BAŞKENT UNIVERSITY
JOURNAL OF EDUCATION

2018, 5(1), 40-50 ISSN 2148-3485

İki ve Çok Kategorili Puanlanan Maddelerde Değişen Madde Fonksiyonlarının Karşılaştırılması

Comparison of Differential Item Functioning for Two-category Scored and Multi-Category Scored Items

Emine Burcu Tunç[a*], Ömer Kutlu[b]

[a]*Marmara University,İstanbul,Turkey*
[b]*Ankara University,Ankara,Turkey*

Öz

Bu araştırmanın genel amacı, iki kategorili ve çok kategorili puanlanan maddelerde Değişen Madde Fonksiyonlarının (DMF) karşılaştırılmasıdır. Bu amaç doğrultusunda simülasyon çalışması gerçekleştirilmiş, I. tip hata ve istatistiksel güç oranları üzerinde çalışılmıştır. 20 madde için hem iki kategorili (1-0) hem çok kategorili (4-3-2-1-0) puanlama yapılmış ve böylelikle iki ayrı veri seti oluşturulmuştur. İki kategorili puanlama için, çok kategorili puanlamada beşinci adım olan 4'e 1 puan verilmiş, 3-2-1-0'a ise 0 puan verilmiştir. Simülasyon kapsamında örneklem büyüklüğü (600, 1200, 2400), örneklem büyüklüğü oranı (1:1, 1:2), DMF içeren madde yüzdesi (%10, %30, %50) ve DMF büyüklüğü (0.25, 0.50, 1.00, 1.50) manipüle edilen koşullar olarak ele alınırken, DMF biçimi (Tek Biçimli DMF) ve toplam madde sayısı (20) sabit koşullar olarak ele alınmıştır. Böylelikle 72 koşul kapsamında gerçekleştirilen araştırma için 100 tekrar gerçekleştirilmiştir. Verilerin türetilmesinde, iki kategorili veriler için Rasch, çok kategorili veriler için ise Kısmi Puan Modeli kullanılmıştır. WinGen programında madde parametreleri hesaplanmış, R programında "eRm" paketi yardımıyla veriler türetilmiştir. Hem iki kategorili hem çok kategorili puanlama kapsamında DMF belirlemek için LORDIF analizi kullanılmıştır. Araştırmanın temel amacı olan iki kategorili ve çok kategorili puanlama modelleri kapsamında DMF karşılaştırıldığında, genel olarak çok kategorili puanlama yapılması durumunda I. Tip hata oranlarının daha düşük, istatistiksel güç oranlarının ise daha yüksek olduğu belirlenmiştir. Bu doğrultuda DMF sonuçlarında puanlama modellerinin etkisi olduğu ve kısmi puan dikkate alındığında DMF sonuçlarının değişebileceği ortaya konulmuştur.

Anahtar Kelimeler: Değişen Madde Fonksiyonu, puanlama modelleri, Rasch, Kısmi Puan Modeli, LORDIF, I. Tip hata, istatistiksel güç.

Abstract

The aim of this study was to compare Differential Item Functioning for two-category scored and multi-category scored items. For this purpose, simulation studies were performed; Type I error and statistical power ratios were studied. For 20 items, both two category (1-0) and multi category (4-3-2-1-0) scoring was done and thus two data sets were created. Two category scoring was done by scoring 4, which was the fifth step of multi-category scoring as 1 and scoring the other steps 3-2-1-0 as 0. Whereas sample size (600, 1200, 2400), sample size ratio (1:1, 1:2), percentage of items containing DIF (%10, %30, %50), and DIF magnitude (0.25, 0.50, 1.00, 1.50) were taken as manipulated conditions, DIF format (Uniform DIF) and total item number (20) were considered as stable conditions as part of simulation process. Hereby 100 repetitions were carried out for the research conducted under 72 conditions. In the process of data derivation, Rasch was used for two-category data, and Partial Credit Model was used for multiple-category data. Item parameters were calculated with WinGen program, and data was derived with "eRm" package of R program. d LORDIF analyses were used for two category and multiple category data. When DIF was compared within the scope of two category and multiple category scoring models, which is the main purpose of this study, it was seen that when multiple category scoring was done, ratio of Type I error was lower, but statistical power ratio was higher. In this context, it was observed that scoring models effect DIF results and DIF results may vary considering partial credit scores.

Keywords: Differential Item Functioning, Scoring Models, Rasch, Partial Credit Model, LORDIF, Type I Error, statistical power.

[*]*ADDRESS FOR CORRESPONDENCE*: *Dr. Emine Burcu TUNÇ, Marmara University, İstanbul, Turkey. E-mail address: burcu.tunc@marmara.edu.tr.ORCID ID: 0000-0002-8225-9299.*
[b]*Asst. Prof. Dr. Ömer KUTLU, Department of Educational Measurement and Evaluation, Faculty of Educational Sciences, Ankara University, Ankara, Turkey. E-mail address: omerkutlu@ankara.edu.tr. ORCID ID: 0000-0003-4364-5629.*
Received Date: *December 11th,2017. Acceptance Date*:*January 28th,2018.*

1. Giriş

Psikometri bilimi yaklaşık yüz elli yıllık geçmişiyle zekâ, ilgi, tutum, algı, başarı gibi örtük özellik gösteren psikolojik yapıları, geliştirdiği testler aracılığıyla ölçmeye ve insanların sahip olduğu bu örtük davranışlar hakkında kararlar vermeye çalışmaktadır. Ancak testlerde yer alan maddeler bazı durumlarda testin ölçmek istediği amaçtan uzaklaşmakta ve geçerlik sorunları ortaya çıkmaktadır. Bu durumlardan biri madde yanlılığıdır.

Madde yanlılığı, test maddesine, aynı yetenek düzeyinde olan fakat farklı alt gruplardan gelen bireylerin doğru yanıt verme olasılıklarının aynı olmaması durumu olarak tanımlanmaktadır (Osterlind, 1990). Madde yanlılığı test sürecinin yapay bir eseri olup; sosyal, politik ve etnik uygulama kararları için çok önemlidir (Zumbo ve Gelin, 2005). Madde yanlılığı belirleme çabası, yeni ölçekler geliştirmek, var olan ölçekleri yeni durumlara, bireylere, yeni bir dile ya da kültüre uyarlamak için, kısacası daha geçerli test puanları elde etmek için çok önemlidir (Zumbo, 2007).

Eğer bir grup, ilgili madde üzerinde düşük performans gösteriyorsa ve bu düşük performans madde üzerindeki bazı haksızlıklardan dolayı ise, bu maddenin bu grup için yanlı olduğu ifade edilebilir. Bu durum ters etki olarak da adlandırılır. Düşük performans maddedeki yanlılıktan mı yoksa ilgili grubun gerçek başarısızlığından mı kaynaklı? (Penfield ve Lam 2000). Bu noktada, madde etkisi ve Değişen Madde Fonksiyonu (DMF) terimlerine ayrıntılı olarak değinmek gerekmektedir.

Madde etkisi, farklı gruplardaki bireylerin maddeyi farklı yanıtlama olasılıklarına sahip olması, ancak bu olasılığın ölçülmesi istenen gerçek farktan kaynaklanmasıdır (Camilli ve Shepard, 1994; Clauser ve Mazor, 1998; Zumbo, 1999). Eğer test yanıtlayıcıları bilgileri açısından farklılaşıyorsa, madde yanıtlarının da farklılaşması beklenir. Dolayısıyla bu durumdan kaynaklanan farklılık yanlılık değil madde etkisidir (Perrone, 2006).Testle ölçülen davranışlar açısından bireyler arasında farklılıklar olduğu için, etki de alışılagelmiş bir durumdur. Ancak amaç yanlılıktan kaynaklanan farklılıkların açıklanmasıdır (Ong, Williams ve Lamprianou, 2011).

DMF ise ölçülmesi istenen değişken açısından bireylerin yeteneklerine göre eşleştirilmesi ve daha sonra farklı gruplardaki bu bireylerin maddeyi farklı yanıtlama olasılıklarına sahip olduklarının, istatistiksel olarak belirlenmesidir (Camilli ve Shepard, 1994; Clauser ve Mazor, 1998; Zumbo, 1999). DMF, Educational Test Service –ETS-tarafından 1986'da geliştirilmiş ve psikometrik yanlılık analizlerinde bir standart haline gelmiştir (Roever, 2005). DMF, madde yanlılığı ve madde etkisini belirlemek için bir ön adımdır.

Yirmi yıldır DMF belirlemek için kullanılan tekniklerde, matematiksel algoritmada ve kullanılan bilgisayar programlarında büyük gelişmeler olmuştur. Ancak merak edilen sorular hâlâ yanıtlanmamıştır. Bu soruların bazıları şu şekilde belirtilmiştir: Neden bazı maddeler DMF göstermektedir? DMF'li maddeler belirlendikten sonra ne yapılması gerekir? Test geliştiriciler DMF'li maddelerin yerine başkalarını koymayı, DMF'li maddeler yerine yeni maddeler geliştirmeyi ve tüm bu süreci DMF'li madde kalmayana kadar devam ettirmeyi önermişlerdir. Fakat bu süreç hem pahalı hem de zaman alıcıdır. Örneğin çok sayıda DMF'li madde çıkmışsa, bunları yenileriyle değiştirmek kapsam geçerliğinin de değişmesine neden olabilecektir. Diğer bir öneri DMF'li maddeleri düzeltmektir. Ancak düzeltilmiş maddelerin de yeni bir gruba uygulanması ve tekrar DMF analizi yapılması gerekmektedir (Ellis ve Raju, 2003). Söz edilen tüm bu nedenlerden dolayı DMF'nin kaynaklarının belirlenmesi, maddelerin faklı grupların ölçümlerini nasıl etkilediğinin belirlenmesinde ve yanlılığı azaltmak için maddelerin atılması ya da yeniden gözden geçirilmesi kararlarının alınmasında etkili olacağından son derece önemlidir. Söz edilen bu neden sorusuna yanıt bulabilmek ve daha güvenilir DMF kestirimleri yapabilmek için, DMF kapsamında belirli koşullar altında I. Tip hata ve istatistiksel güç çalışmaları da önem kazanmıştır.

DMF için I. Tip hata; gerçekte bir maddenin odak ve referans gruplar için farklılık göstermemesi, yani maddenin DMF içermemesi, ancak yapılan çözümlemeler sonucunda ilgili maddenin DMF'li olarak belirlenmesini; istatistiksel güç ise gerçekte bir maddenin gruplar için farklılık göstermesi yani maddenin DMF içermesi ve yapılan çözümlemeler sonucu da ilgili maddenin DMF'li olarak belirlenmesini ifade etmektedir (Bilican, 2014; Kim, 2010; Wyse ve Mapuranga, 2009). Belirli koşullar altında gerçekleştirilen I. Tip hata ve istatistiksel güç çalışmaları, sadece DMF'li madde belirlemenin ötesinde, DMF'ye neden olan koşullar hakkında fikir vermektedir. Bu çalışmada da DMF'nin olası kaynaklarından biri olarak puanlama modelleri ele alınmış ve puanlama modellerinin DMF üzerinde etkisi olup olmadığı I. Tip hata ve istatistiksel güç kapsamında araştırılmıştır.

Puanlama modelleri genel olarak iki kategorili puanlama ve çok kategorili puanlama olarak ele alınmaktadır. Tarihsel sürece bakıldığında yanıtlayıcıların bir testte yer alan maddelere verdiği yanıtların iki kategorili puanlandığı yani doğru ya da yanlış olarak sınıflandırıldığı modeller üzerinde durulmuştur. Bu modelde yanıtlayıcıların yalnızca tek bir seçeneği değerlendirilmekte ve doğru yanıtlara 1, yanlış ve boş bırakılan yanıtlara 0 puan verilmektedir (Ben-Simon, Budescu ve Nevo, 1997; Frary, 1989; Kurz, 1999).

Geleneksel iki kategorili puanlama dolaylı olarak, bütün maddelerin başarıyı eşit olarak temsil ettiğini ve bütün seçeneklerin eşit bilgi verdiğini varsaymaktadır (Haladyna, 1990). Puanlamada kısmi bilgi ve yanlış bilgiyi ayırt etmek genellikle gözden kaçırılmaktadır. Bu durum çoktan seçmeli maddelerin birçok türünde bile böyledir. Örneğin doğru yanıtın D seçeneği olduğu çoktan seçmeli bir madde için, yalnızca D seçeneğini işaretleyen öğrenciler tam puan

alabileceklerdir. Diğer seçenekleri işaretleyenler ise 0 puan alacaklardır. Ancak B seçeneğini işaretleyen öğrenciler de kısmi doğruluk içeren bilgiye sahip oldukları halde 0 puan alacaktır. Kısacası iki kategorili puanlama için öğrenciler tam bilgiye sahip olmadığı sürece ya da ilgili alanla ilgili tam yetkinlik düzeyinde olmadığı sürece puan alamayacaklardır (Wongwiwatthananukit, Popovich ve Bennett, 2000).

Mevcut seçenekler arasından rastgele seçim yapmamış olmalarına rağmen doğru yanıtı belirleyememiş olan yanıtlayıcıların bu durumu bilgi eksikliğini göstermektedir (Lau, Lau, Hong ve Usop, 2011). Bu nedenle, yanıtlayıcıların özdeki bilgi eksikliği, bilgisi olmayan yanıtlayıcılar tarafından işaretlenmemiş bir çeldiricinin tercihini gösteriyor olabilir. Özetle iki kategorili puanlama modeli, çeldirici seçimlerini ayırt etmemekte ve bu nedenle kısmi bilgi vermemektedir (Diedenhofen ve Musch, 2015).

İki kategorili puanlama modeline göre, şanslı ya da test bilgeliği olan öğrenciler yanlılığa neden olmaktadır. Bu nedenle öğrencilerin yetenek düzeyi kestirimlerini daha iyi yapabilmek için kısmi bilgi, yanlış bilgi ve şansla elde edilmiş bilgi ayrımlarını yapmak gerekmektedir. Kısmi puanın temel mantığı, bireyin performansının orta düzeyinde bulunması, yani doğru yanıt ve yanlış yanıt arasında yer almasıdır (Abu-Sayf, 1979). Kısmi bilgi, probleme doğru başlamak ancak ya bir hata yapmak ya da bir adımı atlayarak doğru yanıta ulaşamamak olarak da tanımlanmaktadır (Grunert, Raker, Murphy ve Holme, 2013). Yapılan çalışmalarda kısmi bilgiye sahip öğrencilerin de ödüllendirilmesi gerektiği ve test puanlarında kısmi puanlamanın önemi belirtilmiş ve kısmi bilginin daha güvenilir ve geçerli puanlar verdiği ifade edilmiştir (Wongwiwatthananukit, Popovich ve Bennett, 2000).

İki kategorili puanlama modelinde, dikkatli olup risk almayanlar ve maddeyi boş bırakanlar, risk alanlara göre cezalandırılmış olmaktadır ve hatta şansla elde edilmiş puanlar da ödüllendirilmektedir. Bu problemlerin üstesinden gelmek için çok kategorili puanlama modelleri geliştirilmiştir. Bu puanlama modelleri çoktan seçmeli maddelerin zayıflıklarını kırmak ve yanıtlayıcıların yetenek kestirimleriyle ilgili daha sağlıklı bilgiler elde etmek ve özellikle kısmi bilgiyi ortaya çıkarmak için kullanılmaktadır. Aynı zamanda bu modellerin geçerliği ve güvenirliği arttırdığı ve de risk alma davranışını daha az sergileyen yanıtlayıcıları da cezalandırmadığının üstünde durmak gerekmektedir (Kurz, 1999).

Bu çalışmada iki kategorili ve çok kategorili puanlanan maddeler kapsamında DMF'yi incelemek için, iki kategorili puanlama için Rasch model kullanılmıştır. Çok kategorili model için ise, doğrudan yanıtlama modelleri içerisinde yer alan MTK'ya dayalı modellerden Kısmi Puan Modeli (KPM) kullanılmıştır. MTK'nın avantajlarını ve Rasch modelinin özelliklerini taşımasından dolayı bu model tercih edilmiştir. Bu noktada KPM'ye ayrıntılı olarak değinmeden önce Rasch modelin özelliklerine değinmek gerekmektedir.

Rasch model Danimarkalı Matematikçi Georg Rasch tarafından 1960 yılında iki kategorili maddeler için geliştirilmiştir. MTK'da çokça kullanılan modellerden biri Rasch modeldir. Modelin temel mantığı; bireyin bir maddeye doğru yanıt verme olasılığının, bireyin θ düzeyi ile madde güçlüğü arasındaki farkın lojistik fonksiyonu olarak tanımlanmasıdır (Pallant ve Tennant, 2007).KPM ise Masters tarafından 1982'de geliştirilmiştir ve iki kategorili maddeler için geliştirilmiş olan Rasch modelin uzantısıdır. Bu model, çözümleme sürecinde farklı aşamaların tamamlanması durumunda kısmi puan vermenin önemli olduğu veya Likert tipi maddelerde yanıt kategorileri arasındaki uzaklıkların maddeden maddeye farklılık gösterdiği durumlar için geliştirilmiştir. Modelin önemli özelliklerinden biri θ düzeyi orta derecede olan kişilerin de puanlandırılmasının mümkün olmasıdır (Koch ve Dodd, 1989). Bir başka anlatımla kısmi puanlamanın temel amacı, kısmi başarılara da puan verebilmektir.

İlgili araştırmalar incelendiğinde, Türkiye'de DMF ve puanlama modellerinin birlikte ele alındığı bir simülasyon çalışmasının olmadığı görülmektedir. Gerçek veriyle yapılmış olan araştırmalar ise puanlama modellerine göre DMF'nin farklılaştığını ortaya koymuştur. DMF kapsamında yapılmış olan simülasyon çalışmaları ise tekniklerin I. Tip hata ve istatistiksel güçlerini karşılaştırmaya yönelik olmuştur. Yurt dışında yapılmış olan çalışmalar dikkate alındığında, puanlama modelleri ve DMF'nin birlikte ele alındığı çalışmaların olduğu ancak iki kategorili puanlama yapılması durumunda ortaya çıkan bilgi kaybının DMF kapsamında araştırıldığı bir simülasyon çalışmasının olmadığı görülmektedir. Bu doğrultuda bireylerin kısmi bilgilerini dikkate alan çok kategorili puanlama modelleri ve iki kategorili puanlama modellerinin I. Tip hata ve istatistiksel güç kapsamında DMF sonuçlarını etkileyip etkilemediğinin belirlenmesine yönelik duyulan gereklilik, bu araştırmanın problemini oluşturmuştur.

Bu araştırmanın genel amacı iki kategorili ve çok kategorili puanlanan maddelerde DMF'yi belirlenen koşullar altında karşılaştırmaktır. Bu amaç doğrultusunda aşağıdaki sorulara yanıt aranmıştır.

1. 600 örneklem büyüklüğü için, farklılaşan örneklem büyüklüğü oranları (1:1 ve 1:2), DMF'li madde oranları (%10, %30 ve %50) ve DMF büyüklükleri (0.25, 0.50, 1.00 ve 1.5) koşullarında, iki kategorili ve çok kategorili puanlama modelleri kapsamında I. Tip hata oranları nasıl değişmektedir?
2. 600 örneklem büyüklüğü için, farklılaşan örneklem büyüklüğü oranları (1:1 ve 1:2), DMF'li madde oranları (%10, %30 ve %50) ve DMF büyüklükleri (0.25, 0.50, 1.00 ve 1.5) koşullarında iki kategorili ve çok kategorili puanlama modelleri kapsamında istatistiksel güç oranları nasıl değişmektedir?

3. 1200 örneklem büyüklüğü için, farklılaşan örneklem büyüklüğü oranları (1:1 ve 1:2), DMF'li madde oranları (%10, %30 ve %50) ve DMF büyüklükleri (0.25, 0.50, 1.00 ve 1.5) koşullarında iki kategorili ve çok kategorili puanlama modelleri kapsamında I. Tip hata oranları nasıl değişmektedir?
4. 1200 örneklem büyüklüğü için, farklılaşan örneklem büyüklüğü oranları (1:1 ve 1:2), DMF'li madde oranları (%10, %30 ve %50) ve DMF büyüklükleri (0.25, 0.50, 1.00 ve 1.5) koşullarında iki kategorili ve çok kategorili puanlama modelleri kapsamında istatistiksel güç oranları nasıl değişmektedir?
5. 2400 örneklem büyüklüğü için, farklılaşan örneklem büyüklüğü oranları (1:1 ve 1:2), DMF'li madde oranları (%10, %30 ve %50) ve DMF büyüklükleri (0.25, 0.50, 1.00 ve 1.5) koşullarında iki kategorili ve çok kategorili puanlama modelleri kapsamında I. Tip hata oranları nasıl değişmektedir?
6. 2400 örneklem büyüklüğü için, farklılaşan örneklem büyüklüğü oranları (1:1 ve 1:2), DMF'li madde oranları (%10, %30 ve %50) ve DMF büyüklükleri (0.25, 0.50, 1.00 ve 1.5) koşullarında iki kategorili ve çok kategorili puanlama modelleri kapsamında istatistiksel güç oranları nasıl değişmektedir?

2. Yöntem

2.1. Araştırmanın Modeli

Bu araştırmada, belirlenen sabit ve manipüle edilen koşullar altında iki kategorili ve çok kategorili puanlama modellerine göre DMF için I. Tip hata ve istatistiksel güç oranlarının farklılaşıp farklılaşmadığı incelenmiştir. Farklı koşullar altında veriler türetildiği için bu araştırma bir simülasyon araştırmasıdır. Simülasyon araştırmasıyla, belirlenen koşullar kapsamında puanlama modellerinin birbirleriyle karşılaştırılması yapılmış, üstünlükleri ve sınırlılıkları ortaya konulmuştur. Bu doğrultuda DMF'nin geliştirilmesine katkı sağlanacağı düşünüldüğünden, araştırmanın temel araştırma niteliğinde olduğu belirlenmiştir.

2.2. Simülasyon Deseni

Tablo 1'de bu araştırmada kullanılan simülasyon deseni için sabit ve manipüle edilen koşulların özeti sunulmuştur.

Tablo 1
Simülasyon Deseni

Örneklem Büyüklüğü	Örneklem Büyüklüğü Oranı	DMF İçeren Madde Yüzdesi	DMF Büyüklüğü
600	(1:1) 300:300	10, 30, 50	0.25, 0.50, 1, 1.5
600	(1:2) 200:400	10, 30, 50	0.25, 0.50, 1, 1.5
1200	(1:1) 600:600	10, 30, 50	0.25, 0.50, 1, 1.5
1200	(1:2) 400:800	10, 30, 50	0.25, 0.50, 1, 1.5
2400	(1:1)1200:1200	10, 30, 50	0.25, 0.50, 1, 1.5
2400	(1:2) 800:1600	10, 30, 50	0.25, 0.50, 1, 1.5

Tablo 1'de görüldüğü gibi, 3*(örneklem büyüklüğü)* X 2*(örneklem büyüklüğü oranı)* X 3*(DMF içeren madde yüzdesi)* X 4*(DMF büyüklüğü)* olmak üzere toplam 72 simülasyon koşulu ortaya çıkmıştır. Her bir koşul için ise 100 tekrar yapılmıştır. Tekrar sayısını etkileyen durumlardan biri koşul sayısı olduğundan, 72 koşul için 100 tekrar uygun görülmüştür. Böylelikle her puanlama modeli ve her analiz için 7200 veri dosyası oluşturulmuş ve 7200 X 2 (iki kategorili ve çok kategorili puanlama modeli) olmak üzere toplam 14.400 veri dosyası elde edilmiştir.

2.3. Verilerin Türetilmesi

Veriler, hem iki kategorili hem de çok kategorili veri türetebilmek için uygun olan R programında elde edilmiştir. R programında veri türetmeden önce WinGen programında madde parametreleri hesaplanmıştır. Madde parametrelerini türetmek için WinGen programında simülasyon desenine ilişkin toplam madde sayısı (20 madde), yanıt kategorilerinin sayısı (beş kategori), kullanılan model (KPM) ve parametreler için ele alınan minimum ve maksimum değerler [-3, +3] belirtilmiştir. Elde edilmiş olan bu parametreler R programında kullanılarak veri türetilmeye başlanmıştır. Elde edilen madde adım güçlüğü parametreleri çok kategorili puanlama içindir. Araştırmanın amacı iki kategorili ve çok kategorili puanlama yapılması durumunda DMF'nin farklılaşıp farklılaşmadığını araştırmak olduğundan, belirlenmiş olan beş kategori (4-3-2-1-0), iki kategorili puanlama için iki kategoriye (1-0) çevrilmiştir. Böylelikle iki ayrı veri seti oluşturulmuştur. Tablo 2'de çok kategorili-iki kategorili puanlama durumları için birinci koşul, birinci tekrar

kapsamında elde edilmiş olan yanıt örüntüsünden örnek sunulmuştur. Tabloda görüldüğü üzere sadece 5. kategori olan 4'e 1 puan verilmiş, 3-2-1-0 kategorilerine ise 0 puan verilmiştir.

Tablo 2
Çok Kategorili-İki Kategorili Puanlama için Yanıt Örüntüsü Örneği

İki Kategorili Puanlama	Çok Kategorili Puanlama
11011100000000000000	44344433222012210010
10000101010000000000	43333434341021210100
01001001000000000000	24334334321002123000
00101100010000000000	33424413341012321121
11110101000000000000	44443434221222301001
11000000010000100000	44333222341032412101
11110100000000000000	44443423232022221102
11011100010000000000	44344413342001011110
10011100100000000000	43244423430032131110
11010100000100100000	44343413331421420110
11110000000000000000	44443313332132221100
11011100000000000000	44344423131022322110
11111000000010000000	44444323231042222100
11111100010000000000	44444432242122331000
11110000000000000000	44442333231221120010
00010100000000000000	23343422231012212102
01110000000000000000	34443322233212122200
11110100000000000000	44443432232032212010
11010000000000100000	44343323120112432011
11010101010000000000	44343404242132303110

Bu araştırmada simülasyon çalışması kapsamında DMF için I. Tip hata ve istatistiksel güç çözümlemeleri yapılmıştır. I. Tip hata ve istatistiksel güç oranlarının değerlendirilmesi kapsamında, Bradley'in referansı (Bradley's criterion) dikkate alınmıştır. Bradley (1978)'in referansına göre (akt., Bilican, 2014); I. Tip hata oranı 0.075'ten küçük olduğunda, hatanın kontrol altına alınabildiği sonucuna varılmıştır. İstatistiksel güç oranı için ise 0.80 ölçüt olarak ele alınmış ve 0.80'den büyük olan oranlar için yeterli istatistiksel güç yorumu yapılmıştır.

2.4. Verilerin Çözümlenmesi

Bu araştırma DMF belirleme tekniği olarak LORDIF analizi kullanılmıştır. LORDIF alan yazında "Logistic regression differential item functioning using IRT" ya da ''Logistic (ordinal) regression (LOR)-plus-IRT" olarak geçmektedir. Aslında bu tekniğin LR ve MTK'nın özelliklerini birleştirerek DMF belirlediği ifade edilebilir. LR/MTK karışımı bir teknikle DMF belirleyebilmek için, R programında lordif paket programı geliştirilmiştir, analizin adı da LORDIF olarak kullanılmaktadır. Hem iki kategorili hem de çok kategorili maddeler için kullanılabilmesi bu tekniğin önemli avantajıdır.

LORDIF diğer teknikler gibi toplam puanlar üzerinden çalışmak yerine, MTK ve yetenek puanları temelli bir model içermektedir. Temel mantık model karşılaştırması yapılarak DMF belirlenmesidir. Her bir madde için, bir sabit terim modeli ve üç tane iç içe (nested) model, ek bağımsız değişkenlerle birlikte aşamalı olarak aşağıda belirtildiği gibi oluşturulmaktadır (Choi, Gibbons ve Crane, 2011).

$$\textit{Model 0: logit } P(u_i \geq k) = \alpha_k$$
$$\textit{Model 1: logit } P(u_i \geq k) = \alpha_k + \beta_1 * \textit{yetenek}$$
$$\textit{Model 2: logit } P(u_i \geq k) = \alpha_k + \beta_1 * \textit{yetenek} + \beta_2 * \textit{grup}$$
$$\textit{Model 3: logit } P(u_i \geq k) = \alpha_k + \beta_1 * \textit{yetenek} + \beta_2 * \textit{grup} + \beta_3 * \textit{yetenek} * \textit{grup}$$

u_i, madde *i*'ye verilen yanıt, α_k ise kesişim terimidir ve kategori *k*'ya bağlıdır. Buna ek olarak, $P(u_i \geq k)$ kategori *k* ve üstünde yanıt olasılığıdır. Terim olarak "yetenek", gizil bir değişken ya da gözlenen toplam puan olarak test tarafından ölçülen özelliği temsil eder (Choi, Gibbons ve Crane, 2011). Bu modellerden Model 1 madde performansını kestirebilmek için yalnızca yeteneği kullanmaktadır. Model 2 madde performansını kestirebilmek için yetenekle birlikte grup değişkenini de kullanmaktadır. Model 3 ise madde performansını kestirebilmek için yetenek, grup ve yetenek-grup etkileşimini bir arada kullanmaktadır (Hahn ve diğ., 2014). Örneğin; Model 1 matematik başarı

düzeyinin madde yanıtlarını yordadığı model, Model 2 matematik başarı düzeyi ve cinsiyetin madde yanıtlarını yordadığı model, Model 3 ise matematik başarı düzeyi, cinsiyet ve matematik başarı düzeyi ve cinsiyet etkileşiminin madde yanıtlarını yordadığı modeldir. Modeller arasında manidar farklılıkların bulunması DMF'ye işaret etmektedir.

3. Bulgular

Tablo 3
600 Örneklem Büyüklüğü için, İlgili Koşullarda, İki Kategorili ve Çok Kategorili Puanlama Modelleri Kapsamında I. Tip Hata Oranları

600 örneklem büyüklüğü		***İki Kategorili Puanlama***				***Çok Kategorili Puanlama***			
Örneklem büyüklüğü oranı	*DMF'li madde oranı*	*DMF büyüklüğü*				*DMF büyüklüğü*			
		0.25	*0.50*	*1.00*	*1.50*	*0.25*	*0.50*	*1.00*	*1.50*
300/300	*%10*	0.05	0.07	0.09	0.14	0.02	0.04	0.04	0.04
	%30	0.06	0.11	0.29	0.55	0.04	0.05	0.04	0.12
	%50	0.08	0.19	0.55	0.79	0.09	0.35	0.31	0.84
200/400	*%10*	0.06	0.06	0.09	0.13	0.03	0.04	0.03	0.03
	%30	0.07	0.11	0.30	0.52	0.04	0.04	0.03	0.06
	%50	0.08	0.16	0.55	0.78	0.08	0.30	0.23	0.72

600 örneklem büyüklüğü kapsamında iki kategorili ve çok kategorili puanlama modellerine göre karşılaştırma yapıldığında, 24 koşuldan 19'unda çok kategorili puanlama modeli kullanıldığında I. Tip hata oranının daha düşük olduğu belirlenmiştir. DMF'li madde oranı %10 olduğunda çok kategorili puanlama modelinde hata kontrolü sağlanırken, iki kategorili puanlama modelinde sağlanamamış olduğu görülmektedir. Özellikle DMF'li madde oranı %30 olduğunda iki kategorili puanlama modeline göre çok kategorili puanlama modelinde I. Tip hata oranlarında önemli düşüşlerin olduğu dikkati çekmektedir.

Tablo 4
600 Örneklem Büyüklüğü için, İlgili Koşullarda, İki Kategorili ve Çok Kategorili Puanlama Modelleri Kapsamında İstatistiksel Güç Oranları

600 örneklem büyüklüğü		***İki Kategorili Puanlama***				***Çok Kategorili Puanlama***			
Örneklem büyüklüğü oranı	*DMF büyüklüğü*	*DMF büyüklüğü*				*DMF büyüklüğü*			
		0.25	*0.25*	*0.25*	*1.50*	*0.25*	*0.50*	*1.00*	*1.50*
300/300	*%10*	0.14	0.54	0.99	1.00	0.29	0.97	1.00	1.00
	%30	0.10	0.30	0.72	0.83	0.18	0.80	1.00	1.00
	%50	0.08	0.17	0.49	0.67	0.10	0.38	0.93	1.00
200/400	*%10*	0.13	0.52	0.99	1.00	0.30	0.90	1.00	1.00
	%30	0.09	0.25	0.72	0.83	0.15	0.83	1.00	1.00
	%50	0.09	0.14	0.49	0.65	0.09	0.32	0.93	0.99

600 örneklem büyüklüğü için iki kategorili ve çok kategorili puanlama modellerine göre karşılaştırma yapıldığında, 24 koşuldan 22'sinde çok kategorili puanlama yapıldığında daha yüksek istatistiksel gücün elde edilmiş olduğu belirlenmiştir. İki koşulda ise aynı sonuçlar elde edilmiştir. Özellikle DMF büyüklüğü 0.50, 1.00 ve 1.50 olduğunda çok kategorili puanlama modeli kapsamında yapılan analiz sonuçlarında büyük farklılıklar belirlenmiştir. İki kategorili puanlama yapıldığında, %50 DMF'li madde oranı ve 1.00/1.50 DMF büyüklüğü koşullarında yeterli istatistiksel güce ulaşılamamışken, çok kategorili puanlama yapıldığında ilgili koşullarda yeterli istatistiksel güce ulaşıldığı ortaya konulmuştur. Özetle 600 örneklem büyüklüğü için gerçekte DMF'li olan maddelerin analizler sonucu DMF'li olarak ortaya çıkmasının, çok kategorili puanlama yapıldığında daha güçlü olarak belirlendiği ifade edilmektedir.

Tablo 5
1200 Örneklem Büyüklüğü için, İlgili Koşullarda, İki Kategorili ve Çok Kategorili Puanlama Modelleri Kapsamında I. Tip Hata Oranları

1200 örneklem büyüklüğü		***İki Kategorili Puanlama***				***Çok Kategorili Puanlama***			
Örneklem büyüklüğü oranı	*DMF büyüklüğü*	*DMF büyüklüğü*				*DMF büyüklüğü*			
		0.25	*0.25*	*0.25*	*1.50*	*0.25*	*0.50*	*1.00*	*1.50*
600/600	*%10*	0.06	0.07	0.13	0.23	0.03	0.03	0.03	0.04
	%30	0.09	0.17	0.50	0.73	0.05	0.05	0.05	0.83
	%50	0.13	0.33	0.76	0.87	0.19	0.34	0.84	1.00
400/800	*%10*	0.06	0.07	0.13	0.23	0.03	0.03	0.04	0.04
	%30	0.07	0.17	0.47	0.74	0.06	0.05	0.06	0.66
	%50	0.12	0.30	0.76	0.87	0.17	0.42	0.76	1.00

1200 örneklem büyüklüğü için iki kategorili ve çok kategorili puanlama modellerine göre karşılaştırma yapıldığında 24 koşuldan 15'inde çok kategorili puanlama durumunda I. Tip hata oranının daha düşük olduğu belirlenmiştir. Özellikle DMF'li madde oranı %10 olduğunda iki kategorili puanlama modeline göre çok kategorili puanlamada I. Tip hata oranları için önemli düşüşlerin olduğu dikkati çekmektedir. Bu koşul altında iki kategorili puanlama modeli kullanıldığında, I. Tip hata kontrolü sağlanamazken, çok kategorili puanlama modeli kullanıldığında hata kontrolünün sağlanabildiği ifade edilebilir. DMF'li madde oranının %50 olduğu koşul için, iki kategorili puanlama modeli kullanıldığında I. Tip hata oranları düşmekle birlikte, her iki puanlama modeline göre hata kontrolünün sağlanamadığı görülmektedir.

Tablo 6
1200 Örneklem Büyüklüğü için, İlgili Koşullarda, İki Kategorili ve Çok Kategorili Puanlama Modelleri Kapsamında İstatistiksel Güç Oranları

1200 örneklem büyüklüğü		***İki Kategorili Puanlama***				***Çok Kategorili Puanlama***			
Örneklem büyüklüğü oranı	*DMF büyüklüğü*	*DMF büyüklüğü*				*DMF büyüklüğü*			
		0.25	*0.50*	*1.00*	*1.50*	*0.25*	*0.50*	*1.00*	*1.50*
600/600	*%10*	0.29	0.86	1.00	0.99	0.65	1.00	1.00	0.99
	%30	0.15	0.47	0.83	0.83	0.42	0.98	1.00	1.00
	%50	0.10	0.29	0.70	0.75	0.18	0.76	1.00	1.00
400/800	*%10*	0.32	0.81	1.00	1.00	0.62	1.00	1.00	1.00
	%30	0.14	0.45	0.83	0.83	0.38	0.96	1.00	1.00
	%50	0.09	0.27	0.70	0.76	0.17	0.64	1.00	1.00

Her iki puanlama modeli karşılaştırıldığı zaman 24 koşuldan 20'sinde çok kategorili puanlama modeli kullanıldığında daha yüksek istatistiksel gücün elde edilmiş olduğu belirlenmiştir. Dört koşulda ise aynı sonuçlar elde edilmiştir. Özellikle DMF büyüklüğü 0.50 olduğunda çok kategorili puanlama modeli kullanılan analiz sonuçlarında büyük farklılıklar belirlenmiştir. İki kategorili puanlama yapıldığında, %50 DMF'li madde oranı ve 1.00/1.50 DMF büyüklüğü koşullarında yeterli istatistiksel güce ulaşılamamışken, çok kategorili puanlama yapıldığında ilgili koşullarda yeterli istatistiksel güce ulaşıldığı ortaya konulmuştur. Özetle 1200 örneklem büyüklüğü için gerçekte DMF'li olan maddelerin analizler sonucu DMF'li olarak ortaya çıkması, çok kategorili puanlama yapıldığında daha güçlü olarak belirlenmektedir.

Tablo 7
2400 Örneklem Büyüklüğü İçin, İlgili Koşullarda, İki Kategorili ve Çok Kategorili Puanlama Modelleri Kapsamında I. Tip Hata Oranları

2400 örneklem büyüklüğü		***İki Kategorili Puanlama***				***Çok Kategorili Puanlama***			
Örneklem büyüklüğü oranı	*DMF büyüklüğü*	*DMF büyüklüğü*				*DMF büyüklüğü*			
		0.25	*0.50*	*1.00*	*1.50*	*0.25*	*0.50*	*1.00*	*1.50*
1200/1200	*%10*	0.07	0.09	0.24	0.42	0.04	0.03	0.04	0.04
	%30	0.10	0.28	0.70	0.86	0.04	0.04	0.74	1.00
	%50	0.20	0.58	0.87	0.91	0.37	0.36	0.99	1.00
800/1600	*%10*	0.06	0.09	0.21	0.38	0.03	0.03	0.03	0.04
	%30	0.10	0.27	0.70	0.86	0.04	0.03	0.67	1.00
	%50	0.17	0.52	0.87	0.90	0.30	0.30	0.99	1.00

2400 örneklem büyüklüğü için iki kategorili ve çok kategorili puanlama modellerine göre karşılaştırma yapıldığında 24 koşuldan 15'inde çok kategorili puanlama modelinde I. Tip hata oranlarının daha düşük olduğu belirlenmiştir.

Tablo 8
2400 Örneklem Büyüklüğü için, İlgili Koşullarda, İki Kategorili ve Çok Kategorili Puanlama Modelleri Kapsamında İstatistiksel Güç Oranları

2400 örneklem büyüklüğü		***İki Kategorili Puanlama***				***Çok Kategorili Puanlama***			
Örneklem büyüklüğü oranı	*DMF'li madde oranı*	*DMF büyüklüğü*				*DMF büyüklüğü*			
		0.25	*0.50*	*1.00*	*1.50*	*0.25*	*0.50*	*1.00*	*1.50*
1200/1200	*%10*	0.62	0.99	1.00	1.00	0.99	1.00	1.00	1.00
	%30	0.29	0.75	0.83	0.83	0.83	1.00	1.00	1.00
	%50	0.20	0.51	0.78	0.82	0.39	0.95	1.00	1.00
800/1600	*%10*	0.52	0.96	1.00	1.00	0.94	1.00	1.00	1.00
	%30	0.28	0.71	0.83	0.83	0.79	1.00	1.00	1.00
	%50	0.16	0.50	0.79	0.84	0.35	0.89	1.00	1.00

2400 örneklem büyüklüğü için iki kategorili ve çok kategorili puanlama modellerine göre 24 koşuldan 20'sinde çok kategorili puanlama modeli için daha yüksek istatistiksel gücün elde edilmiş olduğu belirlenmiştir. Dört koşulda ise aynı sonuçlar elde edilmiştir. Özellikle DMF büyüklüğü 0.25 ve 0.50 olduğunda çok kategorili puanlama modeli kullanılan analiz sonuçlarında dikkate değer istatistiksel güç artışları olmuştur.

Genel olarak örneklem büyüklüğü için karşılaştırma yapıldığında tüm koşullar altında örneklem büyüklüğü 600'den 2400'e doğru arttıkça I. Tip hata oranlarının arttığı belirlenmiştir. LORDIF analizi için iki kategorili puanlama modelinde örneklem büyüklüğü arttıkça I. Tip hata oranları artmışken, çok kategorili puanlama modelinde %10 DMF'li madde oranı ve 0.25/0.50 DMF büyüklükleri için örneklem büyüklüğü arttıkça I. Tip hata oranının azaldığı ya da aynı kaldığı belirlenmiştir.

4. Sonuç, Tartışma ve Öneriler

Araştırmanın amacı doğrultusunda iki kategorili ve çok kategorili puanlama modelleri karşılaştırıldığında, çok kategorili puanlama yapılması durumunda I. Tip hata oranlarının daha düşük, istatistiksel güç oranlarının ise daha yüksek olduğu belirlenmiştir. Bu doğrultuda çok kategorili puanlamanın, yetenek kestiriminde daha başarılı olduğu ve DMF sonuçlarında puanlama modellerinin etkisi olabileceği ifade edilmiştir.

İki kategorili ve çok kategorili puanlama kapsamında DMF incelendiğinde, genel olarak çok kategorili puanlama yapılması durumunda I. Tip hata oranlarının daha düşük, istatistiksel güç oranlarının ise daha yüksek olduğu belirlenmiştir. Daha önce puanlama modelleri ve DMF kapsamında yapılmış bir simülasyon çalışmasına rastlanmamıştır. Ancak gerçek veriyle yapılan çalışmalarda puanlama modelleri ve DMF arasında ilişkinin olabileceği belirtilmiştir. Bu doğrultuda yapılmış olan çalışmaların bir kısmında, çoktan seçmeli ve açık uçlu maddeler cinsiyete göre DMF açısından karşılaştırılmış ve farklılaşan sonuçlar elde edilmiştir. Henderson (2001) iki kategorili puanlanan maddelerin çoğunun erkek öğrenciler lehine, çok kategorili puanlanan maddelerin ise tamamının kız öğrenciler lehine olduğunu ve elde edilen bu sonuçların, madde türü ve cinsiyet arasında bir etkileşim olduğunu gösterebileceğini ifade etmiştir. Feng (2008) de açık uçlu maddelerin kızların lehine, Zenisky, Hambleton ve Robin (2003) çoktan seçmeli

maddelerin erkeklerin, açık uçlu maddelerin ise kız öğrencilerin lehine DMF göstermeye eğilimli olduğunu belirtmiştir. Chaimongkol, Huffer ve Kamata (2007) ise çoktan seçmeli maddelerin açık uçlu maddelere göre daha fazla DMF'ye sahip olduğuna bulguları arasında yer vermiştir. Qian (2011) da benzer şekilde çoktan seçmeli maddelerde daha fazla DMF'ye sahip madde belirlemiştir.

Puanlama modellerinin DMF'den bağımsız ele alındığı çalışmalarda da genel olarak çok kategorili puanlama modellerinin iki kategorili puanlama modellerine göre daha iyi sonuçlar verdiği ortaya konulmuştur. Wongwiwatthananukit, Popovich ve Bennett (2000) çok kategorili puanlama ve geleneksel puanlama için madde yanıt verileri; madde güçlüğüne, madde ayırt ediciliğine, güvenirliğe, yeterliğe ve bilişsel alana göre karşılaştırılmıştır. Sonuçlar çok kategorili puanlamanın madde güçlüğünü arttırdığını ve iki kategorili puanlamaya göre öğrencilerle ilgili daha fazla bilgi verdiğini göstermiştir. Bauer ve diğ. (2011) de kısmi puanlama modellerinin iki kategorili puanlamaya göre daha iyi sonuçlar verdiğini ve bu doğrultuda kısmi bilginin ödüllendirilmesi gerektiğini ifade etmiştir. Diedenhofen ve Musch (2015) ise benzer şekilde iki kategorili puanlamayla karşılaştırıldığında testin geçerlik ve güvenirliğini geliştiren deneysel tercih ağırlıklandırmasını önermektedirler. Aynı zamanda bu çalışma doğrultusunda doğru sınıflandırma oranının, deneysel tercih ağırlıklandırması için daha yüksek olduğu ifade edilmiştir. Test sonuçlarının, deneysel tercih ağırlıklandırması kullanılarak hesaplandığında bilginin derecelerini daha iyi yansıttığı ortaya konulmuştur.

Chang (2007) Rasch ve KPM modelini karşılaştırmış ve kısmi puanlamanın iki kategorili puanlamaya göre daha iyi sonuçlar verdiğini belirtmiş ve kısmi bilginin ödüllendirilmesi gerektiği yorumunu yapmıştır. Gözen Çıtak (2007) ise bu çalışmalardan farklı olarak çalışmasında MTK kapsamında "1-0" puanlamanın kullanıldığı durumda yetenek ölçeği üzerindeki parametrelerin ağırlıklı puanlamaların kullanıldığı duruma göre daha doğru kestirildiğini göstermiş, bu puanlama yönteminin test geçerliği açısından da daha etkili olduğu sonucuna ulaşmıştır.

Daha önce de ifade edildiği üzere puanlama modelleri ve DMF kapsamında yapılmış bir simülasyon çalışmasına rastlanmamış, gerçek veriyle yapılan çalışmalarda puanlama modelleri ve DMF arasında ilişkinin olabileceği belirtilmiştir. Bu çalışmanın bulgularına paralel olarak Selvi (2013) de puanlama modelleri ve DMF arasında ilişki olduğunu belirtmiştir. Yaptığı çalışmada çok kategorili puanlama yapılması durumunda iki kategorili puanlama durumundakine oranla kullanılan bütün teknikler için belirlenen DMF'li madde sayılarında iki katın üzerinde bir artış olduğu görülmüştür. Bu durumun maddelerin ağırlıklı puanlanmasının bireylerin kısmi bilgilerini daha iyi yansıtmasından, yetenek boyutunun daha büyük bir ranjı hakkında bilgi sağlamasından kaynaklandığı ifade edilmiştir. Wetzel, Böhnke, Carstensen, Ziegler ve Ostendorf (2013) kişilik envanterinin maddelerine ait kategorilerin farklı şekillerde puanlanmasıyla elde edilen veriler üzerinde cinsiyet değişkeni açısından DMF analizleri yapmışlar ve maddelerin yanıt şekline göre DMF bulgularının değiştiğini bulmuşlardır. Gelin ve Zumbo (2003) da yine bu araştırmanın bulgularına paralel olarak bir ölçeği hem iki kategorili hem de çok kategorili olmak üzere farklı modellerle puanlamış ve puanlama modellerinin DMF sonuçlarını etkilediğini ortaya koymuşlardır. İki kategorili puanlama yapıldığında erkek ve kadınlar arasında DMF oluşmadığı, çok kategorili puanlama yapıldığında ise oluştuğu belirtilmiş ve bunun gerekçesi olarak ise ikili puanlamanın, bireyler arasında yeterli değişkenlik oluşturmadığı, bir başka ifadeyle bireylerin gerçek durumunu yansıtmakta yetersiz kaldığı gösterilmiştir.

Araştırma kapsamında genel olarak çok kategorili puanlama modeli kapsamında yapılmış olan analizler sonucu I. Tip hata oranlarının daha düşük, istatistiksel güç oranlarının ise daha yüksek olduğu belirlenmiştir. Yani gerçekte DMF'li olarak belirlenen maddelerin DMF'li olarak ortaya çıkma olasılığı çok kategorili puanlama modelinde daha yüksektir. Benzer şekilde gerçekte DMF'li olmayan maddelerin DMF'li olarak belirlenme olasılığı da çok kategorili puanlama modelinde daha düşüktür. Bu doğrultuda DMF kapsamında çok kategorili puanlama modelinin daha sağlıklı sonuçlar verdiği ifade edilebilir. Bu nedenle, bireylerin yeteneğine ilşikin bilgi kaybını daha aza indiren, tam bilgi, kısmi bilgi ve yanlış bilgi ayrımlarını yapan puanlama modellerinin kullanılması önerilmektedir.

Kaynakça

Abu-Sayf, F.K. (1979). The scoring of multiple-choice tests: A closer look. *Educational Technology*, 19, 5-15.

Bauer, D., Holzer, M., Kopp, V. and Fischer, M. R. (2011). Pick-N multiple choice-exams: a comparison of scoring algorithms.*Advances in health sciences education,16*(2), 211-221.

Ben-Simon, A., Budescu, D.V. and Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *AppliedPsychological Measurement,* 21 (1), 65-88.

Bilican, S. (2014). *Çok kategorili puanlanan maddelerde madde işlev farklılığının mantel test ve olabilirlik oran testi ile karşılaştırılması* (Doktora Tezi). Ankara Üniversitesi, Ankara.

Camilli G. and Shepard L. A. (1994). *Methods for identifying biased test items* (volume 4). California: SAGE Publications. Inc.

Chaimongkol, S., Huffer, F. W. and Kamata, A. (2007). An explanatory differential item functioning (DIF) model by the WinBUG 1.4. *Songklanakarin Journal of Science and Technology*, 29(2), 449-458.

Chang, S. H., Lin, P. C. and Lin, Z. C. (2007). Measures of partial knowledge and unexpected responses in multiple-choice tests.*Educational Technology & Society*,*10*(4), 95-109.

Choi, S. W., Gibbons, L. E. and Crane, P. K. (2011). LORDIF: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, 39(8), 1–30.

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

Diedenhofen, B. and Musch, J. (2015). Empirical option weights improve the validity of a multiple-choice knowledge test.*European Journal of Psychological Assessment*.

Diedenhofen, B. and Musch, J. (2015). Empirical option weights improve the validity of a multiple-choice knowledge test.*European Journal of Psychological Assessment*.

Ellis, B. B. and Raju, N.S.(2003). Test and item bias: What they are, What they aren't, and How to detect them. Web: http://files.eric.ed.gov/fulltext/ED480042.pdf adresinden 17 Mayıs 2016'da erişilmiştir.

Frary, R. (1989). Partial credit scoring methods for multiple choice tests. *Applied Measurement in Education*, 2 (1), 79-96.

Gelin, M.N. and Zumbo, B.D. (2003). Differential item functioning results may change depending on how an item is scored: an illustration with the center for epidemiologic studies depression scale. *Educational and PsychologicalMeasurement*, 63(1), 65-74.

Gözen Çıtak, G. (2007). *Klasik test ve madde-tepki kuramlarına göre çoktan seçmeli testlerde farklı puanlama yöntemlerinin karşılaştırılması* (Doktora Tezi). Ankara Üniversitesi, Ankara.

Grunert, M. L., Raker, J. R., Murphy, K. L. and Holme, T. A. (2013). Polytomous versus dichotomous scoring on multiple-choice examinations: development of a rubric for rating partial credit.*Journal of Chemical Education*,90(10), 1310-1315.

Hahn, E. A., Kallen, M. A., Jacobs, E. A., Ganschow, P. S., Garcia, S. F. and Burns, J. L. (2014). English-Spanish equivalence of the health literacy assessment using talking touchscreen technology (Health LiTT).*Journal of Health Communication*: *International Perspectives*, *19*(2), 285-301.

Haladyna, T. M. (1990). Effects of empirical option weighting on estimating domain scores and making pass/fail decisions. *Applied Measurement in Education*, 3, 231–244.

Henderson D. L. (2001). *Prevalence of gender DIF in mixed format high school exit examinations*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle.

Kim, J. (2010). *Controlling type 1 error rate in evaluating differential item functioning for four DIF methods: Use of three procedures for adjustment of multiple item testing* (DoctoralDissertation). Georgia State University, ABD.

Koch, W. R. and Dodd, B. G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education*, 2(4), 335-357.

Kurz, T.B. (1999). A review of scoring algorithms for multiple-choice test items. *EDRS Publications*, Report No: ED 428 076.

Lau, P. N. K., Lau, S. H., Hong, K. S. and Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple- choice tests. *Educational Technology and Society*, 14, 99–110.

Ong, Y.M., Williams, J. S. and Lamprianou, I. (2011). Exploration of the validity of gender differences in mathematics assessment using differential bundle functioning. *International Journal of Testing*, 11, 271-293.

Osterlind, S. (1990). *Test item bias*. Newbury Park: Sage Publications.

Pallant, J. F. and Tennant, A. (2007). An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, 46(1), 1-18.

Penfield, R. D. and Lam, T. (2000). Assessing differential item functioning in performance assessment: Review and recommendations.*Educational Measurement: Issues and Practice*,*19*(3), 5-15.

Perrone, M. (2006). Differential item functioning and item bias: Critical considerations in test fairness. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics*, 6, 1-3.

Qian. X. (2011). *A Multi-Level Differential Item Functioning Analysis of Trends In International Mathematics And Science Study: Potential Sources Of Gender And Minority Difference Among U.S. Eighth Graders' Science Achievement.*

Roever, C. (2005). That's not fair! Fairness, bias, and differential item functioning in language testing. *SLS Brownbag*, 9(15), 1-14.

Selvi, H. (2013). *Klasik test ve madde tepki kuramlarına dayalı değişen madde fonksiyonu belirleme tekniklerinin farklı puanlama durumlarında incelenmesi* (Doktora Tezi). Mersin Üniversitesi, Mersin.

Wetzel, E., Böhnke, J.R.,Carstensen, C.H.,Ziegler, M. and Ostendorf, F. (2013). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences,* 34(2), 69–81.

Wongwiwatthananukit, S., Popovich, N. G. and Bennett, D. E. (2000). Assessing pharmacy student knowledge on multiple-choice examinations using partial-credit scoring of combined-response multiple-choice items.*American Journal of Pharmaceutical Education*,*64*(1), 1.

Wongwiwatthananukit, S., Popovich, N. G. and Bennett, D. E. (2000). Assessing pharmacy student knowledge on multiple-choice examinations using partial-credit scoring of combined-response multiple-choice items.*American Journal of Pharmaceutical Education*,*64*(1), 1.

Wyse, A. E. and Mapuranga, R. (2009). Differential item functioning analysis using Rasch item information functions. *International Journal of Testing*, 9(4), 333-357.

Zenisky A. L., Hambleton R. K. and Robin F. (2003). DIF detection and interpretation in large scale science assessments: Informing item writing practices. *Center for Educational Assessment MCAS Validity Report,* 1. (CEA-429).

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 ResourcesResearch and Evaluation, Department of National Defense.

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.

Zumbo, B. D. and Gelin, M. N. (2005). A matter of test bias in educational policy research: bringing the context into picture by investigating sociological community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies, 5*, 1-23.