıştır. Yine aynı koşul altında Roussoss (1992) SIBTEST'in madde işlev farklılığını belirlemede I. tip hata oranının MH yönteminden daha kabul edilebilir olduğunu göstermiştir. Bir diğer koşul olan benzer yetenek dağılımına sahip büyük ve küçük örneklemler için madde işlev farklılığını belirlemede I. tip hata oranının MH ve SIBTEST için düşük fakat yetenek dağılımları arasındaki fark arttığında I. tip hata oranının da yüksek olduğu sonucuna ulaşılmıştır (Roussos ve Stout, 1996; Shealy ve Stout 1993). Uttaro ve Millsap (1994) MH yönteminin farklı test uzunluklarında farklı I. tip hata oranları gösterdiğini bulmuşlardır. Chang ve Mazzeo (1993), poly-SIBTEST yöntemini geliştirmek amacıyla simülasyon veri üzerinden yaptıkları çalışmalarında poly-SIBTEST yöntemi ile GMH ve Mantel yöntemlerini karşılaştırmışlar ve poly-SIBTEST yönteminin MİF belirlemede Mantel Test ve GMH yöntemi kadar başarılı olduğu sonucuna ulaşmışlardır. Mellor ise 1995 yılında yaptığı çalışmasında çok kategorili maddelerde MİF belirlemek için GMH, poly- SIBTEST, OLR, LDFA yöntemlerini karşılaştırmış ve çalışmada 4 yöntemin farklı yetenek dağılımlarında tek biçimli olan ve olmayan MİF'i belirleme gücü ve yöntemlerin 1. tip hata oranlarını incelemiştir. Araştırma hem simülasyon veri hem de gerçek veri üzerinden yürütülmüş ve iki grubun yetenek dağılımlarının aynı olduğu durumlarda dört yöntemin de her kategoride tek boyutlu MİF'i belirlemede başarılı fakat GMH ve poly-SIBTEST'in diğer yöntemlerden daha hassas olduğu sonucuna ulaşmıştır. Roussos ve Stout (1996) similasyon veriyle yaptığı çalışmasında, küçük örneklemlerde MH ve SIBTEST yöntemlerinin I. tip hata oranları arasında büyük farklılıklar tespit etmemiştir. Fakat 3000 kişilik büyük örneklemde SIBTEST ve MH yöntemleri MİF içermeyen maddeler için orta ve önemli düzeyde MİF içerdiği yönünde isabetli olmayan kararlar vermiştir. Henderson (1999), çok kategorili verilerde MİF belirlemek için kullanılan GMH, poly-SIBTEST ve LDFA yöntemlerini gerçek veri kullanarak cinsiyet açısından karşılaştırmıştır. Araştırmanın sonucunda, Poly – SIBTEST yöntemi en çok sayıda MİF gösteren

maddeyi belirlemiştir. Ayrıca bu maddelerin büyük bir çoğunluğu GMH ve LDFA analizleri sonuçlarıyla da uyumlu bulunmuştur. Gierl, Jodoin ve Ackerman (2000)'nın farklı örneklem büyüklükleri ile yaptıkları çalışmada örneklem büyüklüğü arttıkça SIBTEST yöntemi ile MİF gösteren madde sayısının arttığı sonucuna ulaşmışlardır. Ayrıca bu çalışmaya ek olarak bir çok çalışmada SIBTEST için örneklem büyüklüğü arttıkça yöntemlerin I. tip hata oranının arttığı bulunmuştur (Narayanan ve Swaminathan, 1994; Rogers ve Swaminathan, 1993; Rousses ve Stout, 1996). Awuor, 2008 ise çalışmasında eşit olmayan örneklem grupları için MH yönteminin I. Tip hatayı SIBTEST yönteminden daha isabetli olarak kontrol ettiği sonucuna ulaşmıştır. Taylor ve Lee (2012) çalışmalarında cinsiyete ilişkin yanılılığı Poly-SIBTEST ve Rash modeli yöntemleri ile incelemişlerdir. Araştırmanın sonucunda Rasch modeli SIBTEST yönteminden daha çok maddeyi MİF'li olarak tespit etmiştir.

Atalay, Gök, Kelecioğlu ve Arsan (2012), örneklem büyüklüğü (odak ve referans gruplar için eşit örneklem büyüklükleri 400-400; 1500- 1500), yetenek dağılımı [(N(0,1) ve N(0,1); (N(0,1) ve N(0.5,1)] ve testteki MİF'li madde oranı (%5 ve %10)'nın değiştiği koşullarda SIBTEST, MTK-OO, MH ve LR yöntemlerini karşılaştırdıkları bir simülasyon çalışması yapmışlardır. Araştırma sonucunda, MİF'li maddeleri belirlemede MTK-OO yönteminin SIBTEST yönteminden; SIBTEST yönteminin LR ve MH yöntemlerinden, MH yönteminin ise LR yönteminden daha duyarlı olduğu sonucuna varılmıştır. Ayrıca MH yönteminin diğer yöntemlerle uyum yüzdesinin genel olarak düşük olduğu sonucuna ulaşmışlardır. Ulutaş (2012) PISA 2006 fen okuryazarlığı testinde yer alan maddelerin cinsiyet açısından MİF gösterip göstermediğini incelemiş, Türkiye ve Amerika arasında kültürlerarası eşdeğerliğini araştırmıştır. Araştırmacı MİF gösteren maddelerin belirlenmesi için klasik test kuramına dayanan Mantel-Haenszel, SIBTEST ve madde tepki kuramına dayanan olabilirlik oran analizi yöntemlerini kullanmıştır. Türkiye örnekleminde cinsiyete yönelik MİF belirlemek için yapılan analizlerde üç yöntemin tutarlı olarak MİF gösterdiğini belirlediği madde bulunmamıştır. Arıkan, Uğurlu, Atar (2016) MIMIC, SIBTEST, Lojistik Regresyon ve Mantel-Haenszel Yöntemleriyle Gerçekleştirilen MİF ve Yanlılık Çalışması isimli araştırmalarında örneklem büyüklüğü arttıkça SIBTEST yönteminin belirlediği DMF'li madde sayısı arttığını göstermiştir. Ayrıca SIBTEST ve MH yöntemlerine göre DMF içeren maddelere baktığımızda genel olarak iki yöntemde de MİF içeren ortak maddelerin tutarlı olduğu sonucuna ulaşmışlardır.

Yapılan araştırmalarda madde işlev farklılığı belirlemede kullanılan yöntemlerin üstün yanları ve sınırlılıkları tartışılmaktadır. Her bir yöntemin dayandıkları varsayımlar ve matematiksel modeller açısından madde işlev farklılığı belirlemede avantajlı ve dezavantajlı yanları vardır. Bu nedenle bu çalışmada birden çok yöntemin bir arada kullanılıp sonuçların karşılaştırılmasına karar verilmiştir.

Araştırmanın Amacı

Araştırmada, madde işlevinin farklılaşması hakkında istatistiki kanıt bulmak üzere farklı özelliklere sahip olan altı MİF belirleme yöntemine (SIBTEST, Mantel-Haenszel, Rasch Modeli Poly-SIBTEST, Mantel Test, Kısmi Puan Modeli) başvurulmuştur. MİF çalışmalarında birden çok yönteme dayalı inceleme yapılması önerilmektedir (Holland ve Wainer, 1993; Osterlind ve Everson, 2009). Söz konusu yöntemler kullanılarak yapılan analiz sonuçlarının tutarlılığının belirlenmesi alan yazında önem teşkil etmektedir. Türkiye'de MİF belirlemeye yönelik yapılan çalışmalar incelendiğinde, bu çalışmaların daha çok simülatif veri kullanılarak çeşitli koşullar altında MİF belirleme yöntemlerinin nasıl işlediğini gösteren çalışmalar olduğu görülmektedir. Bu araştırma, Türkiye'de psikolojik testlerde, özellikle zekâ testlerinde MİF belirlemeye yönelik yapılan, gerçek veri kullanılan ilk çalışmadır. Bu açıdan bu çalışma daha önce çeşitli koşullar altında denemesi yapılan MİF belirleme yöntemlerinin gerçek veri üzerinde ve psikolojik testlerde test edilmesi açısından önemlidir.

Yukarıda yapılan açıklamalara dayalı olarak bu araştırmanın problemini, Wechsler Çocuklar için Zekâ Ölçeği IV Türkçe formundaki maddelerin sosyo-ekonomik düzey ve cinsiyete göre yanlılık taşıyıp taşımadığını birden fazla yöntemle araştırarak yöntemleri karşılaştırmak oluşturmaktadır. Bu

problem çercevesinde bu araştırmaın amacı, Wechsler Çocuklar için Zekâ Ölçeği IV Türkçe formundaki maddelerin sosyo-ekonomik düzey ve cinsiyete göre madde işlev farklılığını belirlemede kullanılan istatistiksel yöntemlerin (Mantel-Haenszel, SIBTEST ve Rasch MODELİ, Mantel Test, Poly-SIBTEST ve Kısmi Puan Modeli) birbiri ile uyumlu sonuçlar verip vermediğinin karşılaştırılmasıdır.

YÖNTEM

Çalışma Grubu

Bu çalışmada, Wechsler Çocuklar için Zekâ Ölçeği IV Türkiye Uyarlama ve Standardizasyon Çalısması ön uygulama verileri kullanılmıştır. WÇZÖ-IV standardizasyon ön uygulama çalışması doğrultusunda, stardardizasyon çalışmasını yürüten proje ekibinin eğitim verdiği psikologlar tarafından 819 kişiden veri toplanmıştır. Ancak uç değerler nedeniyle 2 kişiye ait veriler analizlere dâhil edilmemiş, analizler 817 kişi üzerinden yürütülmüştür. Testi alan bireylerin sosyo-ekonomik düzeye göre gruplandırılmaları uyarlama çalışmasındaki ölçüte uygun olarak anne eğitim durumları dikkate alınarak yapılmıştır. Okuryazar değil, okuryazar, ilkokul ve ortaokul mezunları düşük; lise mezunları orta; üniversite, yüksek lisans ve doktora mezunları ise yüksek sosyoekonomik düzey olarak gruplandırılmıştır. Testi alanların cinsiyet ve sosyoekonomik düzeye göre dağılımları Tablo 1'de verilmiştir.

Tablo 1. Testi Alan Çocukların Cinsiyet ve Sosyoekonomik Düzeye Göre Dağılımları

Cinsiyet	Frekans	Yüzde
Kadın	443	54.22
Erkek	374	45.78
Toplam	817	100.00
Sosyo-Ekonomik Düzey	Frekans	Yüzde
Yüksek SED	156	19.09
Orta SED	254	31.09
Düşük SED	407	49.82
Toplam	817	100.00

Verilerin Elde Edilmesi

Araştırmada kullanılan veriler, Türkiye'de Wechsler Çocuklar için Zekâ Ölçegi IV (WÇZÖ-IV) Türkiye Norm, Uyarlama ve Standardizasyon Çalışmaları 107K493 ve 109K533 projeleri kapsamında TÜBİTAK destekli olarak Öktem, Gençöz, Erden, Sezgin ve Uluç tarafından (2007-2011 tarihleri arasında) tamamlanan çalışma kapsamında proje ekibinin eğitim verdiği psikologlar tarafından toplanmıştır (Öktem ve diğerleri, 2013). Araştırmada, bu projeler kapsamında toplanan verilerden, Wechsler Çocuklar için Zekâ Ölçeği IV Türkiye Uyarlama ve Standardizasyon Çalısması ön uygulama verileri, proje ekibinin izni alınarak kullanılmıştır.

Veri Toplama Aracı

WÇZÖ-IV David Weschsler'in teorisi temel alınarak geliştirilmiş, bireysel olarak uygulanan bir zekâ testidir. WÇZÖ-IV önceki versiyonlarla karşılaştırıldığında yapısının büyük ölçüde değişime uğradığı görülmektedir. WÇZÖ-R'dan Yapal Zekâ Puanı, Sözel Zeka Puanı ve Tüm Test Zeka Puanı olmak üzere üçtür birleşik puan elde edilmektedir. WÇZÖ-IV için ise Sözel Kavrama Birleşik Puanı (SKBP), Algısal Akıl Yürütme Birleşik Puanı (AAYBP), Çalışma Belleği Birleşik Puanı (ÇBBP), İşlem Hızı Birleşik Puanı (İHBP), Tüm Test Zekâ Puanı (TTZP) olmak üzere beş ayrı birleşik puan elde edilmektedir (Flanagan ve Kaufman, 2004).

Wechsler Çocuklar için Zekâ Ölçeği-IV 10' u asıl, 5'i yedek olmak üzere 15 alttesten oluşmaktadır. Arastırma kapsamında 3 alttest hız testi (Şifre, Simge Arama, Çiz Çıkar) olduğundan analizlere dâhil edilmemiştir. Analizler 6 tanesi çoklu puanlanan, 6 tanesi ikili puanlanan maddelerden oluşan toplam 12 alttest, 315 madde üzerinden yürütülmüştür.

Verilerin Analizi ve Yorumlanması

Öncelikle veriler, cinsiyet ve sosyoekonomik statü göz önüne alınarak düzenlenmiştir. Ardından, Wechsler Çocuklar için Zeka Ölçeği IV Türkçe formunda ikili puanlanan maddeler açısından Rasch modeli, çoklu puanlanan maddeler açısından da Kısmi Puan Modeli ile yapılan analizlerde cinsiyete ve sosyoekonomik düzeye göre MİF gösteren maddelerin tespiti için Rasch modeli kestirimleri yapılmıştır. Rasch modeli doğrultusunda hesaplanan MİF istatistikleri WINSTEP programı ile gerçekleştirilmiştir.

Maddelerin cinsiyete ve sosyoekonomik düzeye göre MİF içerip içermediği MİF kontrast, t ve p değerlerine göre tespit edilmektedir. Bond ve Fox (2001, 2007) maddenin MİF içerdiğinin göstergesi olarak, MİF kontrast±0.5 (MİF Kontrast ≥ |0.5|)'den büyük ve MİF'in istatistiksel olarak anlamlılığının tespiti için t değerinin ±2.0(t≥|2.0|, p<0.05)' den büyük olmak üzere kriterlerin bir arada kullanılmasını önermektedir. Negatif DMF Kontrast değeri maddenin odak gruba kolay geldiğini ve maddenin odak gruba avantaj sağladığını göstermektedir. Bu çalışmada ikili puanlanan maddeler açısından Rasch modeli, çoklu puanlanan maddeler açısından da Kısmi Puan Modeli ile yapılan analizlerde cinsiyete ve sosyoekonomik düzeye göre MİF gösteren maddelerin tespiti için bu kriterler kullanılmıştır.

Son olarak, Wechsler Çocuklar için Zeka Ölçeği IV Türkçe formundaki maddelerin sosyo-ekonomik düzey ve cinsiyete göre işlev farklılığı gösterip göstermediğine karar vermek için analizler WÇZÖ-IV'ün ikili puanlanan maddelerden oluşan alt testleri için Mantel-Haenszel, SIBTEST yöntemleri kullanılarak, WÇZÖ-IV'ün çoklu puanlanan maddelerden oluşan alt testleri için Mantel Test ve Poly-SIBTEST yöntemleri kullanılarak incelenmiştir. Mantel-Haenszel ve Mantel Test'e ilişkin analizler DIFAS programı ile SIBTEST ve Poly-SIBTEST'e ilişkin analizler ise SIBTEST programı ile gerçekleştirilmiştir.

Mantel-Haenszel ve Mantel Test analizlerinde ilgili maddelerin MİF gösterip göstermediğine karar vermek için χ^2_{Mantel} ve $\chi^2_{Mantel\text{-}Haenszel}$ istatistiği kullanılmıştır. Analiz sonucu elde edilen ki-kare istatistiği bir serbestlik derecesinde ki-kare dağılımı göstermektedir (Mantel, 1963; Zwick ve diğerleri,1993; Zwick ve diğerleri, 1997). Bu istatistik için 0.05 anlamlılık düzeyinde kritik değer 3.84, 0.01 anlamlılık düzeyinde kritik değer ise 6.63'tür. Ayrıca Mantel-Haenszel ve Mantel Test analizleri için $\chi^2_{Mantel\text{-}Haenszel}$ ve χ^2_{Mantel} istatistiğine göre MİF gösteren maddelerin hangi alt grup lehine MİF gösterdiğinin belirlenmesi için ikili pualanan maddeler için Mantel-Haenszel-Lojistik Odds Oranı, çoklu puanlanan maddeler için Liu-Agesti- Lojistik Odds Oranı istatistiği kullanılmıştır. Mantel-Haenszel-Lojistik Odds Oranı ve Liu-Agesti- Lojistik Odds Oranının pozitif değerleri o maddenin referans grubun lehine, negatif değerleri ise o maddenin odak grubun lehine çalıştığını gösterir (Penfield, 2013). SIBTEST ve Poly-SIBTEST yöntemi ile MİF gösteren maddeler ve MİF miktarı $|\beta_{UNI}|<0{,}059$ ise maddede ihmal edilebilir düzeyde (A düzeyi); $0{,}059\leq |\beta_{UNI}|<0{,}088$ ise maddede orta düzeyde (B düzeyi); $|\beta_{UNI}|\geq 0{,}088$ ise maddede önemli düzeyde (C düzeyi) ölçütü dikkate alınarak belirlenmiştir. Ayrıca, pozitif β_{UNI} değeri için o maddenin referans grubun lehine, negatif β_{UNI} değeri ise o maddenin odak grubun lehine çalıştığını göstermektedir (Roussos ve Stout, 1996).

BULGULAR

Bu çalışmada verilerin analizi bölümünde belirtilen ölçütlere göre bir maddenin MİF (Madde İşlevsel Farklılığı) içerip içermediğine karar vermek için ikili puanlanan maddeler açısından Rasch

Modeli(MİF Kontrast), Mantel-Haenszel($\chi^2_{\text{Mantel-Haenszel}}$), SIBTEST($\beta_u$) yöntemlerinden, çoklu puanlanan maddeler açısından da Kısmi Puan Modeli(MİF Kontrast), Mantel Test(χ^2_{Mantel}) ve Poly-SIBTEST(β_U) ile yapılan analizlere ilişkin bulgulara yer verilmiştir.

Yukarıdaki ölçütlere göre madde işlev farklılığını belirlemede kullanılan istatistiksel yöntemler (Mantel-Haenszel($\chi^2_{\text{Mantel-Haenszel}}$), SIBTEST($\beta_U$) ve Rasch Modeli(MİF Kontrast), Mantel Test(χ^2_{Mantel}), Poly-SIBTEST (β_U) ve Kısmi Puan Modeli (MİF Kontrast)) açısından her bir alt teste ait cinsiyet açısından MİF içeren maddeler Tablo 2'de, sosyo-ekonomik düzey açısından MİF içeren maddeler ise Tablo 3'de gösterilmiştir.

Tablo 2. İki Koşul İçin Üç Yönteme Göre Her Bir Alt Testteki Cinsiyet Açısından MİF Gösteren Maddeler

Küplerle Desen		
χ^2_{Mantel}	β_u	MİF Kontrast
7	7	-
Benzerlikler		
χ^2_{Mantel}	β_u	MİF Kontrast
5-8-9-21-23	5-8-9	5-23
Sayı Dizisi		
χ^2_{Mantel}	β_u	MİF Kontrast
-	-	-
Resim Kavramları		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
12-15	12-15	6-9
Sözcük Dağarcığı		
χ^2_{Mantel}	β_u	MİF Kontrast
20-21-25	17-20-21-25	-
Harf-Rakam Dizisi		
χ^2_{Mantel}	β_u	MİF Kontrast
-	6	2-9
Mantık Yürütme Kareleri		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
26	26	-
Kavrama		
χ^2_{Mantel}	β_u	MİF Kontrast
16	16	1
Resim Tamamlama		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
3-6-8-10-15-17-18-19-22-24-32-33	10-15-17-18-19-22-24-32-33	3-6-10-15-17-22-32
Genel Bilgi		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
10-14-17-21	14-17	10-14-17-21
Aritmetik		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
8-10-27	27	8-27

Sözcük Bulma		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
2-4-8-13	-	2-4-8

Madde işlev farklılığını belirlemede kullanılan istatistiksel yöntemler (ikili puanlanan maddeler açısından Mantel-Haenszel, SIBTEST, Rasch MODELİ ve çoklu puanlanan maddeler açısından Mantel Test, Poly-SIBTEST ve Kısmi Puan Modeli) ile yapılan analizlerde, WÇZÖ-IV zeka testinin *ana alt testlerinden; küplerle desen* alt testinde 7. maddede kızlar lehine; *benzerlikler* alt testinde 5., 9., ve 23. maddelerde kızlar lehine; 8. ve 21. maddeler ise erkekler lehine; *resim kavramları* alt testinde 9. ve 12. maddelerde erkekler lehine, 6. ve 15. maddelerde ise kızlar lehine; *sözcük dağarcığı* alt testinde 20 ve 25. maddelerde kızlar lehine, 17 ve 21. maddelerde erkekler lehine; *harf-rakam dizisi* alt testinde 6. maddede erkekler lehine, 2. ve 9. maddelerde kızlar lehine; *mantık yürütme kareleri* alt testinde 26. maddede kızlar lehine; *kavrama* alt testinde 1. ve 16. maddelerde erkekler lehine madde işlev farklılığı tespit edilmiştir.

WÇZÖ-IV zeka testinin *yedek alt testlerinden* yine kullanılan yöntemlerin en az ikisine göre; *resim tamamlama* alt testinde 3., 6., 8., 10., 15., 22., 24. ve 33. maddelerde kızlar lehine, 17., 18., 19., 22. ve 32. maddelerde erkekler lehine; *genel bilgi* alt testinde 10., 14., 17. maddelerde erkekler lehine, 21. maddede kızlar lehine; *aritmetik* alt testi için 8. ve 10. maddeler için kızlar lehine, 27. madde için erkekler lehine; *sözcük bulma* alt testi için 2 ve 8. maddeler için kızlar lehine, 4. ve 13. maddeler için erkeler lehine madde işlev farklılığı bulunmuştur. Ana alt testlerden *sayı dizisi* alt testinde madde işlev farklılığı gösteren maddeye rastlanmamıştır.

Tablo 3. İki Koşul İçin Üç Yönteme Göre Her Bir Alt Testteki Sosyoekonomik Düzey Açısından MİF Gösteren Maddeler

YÜKSEK SED-DÜŞÜK SED			DÜŞÜK SED-ORTA SED			YÜKSEK SED-ORTA SED		
Küplerle Desen			Küplerle Desen			Küplerle Desen		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
14	3-14	-	-	3-4	-	8	4-8	-
Benzerlikler			Benzerlikler			Benzerlikler		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
-	19-23	1-2-4-22	23	-	23	16	-	4
Sayı Dizisi			Sayı Dizisi			Sayı Dizisi		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
-	13	-	6	6	-	6	6	6
Resim Kavramları			Resim Kavramları			Resim Kavramları		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast	$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast	$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
16	16	16	6	-	6	13	-	13
Sözcük Dağarcığı			Sözcük Dağarcığı			Sözcük Dağarcığı		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u		χ^2_{Mantel}	β_u	MİF Kontrast
7-8-20-27-32	27-32	1-2-3-4-7-32-33	28	20-28	-	27	27-32	1-2-3-4-7-32
Harf-Rakam Dizisi			Harf-Rakam Dizisi			Harf-Rakam Dizisi		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
-	3-4	1	3	3	9	-	-	-

Mantık Yürütme Kareleri			Mantık Yürütme Kareleri			Mantık Yürütme Kareleri		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast	$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast	$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
-	-	23	22	22	22	22	22	-
Kavrama			Kavrama			Kavrama		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
7-13-16	7-13	3-7-22	2-5-13-22	5-13	22	-	-	3-20
Resim Tamamlama			Resim Tamamlama			Resim Tamamlama		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
11-24-31	11-24-31	5-11-24-26-31-35-36	26	26	26-38	38	38	4-11-36-38
Genel Bilgi			Genel Bilgi			Genel Bilgi		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
8-20-25	20-25	8-15-20-22-25-27	-	-	6-8-20-28	-	-	13-15
Aritmetik			Aritmetik			Aritmetik		
χ^2_{Mantel}	β_u	MİF Kontrast	$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast	$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
19-23	19-21-23	12-20-23	19-22-28	5-11-19-22-28	7-8-9-11-17-18-19-20-27-28-30-31-32	21-28	21-28	20-21-28
Sözcük Bulma			Sözcük Bulma			Sözcük Bulma		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u		χ^2_{Mantel}	β_u	MİF Kontrast
13	12	8-17-21	12-17	12-17	1-12-17	-	-	-

Madde işlev farklılığını belirlemede kullanılan istatistiksel yöntemler (ikili puanlanan maddeler açısından Mantel-Haenszel, SIBTEST, Rasch MODELİ ve çoklu puanlanan maddeler açısından Mantel Test, Poly-SIBTEST ve Kısmi Puan Modeli) ile yapılan analizlerde yüksek-düşük sosyo-ekonomik düzey karşılaştırıldığında; WÇZÖ-IV zeka testinin *ana alt testlerinden; küplerle desen* alt testinde 14. maddede düşük sed lehine, 3. maddede yüksek sed lehine; *benzerlikler* alt testinde 4., 19. ve 22. maddelerde düşük sed lehine, 1., 2. ve 23. maddelerde yüksek sed lehine, sayı dizisi alt testinde 13. maddede düşük sed lehine; *resim kavramları* alt testinde 16. maddede yüksek sed lehine; *sözcük dağarcığı* alt testinde 1., 2., 3., 4., 7. ve 32. maddelerde yüksek sed lehine, 8., 20., 27. ve 33. maddelerde düşük sed lehine; *harf-rakam dizisi* alt testinde 3. maddede yüksek sed lehine, 1. ve 4. maddelerde düşük sed lehine; *mantık yürütme kareleri* alt testinde 23. maddede yüksek sed lehine; *kavrama* alt testinde 3., 7., 16. ve 22. maddelerde düşük sed lehine, 13. maddede de yüksek sed lehine madde işlev farklılığı tespit edilmiştir.

WÇZÖ-IV zeka testinin *yedek alt testlerinden* yine kullanılan yöntemlerin en az ikisine göre; *resim tamamlama* alt testinde 11., 26., ve 31. maddelerde yüksek sed lehine, 5., 24., 35. ve 36. maddede ise düşük sed lehine; *genel bilgi* alt testinde 8., 20., 22., 25. ve 27. maddelerde düşük sed lehine; 15. madde de ise yüksek sed lehine; *aritmetik* alt testi için 19., 20. ve 23. maddelerde yüksek sed lehine, 21. maddede düşük sed lehine; *sözcük bulma* alt testinde 8., 12., 13. ve 17. maddelerde ise yüksek sed lehine, 21. maddede düşük sed lehine madde işlev farklılığı bulunmuştur.

Orta-düşük sosyo-ekonomik düzey karşılaştırıldığında; WÇZÖ-IV zeka testinin ana alt testlerinden; *benzerlikler* alt testinde 23. maddede düşük sed lehine; *sayı dizisi* alt testinde 6. maddede düşük sed

lehine; *resim kavramları* alt testinde 6. maddede orta sed lehine; *sözcük dağarcığı* alt testinde 20. ve 28. maddelerde düşük sed lehine; *harf ve rakam dizisi* alt testinde 3. ve 9. maddelerde orta sed lehine; *mantık yürütme kareleri* alt testinde 22. maddede düşük sed lehine; *kavrama* alt testinde 5., 13. ve 22. maddelerde orta sed lehine, 2. maddede düşük sed lehine madde işlev farklılığı tespit edilmiştir.

WÇZÖ-IV zeka testinin *yedek alt testlerinden* yine kullanılan yöntemlerin en az ikisine göre; *resim tamamlama* alt testinde 26. maddelerde ise düşük sed lehine, 38. maddede orta sed lehine; *genel bilgi* alt testinde 6., 8., 20. ve 28. maddelerde düşük sed lehine; *aritmetik* alt testi için 5., 9., 11., 17., 18., 20., 22., ve 28., maddelerde düşük sed lehine, 7., 8., 19., 27., 30., 31. ve 32. maddelerde orta sed lehine; *sözcük bulma* alt testinde 12. maddede orta sed, 1. ve 17. maddelerde ise düşük sed lehine madde işlev farklılığı bulunmuştur. Ana alt testlerden *küplerle desen,* yedek alt testlerden ise *genel bilgi* alt testinde madde işlev farklılığı gösteren maddeye rastlanmamıştır.

Yüksek-orta sosyo-ekonomik düzey karşılaştırıldığında ise WÇZÖ-IV zeka testinin ana alt testlerinden; *küplerle desen* alt testinde 8. maddede orta sed lehine; *benzerlikler* alt testinde 4. ve 16. maddelerde yüksek sed lehine; *sayı dizisi* alt testinde 6. maddede yüksek sed lehine; *resim kavramları* alt testinde 13. maddede orta sed lehine; *sözcük dağarcığı* alt testinde 1., 2., 3., 4. ve 27. maddelerde orta sed lehine, 7 ve 32. maddelerde yüksek sed lehine; *mantık yürütme kareleri* alt testinde 22. maddede yüksek sed lehine; *kavrama* alt testinde 3. ve 20. maddelerde orta sed lehine madde işlev farklılığı tespit edilmiştir.

WÇZÖ-IV zeka testinin *yedek alt testlerinden* yine kullanılan yöntemlerin en az ikisine göre; *resim tamamlama* alt testinde 4. maddede yüksek sed lehine, 11., 36. ve 38. maddede orta sed lehine; *aritmetik* alt testi için 21. maddede orta sed lehine, 20. ve 28. maddelerde yüksek sed lehine madde işlev farklılığı bulunmuştur. Ana alt testlerden *harf-rakam dizisi,* alt testinde, yedek alt testlerden ise *genel bilgi, sözcük bulma* alt testinde madde işlev farklılığı gösteren maddeye rastlanmamıştır.

SONUÇLAR ve TARTIŞMA

WÇZÖ-IV'ün maddelerin cinsiyet ve sosyo-ekonomik düzeye göre işlev farklılığı gösterip göstermediğine karar vermek için literatür (Bond ve Fox 2001; Mantel, 1963; Penfield, 2013; Roussos ve Stout, 1996; Zwick ve diğerleri, 1993; Zwick ve diğerleri, 1997) doğrultusundaki ölçütlere göre, her bir alt teste ait cinsiyet ve sosyo-ekonomik düzey açısından üç yönteme göre MİF içeren maddeler gösterilmiştir.

Yöntemler arasında MİF içeren madderin tutarlılığına bakılacak olursa cinsiyet açısından MİF içeren maddelerin tespitinde Mantel-Haenszel, SIBTEST ve Mantel Test, Poly-SIBTEST istatistikleri daha tutarlı sonuçlar verirken sosyoekonomik düzeye göre MİF içeren maddelerin tespitinde Mantel-Haenszel, Rasch MODELİ ve Mantel Test, Kısmi Puan Modelinin daha tutarlı sonuçlar verdiği görülmektedir. Örneklem büyüklükleri ve örneklem eşitsizlikleri aralardaki tutarsızlıkların bir nedeni olarak görülebilir. Bu çalışmanın örnekleminde cinsiyet açısından 443 kız, 374 erkek sosyo-ekonomik düzey değişkeni açısından yüksek sed'den gelen 156, orta sed'den gelen 254, düşük sed'den gelen ise 407 kişi bulunmaktadır. Örnekleme bakıldığında cinsiyet açısından ayrılan grubun örneklem büyüklüğü daha fazla ve gruplar açısından daha dengelidir. Ancak sosyo-ekonomik düzey açısından Türkiye'deki dağılıma da benzer olarak daha dengesiz bir dağılım görülmektedir ve örneklem büyüklüğü daha küçüktür. Bu açıdan bulgular incelendiğinde her üç yöntemin en tutarsız sonuçlar verdiği grubun, örneklem eşitsizliğinin en yüksek olduğu Yüksek SED düşük sed'in karşılaştırıldığı grup olduğu görülmektedir. Bu grupta en çok MİF'li maddeyi Rasch modeli ile tespit edilmiştir. Bu bulgu Taylor ve Lee (2012)'nin karışık madde formatlarında matematik ve okuma testinde cinsiyet yanlılığını Poly-SIBTEST ve Rash modeli ile inceleyen çalışmasının sonuçlarıyla paraleldir. Yine bu grupta SIBTEST, Poly-SIBTEST yöntemi Mantel-Haenszel, Mantel Test yönteminden daha fazla MİF'li madde belirlemiştir.

Awuor'un MİF belirleme tekniklerinin eşit olmayan örneklemlerde gücünü SIBTEST ve Mantel-Haenszel yöntemleriyle test ettiği çalışmasında eşit olmayan örneklem grupları için MH yönteminin I. Tip hatayı SIBTEST yönteminden daha iyi kontrol ettiği sonuçuna ulaşmıştır. Bu sonuca paralel olarak SIBTEST yöntemi ile daha fazla MİF'li maddenin belirlenmesinin I. tip hatadan kaynaklandığı düşünülebilir.

Arıkan, Uğurlu, Atar (2016) MIMIC, SIBTEST, Lojistik Regresyon ve Mantel-Haenszel Yöntemleriyle Gerçekleştirilen MİF ve Yanlılık Çalışmasında SIBTEST ve MH yöntemlerine göre genel olarak MİF içeren ortak maddelerin tutarlı olduğu sonucuna ulaşmışlardır. Benzer şekilde bu çalışmada da yüksek ve düşük sed'in karşılaştırıldığı grup dışındaki gruplarda yöntemler genel olarak tutarlı sonuçlar vermiştir.

Madde işlev farklılığı testle ölçülen özellik bakımından benzer olup da cinsiyet, sosyoekonomik düzey gibi değişkenler açısından birbirinden farklı alt gruplarda yer alan bireylerin, bir maddeyi doğru cevaplandırma olasılıklarının farklılaşması olarak tanımlanabilir (Hambleton, Swaminathan ve Rogers, 1991). Ancak aynı yetenek düzeyinde olan fakat farklı gruplardan gelen bireylerin, test maddelerine doğru cevap verme olasılıklarının değişmesi maddenin ölçtüğü özelliğin gerçekten bu iki grup arasında farklı olmasından da kaynaklanıyor olabilir. Bu durum madde etkisi olarak adlandırılmaktadır (Clauser, Mzaor, 1998; Mellenberg, 1983; Osterlind, 1983; Shepard, Camilli ve Williams, 1985; Zumbo, Hubley, 1998). Daha önce de bahsedildiği gibi maddelerin, madde yanlılığı içerip içermediğine ilişkin karar verebilmek için maddelerin MİF göstermesi bir gereklilik olmakla birlikte bu kararı vermek için yeterli değildir. MİF olduğu gözlenen maddelerin yanlı olup olmadığının ve yanlı ise bunun nedenlerinin uzmanlar tarafından incelenmesi gereklidir. Bu incelemede uzmanlar tarafından maddenin ölçülmek istenen yapıyla ilişkisiz olarak, bazı alt gruplar için adil olmayan bir avantaj sağlayıp sağlamadığının belirlemesi gerekir (Camilli ve Shepard, 1994; Zumbo, 1999).

Wechsler Çocuklar için Zeka Ölçeği IV (WÇZÖ-IV) 2003 yılında Amerika'da geliştirilip kullanıma sunulmuş, 2011 yılında ise Türkiye'de uyarlama çalışması tamamlanarak kullanılmaya başlanmıştır. Hambleton ve arkadaşlarının editörlüğünü yaptığı Adapting Educational and Psychological Tests for Cross-Cultural Assessment isimli kitabında uyarlanmış testlere ilişkin hataların ve geçerlik sorunlarının kaynağını kültürel ve dilsel farklılıklar, dizayn, metot gibi teknik konular ve sonuçların yorumlanması olmak üzere üç başlık altında toplamıştır. Kültürel ve dilsel farklılıklar yapıların eşdeğer olmaması, testin yönergesinin verilmesi ve testin uygulanışı ile ilgili sorunlar, madde formatları ve testi alanların hızlarının etkisi faktörlerinden etkilenmektedir. Dizayn, metot gibi teknik konular ise çevirmenlerin seçimi ve eğitimi, çeviri süreci, veri toplama süreci faktörleriyle ilişkilidir. Sonuçların yorumlanması ise eğitim programının benzerliği, öğrencinin motivasyonu, sosyo-politik özellikler gibi faktörlerden etkilenmektedir (Hambleton, 2005). Görüldüğü üzere bir testin uyarlama sürecinde testin geçerliğini etkileyen bir çok faktör vardır. Testin uyarlandıktan sonra yanlılık içermesi de testin geçerliğini tehdit eden sorunlardan biridir. Bir teste ilişkin yanlılık türleri yapı yanlılığı, metot yanlılığı ve madde yanlılığı (madde işlev farklılığı) olarak ayrılabilir. Bu çalışmada ele alınan madde yanlılığının nedenleri ise zayıf çeviriler, muğlak birden çok anlama gelen ifadeler, maddenin ilişkili olduğu istenmeyen faktörler (örnek olarak maddenin ölçülmek istenen yapı dışında başka bir yapıyı ölçmesi) olarak sıralanabilir (Vijver, Poortinga, 2005). Bu araştırmada, belirlenen farklı yöntemlere göre MİF gösteren maddelerin saptanıp bu yöntemler arasındaki uyum düzeyleri ortaya konulmuştur. Bundan sonraki çalışmalar adına MİF bulunan maddelerin uzmanlarca incelenerek yanlılık içerip içermediği belirlenmeli, cinsiyet ve sosyo-ekonomik düzeye göre MİF içeren maddeler için yanlılık içerdiği tespit edilir ise bu yanlılığın kaynağı ve nedenlerine ilişkin yanlılık gösteren maddeler, yukarıda belirtilen tüm faktörler göz önüne alınarak yorumlanmalı uzmanlar tarafından ayrıntılı olarak incelenmelidir.

KAYNAKÇA

Ackerman, T. A., & Evans, J. A. (1992). *An investigation of the relationship between reliability, power, and the Type I error rate of the Mantel-Haenszel and simultaneous item bias detection procedures.* Annual Meeting of the National Council on Measurement in Education, San Francisco.

Adams, R. J., & Rowe, K. J. (1988). Test bias. In J. P. Keeves (Ed), *Educational teearch, methodology, andmeasurement: An international handbook* (p. 398-403). Oxford: Pergamon Pres.

Arıkan, Ç., Uğurlu, S. ve Atar, B. (2016). MIMIC, SIBTEST, Lojistik Regresyon ve Mantel-Haenszel yöntemleriyle gerçekleştirilen DMF ve yanlılık çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(1), 34-52 .

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

Awuor, R. (2008). *Effect of unequal sample sizes on the power of DIF detection: An IRT-Based Monte Carlo study with SIBTEST and Mantel-Haenszel procedures* (Unpublished doctoral dissertation), Virginia Polytechnic Institute and State University Blacksburg, Virginia, US.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* New Jersey: Lawrence Erlbaum Associates.

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

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

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

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th Ed.). New York: Harper Collins Publishers.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists*. Mahwah, NJ: Erlbaum.

Gierl, M. J., Jodoin, M., & Ackerman, T. (2000). *Performance of Mantel-Haenszel, simultaneous item bias test and Logistic Regression when the proportion of DIF items is large.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana, USA.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. CA: Sage.

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

Holland P. W., & Wainer, H. (1993). *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel- Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (p. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Horst, P. (1966). *Psychological measurement and prediction.* Belmont: Wadsworth Pub. Co. Institute for Education Research.

Kamata, A., & Vaughn, K. B. (2004). An introduction to Differential Item Functioning analysis. *Learning Disabilities: A Contemporary Journal, 2*(2), 49-69.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

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

Mellenberg, G. J. (1983). Conditional item bias methods. In S. H. Irvine & W. J. Barry (Eds.), *Human assesment and cultural factors* (p. 293–302). New York: Plenum Pres.

Mellor, T. L. (1995). *A comparison of four differantial item functioning methods for polytomously scored items* (Unpublished doctor dissertation). The university of Texas, Austin.

Murphy, K., & Davidshofer, C. (1994). *Psychological testing: Principles and applications* (3th Ed.). Englewood Cliffs, NJ: Prentice Hall.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform bias. *Applied Psychological Measurement, 20*(3), 257-274.

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

Osterlind S. J., & Everson H. T. (2009). *Differential Item Functioning* (2nd Ed.). California: Sage.

Öktem, F., Gençöz, T., Erden, G., Sezgin, N. ve Uluç, S. (2013). *Wechsler çocuklar için zeka ölçeği-IV (WÇZÖ-IV) uygulama ve puanlama el kitabı Türkçe sürümü.* Türk Psikologlar Derneği Yayınları, Ankara.

Penfield, R. (2013). *DIFAS 5.0 Differential item functioning analysis system: User's manual*. http://soe.uncg.edu/wp-content/uploads/2015/12/DIFASManual_V5.pdf

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish

Reynolds. C. R., Livingston R. B., & Willson, V. (2006). *Measurement and assessment in education.* Boston: Pearson Education. Inc.

Rodney. G. L., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology. 75*, 164- 174.

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

Roussos, L. A. (1992). *Hierarchical agglomerative clustering computer program user's manual.* University of Illinois at Urbana-Champaign.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215-230.

Shealy, R., & Stout, W. (1993). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 197–240). Hillsdale, NJ: Earlbaum.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*, 77–105.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25*(3), 246-280.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin Company.

Tittle, C. K. (1988). Test bias. In J. Keeves (Ed.), Educational teearch, methodology, and measurement: An international handbook (p. 392- 398). Pergamon Press: UK.

Ulutaş, S. (2012). PISA 2006 fen okuryazarlığı testindeki maddelerin yanlılık bakımından araştırılması (Yüksek Lisans Tezi). Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü.

Uttaro, T., & Millsap, R. E. (1994). Factors in uencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*, 15-25.

Vijver, F. ve Poortinga, Y. (2005). Conceptual and methodological issues in adapting tests. In Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (p. 39-63). Mahwah, NJ: Lawrence Erlbaum.

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.

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

Zwick, R., Donoghue, J. R., Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233- 251.

EXTENDED ABSTRACT

Introduction

Bias in statistical terms is a systematic and constant error of measurement as opposed to chance error or a systematic under or over estimation of a population parameter by a statistic based on samples which are drawn from the population (Camilli ve Shepard, 1994; Zumbo 1999). It is unacceptable for educational and psychological tests to contain biased items. If some items provide an advantage to a specific subgroup of examinees that increases their ability to respond correctly, that is, if some items exhibit systematic bias, it is impossible to say that the test was equally fair for all examinees. A typical approach to investigate bias at the item level is called differential item functioning (DIF)

analysis, defined as a difference in the measurement properties of an item for demographic subgroups (Camilli, Shepard, 1994).

Differential item functioning identifies differences in the probability of answering an item correctly accordingly for identifiable subgroups, at every ability level of psychological structure that is aimed to be measured with an item (Embretson & Reise, 2000; Lord, 1980). Statistical procedures that are currently used by test publishers to identify items that function differentially tend to focus on such subgroups as gender, race, language, or socioeconomic status groups. DIF analyses are useful for flagging items that may need to be eliminated or, at least, submitted to additional review.

In the beginning of 1900's, it was recognized that some items that were used to measure IQ were also measuring the effects of cultural training instead of mental capacity. Group differences in IQ tests have been the major research area for many researchers.

The Wechsler Intelligence Scale for Children – Fourth Edition (WISC-IV; Wechsler, 2003) is an individually administered IQ test for children. The Wechsler scales are among the most adapted tests in the world. In developing or adapting a standardized test that is fair for all test takers requires the removal or the revision of potentially biased items. When practitioners are developing or adapting instruments, they should conduct DIF analysis to investigate possible bias at the item level.

The purpose of this study is to investigate potential gender and socio-economic status bias in the WISC-4 by using several differential item functioning detection techniques. In addition, this study will show similarities and differences in practice by comparing three DIF-detection techniques: the Rasch item response theory (IRT) estimations, SIBTEST (Shealy, Stout, 1993), and the Mantel-Haenszel techniques (Holland, Thayer, 1988).

Method

In this study, Wechsler Intelligence Scale for Children: Fourth Edition Turkish standardization pilot test data were used. In accordance with the Wechsler Intelligence Scale for Children: Fourth Edition standardization study, pilot data have been collected from 817 children by psychologists. Table 1 shows descriptive statistics broken down by gender and socio-economic status.

Table 1. Descriptive Statistics for Gender and Socio-Economic Status

Gender	Frequency	Percentage
Female	443	54.22
Male	374	45.78
Total	817	100
Socio-Economic Status	Frequency	Percentage
High	156	19.09
Moderate	254	31.09
Low	407	49.82
Total	817	100

The WISC-4 is an individually administered intelligence test, based on the theory of David Wechsler. Different from previous versions of the Wechsler tests, the structure of WISC-IV is observed to have undergone a significant change, approximately 45% of the entire test has been revised (Flanagan and Kaufman, 2009). WISC-R gives three types of combined scores: a Performance Intelligence Score, a Verbal Intelligence Score, and a Full Test Intelligence Score. However, the WISC-IV also provides additional indexes: namely a Verbal Comprehension Index (VCI), a Perceptual Reasoning Index (PRI), a Working Memory Index (WMI), a Processing Speed Index (PSI), and a Full Scale IQ (FSIQ).
In accordance with the purpose of the study, 315 items were used both in polytomously scored subtests such as *Block Design, Similarities, Digit Span, Vocabulary, Letter-Number Sequencing,*

Comprehension, and dichotomously scored subtest such as *Picture Concepts, Matrix Reasoning, Picture Completion, Information, Arithmetic*, and *Word Reasoning.* Three subtests were excluded from this research because they involved a speeded format.

First, data were prepared for analyses. Then Rasch and Partial Credit Model analysis were performed to detect differential item functioning. DIF values have been estimated using the IRT estimation software Winsteps. Bond and Fox (2001, 2007) suggest t value ± 2.0 ($t \geq |2.0|$, $p<0.05$) and DIF Contrast ± 0.5 ((DIF Contrast ≥ |0.5|) asDIF indicators based on the studied groups. In this research, these criterions were used for evaluating the items if they have DIF or not.
In addition to Rasch DIF analysis, DIF analyses were performed using two different techniques: SIBTEST-Poly-SIBTEST and Mantel-Haenszel Test-Mantel Test. SIBTEST-Poly-SIBTEST analyses have been computed using the SIBTEST software. To decide DIF according to SIBTEST results, β_{UNI} can be interpreted as the magnitude of DIF for each item. Positive values of β_{UNI} indicate DIF favoring the reference group, and negative values indicate DIF favoring the focal group. Roussos and Stout (1996) proposed guidelines for interpreting DIF by combining the SIBTEST statistical results with values for the β_{UNI} parameter estimate to classify DIF on a single item: (a) negligible DIF ($\beta_{UNI} < 0.059$ and H_0: $B = 0$ is rejected), (b) moderate DIF ($0.059 \leq \beta_{UNI} < 0.088$ and H_0: $B = 0$ is rejected), (c) large DIF, ($\beta_{UNI} > 0.088$ and H_0: $B = 0$ is rejected).
Mantel-Haenszel and Mantelanalyses had been computed using the DIFAS software. To decide DIF according to Mantel-Haenszel Test and Mantel Test results, critical values of this statistic are 3.84 for a Type I error rate of 0.05 and 6.63 for a Type I error rate of 0.01. The Mantel chi-square statistic (Mantel, 1963, Zwick, Donoghue & Grima, 1993; Zwick, Thayer & Mazzeo, 1997) is distributed as chi-square with one degree of freedom. To indicate DIF favoring group, Liu-Agresti cummulative common log-odds ratio (Liu & Agresti, 1996; Penfield & Algina, 2003) and *Mantel-Haenszel Common Log-Odds Ratio* (Camilli & Shepard, 1994; Mantel & Haenszel, 1959) statistics were used. This statistic is asymptotically normally distributed. Positive values indicate DIF in favor of the reference group, and negative values indicate DIF in favor of the focal groups.

Results and Discussion

In this reseach, for investigating potential gender and SES bias in the WISC-4, three DIF detection techniques were used. The items that are detected as DIF items based on gender and SES are summarized on Tables 2 and 3. In addition, this study has shown similarities and differences in practice by comparing three DIF-detection techniques: The Rasch item response theory (IRT) estimations, SIBTEST (Shealy, Stout, 1993), and the Mantel-Haenszel techniques (Holland, Thayer, 1988). In conclusion, as it is shown in Table 2 and Table 3, potential gender and SES biased items were identified according to three DIF-detection techniques. Even though items detected as DIF items with three techniquess were quite similar, there are also differences based on these three techniques. There are many reasons for differences based on techniques in the detection of DIF items. Also in order to detect and/or prevent bias, we need to recognize factors that can induce bias. It is important to understand from a substantive, cognitive perspective why these gender and SES differences and also differences based on techniques in the detection of DIF items are occurring.