**ANKARA ÜNİVERSİTESİ**

**EĞİTİM BİLİMLERİ ENSTİTÜSÜ**

**ÖLÇME VE DEĞERLENDİRME ANABİLİM DALI**

**EĞİTİMDE ÖLÇME VE DEĞERLENDİRME BİLİM DALI**

# ÇOK KATEGORİLİ PUANLANAN MADDELERDE MADDE İŞLEV FARKLILIĞININ MANTEL TEST VE OLABİLİRLİK ORAN TESTİ İLE KARŞILAŞTIRILMASI

**DOKTORA TEZİ**

**Safiye Bilican Demir**

**Ankara**

**Ocak, 2014**

**ANKARA ÜNİVERSİTESİ**

**EĞİTİM BİLİMLERİ ENSTİTÜSÜ**

**ÖLÇME VE DEĞERLENDİRME ANABİLİM DALI**

**EĞİTİMDE ÖLÇME VE DEĞERLENDİRME BİLİM DALI**

# ÇOK KATEGORİLİ PUANLANAN MADDELERDE MADDE İŞLEV FARKLILIĞININ MANTEL TEST VE OLABİLİRLİK ORAN TESTİ İLE KARŞILAŞTIRILMASI

**DOKTORA TEZİ**

**Safiye Bilican Demir**

**Danışman: Prof. Dr. Nükhet Çıkrıkçı Demirtaşlı**

**Ankara**

**Ocak, 2014**

Eğitim Bilimleri Enstitüsü Müdürlüğü'ne,

Bu çalışma jürimiz tarafından Ölçme ve Değerlendirme Anabilim Dalında DOKTORA TEZİ olarak kabul edilmiştir.

Başkan .....

Prof. Dr. Mehmet Ali KISAKÜREK

Üye .....

Prof. Dr. Ezel TAVŞANCIL

Üye .....

Prof. Dr. Nükhet ÇIKRIKÇI DEMİRTAŞLI (Danışman)

Üye .....

Prof. Dr. Nizamettin KOÇ

Üye .....

Doç. Dr. Halil YURDUGÜL

Onay

Yukarıdaki imzaların, adı geçen öğretim üyelerine ait olduğunu onaylarım.

.../.../20..

Prof. Dr. İsmail Güven

Enstitü Müdürü

# ÖNSÖZ

Eğitim sürecinde testlerden elde edilen sonuçlar özellikle öğrenci başarısının değerlendirilmesinde, sınıf geçme veya mezuniyet ile ilgili kararlarda, okul yöneticilerinin ve çalışanlarının değerlendirilmesinde giderek daha sıkılıkla kullanılmaktadır. Bu durumda testlerden elde edilen puanların geçerlik ve güvenirlik gibi temel psikometrik özelliklerinin kabul edilebilir seviyelerde olması beklenir. Eğer bir testin psikometrik özellikleri uygun standartları/ölçütleri karşılayamıyorsa, bu test sonuçlarına dayanarak verilecek kararların ve çıkarımların geçerliğiyle ilgili kuşkular ortaya çıkacaktır.

Eğitim ve psikoloji alanında kullanılan testlerden elde edilen puanların geçerliliği için en büyük tehditlerden birisinin yanlılık olduğu vurgulanmaktadır. Yanlılık, testle ölçülmesi amaçlanmayan yapıların test sonuçlarını olumsuz etkileyerek, test puanlarına dayalı olarak verilen kararların geçerliğini azaltması olarak tanımlanabilir. Bu araştırmada ise yanlılık kavramı ele alınarak, geçerlik çalışmaları kapsamında yapılan yanlılık analizlerinin önemine dikkat çekilmiştir. Bu bağlamda madde düzeyinde yanlılık analizlerinin bir parçası olan Madde İşlev Farklılığı (MİF) ve bunu test etmenin yolları üzerinde durulmuştur.

Test geliştirme sürecinde ve ölçme araçlarından elde edilen sonuçların ne kadar adil ve geçerli olduğu konusunda MİF analizlerinin önemi dikkate alındığında, performans, başarı, tutum, kişilik vb. psikolojik yapıların ölçülmesinde artık daha sıklıkla kullanılmaya başlanan çok kategorili puanlanan maddeler için de MİF analizlerinin yapılması gerekmektedir. Çok kategorili puanlanan maddeler için pek çok MİF belirleme tekniği bulunmasına karşın, pratik test uygulamalarında değişen farklı koşullar için hangi tekniklerin daha tutarlı sonuçlar verdiğinin belirlenmesi önemli hale gelmektedir. Bu araştırmada sıklıkla kullanılan MİF belirleme tekniklerinden Mantel Test ve

Olabilirlik Oran Testi'nin farklı test koşullarında MİF belirleme performansları yapay veriler üzerinden karşılaştırılmıştır. Böylece bu araştırma sonuçlarının özellikle uygulayıcılar için değişen test koşullarında daha geçerli sonuçlar elde etmek üzere hangi MİF belirleme tekniğini kullanacakları konusunda önemli bilgiler sunması beklenmektedir.

Öncelikle, lisansüstü eğitimin boyunca yönlendirmeleri ile bana sürekli rehber olan, araştırmamım her aşamasında yardımları ve fikirleri ile yol gösteren sevgili Hocam ve danışmanım Prof. Dr. Nükhet Çıkrıkçı Demirtaşlı'ya motivasyonu, emeği ve sabrı için çok teşekkür ediyorum.

Lisansüstü eğitimim boyunca yetişmemde emeği bulunan değerli hocalarım Prof. Dr. Nizamettin Koç, Prof. Dr. Ezel Tavşancıl, Yrd. Doç. Dr. Ömer Kutlu ve Doç. Dr. Ömay Çokluk'a; tez jürimde yer alarak tezimin gelişmesine katkıda bulunan saygı değer hocalarım Prof. Dr. Mehmet Ali Kısakürek ve Doç. Dr. Halil Yurdugül'e; araştırma süreci boyunca sürekli fikir alışverişinde bulunduğumuz sevgili arkadaşım Özen Yıldırım'a; doktora öğrenimim boyunca yurt içi doktora burs olanağı sağlayan TÜBİTAK'a teşekkür ederim.

Her zaman desteklerini hissettiğim canım anneme ve babama; bu çalışma sürecinde özveri ve anlayışı için sevgili eşim Hasan Basri Demir'e çok teşekkür ediyorum.

Safiye Bilican Demir

# ÖZET


ÇOK KATEGORİLİ PUANLANAN MADDELERDE MADDE İŞLEV FARKLILIĞININ MANTEL TEST VE OLABİLİRLİK ORAN TESTİ İLE KARŞILAŞTIRILMASI

Bilican Demir, Safiye
Doktora, Ölçme ve Değerlendirme Anabilim Dalı
Tez Danışmanı: Prof. Dr. Nükhet Çıkrıkçı Demirtaşlı
Ocak, 2014, xiii+ 109 sayfa

Bu araştırmada, çok kategorili tepki gerektiren maddelerde MİF belirleme testlerinden Mantel Test ve MTK-Olabilirlik Oran Testi'nin farklı test koşullarında I.Tip hata ve istatistiksel güç (power) oranları karşılaştırılmıştır. Çalışmada, Monte Carlo simülasyon tekniği yaklaşımıyla araştırma koşullarına uygun yapay veri setleri elde edilmiştir. Bu çalışmada grupların yetenek dağılımı, örneklem büyüklüğü, MİF miktarı ve MİF örüntüsü manipüle edilen değişkenler olarak belirlenmiş; çok kategorili MTK modeli, madde sayısı, MİF içeren madde sayısı ve MİF türü (tek biçimli) bütün koşullar altında sabit tutulmuştur. Son durumda I. Tip hata çalışması için 18 [3 (yetenek dağılımı) x 3 (örneklem büyüklüğü) x 2 (MİF belirleme testi)] simülasyon koşulu ortaya çıkmıştır. İstatistiksel güç çalışmaları için 76 [3 (yetenek dağılımı) x 3 (örneklem büyüklüğü) x 2 (MİF miktarı) x 2 (MİF örüntüsü) x 2 (MİF belirleme tekniği)] simülasyon koşulu ortaya çıkmıştır. Her bir koşul için 100 tekrar yapılmıştır.

I. Tip hata oranları MİF içermeyen 20 madde için hesaplanırken güç oranları MİF içerecek biçimde modellenen üç madde üzerinden hesaplanmıştır. Çalışmada veri üretmek için WinGen1, MTK-OOT karşılaştırmaları için MULTİLOG ve Mantel Test analizleri için DIFAS programı kullanılmıştır.

Araştırma sonuçları, referans ve odak grubun yetenek dağılımı birim normal dağılım gösterdiği koşulda her iki MİF belirleme testin de I. Tip hatayı iyi kontrol ettiğini göstermiştir. Grup yetenek dağılımlarının benzer olduğu koşul için artan örneklem büyüklüğüne bağlı olarak I. Tip hata oranları Mantel Test için yükselirken MTK-OOT için düşme eğilimi göstermiştir. Her iki MİF belirleme testi için de artan örneklem büyüklüğü ve odak grup yetenek dağılım ortalamasındaki sapmalara bağlı olarak I. Tip hata değerleri artma eğilim göstermiştir. Mantel Test ile karşılaştırıldığında, artan örneklem büyüklüğü ve grupların yetenek ortalamasındaki sapmaya bağlı olarak MTK-OOT için I. Tip hata değerlerindeki artış daha yüksek olmuştur.

Her iki MİF belirleme testi için de, artan MİF miktarı ve örneklem büyüklüğüne bağlı olarak ilgili testlerin istatistiksel güç oranlarını yükselmiştir. Araştırma bulguları tüm örneklem büyüklüğü ve MİF miktarı koşullarında MTK-OOT'in istatistiksel güç oranlarının, Mantel Test'e göre daha yüksek olduğunu göstermektedir. Yüksek MİF örüntüsü koşullarında ilgili testlerin MİF'i belirlemedeki performansı zayıftır. Mantel Test ve MTK-OOT'nin MİF belirlemedeki gücü odak grubun yetenek ortalamasındaki sapmaya bağlı olarak bir miktar yükselmiştir. Ancak genel olarak farklı sapma koşulları için ilgili testlerin istatistiksel gücü birbirine yakın değerler almıştır.

**Anahtar kelimeler:** Çok kategoride puanlanan maddeler, Madde İşlev Farklılığı, Mantel Test, Olabilirlik Oran Testi, I. Tip Hata, istatistiksel güç.

## ABSTRACT

COMPARISION OF MANTEL TEST AND LIKELIHOOD RATIO TEST FOR DETECTION OF DIFFERENTIAL ITEM FUNCTIONING IN POLYTOMOUS ITEM RESPONSES

Bilican Demir, Safiye
Dissertation, Department of Measurement and Evaluation
Advisor: Professor Dr. Nükhet Çıkrıkçı Demirtaşlı
January, 2014, xiii + 109 pages

The purpose of this study was to investigate the power and Type I error rate of the likelihood ratio goodness-of-fit (LR) statistic and Mantel Test in detecting differential item functioning (DIF) under Master's (1969, 1972) Partial Credit Model. A multiple replication Monte Carlo study was utilized for simulated data sets.

Several variables were manipulated in this study, including the sample size, group mean difference, DIF condition and DIF magnitude. On the other hand some variables were held constant, including polytomous IRT model, test length, Percent of Items with DIF and Type of DIF. In final study design, there were 18 conditions [3 (sample size) x 3 (group mean difference) x 2 (methods of DIF detection)] for Type I error rate study and 76 conditions [3 (sample size) x 3 (group mean difference) x 2 (DIF magnitude) x 2 (DIF pattern) x 2 (methods of DIF detection)] for power study. Simulation was replicated for 100 times for each simulation condition.

The conditions investigating Type I error had twenty items with no DIF, whereas the conditions investigating power had three items with DIF. In this study, WinGen3 was used to simulate ability estimates and to generate

response data sets. MULTİLOG and DIFAS were used to conduct the Mantel and IRT likelihood-ratio test DIF analyses.

Results indicated that with equal group distribution, Mantel Test and IRT-LR Test performed similarly under all testing conditions and had better Type I error rate control. Type I error rate for Mantel Test increased as sample size increased. Conversely, as sample size increased, the Type I error rate for IRT-LR decreased under equal group distribution condition. The presence of group mean difference affected the Type I error results of both DIF detection tests. The results showed that large sample size and presence of group mean difference tended to inflate the Type I error rates of both DIF detection tests. IRT-LR had higher Type I error rates than Mantel Test when large sample size and when group mean difference conditions.

For both DIF detection tests, the power to detect DIF increased as the DIF Magnitude and sample size increased. The results also showed that IRT-LR had higher DIF detection rates than Mantel test under all test conditions. For both DIF detection tests, conditions with high-shift DIF pattern had the poorest power. The presence of group mean difference had minimal effect on the power results of both DIF detection tests.

**Key words**: Polytomous items, Differential item functioning, Mantel test, Item Response Theory- Likelihood Ratio test, Type I error, power

## İÇİNDEKİLER

**Sayfa**

## ÇİZELGELER LİSTESİ

## ŞEKİLLER LİSTESİ

EKLER LİSTESİ

# BÖLÜM 1

# GİRİŞ

Bu bölümde araştırma problemi, araştırmanın amacı, önemi, tanımlar ve sınırlılıklar yer almaktadır.

## Problem

Pek çok alanda bireyler hakkında karar vermek üzere testlerden yararlanılmaktadır. Test sonuçlarına dayalı olarak bireyleri bir öğretim kurumuna yerleştirme, bireylerin akademik veya iş performansının belirlenmesi, personel seçimi, öğrenmelerin takibi, geri bildirim verme, davranış bozukluklarının teşhis edilmesi veya mesleki yönlendirme gibi çok çeşitli konularda önemli kararlar alınmaktadır. Eğitim sürecinde ise testlerden elde edilen sonuçlar özellikle öğrenci başarısının değerlendirilmesinde, sınıf geçme veya mezuniyet ile ilgili kararlarda, okul yöneticilerinin ve çalışanlarının değerlendirilmesinde giderek daha sıkılıkla kullanılmaktadır (Kubiszyn ve Borich, 2000; Popham, 1999). Alınan bu kararlar bireylerin kişisel, toplumsal, sosyal ya da politik durumları üzerinde önemli etkilere sahip olabilmektedir (Clauser ve Mazor,1998; Messisck, 1995). Bu durumda ölçme araçlarının nitelikleri alınan bu karaları etkileyeceği için, ölçme araçlarından elde edilen puanların geçerlik ve güvenirlik gibi temel psikometrik özelliklerinin kabul edilebilir seviyelerde olması beklenir (Horst 1966).

Bir ölçme aracı, o aracı alan tüm bireyler için aynı özelliği ölçmeli ve her bir bireye ölçülen psikolojik özelliği göstermeleri konusunda eşit fırsat sunmalıdır (Roever, 2005). Benzer şekilde Messick (1995) de bir ölçme aracından elde edilen puanların geçerli, güvenilir, karşılaştırılabilir ve adil

olması gerektiğini vurgulamaktadır. Bu durum eğitim ve psikolojide kullanılan ölçme araçlarına ilişkin standartlarda da açıkça dile getirilmektedir.

Amerikan Eğitim Araştırmaları Derneği (American Educational Research Association), Amerikan Psikoloji Birliği (American Psychological Association) ve Eğitimde Ölçme Ulusal Konseyi'nin (National Council on Measurement in Education) (1999) *Eğitimde ve Psikolojide Ölçme Standartları* adlı ortak yayını test kullanıcıları, test yayıncıları ve araştırmacılar için testlerin kullanımı, test puanlarının yorumlanması, basım ve dağıtım vb. birçok konuda standartları içermektedir. İlgili yayınında test puanlarının adil (fairness) olması gerektiğiyle ilgili olarak 12 standart sunulmaktadır. Bu standartlarla test ya da madde puanlarının alt gruplardan gelen bireyler için aynı anlam ifade etmesinin puanların karşılaştırılabilir olması açısından önemini vurgulamış; aksi durumda ilgili test ya da maddeler için adilliğin sağlanamayacağı uyarısı yapılmıştır. Örneğin ilgili yayında 7.3 nolu standartta bu durum şöyle vurgulamaktadır:

> Standart 7.3: "Daha önceki araştırmalar, test katılımcıları evrenindeki farklı yaş, etnik köken, kültürel yapı ve cinsiyet grupları için testle ölçülen özellik bakımından madde veya test performansı farklılıklarını inceleme ihtiyacını işaret ediyorsa mümkün olduğu oranda bu tür çalışmalar yapılmalıdır. Bu tür araştırmalar, belirli bir grup için test puanlarında yanlılığa yol açabilen test tasarımı, içeriği veya test formatını belirlemeli ve bunları azaltmak için düzenlenmelidir" (syf: 81).

Eğitim ve psikolojide çeşitli amaçlarla kullanılan testlerin belirtilen standartları karşılaması gerekmektedir. Eğer bir ölçme aracı tüm bireyler için eşit ölçme yapmıyorsa, bireyler arasında gözlenen puan farklılıkları sadece ölçülmesi amaçlanan özellikle ilgili farkları değil aynı zamanda ölçme hatalarını da içerecektir. Diğer bir ifadeyle bir öğrencinin bu testten aldığı puan grup üyeliği ve testle ölçülen özelliğinin bir fonksiyonu olmaktadır. Bu durumda aynı yeteneğe sahip fakat farklı gruptaki bireylerin test puanları ve başarı sıralamaları farklı olacaktır. Örneğin, özel bir orta öğretim

kurumunun öğrenci seçmek üzere kullandığı genel yetenek testinde alt sosyo ekonomik düzeyden gelen katılımcıların üst sosyo ekonomik düzeydekilere göre tutarlı olarak daha düşük puanlar alsın. Bu durumda sadece test puanlarındaki farkları dikkate alarak alt sosyo ekonomik düzeyden gelen öğrencilerin daha düşük yetenek düzeyine sahip olduğu sonucuna varılmaktadır. Bu sonuca uygun olarak üst sosyo ekonomik düzeyden gelen öğrencilerin bu okulda eğitim hakkı kazanması daha olası bir durum olacaktır. Benzer şekilde uluslararası düzeyde öğrenci başarısının belirlendiği PISA ve TIMSS gibi projelerden elde edilen sonuçlara dayalı olarak ülkeler başarıları bakımından sıralanmakta veya eğitim politikalarına ilişkin önemli kararlar alınmaktadır. Bu durumda bu projelerde kullanılan testlerde yer alan maddelerin katılımcılar için eşit ve yansız ölçme yaptığını gösteren kanıtlara ihtiyaç duyulmaktadır.

Araştırmacılar veya test geliştiriciler, kullandıkları ölçme araçlarından elde ettikleri puanların testi alan tüm bireyler için adil ve test puanlarının karşılaştırılabilir olduğunu göstermek için ampirik kanıtlar sunacak geçerlik çalışmaları yapmaktadırlar. Messick (1995), geçerlik kavramını test puanlarını anlamlandırmak veya yorumlamak amacıyla bilimsel sorgulama ve mantığa dayalı karar verme sürecin bir araya getirmek olarak tanımlamaktadır. Geçerlik test ya da değerlendirmelerin bir özelliğinden ziyade test puanlarının anlamıyla ilgilidir. Daha teknik ifadeyle test puanı, ilgili yapıya uygun olarak bir test, ölçek, gözlem veya başka bir ölçme aracı yoluyla gözlenen davranışın kodlanmasını veya özetlenmesini ifade eder. Bu anlamından dolayı test puanları sadece bireylerin davranışlarındaki değişimleri/tutarlılıkları değil, kişilerin/grupların özelliklerini, koşul, çevre vb. sosyal göstergeleri de kapsamaktadır.

Bir ölçme aracından elde edilen puanların geçerliğini belirlemek için yapı geçerliği çalışmaları yapılmaktadır. Bu çalışmalar test puanlarının ölçme amacını ne kadar yerine getirebildiğine ilişkin kanıt toplamak için yapılır (Crocker ve Algina, 1986). Messick (1995), yapı geçerliğini merkeze koyarak diğer geçerlik türlerinin (ölçüt ve kapsam) yapı geçerliğiyle yakından ilişkili olduğunu ve yapı geçerliği çalışmalarında test puanlarını yorumlamak ve anlamlandırmak üzere ölçüt ve kapsamla bağlantılı kanıtların bir araya

getirildiğini belirtmektedir. Messcik (1995) ve Kristanjansonn, Aylesworth, McDowell ve Zumbo (2005) eğitim ve psikoloji alanında kullanılan testlerden elde edilen puanların yapı geçerliği için en büyük tehditlerden birisinin yanlılık olduğuna dikkat çekmektedir. Bu kapsamda ölçmelerdeki yanlılık kavramı geçerlik kapsamında tartışılmış; test ve madde yanlılığı ve Madde İşlev Farklılığı kavramları açıklanmıştır.

## Ölçmede Yanlılık

Yanlılık, yapı geçerliği kavramı kapsamında, testle ölçülmesi amaçlanmayan yapıların açıkladığı varyansın test sonuçlarını olumsuz etkilemesi ve test puanlarına dayalı olarak verilen kararların geçerliğini düşürmesi olarak ele alınabilir. Daha teknik bir tanımla yanlılık, farklı alt gruplardaki bireylere ait test puanlarının bulundukları gruba bağlı olarak sistematik hata içermesidir (Angoff, 1993; Clauser ve Mazor,1998). Böyle bir durumda herhangi bir grup üyeliği, testle ölçülmek istenen yapı dışında bir varyans kaynağı olarak test puanlarına karışmaktadır.

Yanlılık aynı zamanda maddenin çok boyutlu olması kapsamında da tartışılabilir (Camilli ve Shepard, 1994). Eğer bir madde eş zamanlı olarak iki ya da daha fazla yapıyı ölçüyorsa o maddenin çok boyutlu olduğundan bahsedilir. Shealy ve Stout (1993), bir maddenin birden fazla yapıyı ölçtüğü durumlarda yanlılığın ortaya çıkabileceğini belirtmektedir. Çünkü eşleştirme değişkeni boyutunda bireyler aynı yetenek ya da katılma düzeyindeyken, ölçülen diğer boyut için bireylerin yetenekleri veya katılma düzeyleri farklı olacaktır (Gierl, 2005). Ackerman (1992), özel bir ölçekleme yapmadan bireyleri iki farklı boyutta eşitlemenin mümkün olmayacağını vurgulamaktadır. Örneğin okuma yükü gerektiren bir matematik problemi çözmek için, bireyin maddede ölçülen matematik becerisi dışında okuduğunu anlama becerisini de gerektirebilir. Okuma becerisi ilgili maddede bireylerin performansını ölçmek üzere düşünülmeyen bir değişken olduğu için, bireyler matematik becerisi bakımından eşleştirilse bile okuma becerileri farklı olduğu için yanlılık ortaya çıkacaktır. Bu durumda okuma becerisi, matematik becerisinden bağımsız olarak dilbilgisi veya kelime dağarcığı zayıf olan öğrencilerin test puanlarını

düşüren bir faktör olacaktır. Messcik (1995), ölçülmesi amaçlanmayan bu yapıların bazı grup ya da bireyler için testi daha zor hale getirmesini yapıyla ilişkisiz zorluk (construct-irrelevant difficulty) olarak tanımlamış ve bu durumun testlerin puanlanması, puanların yorumlanması ve kullanımıyla ilgili yanlılığın ve haksızlığın (unfairness) temel kaynağı olduğunu belirtmiştir.

Genel olarak yanlılık için olası kaynakların daha çok dil ve kültür kapsamında tartışıldığı görülmektedir. Bu konuda yapılan araştırma bulgularına göre olası yanlılık kaynaklarından bazıları şöyle sıralanabilir (Bakan Kalaycıgolu ve Berberoglu, 2010; Ercikan, 1998; Uiterwijk ve Valen, 2007):

- Soyo-ekonomik farklıklar
- Çeviriden kaynaklı anlam karmaşası,
- Bazı kavramlar için farklı kültürlerde aynı karşılıklarının bulunmaması,
- Farklı kültürden veya sosyal gruplardan gelen öğrencilerin bazı kavramlara farklı düzeylerde aşina olması,
- Kültüre özgü deyim, ima ve kavramların kullanılması
- Negatif, edilgen ve mecazi anlam içeren kavram veya cümle yapılarının kullanılması,
- Kullanım sıklığı yaygın olmayan, soyut ve anlam belirsizliği gösteren kavram ya da cümlelerin kullanımı,
- Dini, siyasi, etnik vb. bakış açıları
- Eğitim programlarındaki (kapsam, öğretim yöntemleri, kazanımlar vs.) farklılıklar

Dil ve kültüre bağlı yanlılık kaynaklarına ek olarak çalışmalarda ayrıca cinsiyet de yanlılık bağlamında incelenmiş; maddenin yapısı/formatı, kapsamı, bilişsel karmaşıklık düzeyi gibi etmenler olası yanlılık kaynağı olarak ortaya konmuştur (Bakan Kalaycıgolu ve Berberoglu, 2010; Mendes-Barnett ve Ercikan, 2006).

**Test ve Madde Yanlılığı.** Yanlılığın ölçme araçlarından elde edilen puanların geçerliği için en büyük tehditlerden biri olduğu dikkate alındığında, ölçme sonuçlarındaki yanlılık kaynaklarının azaltılması için test (test bias) ve madde düzeyinde yanlılık (item bias) analizleri yapılmaktadır. Böylece olası yanlılık kaynaklarının azaltılarak test puanlarının geçerliğinin artırılması amaçlanmaktadır.

Test yanlılığı incelemeleri, testteki maddeleri tek tek değerlendirmekten ziyade, toplam test puan ortalamaları arasındaki farklılık bakımından alt grupları karşılaştırmaktadır. Bu durumda alt gruplara ait test puanlarının gerçek puanları yordadığı dikkate alınarak test yanlılığı değerlendirilmektedir (Camilli ve Shepard, 1994). Test yanlılığı, herhangi bir gruba üye olmaktan kaynaklı test puanlarındaki sistematik farklılıkları ifade etmektedir. Bu durum testin bütünün bir özelliğidir ve toplam test puanları düzeyinde analiz edilir. Çoğu durumda maddelerin toplamının sahip olduğu özelikler, tek tek maddelerin sahip olduğu özelliklerden farklı olabilir. Ayrıca, herhangi bir grup için avantaj ve dezavantaj gösteren madde sayıları birbirine yakın olduğunda, yanlılıklar toplam puan düzeyinde birbirinin etkisini yok edebilir. Bu durumda toplam puanların çok az veya hiç yanlılık taşımadığı sonucuna ulaşılabilir (Hong ve Roznowski, 2001).

Çoğu durumda araştırmacılar madde düzeyinde yanlılık ile ilgilenmektedir. Bu özellikle test geliştirme sürecinde yararlıdır. Böylece yanlı maddeler gözden geçirilebilir ya da testten çıkarılabilir. Bu durum test denkliğini sağlamada önemli ve gerekli bir süreçtir (Kamata ve Vaughn, 2004). Madde yanlılığı testteki her madde için ayrı ayrı geçerlik kanıtı toplama sürecini işaret etmektedir. Bu durumda maddelerin ilgili alt gruplarda benzer işlev gösterdiğine ilişkin veriler elde edilmeye çalışılır.

Madde yanlılığı kavramına iki anlam yüklenebilir: Yargısal ve istatistiksel anlam. Bu kavramın yargısal anlamı, bir maddenin bazı gruplara avantaj/dezavantaj sağlamasını; istatistiksel anlamı ise, karşılaştırılabilir yetenek düzeyindeki grup üyelerinin ilgili maddedeki performansları bakımından istatistiksel bakımdan anlamlı farkların ortaya çıkmasını ifade etmektedir. Madde yanlılığı kavramı, yargısal ve istatistiksel anlamı bakımından bir karmaşaya yol açtığı için, madde işlev farklılığı (MİF) kavramı

yanlılık yerine daha yaygın olarak kullanılmakta, böylece madde yanlılığının ele alınmasında yargısal yaklaşım ile ampirik kanıt bulma yaklaşımı arasındaki ayrım yapılmaktadır (Angoff,1993).

Hambleton, Swaminathan ve Rogers (1991) yanlılık çalışmalarında, odak ve referans gruplarının test maddelerindeki performanslarına ilişkin deneysel kanıtların toplanması gerektiğini vurgulamaktadır. Araştırmacılar yanlılık kararına varmak için deneysel kanıtlar yeterli olmadığını ve bu araştırmalarda elde edilen deneysel kanıtı tanımlamak için madde yanlılığı ile madde işlev farklılığı (MİF) kavramının farklılığını ayırt etmek gerektiğini belirtmektedir. Bu tartışmalara paralel olarak son yıllarda yapılan çalışmalarda "madde yanlıdır" demek yerine MİF kavramının kullanımı tercih edilmektedir (Ellis ve Raju, 2003).

MİF, testle ölçülen psikolojik özellik bakımından denk olan ancak cinsiyet, sosyo-ekonomik düzey gibi değişkenler açısından alt gruplarda yer alan bireylerin, bir maddeyi doğru yanıtlama olasılıklarının farklılaşması olarak tanımlanabilir (Camilli ve Shepard,1994). Belli istatistiksel sınamalar sonrası MİF gösteren bir madde olası yanlı madde olarak değerlendirilir (Kamata ve Vaughn, 2004). Sadece ampirik kanıtlara dayanarak bir maddenin yanlı olduğu kararı verilemez. Madde yanlılığının varlığını belirlemek için sonraki madde yanlılık analizlerinin (içerik analizi, deneysel değerlendirme) yapılması gerekmektedir. Yani, maddenin işlev farklılığı göstermesinin olası sebepleri tespit edilerek, maddenin ölçülmek istenen yapıdan ilişkisiz olarak bazı alt gruplar için adil olmayan bir avantaj /dezavantaj sağlayıp sağlamadığının uzman kişiler tarafından belirlenmesi gerekmektedir (Camilli ve Shepard, 1994; Hambelton ve diğerleri, 1991;Zumbo, 1999).

Yanlılık çalışmalarında, madde etkisi (item impact) ve MİF kavramları arasındaki farkın açıklanması 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 ile ölçülen psikolojik özellik bakımından gruplar arasındaki gerçek farklardan (true differences) kaynaklanmaktadır. Oysa MİF için alt gruplardan gelen bireylerin bir maddeyi doğru yanıtlama olasılıklarının farklılaşması koşulludur; yani alt gruplarda yer alan bireylerin

puan ortalamaları hesaplanmadan önce gruplar testle ölçülen özellik bakımından eşleştirilir. Böylece grupların puan ortalamaları arasındaki fark, grupların yetenek düzeylerindeki farklılaşmadan değil, MİF'ten kaynaklanmaktadır (Camilli ve Shepard, 1994).

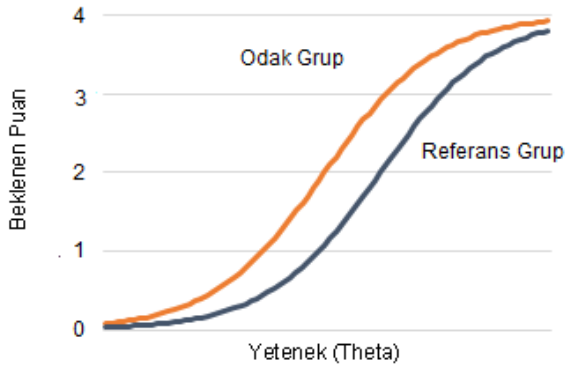
MİF çalışmalarında, ölçülen psikolojik özellik bakımından benzer bireylerin ait oldukları alt gruptan bağımsız olarak test maddelerinde benzer performans göstermesi gerektiği varsayımından yola çıkılarak ilgili özellik bakımından karşılaştırılabilir iki grup kullanılır. Bu iki grup, testle ölçülen özellik bakımından eşleştirilmiş ancak belli bir özellik bakımından (cinsiyet, etnik köken, meslek, vb.) farklı alt gruplardan gelmektedir. Odak (focal) grup ve referans (reference) biçiminde adlandırılan bu gruplardan odak grup, referans gruba göre dezavantajlı olduğu düşünülen gruptur.

Karşılaştırılabilir iki grup elde etmek amacıyla bireylerin testle ölçülen psikolojik özellikleri –yetenekleri- iki ölçüte göre eşleştirilir: iç ölçüt ve dış ölçüt. İç ölçüt, bir testte üzerinde çalışılacak maddelerden elde edilen toplam test puanları veya kestirilen yetenek düzeyi iken; ilgili test maddeleriyle aynı yapıyı ölçen diğer bir testteki performans değişken olarak kullanılırsa, bu durumda kullanılan dış ölçüttür. Her durumda, eşleştirme ölçütünün, yapı geçerliği yüksek olmalıdır; yani bu ölçüt ölçmeyi amaçladığı özelliklerin dışında bir yapı göstermemelidir. Yetenek düzeyi eşleştirilerek odak grubun performansı referans grup ile karşılaştırılır. Karşılaştırma sonuçlarına göre MİF gösteren maddeler için, maddenin doğru cevaplandırılma olasılığının odak ve referans grup üyeleri için farklı olduğu kararı verilir (Camilli ve Shepard, 1994; Zumbo 1999). MİF çalışmalarında odak ve referans grup bireyleri belirli yetenek aralıklarında eşleştirildikten sonra, ilgili maddedeki performansları karşılaştırılır.

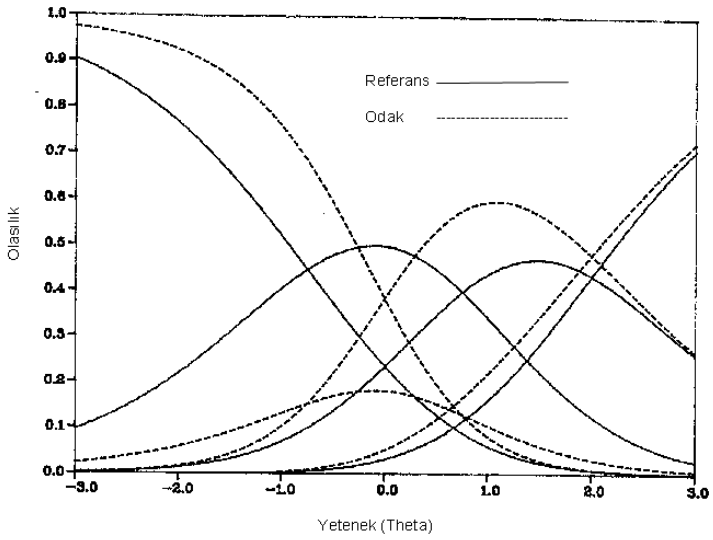
Maddenin referans ve odak grupta MİF göstermesi iki farklı biçimde ortaya çıkabilir: tek biçimli (uniform) ve tek biçimli olmayan (non-uniform) MİF. Su ve Wang (2005), tek biçimli ve tek biçimli olmayan MİF'i testle ölçülmek istenen psikolojik özellik (yetenek) ve grup üyeliği kavramlarını kullanarak açıklamaktadır. Buna göre tek biçimli MİF, bireylerin yetenekleri veya katılma düzeyi ile bireylerin ait oldukları grup arasında bir ilişki olmama durumunu ifade etmektedir. Çok kategoride puanlanan bir madde için tek biçimli MİF,

herhangi bir tepki ya da puan kategorisinde, bir gruptaki bireylerin tutarlı olarak diğer gruba göre daha iyi performans göstermesi ya da daha olumlu tepkiler göstermesi durumunda ortaya çıkmaktadır.

Tek biçimli MİF için örnek olarak, beş kategoride puanlanan bir maddeden elde edilen referans ve odak grup için beklenen tepki kategori puanları ve her bir tepki kategorisindeki olasılıkları gösteren grafikler Şekil 1 ve 2'de verilmektedir.

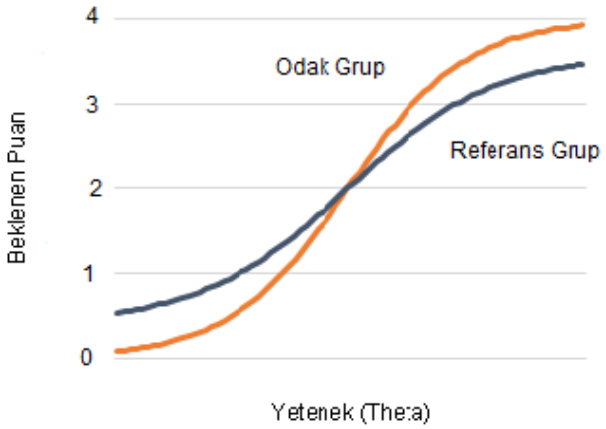



Şekil 1. Tek Biçimli MİF İçin Grupların Beklenen Puan Dağılımı

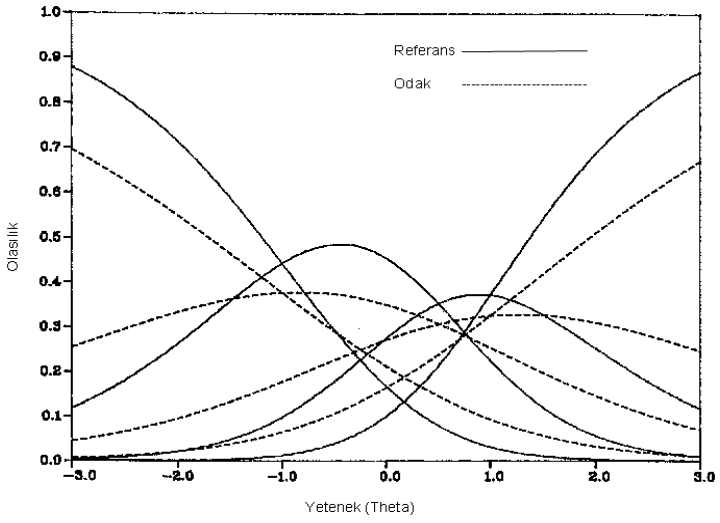



Şekil 2. Tek Biçimli MİF İçin Gruplardan Elde Edilen Kategori Tepki Fonksiyonları

Tek biçimli olmayan MİF, bireylerin yetenekleri veya katılma düzeyi ile bireylerin ait oldukları grup arasında bir ilişki olma durumunu ifade etmektedir. Bu durumda, farklı yetenek düzeyleri için herhangi bir tepki ya da puan kategorisinde yanıt verme olasılıkları herhangi bir grubun lehine/aleyhine değişmektedir. Aynı madde için tek biçimli olmayan MİF durumunu gösteren grafikler Şekil 3 ve 4'te verilmektedir.




Şekil 3. Tek Biçimli Olmayan MİF İçin Grupların Beklenen Puan Dağılımı




Şekil 4. Tek Biçimli Olmayan MİF İçin Gruplardan Elde Edilen Kategori Tepki Fonksiyonları

Eğitim ve psikoloji alanında kullanılan ölçme araçlarından elde edilen puanların geçerlik kanıtları test ve madde yanlılığı analizlerini de içermektedir. Geçtiğimiz 20 yılda madde ve test yanlılığını belirleme çalışmaları öncelikli olarak iki kategorili puanlanan (dikotomus) maddeler üzerinde yoğunlaşmıştır. Ancak, çok kategorili puanlanan (polytomus) madde formatının tutum, kişilik vb. psikolojik yapıların ölçülmesinde daha sıklıkla tercih edilmesi ve performansa dayalı durum belirleme, otantik durum belirleme, portfolyo vb. ölçme yaklaşımlarının gelişimine bağlı olarak bu tür maddeler için de yanlılık belirleme araştırmalarının arttığı dikkat çekmektedir (Kim, Cohen, Alagöz ve Kim, 2007). Bu kısımda çok kategorili puanlanan madde formatı detaylı olarak ele alınmıştır.

**Çok Kategorili Puanlanan (Polytomus) Maddeler**

Çok kategorili puanlanan madde formatı bilişsel ve duyuşsal özelliklerin değerlendirilmesiyle ilgili ölçme uygulamalarında sıklıkla kullanılmaktadır. Bilişsel becerilerle ilgili durum belirlemeler okuduğunu anlama, yazılı anlatım ve matematik gibi alanlarda bilişsel yapıların ölçülmesini içerir. Bu ölçmelerde tamamen çok kategorili puanlanan maddeler kullanılacağı gibi, ikili ve çoklu puanlanan maddeler birlikte de kullanılabilir (Shaeffer, Henderson-Montero, Julian, ve Bené, 2002; Thissen ve Wainer, 2001). Böylece test geliştiriciler her iki tür madde formatının da avantajlarını kullanabilirler. Çünkü iki kategorili madde formatı bilişsel alandaki birçok konu için kullanılabilir ve puanlanması daha az zaman alıcıdır. Buna karşın, çok kategorili puanlanan maddeler belirli bir kavram ya da beceri ile ilgili olarak bireyin karmaşık öğrenmelerini değerlendirme konusunda daha etkilidir; ancak bu tür maddelerin puanlanması zaman alıcıdır ( Sykes ve Hou, 2003; Thissen ve Wainer, 2001).

Bilişsel becerilerle ilgili durum belirlemelerde çok kategorili puanlanan maddeler genellikle yanıtı testi alan tarafından yapılandırılan maddeler veya performans görevleri biçiminde kullanılabilir. Yapılandırılmış yanıt içeren maddelerde bireylerin bir soruya ya da ifadeye yanıt vermek üzere yanıtlarını yazmaları beklenir. Performans görevi kapsamındaki maddelerde ise

öğrencilerin değerlendirme, analiz etme ve sentez yapma becerilerini göstererek kendilerine özgü yanıtlarını oluşturmaları gerekmektedir. Kompozisyon türü sınavlar ve öğrenci portfolyoları bu tür maddelere örnek olarak verilebilir.

Bilişsel beceriler kapsamında çok kategorili puanlanan madde formatı çözümü birden fazla adım gerektiren maddeler (multistep item) biçiminde de olabilir. Bu tür madde formatında, öğrencilerin çözüm sürecindeki her bir adımı göstererek bir problemi çözmeleri beklenir. Çok adımlı maddeler tipik olarak matematik ve fen derslerinde öğrencilerin başarılarının değerlendirilmesinde kullanılır (Dodd, De Ayala ve Koch, 1995). Yapılandırılmış yanıt gerektiren maddeler, performans görevleri ve çok adımlı maddelere ait puanlar, bireylerin yazılı yanıtlarını veya ürünlerini değerlendirilen puanlayıcılar tarafından belirlenir ve bu puanlar bireylerin ilgili beceri ya da kavramları ne kadar doğru anladıklarının derecesini tanımlar.

Bilişsel becerilerden farklı olarak duyuşsal özelliklerin (örn., ilgi, tutum, kişilik) ölçülmesinde genellikle çok kategorili puanlanan maddeler kullanılmaktadır. Örneğin, tutumla ilgili değerlendirmelerde kullanılan maddeler tipik olarak yanıtlayıcıların katılma ya da katılmama düzeylerini gösteren çok kategorili tepki seçenekleri içerir. Bu tür maddelere “tamamen katılıyorum”dan “hiç katılmıyorum”a uzanan bir boyut üzerinde tepki seçenekleri içeren Likert tipi maddeler örnek olarak verilebilir. Ancak iki kategorili puanlanan maddeler bireylerin düşünceleri veya eğilimlerini belirleme konusunda yanıtlayıcılara uygun seçenek aralığı sunamaz (Ostini ve Nering, 2006). Sonuç olarak çok kategoride puanlanan maddeler bu tür özelliklerin değerlendirmelerde tipik olarak kullanılan maddelerdir.

Bilişsel beceriler kapsamında MİF, ilgili beceri –yetenek- bakımından eşleştirilmiş gruplar arasında gözlenen performans farklılıklarını tanımlar. Tutum, algı, ilgi gibi dereceli tepkilerin verilebildiği madde formatının kullanıldığı ölçme uygulamalarında MİF tanımı tamamen farklıdır. Örneğin, tutum ölçeğindeki tepkiler yetenek yerine genel katılma (overall agreement) düzeyinde eşleştirilir (Dodeen, 2004). Dodeen ve Johanson (2003), bilişsel bir özelliğin ölçülmesinde doğru yanıtın, bir tutum maddesine yönelik “katılma” düzeyindeki artıştaki etki ile benzer olduğunu belirtmektedir. Tutum ölçen bir

madde için, MİF belirli bir alt grubun diğer alt gruba göre o maddeye daha pozitif tepkiler verdiğini ifade eder.

Eğitim ve psikoloji alanının konuları düşünüldüğünde, çok kategorili tepki seçenekleri sayesinde daha çok bilgi sunan çok kategorili puanlanan maddeler, iki kategorili puanlanan maddelere göre daha ilgi çekici görünmektedir (Ostini ve Nering, 2006). Çok kategorili tepki modellerine olan ihtiyaç özellikle kişilik, tutum vb. duyuşsal değişkenlerin ölçümünde daha belirgin olarak ortaya çıkmaktadır. Cox (1980) ve Kamakura ve Balasubramanian (1989) bu bağlamdaki ölçümler için iki kategorili puanlama ayırımının çok belirgin olmadığını, yanıtlayıcıları sınırlandırdığını ve daha çok kategoride puanlamaya ihtiyaç olduğunu ifade etmektedir (Akt.Ostini ve Nerig, 2006).

Çok kategorili puanlanan maddelerin kullanımının artmasına (Dodeen, 2004; Zwick, Donoghue ve Grima, 1993) bağlı olarak bu tür maddelerden elde edilen veri setlerini analiz etmek üzere farklı ölçme yaklaşımları kullanılabilir. Bu türden farklı puanlama gerektiren maddelere verilen tepkilerden elde edilen puanların ölçeklenmesinde iki temel ölçme kuramı bulunmaktadır: Klasik Test Kuramı ve Madde Tepki Kuramı (MTK). Bu kuramların temel odak noktası, (1) ölçme araçlarının ya da maddelerin özelliklerini kestirmek, (2) bireyin ölçülen özellikteki yerini kestirmektir (Bei, 1995).

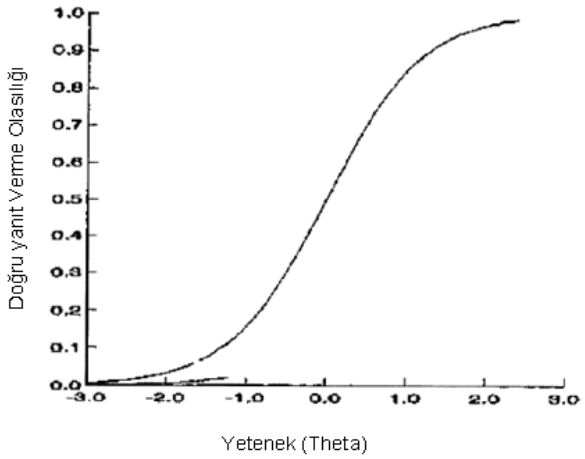
## Madde Tepki Kuramı ve Çok Kategorili Tepki Modelleri

Ölçme tarihi incelendiğinde, ölçme yaklaşımlarıyla ilgili yöntemleri etkileyen esas faktörler arasında test maddelerinin niteliğinin değerlendirilmesinde ve test puanlarının elde edilmesiyle ilgili olarak Klasik Test Kuramından Madde Tepki Kuramına (MTK) doğru değişen eğilim olduğu dikkat çekmektedir (Beretvas, 2000).

MTK, bireyleri ve maddeleri tek bir örtük değişken (yetenek) üzerinde ölçeklemek için bireylerin ilgili maddelere verdikleri yanıtları kullanmaktadır ve herhangi bir yetenek düzeyindeki bireyin maddeyi doğru yanıtlama olasılığının, bireyin yeteneğinin ve maddenin özelliklerinin bir fonksiyonu

olduğunu belirtmektedir. Bu modelin grafiksel gösterimi ise madde karakteristik eğrisini tanımlamaktadır.

Madde karakteristik eğrisi, belirli yetenek düzeyindeki bireyin bir maddeye vereceği tepkiye ilişkin olasılıkları grafiksel olarak göstermektedir. 1-0 puanlanan bir başarı testi için bu eğri herhangi bir yetenek düzeyindeki bireyin ilgili maddeyi doğru yanıtlama olasılığı hakkında bilgi verir (Hambelton ve Swaminathan,1985). Aşağıda bir madde karakteristik eğrisi örneği verilmiştir.




Şekil 5. Madde Karakteristik Eğrisi

Şekil 5'te yatay eksen bireyin test ile ölçülen örtük özelliğini –yetenek-, dikey eksen ise ilgili maddeye doğru yanıt verme olasılığını- P(Ɵ)- göstermektedir. Bu eğriye göre, bireyin yetenek düzeyi arttıkça maddeyi doğru yanıtlama olasılığı da artmaktadır.

Madde Tepki Kuramında bireyin bir test maddesindeki performansı ile örtük değişken arasındaki ilişkiyi doğrusal olmayan matematiksel modelleme yoluyla açıklamaktadır. Bu modeller kullanılan puanlama türüne göre ikili (1-0) veya çoklu modeller biçiminde olabilir. Bu matematiksel modelleri kullanarak bireylerin testteki performansını belirlemek için öncelikle bazı varsayımların karşılanması gerekmektir. MTK uygulamaları için bu varsayımlar tek boyutluluk ve yerel bağımsızlıktır (Hambleton ve diğerleri, 1991).

Tek boyutluluk testin tek bir örtük özelliği ölçmesi yani test puanlarının baskın tek bir özelliğe bağlı olmasıdır. Yerel bağımsızlık ise belirli bir yetenek düzeyindeki bireylerin, belirli bir maddeyi doğru cevaplandırma olasılıklarının testteki diğer maddelerden bağımsız olma durumunu ifade etmektedir. Bu varsayımlara ek olarak seçilen MTK modelinin madde ve yetenek düzeyinde veriyi ne kadar iyi temsil ettiği incelenmelidir. Bu yüzden herhangi bir MTK modeli seçilirken eldeki veriyi en iyi temsil edecek modelin seçilmesi önemlidir. Böylece, ölçülen örtük özellik ile madde performansı arasındaki ilişkiyi en doğru gösteren model elde edilir. Bu durum model-veri uyumunun incelenmesini belirtmektedir.

İlgili varsayımlar ve model-veri uyumu sağlandıktan sonra, MTK'nın Klasik Kurama göre en önemli avantajı kestirilen madde ve yetenek parametrelerinin değişmezlik (invariance) özelliğidir. Yani, kişilerin yetenek parametreleri maddelerin kolay ya da zor olmasından bağımsız olarak kestirilebilmektedir (item-free). Benzer şekilde madde parametreleri de aynı evrenin farklı alt gruplarından bağımsız olarak kestirilebilir (sample-free). Buna ek olarak MTK modelleri her bir yetenek düzeyi için yapılan ölçümlerin doğruluğu konusunda daha çok bilgi vermektedir. Yetenek kestiriminin doğruluğu test bilgi fonksiyonu kullanılarak değerlendirilebilir. MTK belirtilen bu avantajları sayesinde, eğitim alanında kullanılan test ve ölçme problemlerine (madde yanlılığı, test puanlarını eşitleme, amaca uygun test geliştirme vb.) daha iyi çözümler sunmaktadır.

İlk MTK modelleri ikili cevaplar örneğin 0 (yanlış / hayır), 1(doğru / evet) için geliştirilmiştir; ancak son yıllarda eğitim ve psikolojide kullanılan tüm veri türleri için MTK modelleri de bulunmaktadır. MTK uygulamalarına ve bu kuramın gelişmesine önemli bir katkı, kuramı çok kategorili yanıt gerektiren veriler ve çok boyutlu durumlar için de genişleten Samejima tarafından yapılmıştır. Samejima (1969) iki veya daha fazla sıralı yanıt kategorisi içeren maddeler için Dereceli Tepki Modelinin (Graded Response Model) kullanılmasını önermiştir. Diğer çok kategorili tepki modelleri iki parametreli modeller esas alınarak formüle edilmiştir.

Bu modellerden bazıları şunlardır (Ostini ve Nerig, 2006):

- Samejina'nın Dereceli Yanıt Modeli (Graded Response Model)
- Muraki'nin Dereceli Ölçek Modeli (Muraki's Rating Scale Model)
- Bock'un Sınıflamalı Yanıt Modeli (Nominal Response Model)
- Master'ın Kısmi Puan Modeli (Partial Credit Model)
- Muraki'nin Genelleştirilmiş Kısmi Puanlama Modeli (Generalized partial credit Model)

Dereceli tepkileri analiz etmek üzere pek çok MTK modelleri bulunmasına karşın, Kısmi Puan Modeli, Rasch modellerinin özelliklerini taşıdığından çok kategorili MTK modelleri arasında yaygın olarak kullanılan bir modeldir. Çok kategorili Rasch modellerinin iki önemli avantajı bulunmaktadır (Embretson ve Reise, 2000). Bunlardan ilki parametre kestiriminin tek adımda ve tek bir eşitlik (direct models) yardımıyla yapılmasıdır. Bu tür modellere Kısmi Puan Modeli ve Andrich'in Dereceli Tepki Modeli örnek olarak verilebilir. Bu tür doğrudan modeller, bir bireyin herhangi bir kategoride yanıt verme olasılığını iki adımda hesaplayan modellere (indirect models) göre daha az matematiksel hesaplama gerektirmektedir. Çok kategorili Rasch modellerinin diğer avantajı ise, testten alınan toplam puanın bireyin performansını kestirmek için yeterli bir istatistik olarak kabul edilmesidir.

Çok kategorili puanlanan maddeler üzerinde yapılan MİF belirleme çalışmalarında ağırlıklı olarak Dereceli Tepki Modeli'ne uygun verilerin kullanıldığı dikkat çekmektedir (Örneğin: Ankenmann, Witt ve Dunbar, 1999; Artar, 2007; Bolt, 2002, Dooden, 2004; Fidalgo ve Bartman, 2010; Garett, 2009; Wang ve Su, 2004). Uygulama şartlarında, araştırmacıların sahip oldukları verilerin her zaman bu modele uyum göstermesi beklenemez. Bu model dışında sıralı tepki seçenekleri için geliştirilmiş diğer çok kategorili MTK modelleri de bulunmaktadır. Bu modellere uygun veriler üzerinde de MİF belirleme çalışmalarının yapılması araştırmacıların/uygulayıcıların daha geçerli değerlendirmeler yapması konusunda yaralı olacaktır.

Son dönemde yapılan çalışmalarda (örn., Carter, 2011; Wang, Tay ve Drasgow, 2013) parametrik MTK modellerinden açımlamalı (unfolding) modellerin kullanım sıklığının arttığı dikkat çekmedir. Ancak bu modellerde

parametre kestirimi için farklı yazılım programlarının gerekmesi ve bu programları edinme konuları dikkate alınarak, sıralı tepki seçenekleri için çok kategorili MTK modellerinden Kısmi Puan Modeli daha önce bahsedilen avantajlarından dolayı bu çalışmada tercih edilmiştir. Bu model ayrıca diğer araştırmacılar tarafından da tercih edilen bir modeldir (örn.,Garrett, 2009; Kim, Cohen, Distefano ve Kim, 1998; Su ve Wang, 2005; Wang ve Su,2004). Bu modele ilişkin açıklamalar aşağıda sunulmuştur.

**Kısmi Puan Modeli**

Kısmi Puan Modeli (Partial Credit Model), iki veya daha fazla sayıda sıralı tepkilerin analizi için kullanılan tek boyutlu bir MTK modeldir. Master (1982) bu modeli esas olarak, çözümü birden çok adım gerektiren (multistep) test maddelerinin analizi için geliştirilmiştir. Bu model kısmi doğru yanıtların verilebileceği başarı testleri veya ölçülen özelliğin farklı düzeylerde olabileceği kişilik, tutum, inanç vb. değişkenlerin değerlendirildiği ölçeklerdeki maddelerin analizi için de uygundur. Belirli bir yetenek düzeyindeki birey için k veya k-1 kategorisini seçme olasılıkları ikili MTK modelleri yoluyla karşılaştırılır.

De Ayala (1993), kısmi puanlanan bir madde örneğini çözümü birden fazla adım gerektiren bir problem örneği üzerinden şöyle açıklamaktadır: (6:3)+2. Bir öğrencinin bu problemi çözebilmesi ve tam puan alabilmesi için iki adımı tamamlaması gerekmektedir:

1. Adım: Bölme işlemi (6:3)
2. Adım: Toplama işlemi (2+2)

Bu problemin doğru olarak çözülebilmesi için bu adımların sırasıyla tamamlanması gerekmektedir. Yalnızca birinci adımı tamamlayan öğrenci kısmı puan kazanmaktadır; ancak hem birinci hem de ikinci adım tamamlanamazsa öğrenci puan alamamaktadır. Bu durumda bu madde için üç farklı puan değeri vardır: 0,1 ve 2. Kategori 0 öğrencinin puan alamadığını, Kategori 1 kısmi puan aldığını ve Kategori 2 tam puan aldığını ifade etmektedir.

Kısmi Puan Modeli, iki puan kategorisine sahip olduğunda ikili puanlanan maddeler için 1 parametreli lojistik modele (Rasch) denktir. Bilindiği gibi bu model bireyin bir maddeye yanıt verme olasılığının bireyin yeteneğinden ve maddenin güçlük parametresinden etkilendiğini varsaymaktadır. Ayırıcılık parametresi tüm maddeler için sabittir ve yetenek parametrelerinin kestirimi için ham puanlar yeterli bir istatistiktir. Kısmı Puan Modelinde ise madde güçlük parametresi yerine *madde adım güçlüğünü* (item step difficulty) ($\delta$) parametresi bulunmaktadır. Bu model bireyin testteki performansı ile örtük özellik (yetenek) arasındaki doğrusal olmayan ilişkiyi modellemek için sadece "adım güçlüğü parametresi" kullanılır.

Adım güçlüğü parametresi, örtük özellik ölçeğinde ardışık iki tepki kategori eğrisinin kesiştiği noktayı göstermektedir. Bu yüzden madde parametreleri *kategori kesişim parametresi* (category intersection parameter) olarak da adlandırılmaktadır (Embretson ve Reise, 2000). Bu modelde, maddenin sahip olduğu tepki kategori sayısının bir eksiği kadar adım güçlüğü parametresi bulunmaktadır ve tüm maddelerin eşit eğim (ayırıcılık) parametresine sahip olduğu varsayılmaktadır. Madde adım güçlüğü parametresi ne kadar yüksek olursa, bu durum bir sonraki adıma geçmenin o kadar zorlaştığını ifade etmektedir. Bu model için herhangi bir kategoride yanıt verme olasılığı aşağıdaki formülle ifade edilir:

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

| | |
|---|---|
| Θ: Örtük özellik | j: 0, 1,2,…m |
| i: birey | mi: kategori eşik parametre sayısı |
| x: tepki kategorileri | $\delta_{ij}$ : j ile puanlanan kategoriye ait adım güçlüğü parametresi |

Bu formül, i bireyinin x kategorisindeki mi. adımda yanıt verme olasılığının, bireyin yetenek parametresi ile bir kategori eşik parametresi arasındaki farkının fonksiyonu olduğunu göstermektedir. Bu modelde, birey ne kadar çok adımı tamamlarsa, o kadar yüksek tepki kategori puanı almaktadır.

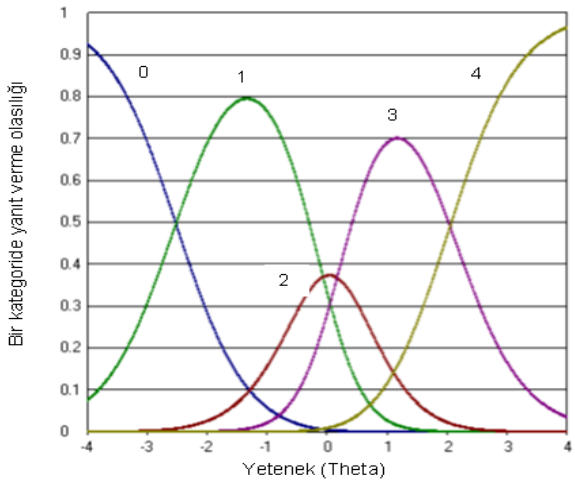
Yüksek kategori tepki puanı ise düşük tepki kategori puanına göre daha yüksek yeteneği ($\theta$) ifade etmektedir. Örneğin beş tepki seçeneği bulunan bir tutum ifadesi için tepki seçenekleri ve tamamlanması gereken adımlar aşağıdaki gibi gösterilebilir:

(Hiç katılmıyorum----------Tamamen katılıyorum)

0----------1----------2----------3----------4

Adım 1 Adım 2 Adım 3 Adım 4

Bu madde için adım güçlüğü parametreleri sırasıyla $\delta_1 : -2.519, \delta_2 : -0.063, \delta_3 : 0.170, \delta_4 : 2.055$ olarak hesaplanmıştır. Bu parametreler örtük özellik boyutunda kategori tepki eğrilerinin kesiştiği yerleri gösterdiği için, bir kategoride tepki verme olasılığının bir önceki kategoriye göre daha olası olduğu noktayı göstermektedir. Bir bireyin bu maddeden en yüksek kategoride puan alabilmesi için tüm adımları tamamlaması gerekmektedir. Bu madde için tamamlanması en zor adım katılıyorum ile tamamen katılıyorum tepki kategorileri arasındaki adımdır. Bu maddeye ait tepki kategori eğrilerinin grafiksel gösterimi Şekil 6'da verilmektedir.




Şekil 6. Beş Kategorili Bir Madde İçin Kategori Tepki Fonksiyonu

$(\delta_1 : -2.519, \delta_2 : -0.063, \delta_3 : 0.170, \delta_4 : 2.055)$

Şekil 6'ya göre, örneğin örtük özellik ölçeğinde yeri -3 olan bir birey bu madde için hiçbir adımı geçemeyerek 0 puan alması daha olasıdır. Buna karşılık örtük özelliği 2.5 olan bir bireyin dört adımı da tamamlayarak 5 puan alması daha olası olacaktır.

İkili puanlanan maddelerde olduğu gibi, çoklu puanlanan maddeler için de maddelerin alt gruplarda farklı işlev gösterip göstermediğinin incelenmesi gerekmektedir. Grupların madde performanslarının farklılaşması yetenek farklılıklarından kaynaklanabilir; ancak aynı yetenek düzeyindeki bireylerin farklı madde performansları göstermesi bu maddeler için MİF göstergesi olabilir. Çok kategorili MTK modelleri kapsamında kullanılabilecek MİF belirleme tekniklerine ilişkin açıklamalar aşağıda sunulmuştur.

**Çok Kategorili Puanlanan Maddelerde MİF Belirleme Teknikleri**

Çok kategorili puanlanan maddelerde MİF belirme teknikleri, iki kategorili puanlanan maddeler için geliştirilen MİF tekniklerinden farklı değildir. Benzer şekilde çok kategoride puanlanan maddelerde MİF belirlemek üzere herhangi bir tekniğin seçilmesi genellikle zordur ve bazı durumlarda oldukça karmaşıktır (Kim ve diğerleri, 2007). Bu karmaşa özellikle MİF tekniklerinin aynı maddelerde MİF göstermemesi durumunda ortaya çıkmaktadır.

Son yıllarda, eğitim ve psikoloji alanında kullanılan testlerde MİF belirlenmesi için birçok istatistiksel yöntem geliştirilmiştir. MİF ile ilgili alan yazın incelendiğinde, istatistiksel bir ölçüte dayalı olarak farklı MİF belirleme yaklaşımları olduğu görülmektedir. Bu yaklaşımlarda esas alınan ölçütler "anlamlılık testlerini, MİF miktarının yorumlanabilir ölçüsü ve kestirime ilişkin standart hata" nın kullanımını kapsamaktadır (Mapuranga,Dorans ve Middleton, 2008; Akt. Wang, 2009).

Çok kategorili puanlanan maddelerde MİF belirleme ile ilgili olarak kullanılan yaklaşımlardan biri, *hipotez testine dayalı anlamlılık sınamalarıdır*. Çok kategorili maddelerde MİF belirlemek üzere anlamlılık testlerini kullanan tekniklerinden bazıları şöyle sıralanabilir (Bolt, 2002; Kim ve diğerleri, 2007):

- Lojistik regresyon (Miller ve Spray,1993; French ve Miller, 1996; Zumbo,1999)
- Olabilirlik Oran Testi (Likelihood ratio test) (Thissen, 2001; Wainer, Sireci ve Thissen, 1991)
- Test ve maddelerin İşlev Farklılığı işlemi (Differential Functioning of Items and Test Prodecure) (DFIT- Flowers, Oshima, ve Raju,1999)
- Mantel Test (Mantel, 1963)
- Poly-SIBTEST (Chang,Mazzeo ve Roussos,1996)

Dorans ve Potenza (1994), bu MİF belirleme tekniklerini bireyin maddedeki performansı ile eşleştirme değişkeni arasındaki ilişkiyi gösteren özel bir modelin olup olmamasına göre şöyle sınıflandırmıştır:

- Parametrik olmayan gözlenen puan tekniği (Mantel Haenszel Test)
- Parametrik gözlenen puan tekniği (Lojistik regresyon)
- Parametrik olmayan örtük özellik tekniği (SIBTEST)
- Parametrik örtük özellik tekniği (Olabilirlik Oran Testi)

Bireyin maddedeki performansı ile eşleştirme değişkeni arasındaki ilişkiyi gösteren özel bir modelin olmadığı (model-free) teknikler parametrik olmayan; bu ilişkinin özel bir modelle tanımlandığı yöntemler ise parametrik teknikler olarak adlandırılmıştır (Dorans ve Potenza, 1994). Parametrik ve parametrik olmayan teknikler için kullanılan eşleştirme değişkeni gözlenen verinin bir fonksiyonu olarak kestirilen örtük özellik veya gözlenen puan olarak değişmektedir.

Bu çalışmada, MİF belirleme testi olarak parametrik örtük özellik tekniğine ait MTK'ya dayalı Olabilirlik Oran Testi (MTK-OOT) kullanılmıştır. MTK kuramsal özellikleri sayesinde, eğitim alanında kullanılan test ve ölçme problemlerine sunduğu çözümlerle araştırmacıların dikkatini çekmektedir. Bilindiği gibi, PISA ve TIMSS gibi uluslararası değerlendirme programlarında MTK'ya göre ölçeklemeler yapılarak bu kuramın avantajlarından yararlanılması amaçlanmaktadır. Bu durumda kuramsal özellikleri dikkate alınarak bireylerin örtük özelliklerini kullanan MİF belirleme tekniklerinin

(örneğin MTK-OOT) öncelikli olarak tercih edilme olasılığı yüksektir. MTK-OOT'ye ilişkin detaylı açıklamalar aşağıda sunulmuştur.

**MTK'ya Dayalı Olabilirlik Oran Testi (MTK-OOT).** MTK-OOT, ikili ve çoklu puanlanan maddelerde MİF belirlemek üzere Thissen, Steinberg ve Gerrard (1988) tarafından önerilmiştir. Dorans ve Potenza (1994) tarafından yapılan sınıflama dikkate alındığında, MTK-OOT, parametrik ve bireylerin gerçek puanlarına (örtük özellik) dayanan MİF belirleme teknikleri içinde yer almaktadır.

MTK-OOT aynı yetenek düzeyindeki bireylerin, maddelere ait tepki kategorilerinin herhangi birinde yanıt verme olasılıklarının birbirinden farklı olup olmadığını test eder. Çok kategorili puanlanan maddeler için, referans ve odak gruplarda madde gerçek puan fonksiyonu (item true function) eşit değilse, maddenin söz konusu iki grup arasında testle ölçülen özellikten ayrı olarak farklı işlev gösterdiği sonucuna ulaşılır (Cohen, Kim ve Baker, 1993). Referans ve odak gruplardan elde edilen sınır tepki fonksiyonu (boundary response function) veya madde parametreleri eşitse, madde gerçek puan fonksiyonu her iki grup için denktir.

MTK-OOT, MİF belirlemek üzere dar ve geniş model olarak tanımlanan iki modelin, olabilirlik oran farklarının anlamlılığının test etmektedir. Bu modellerden ilki, aynı maddeler için parametrelerin iki grup için eşitlendiği *dar* (compact, model C) modeldir. Dar model, referans ve odak grubun denk olduğunu varsayarak madde parametrelerini tüm örneklem grubunu kullanarak kestirmektedir. Bu kestirime ilişkin elde edilen genel uyum istatistiği, parametrelerin var olan verileri ne kadar iyi temsil ettiğini gösteren -2log olabilirlik (likelihood) değeridir. Diğer model ise MİF şüphesi olan madde/lerin (studied item), her iki grupta eşit parametrelere sahip olma sınırlığının bulunmadığı *geniş* (augmented, model A) modeldir. Geniş modelde MİF şüphesi olan madde için madde parametreleri referans ve odak gruptan ayrı ayrı kestirilir; yani ilgili maddeye ait parametrelerin gruplar arasında değişmesine izin verilmektedir. Diğer maddeler için parametre kestirimi dar modelde olduğu gibi yapılır ve geniş modele ilişkin genel uyum istatistiği elde

edilir. Geniş modelde, modele ilave edilen, parametrelere ihtiyaç olup olmadığını ve ilave parametreleri modelin uyum derecesini artırmada katkı sağlayıp sağlamadığı tespit edilir. Burada Ho hipotezi dar modelin K sayıda parametre içerdiğini, H1 hipotezi geniş modelin, L sayıda parametre içerdiği şeklinde kurulur. Olabilirlik oran analizi ile K sayıdaki madde parametresi uyum iyiliğini sağlamakta mıdır? L sayıdaki parametre modelin uyum iyiliğine katkı getirmekte midir? sorularına cevap aranır.

Eğer ilgili madde MİF gösteriyorsa geniş modelin daha iyi uyum istatistiği vermesi beklenir. Çünkü MİF içeren bir madde için gruplardaki bireylerin farklı yanıt davranışları göstermesi beklenir. Bu durumda ilgili maddeye ait parametrenin gruplarda serbest bırakılması bu maddeye ilişkin daha doğru kestirimlerin yapılmasına olanak sağlar ve geniş model için daha iyi uyum değerleri elde edilir. Her iki modelden elde edilen -2log olabilirlik değerlerini karşılaştırmak üzere bir $G^2$ istatistiğini kullanılır. Bu istatistiğe ilişkin formül şöyledir:

$$G^2_{(df)} = -2\log\left[\frac{\text{likelihood [A]}}{likelihood[C]}\right]$$

Likelihood [A] ve [C]: en çok olabilirlik kestirim yöntemiyle kestirilen modeldeki parametrelerin olabilirliği

df: iki model arasındaki parametre sayısındaki farka eşit olan serbestlik derecesi

Bu formülde, $G^2$ iki model arasındaki parametre sayısındaki farka eşit olan serbestlik derecesinde (df), her iki modelden elde edilen uyum istatistikleri arasında farksızlığı savunan yokluk hipotezini (Ho) test etmek üzere ki-kare dağılımı gösterir. Bu değer tablo değerinden büyükse iki gruptan kestirilen madde parametrelerinde farklılık olduğuna ve maddenin MİF içerdiğine karar verilir.

MTK çalışmalarında, farklı gruplardan kestirilen madde parametrelerinin karşılaştırılabilmesi için bu parametrelerin ortak bir ölçek üzerinde ölçeklenmesi gerekmektedir (Stocking ve Lord, 1983). Bunun için parametrelerin dönüşümü veya bağlanması (linking) işlemleri uygulanmaktadır. Örneğin iki grup söz konusu olduğunda, odak grup parametreleri referans grubunun ölçeğine yerleştirilir; yani ortak ölçek için referans grubun verileri kullanılır. Ancak, MULTILOG programı kullanılarak yapılan bir MTK-OOT analizi için bu tür işlemlere gerek yoktur. Çünkü MULTILOG'da parametreleri ortak ölçek üzerinde ölçeklemek üzere bir dizi ankor madde (common anchor items) kullanılmaktadır ve madde parametreleri, eş zamanlı olarak referans ve odak gruptan kestirilmektedir.

Doğru bir ölçekleme yapılması için seçilen ankor maddelerin MİF içermemesi önemlidir. Ankor maddelerin MİF içermediği varsayımı vardır ve bu maddeler için MİF analizi yapılmaz. Ankor maddelerin seçiminde *tekrarlayıcı madde ayıklama* işlemleri sonrası MİF göstermeyeceği tahmin edilen maddeler ankor madde olarak belirlenir.

Bu çalışmada ayrıca bireylere ait gözlenen puanların kullanıldığı MİF belirleme tekniği olan Mantel Test kullanılmıştır. Bu test, ucuz yazılım gerektirmesi, matematiksel hesaplama ve uygulama kolaylığı gibi avantajlara sahip olması, bu testin araştırmacılar tarafından daha sıklıkla tercih edilmesini sağlamaktadır. Böylece bu testin MİF analizlerinde kullanılmasının önemi devam etmektedir. Bu teste ilişkin açıklamalar aşağıda verilmektedir.

**Mantel Test.** Dorans ve Potenza (1994) tarafından yapılan sınıflamaya göre Mantel Test, parametrik olmayan ve bireylerin gözlenen puanlarına (observed score) dayanan MİF belirleme teknikleri arasında yer almaktadır. Mantel Test, ikili puanlanan maddeler için bir MİF belirleme tekniği olan Mantel-Haenszel Testinin bir uzantısı olup Zwick ve diğerleri (1993) tarafından çok kategorili puanlanan maddelerde MİF belirlemede kullanılması önerilmiştir.

Bu test, MİF'in olmadığı yokluk hipotezini test etmek üzere bir serbestlik derecesinde ki-kare dağılımına ilişkin istatistik vermektedir (Meyer, Huynh ve Seaman, 2004; Zwick ve diğerleri, 1993) ve MİF'i test etmek üzere 2XTXK çapraz tablo kullanmaktadır. Burada T bir maddedeki tepki kategori sayısı, K puan aralık sayısını ifade eder. Her puan aralık düzeyi için bir 2XT tablosu oluşturulur. Bu noktada, bireyleri eşleştirmek için farklı puan aralık düzeyleri oluşturulmalı ve MİF analizi için gerekli aralık sayısına karar verilmelidir. Örneğin her biri 4 kategoride (1,2,3 ve 4) puanlanan 20 maddelik bir test için toplam puan en düşük 20 ile en yüksek 80 arasında değişecektir. Bu test için iki farklı yaklaşımla bireyler yetenek düzeylerinde eşleştirilebilir.

Yaklaşımlardan ilki bu aralıktaki her bir toplam puanı, puan aralığı olarak kullanmaktır. Bu durum *dar aralıkta eşleştirmeyi* (thin matching) ifade etmektedir. Diğer bir yaklaşım ise, daha geniş puan aralık düzeyleri elde etmek üzere bazı toplam puanları birleştirerek aralıklar oluşturmaktır. Böyle bir eşleştirme ise *geniş aralıkta eşleştirme* (thick matching) olarak adlandırılmaktadır. Bu puan aralıkları, her bir aralıkta en az bir gözlem olacak biçimde oluşturulur (Donoghue ve Allen, 1993). Çizelge 1'de her bir puan aralığı için oluşturulmuş bir çapraz tablo örneği verilmektedir.

Çizelge 1. Mantel Test İçin Her Bir Puan Aralığında Oluşturulan Çapraz Tablo Örneği

| | Madde Puanları | | | | | |
|---|---|---|---|---|---|---|
| Grup | $Y_1$ | $Y_2$ | $Y_3$ | | $Y_t$ | Toplam |
| Referans | $n_{R1k}$ | $n_{R2k}$ | $n_{R3k}$ | ……. | $n_{Rtk}$ | $n_{R+k}$ |
| Odak | $n_{o1k}$ | $no_{2k}$ | $n_{o3k}$ | ……. | $n_{otk}$ | $n_{o+k}$ |
| Toplam | $n$+1k | $n$+2k | $n$+3k | ……. | $n$+tk | $n$++k |

Not: Çizelge Wang ve Su (2004)'dan alınmıştır.

$y_1$, $y_2$, $y_3$,..,$y_t$: Bir maddeye ait olası kategori puanları

$n_{Rtk}$ ve $n_{otk}$: İlgili puan aralık düzeyinde y maddesinin t tepki kategorisinde yanıt veren referans ve odak gruptaki birey sayısı

*n*+1k: İlk tepki kategorisinde yanıt veren referans ve odak gruptaki toplam birey sayısı

k: Eşleştirme yapılan puan aralık sayısı

Mantel Test istatistiği aşağıdaki gibi hesaplanmaktadır:

$$\chi^2_{Mantel} = \frac{[\sum \mathrm{O_K} - \sum \mathrm{E(O_K)}]^{\,2}}{\sum \mathrm{Var\ (O_K)}}$$

Formülde,

$O_k$: Beklenen odak grup toplam puanları

$E(O_k)$: Beklenen odak grup puanları

Var ($O_k$): Odak grup puanlarının varyansı

Bu teste ait yokluk hipotezi, referans ve odak grubun satır ortalama puanları (row mean scores) arasında ilişki olmadığını ifade eder. Herhangi bir puan aralık düzeyinde satır ortalama puanları arasındaki ilişki incelenen (studied) maddeler için MİF'in varlığını göstermektedir.

Çok kategorili puanlanan maddeler için pek çok MİF belirleme tekniği bulunmasına karşın, pratik test uygulamalarında değişen farklı koşullar için hangi tekniklerin daha tutarlı sonuçlar verdiğinin belirlenmesi önemli hale gelmektedir. Bu konudaki araştırmaların büyük kısmının yapay verilerle yapıldığı dikkat çekmektedir (örn., Ankenmann ve diğerleri,1999; Artar, 2007; Carter, 2011; Fidalgo ve Bartman, 2010; Garett, 2009; Wang ve Su, 2004).

Yapay test verisi üzerinden çalışmak iki açıdan önemlidir. İlk önemli nokta, araştırmacılar yapay veriler yoluyla gerçek madde parametrelerini bilmektedir ve böylece hangi madde ya da maddelerin MİF içerdiğini kontrol edebilmektedir. Gerçek verilerle yapılan MİF belirleme tekniklerinin karşılaştırıldığı durumlarda, gerçek madde parametreleri bilinmediğinden hangi maddenin gerçekten MİF içerip içermediğinin bilinmesi zordur. Böylece değişen test koşulları için hangi MİF belirleme tekniğinin en iyi sonuç verdiğini belirlemek kolay olmayacaktır. Önemli ikinci nokta ise, yapay veri kullanılan çalışmalarda araştırmacılar veri seti ile ilgili örneklem büyüklüğü, örneklemin dağılım özellikleri, madde sayısı vb. diğer değişkenleri araştırma amacına uygun olarak değişimleyebilmektedir. Bu olanaklar araştırmacılara değişen test koşullarına bağlı olarak en iyi sonucu veren MİF belirleme tekniğini bulma konusunda yardımcı olmaktadır. Araştırmada yapay veri kullanımının avantajları ve araştırma amacına uygunluğu dikkate alınarak bu çalışma kapsamında da yapay veriler kullanılmıştır.

Yukarıda belirtilen MİF belirleme teknikleri için uygulamada karşılaşılan büyük sorunlardan bir tanesi, bu tekniklerin örneklem büyüklüğüne olan duyarlılıklarıdır. Bu yüzden MİF belirleme tekniklerinin örneklem büyüklüğüne olan bu duyarlılığını aşmak için farklı yaklaşımlar önerilmiştir. Bunlardan biri, I. Tip Hata ve güç (power) çalışmalarıdır (Ankenmann ve diğerleri, 1999; Zwick ve diğerleri, 1993).

I.Tip Hata, gerçekte doğru olan bir yokluk (null) hipotezinin, örneklem istatistiği kullanılarak yapılan testler sonucunda reddedilmesi ile yapılan hata olarak tanımlanır ve böyle bir hataya düşme olasılığı da α ile gösterilir. MİF belirleme çalışmaları için I.Tip Hata, gerçekte bir maddenin doğru cevaplandırılma olasılığının referans ve odak grup üyeleri için farklı olmadığı yani MİF göstermediği durumda yapılan testler sonucu ilgili maddenin MİF'li

olarak belirlenmesini ifade etmektedir. Gerçekte MİF içermeyen bir madde için 1000 defa yapılan bir tekrarlamada (replikasyon) bu maddenin 48 tekrarda MİF içerdiği kararı verilmişse; yani yokluk hipotezi reddedilmişse, bu test için I.Tip Hata oranı 0.048 olacaktır.

İstatistiksel güç (power), gerçekte yanlış olan bir yokluk hipotezini reddedebilme olasılığını göstermektedir. MİF analizleri için güç çalışmalarında, gerçekte bir maddenin doğru cevaplandırılma olasılığının referans ve odak grup üyeleri için farklı olduğu yani MİF içerdiği durumda, o maddenin ilgili MİF belirleme testleri tarafından MİF'li olarak belirlenip belirlenmediği incelenir. Örneğin 1000 tekrarlamanın yapıldığı yapay verinin kullanıldığı bir çalışmada ilgili nominal α düzeyi için MİF'in doğru olarak belirlendiği tekrarlama oranı o MİF belirleme testinin gücünü tanımlamaktadır.

Sonuçlarına dayanılarak eğitim programlarının değerlendirilmesi, yerleştirme, mezuniyet, geçme/kalma vb çok önemli kararların verildiği test uygulamalarının hukuksal olarak savunulabilmesi, bu test puanları üzerinde MİF belirlemeye yönelik çalışmaların yapılmış olmasına bağlıdır (Haladyna, 1997). Test geliştirme sürecinde ve ölçme araçlarından elde edilen sonuçların ne kadar adil ve geçerli olduğu konusunda MİF analizleri önemli hale gelmektedir.

Eğitim ve psikoloji alanında yapılan ölçmelerde iki kategorili tepki gerektiren maddelerden çok dereceli ya da kısmı olarak puanlanabilen maddeler daha sıklıkla tercih edilmektedir. Günümüzde geniş ölçekli test uygulamaları kapsamında bilişsel performans ve bu performansı etkilediği düşünülen algı, tutum vb. duyuşsal değişkenlerin ölçülmesinde çok kategorili puanlama gerektiren maddeler yaygın olarak kullanılmaktadır. Kullanım sıklığındaki artışa bağlı olarak çok kategorili puanlanan maddeler için sağlam geçerlik ve güvenirlik kanıtlarının sunulması çoktan seçmeli maddelerle yapılan değerlendirmeler kadar önemlidir. Özellikle, bilişsel performansın ölçülmesi kapsamında kullanılan performansa dayalı durum belirlemenin (performance based assesment) çoktan seçmeli maddelerle yapılan değerlendirmeye göre yanlılık problemine daya yatkın olduğu iddia edilmektedir (Zwick ve diğerleri, 1993). Bilindiği gibi performansa dayalı durum

belirleme kapsamında öğrencilerin eleştirel düşünme, karar verme, problem çözme vb. üst düzey zihinsel becerileri yoklanmaktadır. Bu tür becerilere sahip olmak aynı zamanda birden çok beceriyi kullanmayı gerektirebilir. Bu durumda ölçülmesi amaçlanmayan yapılar yanlılık kaynağı olarak ortaya çıkabilir. Bu kapsamdaki ölçmelerde sıklıkla kullanılan çok kategorili puanlanan maddelerde MİF'in belirlenmesi test puanlarının geçerliği bakımından oldukça önemli hale gelmektedir. Çünkü MİF göstermeyen testler tüm katılımcılar için tarafsız ve denk olacaktır (Wyse ve Mapuranga, 2009).

Bilişsel ve duyuşsal özelliklerin belirlenmesiyle ilgili güncel test uygulamaları incelendiğinde MİF analizlerinin önemi daha iyi anlaşılmaktadır. Örneğin De Leo, Van Dam, Hobkirk ve Earleywine (2010), sağlık, psikoloji ve iletişim gibi çok farklı alanlarda bireylerin dürtüsel heyecan arama (impulsive sensation seeking) durumlarını belirlemek üzere sıklıkla kullanılan 19 maddelik Likert tipi ölçekle ilgili yaş, ırk ve eğitim düzeyi bakımından MİF analizleri yapmıştır. Araştırmacılar bu çalışmayla ilgili yapının ölçülmesindeki olası sınırlılıkların belirlenmesi ve yaş, ırk ve eğitim düzeyi bakımından görülen farklılıkların ölçme hatalarından mı yoksa ölçülen özellikten mi kaynaklandığını belirlemeye çalışmıştır.

French, Hand, Therrien ve Vazquez (2012) uygulanması, puanlanması ve yorumlanması kolay olan Cornell Kritik Düşünme Beceri Testi'ni cinsiyet bakımından incelemiştir. Bu aracın kullanıldığı pek çok araştırmada özellikle cinsiyet bakımından çok farklı sonuçlar elde edilmiştir. Araştırmacılar, cinsiyet bakımından bireyler arasında herhangi bir fark olup olmadığını değerlendirmeden önce bu aracın özelliklerinin iyi değerlendirilmesi gerektiğini vurgulamış ve bu aracın kız ve erkekler için aynı yapıyı ölçtüğüne dair kanıt bulmak ve gruplar arası ortalama puanları karşılaştırmayı daha anlamlı hale getirmek amacıyla MİF analizleri yapmıştır. MİF grup ortalamalarını karşılaştıran istatistikleri olumsuz etkileyebilmektedir (Li ve Zumbo, 2009). Bu yüzden grupları karşılaştırmadan önce MİF'in belirlenmesi daha doğru sonuçlara ulaşmak için önemli hale gelmektedir

Wetzel ve Hell (2013), araştırmalarda mesleki ilgiler bakımından kadınlar ve erkekler arasında her zaman büyük farklılıklar bulunduğuna dikkat çekmektedir. Araştırmacılar, Bergmann ve Eder (2005) tarafından geliştirilmiş

Likert tipinde 160 maddeden oluşan Genel Yetenek Testi'i (Allgemeiner Interessen- Struktur-Test: AIST-R) için cinsiyet bakımından gözlenen farkların ölçülen özellikle ilgili geçerli ve güvenilir farkları yansıtıp yansıtmadığını incelemek üzere MİF analizleri yapmıştır.

Wetzel ve diğerleri (2013), kişilik ölçümlerinde kullanılan Costa ve McCrae, (1992) tarafından geliştirilmiş NEO Kişilik Envanteri (NEO Personality Inventory-NEO-PI-R) için cinsiyet bakımından MİF analizleri yapmıştır. Kişilik ölçümleri, bireylerin kişilik özelliklerinin iş performansının önemli bir yordayıcısı olması bakımından personel seçiminde sıklıkla kullanılmaktadır (Barrikc ve Mounth, 1991). Bu özelliklerin ölçümünde genellikle kadın-erkek gibi farklı gruptan bireyler karşılaştırılmaktadır. Örneğin, bu kişilik envanterinin kullanıldığı çalışmalarda özsaygı ve kendine güven boyutlarında erkeklerin kadınlara göre daha yüksek puanlar aldığı belirlenmiştir (Feingold, 1994). Ancak yapılan bu karşılaştırmaların savunulabilir ve yasal olması için kullanılan ölçme araçlarının testi alan bireyler için denk olması önemlidir. Aksi durumda MİF içeren ölçme sonuçlarıyla karşılaştırmalar yapmak anlamlı olmayacaktır.

Örnek test uygulamalarından da görüldüğü gibi farklı psikolojik yapıların ölçülmesinde artık daha sıklıkla kullanılmaya başlanan çok kategorili puanlanan maddeler için de MİF analizlerinin yapılması gerekmektedir. Çok kategorili puanlanan maddeler için pek çok MİF belirleme tekniği bulunmasına karşın, pratik test uygulamalarında değişen farklı koşullar için hangi tekniklerin daha tutarlı sonuçlar verdiğinin belirlenmesi önemli hale gelmektedir. Çok kategorili MİF belirleme tekniklerinin değişen koşullarda ne kadar tutarlı sonuçlar verdiğinin bilinmesi, test uygulayıcıları/ geliştiricileri için test puanlarını daha geçerli hale getirecektir. Bilindiği gibi MİF analiz süreci sonucunda yanlılık belirlenen maddelerin asıl test formundan çıkarılarak ilgili test üzerinde yapılacak bundan sonraki herhangi bir istatistiksel işlemlerde (örneğin test puanlarının eşitlenmesi) kullanılmaması önerilmektedir. Bu durumda, seçilen MİF belirleme tekniklerinin gerçekten yanlı olabilecek maddeleri belirlemesi gerekmektedir. Seçilen MİF belirleme tekniğinin hatalı olması, gerçekte MİF içermeyen maddelerin yanlı olduğu sonucuna ulaşılmasına neden olabilir. Buna bağlı olarak bu madde ya da maddelerin

testten çıkarılması hem testin yapı ve kapsam geçerliğini olumsuz etkileyecek hem de test puanları üzerinde yapılacak diğer işlemlerin hatalı olmasına neden olacaktır.

Çalışmalarda veri setleriyle ilgili olarak örneklem büyüklüğü, grupların yetenek ölçülerinin dağılımı, MİF örüntüsü ve miktarı gibi değişkenlerin MİF belirleme tekniklerinin performansı üzerindeki etkisi incelenen değişkenler olarak öne çıkmaktadır. Bu durumda gerçek test uygulamalarında sıklıkla tercih edilen MTK-OOT ve Mantel Test için değişen test koşullarında I. Tip hata ve istatistiksel güç bağlamında performanslarını karşılaştırarak bu testlerin hangi test koşullarında daha iyi performans gösterdiği veya hangi koşullara daha duyarlı olduğunun belirlenmesi gerekmektedir

## Amaç

Bu çalışmanın genel amacı, çok kategorili tepki gerektiren maddelerde MİF belirleme testlerinden Mantel Test ve MTK-Olabilirlik Oran Testi'nin farklı test koşullarında performanslarının karşılaştırılmaktır. Bu amaçla çalışma, bu testlere ilişkin karşılaştırma ölçütü olarak I.Tip Hata ve istatistiksel güç (power) oranlarının esas alınarak karşılaştırıldığı iki bölüm altında yürütülmüştür.

İlk bölümde, I.Tip hata çalışmaları kapsamında aşağıdaki sorulara yanıt aranmıştır:

1) Mantel Test ve MTK-OOT için,

   1.1 Referans ve odak gruba ait yetenek dağılımlarının birim normal dağılım özelliği gösterdiği koşulda, grupların örneklem büyüklüklerindeki değişime bağlı olarak I.Tip Hata oranları nasıl değişmektedir?

   1.2 Referans ve odak gruba ait yetenek dağılımlarının farklılaştığı koşulda örneklem büyüklüklerindeki değişime bağlı olarak I.Tip Hata oranları nasıl değişmektedir?

Referans gruba ait yetenek dağılımın birim normal ve odak gruba ait yetenek dağılım ortalamasının

a) -0.5

b) -1

olduğu koşulda grupların örneklem büyüklerindeki değişime bağlı olarak I. Tip Hata oranları nasıl değişmektedir?

İkinci bölümde, istatistiksel güç çalışmaları kapsamında aşağıdaki sorulara yanıt aranmıştır:

2) Mantel Test ve MTK-OOT için,

2.1 Referans ve odak gruba ait yetenek dağılımlarının birim normal dağılım özelliği gösterdiği koşulda,

a) düşük MİF örüntüsü

b) yüksek MİF örüntüsü için

değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak güç oranları nasıl değişmektedir?

2.2 Referans gruba ait yetenek dağılımının birim normal dağılım ve odak gruba ait yetenek dağılım ortalamasının -0.5 olduğu koşulda,

a) düşük MİF örüntüsü

b) yüksek MİF örüntüsü için,

değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak güç oranları nasıl değişmektedir?

2.3 Referans gruba ait yetenek dağılımının birim normal dağılım ve odak gruba ait yetenek dağılım ortalamasının -1 olduğu koşulda

a) düşük MİF örüntüsü

b) yüksek MİF örüntüsü için,

değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak güç oranları nasıl değişmektedir?

## Önem

Test geliştirme sürecinde ve ölçme araçlarından elde edilen sonuçların ne kadar adil ve geçerli olduğu konusunda MİF analizleri önemli hale gelmektedir. Bu bakımdan farklı psikolojik yapıların ölçülmesinde artık daha sıklıkla kullanılmaya başlanan çok kategorili puanlanan maddeler için de MİF analizlerinin yapılması gerekmektedir.

MİF analiz tekniklerinin sonuçları çok farklı test koşullarından etkilenebilir. Bu araştırmada gerçek test koşullarına uygun olarak grupların örneklem büyüklüğü, grupların yetenek dağılım ölçüleri, MİF miktarı ve MİF örüntüsü değişkenleri ele alınmış ve bu değişkenlere göre güncel test uygulamalarında sıklıkla tercih edilen Mantel Test ve MTK-OOT'nin performansının nasıl değiştiği incelenmiştir. Bu bakımdan bu araştırma sonuçlarının farklı test koşulları için ilgili MİF belirleme testlerinin zayıf veya güçlü yönlerinin belirlenmesine olanak sağlaması ve test uygulayıcıları için çok kategorili puanlanan maddelerin yer aldığı ölçme araçlarından elde edilecek bilgilerin geçerliğine katkıda bulunması beklenmektedir. Ayrıca bu araştırma bulguları doğru MİF belirleme testlerinin seçilmesi konusunda araştırmacılara ve uygulayıcılara önemli bilgiler sunacak ve test puanları üzerinde yapacakları işlemlerin daha geçerli olmasına katkı sağlayacaktır. Bunlara ek olarak, bu araştırma sonuçlarına dayanarak ilgili MİF belirleme testlerin ortak olan zayıf yanları belirleyerek olası daha yeni MİF belirleme tekniklerinin geliştirilmesine olanak sağlaması beklenmektedir. Bu bakımdan bu araştırmadan elde edilecek sonuçların ilgili alan yazına katkı getirmesi beklenmektedir.

## Sınırlılıklar

Bu araştırmadan elde edilen sonuçlar, araştırmanın amacı doğrultusunda ele alınan değişkenler ve oluşturulan simülasyon koşulları ile sınırlıdır.

## Tanımlar

**Düşük MİF örüntüsü (D-MİF):** Referans ve odak grubun, birinci ve ikinci puan kategorisi arasındaki adım güçlüğü parametresinde ($\delta 1$) ortaya çıkan MİF.

**Yüksek MİF örüntüsü (Y-MİF):** Referans ve odak grubun, üçüncü ve dördüncü puan kategorisi arasındaki adım güçlüğü parametresinde ($\delta 3$) ortaya çıkan MİF.

**MİF miktarı:** Referans ve odak grubun adım güçlüğü parametreleri arasındaki farkın lojit birim ifadesi. Bu lojit birimler 0.42 ve 0.64'tür.

# BÖLÜM 2

# İLGİLİ ARAŞTIRMALAR

Bu bölümde literatürde MİF belirleme tekniklerinin performansını etkileyen değişkenlerle ilgili var olan durumu belirlemek amacıyla ilgili araştırmaların bulguları bu çalışmada kullanılan temel değişkenler kapsamında özetlenmiştir.

## Örneklem Büyüklüğü ve Grupların Yetenek Dağılımı

İlgili araştırmalarda sıklıkla örneklem büyüklüğü ve referans ve odak grubun yetenek dağılımlarıyla ilgili koşullar MİF belirleme tekniklerinin performansına etkisi incelenmiştir. Bu koşullar her iki grubun yetenek dağılımlarının birim normal dağılım göstermesi ve odak grubun yetenek dağılım ortalamasının farklı miktarlarda normal dağılımdan sapması biçiminde olmaktadır. Yapılan çalışmalarda odak grubun yetenek dağılım ortalamasının normalden sapma değerleri -0.25, -0.50, -1 ve -1.5 biçiminde farklılaşmaktadır (Örn. Bolt, 2002; Garrett, 2009; Wood, 2011). Örneklem büyüklüğü ve grupların yetenek dağılımındaki farklılaşmaya bağlı olarak MİF belirleme testlerinin performansını karşılaştıran ilgili araştırmalar aşağıda özetlenmektedir.

Kim ve diğerleri (1998) Kısmı Puan Modeli'ne uygun olarak ürettikleri dört kategoride puanlanan 30 maddelik bir test formu kullanarak sırasıyla referans ve odak grup için üç farklı örneklem büyüklüğü (300:300, 1000:300 ve 1000:1000) ve iki farklı grup yetenek dağılımı (R~N(0,1), O~N(0,1 ve

R~N(0,1), O~N(1,1)) koşulunda MTK-OOT'nin I. Tip hata oranlarını incelemiştir. Araştırma sonuçları tüm örneklem ve yetenek dağılımı koşullarında MTK-OOT için I. Tip hatayı kontrol etmede yeterli olduğunu göstermiştir. Odak grup yetenek dağımı birim normal dağılımdan saptıkça, I. Tip hata oranları bir miktar yükselmiştir. Artan örneklem büyüklüğüne bağlı olarak I. Tip hata oranları azalmıştır.

Ankenmann ve diğerleri (1999) tarafından yapılan çalışmada Samejima'nın Dereceli Tepki Modeli'ne uygun olarak üretilen beş kategoride puanlanan altı madde için üç farklı örneklem büyüklüğü koşulu (2000:2000, 2000:500 ve 500:500) ve iki farklı grup yetenek dağılımı (R~N(0,1), O~N(0,1 ve R~N(0,1), O~N(1,1)) koşulunda MTK-OOT ve Mantel Test'in I.Tip hata oranları ve güç değerleri incelenmiştir. Her iki teste ilişkin güç istatistikleri beklendiği gibi örneklem büyüklüğünden etkilenmiştir. MTK-OOT'i referans ve odak grup büyüklüğünün küçük olduğu (500) durumlarda MİF belirleme konusunda yetersiz kalmıştır. Artan örneklem büyüklüğüne bağlı olarak testlerin I. Tip hata oranları azalma eğilimi göstermiştir. Grupların yetenek dağılımındaki farklılaşma ve artan örneklem büyüklüğü testlerin I. Tip Hata oranlarını yükseltmiştir. I. Tip hata oranlarındaki artış Mantel Test için daha fazla olmuştur.

Bolt (2002), Dereceli Tepki Modelini esas aldığı çalışmasında parametrik MİF belirleme tekniği olarak MTK-OOT ve Test ve Madde-Test işlev Farklılık Fonksiyonu (DFIT-Dıfferential Functioning of Items and Tests); parametrik olamayan teknik olarak ise Poly-SIBTEST'i seçmiştir. İki farklı örneklem büyüklüğü koşulu (300:300 ve 1000:1000) ve iki farklı grup yetenek dağılımı (R~N(0,1), O~N(0,1 ve R~N(0,1), O~N(1,1)) koşulunda MTK-OOT'i ve Mantel Test'in I.Tip hata oranları ve güç değerleri incelenmiştir. Araştırma sonuçları OOT ve Poly- SIBTEST için grup yetenek dağılımlarının benzer olduğu koşulda artan örneklem büyüklüğüne bağlı olarak I. Tip hata oranlarının yükseldiğini göstermiştir. Grup yetenek dağılımlarının farklılaştığı koşulda ise I. Tip hata oranları yükselmiş ve artan örneklem büyüklüğüne göre azalmıştır. DFIT ise tüm test koşullarında birbirine yakın I. Tip hata değerleri vermiştir. Güç sonuçları ise her üç testin güç değerlerinin artan örneklem

büyüklüğüne bağlı olarak yükseldiğini; ancak grupların yetenek dağılımındaki farklılaşmadan etkilenmediğini göstermiştir.

Wang ve Su (2004), Kısmi Puan modeli ve genelleştirilmiş Kısmı Puan Modeli'ne uygun yapay veri kullanarak yaptıkları çalışmada, 500:500 örneklem büyüklüğü ve üç farklı grup yetenek dağılımı (R~*N*(0,1), O~*N*(0,1), R~*N*(0,1), O~*N*(-0.5, 1)) ve R~*N*(0,1), O~*N*(-1.5,1)) koşulunda Mantel Test ve genelleştirilmiş Mantel-Haenszel Test'in (GMH) I. Tip hata ve güç değerlerini farklı test koşullarında karşılaştırmışlardır. Araştırma sonuçları, farklı test koşulları için grup yetenek dağılımlarındaki farklılaşmaya bağlı olarak testlerin I. Tip hata oranlarının yükseldiğini göstermiştir. Güç sonuçları, grupların yetenek dağılımındaki farklılaşmanın testlerin güç değerleri üzerinde az bir etkiye sahip olduğunu göstermiştir. Referans ve odak gruba ait yetenek dağılımları arasındaki farkın arttığı durumlarda testler I.Tip hatanın kontrol edilmesi bakımından iyi sonuçlar vermemiştir. Her üç teknik de farklı test uzunluklarında I.Tip hata ve güç oranları bakımından benzer sonuçlar vermiştir.

Kristjansson ve diğerleri (2005) yaptıkları çalışmada, Genelleştirilmiş Kısmı Puan Modeli'ne uygun olarak dört kategoriden oluşan 26 maddelik bir test verisi kullanarak İki farklı örneklem büyüklüğü koşulu (2000:2000 ve 3200:800) ve iki farklı grup yetenek dağılımı (R~N(0,1), O~N(0,1 ve R~N(0,1), O~N(-0.5,1)) koşulunda MİF belirleme tekniği olarak Mantel, GMH, Lojistik diskriminat fonksiyon analizi (LDFA- logistic discriminant function analysis) ve serbest lojit birikimli sıralı lojistik regresyonu (UCLOLR- unconstrained cumulative logits ordinal logistic regression) karşılaştırmıştır. Araştırma sonuçları grupların yetenek dağılımının birim normal dağım özelliği gösterdiği koşulda tüm testlerin I. Tip hatayı kontol etmede yeterli olduğunu göstermiştir. Farklılaşan yetenek dağılım koşullarında LDFA hariç testlerin I. Tip hata oranları yükselmiştir; bu artış Mantel Test için daha fazla olmuştur. Tek biçimli MİF'i belirleme konusunda her iki yetenek dağılımı ve örneklem büyüklük oranı koşulunda testler birbirine yakın ve kabul edilebilir I. Tip hata değerleri vermiştir.

Su ve Wang (2005) Kısmi Puan ve Dereceli Tepki Modeline uygun 20'si iki kategoride ve altısı çok kategoride puanlanan maddelerden oluşan bir testi kullanarak MİF belirleme testlerinden Mantel Test, GHM ve lojistik diskriminant fonksiyon analizini, 500:500 örneklem büyüklüğü ve üç farklı grup yetenek dağılımı (R~N(0,1), O~N(0,1), R~N(0,1), O~N(-0.5, 1)) ve R~N(0,1), O~N(-1.5,1)) koşulunda veri kullanarak karşılaştırmıştır. Araştırma sonuçları gruplara ilişkin tüm yetenek dağılım koşullarında tüm testlerin I. Tip Hatayı kontrol etmede yeterli olduğunu göstermiştir. R~*N*(0,1), O~*N*(-0.5, 1)) ve R~*N*(0,1) dağılım koşulları için, grupların yetenek dağılımlarının birim normal dağılım özelliği gösterdiği koşullara benzer I. Tip hata oranları elde edilmiştir. R~*N*(0,1), O~*N*(-1.5,1) koşulu için testlerin I. Tip hata değerleri bir miktar yükselmiştir; ancak kabul edilebilir seviyede kalmıştır.

Stark, Chernyshenko ve Drasgow (2006) yaptıkları çalışmada MTK-OOT ve doğrulayıcı faktör analizine dayalı yöntemler (mean and covariance structures method [MACS]) için 15 maddelik veri setinde iki farklı örneklem koşulu (500:500 ve 1000:1000) ve iki farklı grup yetenek dağılımı (R~*N*(0,1), O~*N*(0,1) ve R~*N*(0,1), O~*N*(-0.50, 1)) koşulunda incelemiştir. Araştırma sonuçları, OOT ve MACS'in I. Tip hata kontrolü bakımından benzer sonuçlar verdiğini ve artan örneklem büyüklüğüne bağlı olarak I. Tip hata değerlerinin yükseldiğini göstermiştir. Grupların yetenek dağılımındaki farklılaşma testlerin MİF belirlemedeki performansı üzerinde küçük bir etkiye sahip olmuştur.

Artar (2007) tarafından yapılan çalışmada grupların yetenek dağılımlarının birim normal dağılım özelliği gösterdiği farklı örneklem büyüklüğü (600, 1200 ve 2400) koşulları için MTK-OOT ve lojistik regresyonun I. Tip hata ve güç değerleri incelenmiştir. Araştırmada dört kategoride puanlanan altı madde için Dereceli Tepki Modeline uygun parametre kestirimleri yapılmıştır. Araştırma sonuçları, artan örneklem büyüklüğüne bağlı olarak testlerin I. Tip hata oranlarının yükseldiğini göstermiştir. I. Tip hata oranlarındaki yükseliş MTK-OOT için daha fazla olmuştur. Artan örneklem büyüklüğü testlerin MİF belirlemedeki istatistiksel gücünü artırmıştır.

Fortman-Johnson (2007), Dereceli Tepki Modeli'ne uygun 40 maddelik bir test verisi için iki farklı örneklem koşulu (500:500 ve 1000:1000) ve iki farklı grup yetenek dağılımı (R~*N*(0,1), O~*N*(0,1) ve R~*N*(0,1), O~*N*(-1, 1)) koşulunda MTK-OOT ve parametre tekrarlama (item replication) tekniklerinin I. Tip hata ve güç oranlarını incelemiştir. Teknikler arasındaki fark tek biçimli olmayan MİF için gözlenmiştir. Bu koşulda MTK-OOT küçük miktarlardaki MİF'i belirleme konusunda kabul edilebilir istatistiksel güç oranları vermiştir. Örneklem büyüklüğü ve odak grubun yetenek dağılımı MİF belirme bakımından teknikler arasında çok az etkiler oluşturmuştur.

Garrett (2009), Kısmı Puan Modeline uygun olarak ürettiği 20 maddelik veride, farklı örneklem büyüklüklerinde (700:300, 900:100, 500:500,845:355 ve 1183:317) ve dört farklı grup yetenek dağılımı (R~*N*(0,1), O~*N*(0,1), R~*N*(0,1), O~*N*(-0.25, 1) ve R~*N*(0,1), O~*N*(-0.50, 1) ve R~*N*(0,1), O~*N*(-0.75,1)) koşulunda Mantel Test ve Lojistik Resresyon için I. Tip hata ve güç oranlarını incelemiştir. Araştırma sonuçlarına göre, benzer yetenek dağılımı koşullarında her iki test de I. Tip hatayı kontrol etmede iyi performans göstermiş; Mantel Test için artan örneklem büyüklüğüne bağlı olarak I. Tip hata oranlarında artma eğilimi ortaya çıkarken, Lojistik regresyon için benzer I. Tip hata değerleri elde edilmiştir. Grup yetenek dağılımlarındaki farklılaşmaya bağlı olarak testlerin I. Tip hata oranları bir miktar yükselmiştir. Ancak bu koşulda artan örneklem büyüklüğü testlerin I. Tip hata oranlarını düşürmüştür. Artan örneklem büyüklüğüne bağlı olarak testlerin güç değerleri yükselmiştir.

Fidalgo ve Bartram (2010) yaptıkları çalışmada Wang ve Su (2004) tarafından da kullanılan madde parametrelerini kullanarak eş örneklem büyüklüğünde 1000 (500:500) Mantel-Haenszel test (GMH) ve Mantel Test'e ilişkin I. Tip hata ve istatistiksel güç oranlarını farklı test koşullarında incelemiştir. Araştırma sonuçlarına göre her iki teknik de grup yetenek dağılımlarının benzer olduğu koşullarda birbirine yakın ve 0.05 civarında I. Tip hata değerleri vermiştir. Grupların yetenek dağılımı farklılaştıkça tekniklerin I. Tip hata değerleri de artış göstermiştir. İstatistiksel güç sonuçlarına GMH'nin güç oranları grupların yetenek dağılımındaki farklılaşmadan etkilenmezken, Mantel Test için güç oranları bir miktar yükselmiştir.

Carter (2011), tarafından yapılan çalışmada grupların yetenek dağılımlarının birim normal dağılım özelliği gösterdiği farklı örneklem büyüklüğü (600 ve1000) koşulları için dereceli açımlamalı modeli (Generalized Graded Unfolding Model) kullanarak altı kategoride puanlanan 20 maddelik bir veri setinde MTK_OOT ve DFIT (differential functioning of items and tests) incelemişlerdir. Araştırma bulguları, her iki test için de artan örneklem büyüklüğüne bağlı olarak güç oranlarının yükseldiğini ve grup büyüklüklerinin denk olduğu koşullarda ilgili testlerin daha yüksek istatistiksel güç oranları verdiğini ortaya koymuştur. OOT için I. Tip hata oranları farklı koşullara göre fazla değişiklik göstermemiş ve tüm koşullarda 0.08'in altında kalmıştır.

Wood (2011a), çalışmasında Dereceli Tepki Modeli'ne uygun olarak ürettiği 16 maddelik veride, iki farklı örneklem koşulu (400:400 ve 400:40) ve iki farklı grup yetenek dağılımı (R~*N*(0,1), O~*N*(0,1) ve R~*N*(0,1), O~*N*(-0.5, 1)) koşulunda Mantel Test, Liu-Agresti istatistiği ve HW3'ün performansını incelemiştir. Araştırma sonuçları, oluşturulan koşuların birçoğunda Mantel Test ve Liu-Agresti istatistiğine ilişkin I. Tip hata değerlerinin 0.05 civarında olduğunu göstermiştir. HW3 için çok yüksek I. Tip hata değerleri elde edilmiştir. Artan örneklem büyüklüğüne bağlı olarak testlerin I. Tip hata oranları yükselmiştir. Araştırmacı büyük örneklem (400:40) ve gruplar arası yetenek dağılımlarının farklılaştığı koşullar için daha yüksek istatistiksel güç değerleri elde etmiştir.

Wang ve diğerleri (2013) çalışmalarında dereceli açımlamalı modeli (Generalized Graded Unfolding Model) kullanarak beş kategoride puanlanan 10 ve 20 maddelik bir veri setleri oluşturmuştur. Araştırmacılar, MİF belirlemede hipotez testlerini (MTK-OOT, Akaike bilgi ölçütü ( Akaike information criterion) [AIC] ve Lord ki-kare) ve etki büyüklüğü ölçüsünü iki farklı yaklaşımda (daraltılmış ve serbest bırakılmış yöntem) karşılaştırmıştır. Değişen odak grup yetenek dağılım ortalaması (0, -0.25 ve -0.50 SD) ve örneklem büyüklüğü (250:250, 500:500 ve 1000:1000) koşullarında hipotez testleri ve etki büyüklükleri daraltılmış ve serbest bırakılmış yöntemler bağlamında incelenmiş ve her bir koşul için 100 tekrar yapılmıştır. Araştırma sonuçları odak grubun yetenek dağılımındaki farklılaşmanın özellikle 0.25

standart sapmalık farktan büyük olduğu durumlarda hipotez testleri etkilediğini ancak etki büyüklüğü ölçülerini etkilemediğini göstermiştir. Özellikle odak grup yetenek dağılımı bakımından farklılaşmanın olduğu koşullarda, serbest bırakılmış yöntemlerin (free baseline method) I. Tip hata ve istatistiksel güç bakımından kabul edilebilir sonuçlar elde edilmiştir. AIC de gruplar arası yetenek farklılaşmasının olmadığı ya da çok az olduğu koşullarda MİF belirleme konusunda MTK-OOT ile benzer performans göstermiştir. Ancak grupların yetenek dağılımları arasındaki farklılık büyüdükçe hiçbir hipotez testi MİF belirlemede kabul edilebilir güç oranları vermemiştir. Örneklem büyüklüğü MİF belirleme tekniklerinin performansında önemli etkiye sahip olmuştur.

Yukarıda özetlenen araştırma bulguları genel olarak örneklem büyüklüğü ve grupların yetenek dağılım ortalamalarındaki sapmaların MİF belirleme tekniklerinin performansı üzerinde önemli bir etkiye sahip olduğunu göstermektedir. Artan örneklem büyüklüğü MİF belirleme tekniklerinin performansını olumlu yönde etkilerken, odak grubun yetenek dağılımındaki sapmalar olumsuz yönde etkilediğini göstermektedir. Aratan örneklem büyüklüğü ve sapma miktarları MİF belirleme tekniklerinin I. Tip hata oranlarını yükseltmektedir. İstatistiksel güç sonuçları ise her üç testin güç değerlerinin artan örneklem büyüklüğüne bağlı olarak yükselmekte olduğunu; ancak grupların yetenek dağılımındaki farklılaşmadan etkilenmediğini göstermektedir.

İlgili araştırmalarda MİF belirleme tekniklerinin performansı üzerindeki etkisi sıklıkla incelenen diğer değişkenler MİF miktarı ve MİF örüntüsüdür. MİF miktarı ve MİF örüntüsündeki farklılaşmaya bağlı olarak MİF belirleme testlerinin performansını karşılaştıran ilgili araştırmalar aşağıda özetlenmektedir.

**MİF Miktarı ve MİF örüntüsü**

İlgili araştırmalarda referans ve odak grubun adım güçlüğü parametreleri arasında fark oluşturmak üzere değişik miktarlarda lojit birimlerin kullanıldığı görülmektedir. Araştırmalarda ağırlıklı olarak sabit MİF örüntüsü

ele alınmakla birlikte yakın tarihli araştırmalarda farklı MİF örüntüsü kombinasyonları çalışılmıştır.

Ankenmann ve diğerleri (1999), sabit ve dengeli MİF örüntüsünde ve 0.25 lojit birim MİF miktarında Mantel Test ve MTK-OOT'nin istatistiksel güç oranlarını incelemiştir. Araştırma sonuçları sabit MİF örüntüsü için özellikle küçük örneklemlerde Mantel Test'in MTK-OOT'ye göre daha yüksek istatistiksel güç değerleri verdiğini göstermektedir. Dengeli MİF örüntüsünde ise tüm örneklem koşullarında MTK-OOT, Mantel Test'e göre MİF'i belirlemede daha yüksek güç oranları vermiştir.

Blot (2002) çalışmasında sabit MİF örüntüsünde ve 0.50 MİF miktarında MTK-OOT, Poly-SIBTEST ve Madde-Test İşlev Farklılığı'nın (DFIT) istatistiksel gücünü farklı örneklem büyüklüğü ve model-veri uyumunu bozulduğu durumlar için incelemiştir. Araştırma sonuçları, MTK-OOT için ilgili test koşullarında yeterli güç istatistikleri elde edildiğini göstermiştir. Ayrıca Poly-SIBTEST tüm koşullarda en düşük istatistiksel güç değerleri vermiş ve DFIT model veri uyumsuzluğundan en az etkilenen MİF belirleme testi olmuştur.

Wang ve Su (2004), sabit ve dengeli MİF örüntüsünde 0.10 ve 0.25 lojit birim MİF miktarları için Mantel Test ve Genelleştirilmiş Mantel Test'in (GMT) istatistiksel güç değerlerini incelmiştir. Araştırma sonuçları, artan MİF miktarına bağlı olarak ilgili MİF belirleme testlerinin istatistiksel gücünün yükseldiğini göstermiştir. Sabit MİF örüntüsünde Mantel Test GMH'e göre daha iyi güç değerleri vermiştir. Dengeli MİF örüntüsünde ise Mantel Test GMH'e göre daha düşük güç değerleri vermiştir.

Kristanjansonn ve diğerleri (2005) sabit MİF örüntüsünde 0.25 lojit birim MİF miktarları için Mantel Test, Genelleştirilmiş Mantel Test (GMT), Lojistik Diskriminant Fonksiyon Analizi'nin (LDFA) ve Serbest Birikimli Sıralı Lojistik Regresyon (UCLOLR- unconstrained cumulative logits ordinal logistic regression) için istatistiksel güç değerlerini incelmiştir. Tek biçimli MİF için tüm test koşullarında ilgili MİF belirleme testleri çok yüksek güç değerleri vermiştir. Mantel Test ve LDFA tek biçimli olmayan MİF'i belirlemede zayıf kalmıştır.

GMT ve UCLOLR ise tüm test koşullarında MİF'i belirlemede yüksek performans göstermiştir.

Su ve Wang (2005), sabit, dengeli, düşük ve yüksek MİF örüntüsünde 0.10, 0.25 ve 0.40 lojit birim MİF miktarları için Mantel Test, Genelleştirilmiş Mantel Test (GMT) ve Lojistik Diskriminant Fonksiyon Analizi'nin (LDFA) istatistiksel güç değerlerini incelmiştir. Araştırma sonuçları, artan MİF miktarına bağlı olarak ilgili MİF belirleme testlerinin istatistiksel gücünün yükseldiğini göstermiştir. Sabit MİF örüntüsünde Mantel Test ve LDFA birbirine benzer ve GMH'e göre daha iyi güç değerleri vermiştir. Dengeli MİF örüntüsünde Mantel Test ve LDFA birbirine benzer ancak GMH'e göre daha düşük güç değerleri vermiştir. Yüksek ve düşük MİF örüntüsünde her üç test de birbirine yakın güç değerleri vermiştir. Mantel Test sabit MİF örüntüsünde en yüksek güç değerlerini vermiştir.

Stark ve diğerleri (2006), sabit MİF örüntüsünde 0.25 ve 0.50 lojit birim MİF miktarları için MTK-OOT ve doğrulayıcı faktör analizine dayalı (mean and covariance structures method [MACS])) istatistiksel güç değerlerini incelmiştir. Araştırmacılar aynı zamanda OOT kapsamında daraltılmış modellere (constrained-baseline model) ek olarak serbest model (free baseline model) ve Bonferroni düzeltmesi içeren kritik p değerlerinin kullanılmasını değerlendirmiştir. Araştırma sonuçları OOT kapsamında serbest modellerin alternatif modellere (constrained-baseline model) göre daha etkili sonuçlar verdiğini göstermiştir. Test koşullarının büyük kısmında MACS ve OOT benzer performans göstermiştir. Artan MİF miktarına bağlı olarak testlerin MİF belirlemedeki gücü de artmıştır.

Artar (2007), düşük, yüksek ve dengeli MİF örüntüsünde 0.32, 0.43 ve 0.53 lojit birim MİF miktarları için MTK-OOT ve lojistik regresyon için istatistiksel güç değerlerini incelmiştir. Araştırma sonuçları her iki MİF belirleme testi için artan örneklem büyüklüğü ve MİF miktarına bağlı olarak testlerin istatistiksel güç değerlerinin yükseldiğini göstermiştir. Tüm MİF örüntüsü koşullarında MTK-OOT, lojistik regresyona göre daha yüksek güç değerleri vermiştir. MTK-OOT için en yüksek güç değerleri büyük örneklem ve büyük MİF miktarı koşullarında elde edilmiştir. Özellikle dengeli MİF örüntüsü

koşullarında lojistik regresyon için elde edilen güç değerleri kabul edilir seviyede olmamıştır.

Fortmann-Johnson (2007),sabit MİF örüntüsünde 0.50 ve 1 lojit birim MİF miktarları için MTK-OOT ve çok kategorili puanlanan maddeler için kesim noktası belirlemek üzere parametre tekrarlama (item replication) tekniklerinin etkililiğini test etmiştir. MTK-OOT tüm test koşullarında yüksek istatistiksel güç değerleri vermiştir. Teknikler arasındaki fark tek biçimli olmayan MİF için gözlenmiştir. Bu koşulda MTK-OOT için küçük miktarlardaki MİF'i belirleme konusunda parametre tekrarlama tekniklerine göre daha yüksek güç oranları elde edilmiştir. .

Garrett (2009), sabit MİF örüntüsünde 0.25, 0.50 ve 0.75 lojit birim MİF miktarları için Mantel Test ve lojistik regresyon için istatistiksel güç değerlerini incelmiştir. Araştırma sonuçları, veri setlerinde kayıp değerler olmadığında her iki MİF belirleme testinin istatistiksel güç değerlerinin örneklem büyüklüğü ve MİF miktarına bağlı olarak yükseldiğini göstermiştir. Bu koşulda testlerin istatistiksel güç değerleri birbirine yakın değerler almıştır. Veri setlerinde kayıp değerlerin olduğu koşullarda ise, artan MİF miktarına bağlı olarak testlerin gücü yükselmiştir; ancak artan kayıp veri oranı testlerin istatistiksel gücünü düşürmüştür.

Thurman (2009), MİF belirleme tekniği olarak Mantel Test, GHM ve ordinal Lojistik regresyonu I. Tip hata ve güç istatistikleri bakımından karşılaştırmıştır. Dört kategoride puanlanan 20 maddeye ait yapay verilerin kullanıldığı çalışmada, madde ayırıcılık parametresi, madde eşik parametresi, puan kategorileri arasındaki MİF örüntüsü ve referans – odak grup dağılım özelikleri bakımından farklı koşullar oluşturulmuştur. Araştırma sonuçları, maddelere ait ayırıcılık ve eşik parametresinin MİF belirleme oranlarının yükselmesi ile doğrudan ilişkili olduğunu göstermiştir. Buna ek olarak, madde güçlüğüne bağlı olarak MİF miktarı ve MİF örüntüsü MİF belirleme oranlarını etkilemiştir. MİF belirleme gücü de MİF miktarının artmasına bağlı olarak artmıştır. Araştırma sonunda madde ayırıcılık parametresinin farklılaştığı durumlarda, GHM'nin diğer tekniklere göre daha yüksek güç oranları verdiği için öncelikle tercih edilmesi önerilmiştir.

Fidalgo ve Bartram (2010) yaptıkları çalışmada sabit, düşük, yüksek ve dengeli MİF örüntüsünde 0.25 ve 0.40 MİF miktarları için genelleştirilmiş Mantel-Haenszel test (GMH) ve Mantel Test'e ilişkin istatistiksel güç oranlarını farklı test koşullarında incelemiştir. İstatistiksel güç sonuçlarına göre her iki testin güç oranları artan MİF miktarına bağlı olarak artış göstermiştir. GMH'nin güç oranları grupların yetenek dağılımındaki farklılaşmadan etkilenmezken, Mantel Test için güç oranları bir miktar yükselmiştir. İki test için en iyi güç oranları dengeli MİF örüntüsü koşulunda elde edilmiştir. Genel olarak oluşturulan farklı koşullar için GHM'nin istatistiksel güç oranları Mantel Test'e göre daha yüksek olmuştur.

Carter (2011), çalışmasında sabit MİF örüntüsünde ayırıcılık ve adım güçlüğü parametrelerinde sabit miktarlarda MİF koşulları için MTK-OOT ve DFIT'in (differential functioning of items and tests) istatistiksel gücünü incelemiştir. Araştırma bulguları, her iki test için grup büyüklüklerinin denk olduğu koşullarda ilgili testlerin daha yüksek güç oranları verdiğini ortaya koymuştur. Farklı test koşulları için istatistiksel güç oranları bakımından OOT DFIT'e göre daha yüksek performans göstermiştir.

Wood (2011a), çalışmasında sabit, yakınsak ve ıraksak MİF örüntüsü ve 0.45 MİF miktarı için, çok küçük örneklem durumlarında çok kategorili puanlanan maddeler için MİF belirleme tekniklerinden Mantel Test, Liu-Agresti istatistiği ve HW3'ün performansını incelemiştir. Bu teknikler küçük örneklem koşullarında istatistiksel manidarlığı test etmek üzere herhangi bir modifikasyonun (Bayesian, normalleştirme ve log-linear düzeltme) kullanıldığı ve kullanılmadığı durumlarda karşılaştırılmıştır. Araştırma sonuçlarına göre, testlere ilişkin herhangi bir modifikasyonun yapılmadığı durumda Mantel Test ve Liu-Agresti istatistiği sabit ve yakınsak MİF örüntüsünde en yüksek istatistiksel güç değerlerini almıştır. Modifikasyonun olmadığı duruma göre, Log-linear düzeltmede Mantel Test ve Liu-Agresti istatistiğine ilişkin güç değerleri bir miktar yükselirken Bayesian ve normalleştirme modifikasyonlarında bir farklılık gözlenmemiştir.

Wood (2011b) tarafından yapılan diğer çalışmada ise referans ve odak gruba ilişkin yetenek dağılımlarının normallikten sapma gösterdiği durumlarda OOT, Mantel Test, Genelleştirilmiş Mantel–Haenszel (GMH) Test ve poly-SIBTEST'in performansı karşılaştırılmıştır. Araştırmada MİF örüntüsü (sabit veya empirically observed) ve yetenek dağılım biçimlerinin (normal, pozitif çarpık ve pozitif basık) farklılaştığı test koşulları oluşturulmuştur. I. Tip hatanın kontrol edildiği koşullarda OOT, GMH ve poly-SIBTEST için yüksek istatistiksel güç değerleri elde edilmiştir. Mantel Test ise, sabit MİF örüntüsünde tek biçimli MİF'i belirlemede iyi performans göstermiştir. OOT ve GMH için farklı MİF örüntülerinde benzer ve yüksek güç değerleri elde edilirken, Mantel Test ve Poly- SIBTEST sabit MİF örüntüsünde daha yüksek performans göstermiştir.

MİF miktarı ve MİF örüntüsü bağlamında ilgili çalışmalarda artan MİF miktarına bağlı olarak MİF belirleme tekniklerinin istatistiksel gücünün yükseldiği görülmektedir. MİF örüntüsü kapsamında ise, çoğunlukla sabit MİF örüntüsü için yeterli istatistiksel güç değerleri elde edilirken; araştırmalarda kullanılan ayırıcılık ve adım güçlüğü parametrelerinin özelliklerine bağlı olarak MİF örüntüleri için farklı istatistiksel güç değerleri hesaplanmıştır.

Çok kategorili puanlanan maddelerden elde edilen veriler üzerinde MİF belirleme tekniklerinin farklı test koşullarında karşılaştırıldığı yukarıdaki araştırmalar genel olarak incelendiğinde, son yıllarda yapılan çalışmalarda bağımsız değişken sayısının çeşitlendiği ve farklı MTK modelleriyle elde edilen veriler üzerinden çalışmaların yürütüldüğü dikkat çekmektedir. MİF belirlemede hipotez testlerini kullanan tekniklerin farklı test koşullarına olan duyarlılıklarına dikkat çekilerek özellikle örneklem büyüklüğünden daha az etkilenen tekniklerin kullanımı konusunda öneriler giderek öne çıkmaktadır.

İlgili araştırma bulguları, referans ve odak gruba ait örneklem büyüklüğü ve odak grup yetenek dağılım özellikleri, MİF miktarı, MİF örüntüsü, incelenen maddelere ait ayırıcılık ve eşik parametreleri, kayıp veri miktarı gibi değişkenlerin MİF belirleme tekniklerinin performansında farklar oluşturduğunu ortaya koymaktadır. Bu araştırma bulguları, MİF belirleme

tekniklerinin çok geniş örneklem büyüklüklerine duyarlı olduğunu ve bu tür koşullarda bu tekniklerin geçerli sonuçlar vermediğini göstermektedir. Buna bağlı olarak araştırmacılar tarafından bilinen MİF belirleme tekniklerine ek olarak MİF etki büyüklüğü ölçüsü, parametre tekrarlama yöntemleri gibi alternatif yaklaşımlar önerilmektedir.

MTK-OOT gibi parametrik testler, referans ve odak gruba ait örneklem büyüklüğünün küçük olmasından, Mantel Test gibi parametrik olmayan testlerin performansı ise gruplara ait dağılım özelliklerinin farklılaşmasından olumsuz olarak etkilenmektedir. Artan MİF miktarı ve örneklem büyüklüğü parametrik tekniklerin MİF belirlemedeki performansını olumlu yönde etkilemektedir. Ayrıca farklı test uzunlukları veya MİF'li madde oranı gibi değişkenler karşılaştırılan MİF belirleme tekniklerinin performansı bakımından farklar oluşturmamaktadır.

# BÖLÜM 3

# YÖNTEM

Bu bölümde, araştırma modeli, verilerin üretilmesi, verilerin düzenlenmesi ve verilerin analizi ve yorumlanması açıklanmaktadır.

## Araştırma Modeli

Bu çalışmada, oluşturulan farklı test koşulları için ilgili MİF belirleme testlerinin performansı I. Tip hata ve istatistiksel güç değerleri yoluyla incelenmiştir. Çalışmanın amacına uygun olarak, bootstrap kestirim yöntemine dayalı Monte-Carlo simülasyonu kullanılmıştır. Bu yöntemde herhangi büyüklükteki mevcut veri setindeki gözlemler şansa bağlı olarak yer değiştirerek yeniden örneklenerek yeni veri setleri oluşturulmaktadır (Horowitz,1996).

Naylor ve diğerleri (1968) tarafından tanımlanan, Monte-Carlo çalışmalarındaki adımlar MTK çalışmaları için uyarlandığında şu adımları içermektedir (Akt. Harwell, Stone, Hsu ve Kirisci, 1996): (1) problemin tanımlanması; (2) bağımlı ve bağımsız değişkenlerin tanımlanması, simülasyon deseninin oluşturulması, tekrar sayısı ve MTK modelinin belirlenmesi işlemlerini içeren araştırma deseninin oluşturulması; (3) tepki örüntülerinin elde edileceği ve parametre kestirimlerinin yapılacağı bilgisayar yazılımlarının belirlenmesi; (4) sonuçların analiz edilmesi.

Bu yaklaşımda, birey sayısı ve yetenek dağılımı tanımlanarak bireylerin yetenek kestirimi yapılmaktadır. Daha sonra test uzunluğu, kategori sayısı ve MTK modeli tanımlanarak madde parametreleri ve son olarak tekrar sayısı

belirlenerek bireylere ait tepki örüntüleri (raw data) elde edilmektedir. Simülasyon sırasında kullanılan parametreler daha önceki araştırmalardan elde edilebileceği gibi bilgisayar programı yoluyla da tanımlanabilir. Veri setleri referans ve odak grup için üretilerek MİF analizi için kullanılır.

Simülasyon, bilgisayar yardımı ile yapılabileceği gibi el ile de yapılabilir. Simülasyon sürecinde çok sayıda işlem yapma gereksinimi varsa ve çok uzun bir zaman aralığı için sistem simüle edilecekse bilgisayar kullanmak daha uygun olacaktır. Simülasyonun el ile yapılabilmesi için öncelikle değişkenlerin, olasılıkların ve rasgele sayıların belirlenmesi daha sonra da tablolar halinde gösterilmesi gerekmektedir (Kavcar, 2004).

Monte Carlo, gerçek test koşullarının modellenebilmesine ve ampirik veriler yoluyla elde edilemeyen istatistiklerin hesaplanması ve karşılaştırılmasına olanak sağlaması bakımından araştırmalarda önemi giderek artmaktadır (Boomsma, 2013). Örneğin, 1994-1995 yılları arasında *Journals of Applied Psychological Measurement* (APM), *Psychometrika* ve *Journal of Educational Measurement* (JEM) adlı dergilerdeki makalelerin yaklaşık üçte birinde Monte Carlo simülasyon yaklaşımı kullanılmıştır (Harwell ve diğerleri, 1996). Monte Carlo simülasyon yaklaşımıyla sosyal bilimciler de fen bilimcilerin laboratuar ortamında yaptıkları deneyleri bilgisayar ortamında gerçekleştirme olanağı bulmaktadır.

## Veriler

Araştırmada sınanmasına karar verilen MİF belirleme testlerinin farklı test koşullarında I. Tip hata ve istatistiksel güç oranlarını incelemek üzere, referans ve odak gruba ait çok kategoride puanlanan (polytomous) veriler üretilmiştir. I.Tip hata oranlarını değerlendirmek üzere, tamamı MİF içermeyen 20 maddeden üretilen yanıt örüntüleri elde edilmiştir. Güç çalışmaları için bu maddelerden üçü MİF içerecek biçimde tekrar simüle edilmiştir. I.Tip hata ve güç çalışmaları için oluşturulan koşullarda bazı değişkenler sabit tutulurken bazı değişkenler için manipülasyonlar (değişimlemeler) yapılmıştır. Sabit tutulan ve üzerinde değişimleme yapılan değişkenler aşağıda açıklanmaktadır.

**Sabit Tutulan Faktörler**

**Çok Kategorili MTK Modeli.** Referans ve odak gruba ait yanıt örüntüleri çok kategorili MTK modellerinden Kısmi Puan Modeli (Partial Credit Model) kullanılarak elde edilmiştir. Bu model, hem bilişsel hem de duyuşsal özelliklerin ölçülmesinde kullanılabilir bir model olduğu için hem de Rasch modellerinin özelliklerini taşıdığından parametrelerinin yorumlanmasının kolay olmasından dolayı tercih edilmiştir. Bu model ayrıca MİF ile ilgili diğer birçok yapay veri çalışmalarında tercih edilen bir modeldir (Zwick ve diğerleri, 1993; Chang, Mazzeo ve Roussos, 1996; Zwick ve Thayer, 1996; Su ve Wang, 2005; Garrett, 2009).

**Test Uzunluğu.** Daha önce yapılan çalışmalar test uzunluğunun MİF belirleme teknikleri ile ilgili çalışmalarda çok az etkiye sahip olduğu belirlenmiştir (Flowers, Oshima ve Raju, 1999; Oshima, Raju ve Nanda, 2006). Dodeen (2004) ve Dodeen ve Johanson (2003), tutum maddeleri için yaptıkları MİF analizlerinde gerçek test koşullarına uygunluğunu gerekçe göstererek 20 madde kullanmıştır. Bu çalışma için tekrarlama sayısı ve analizleri yorumlama kolaylığı dikkate alınarak test uzunluğu 20 madde olarak belirlenmiştir.

**MİF İçeren Madde Sayısı.** Yapay veri ile yapılan çalışmalarda MİF belirleme teknikleri bir ya da daha fazla sayıda MİF gösteren maddeler üzerinden yürütülmüştür (Chang ve diğerleri, 1996; Zwick, Thayer ve Mazzeo, 1997; Bolt, 2002; Wang ve Su, 2004; Su ve Wang, 2005; Garrett, 2009). Su ve Wang (2004) gerçeğe daha yakın olması bakımından bu tür çalışmalarda MİF içeren madde sayısının birden fazla olmasını önermiştir. Bu çalışmada üç maddenin MİF içerdiği koşullar oluşturulmuştur.

**Tepki Kategori Sayısı.** Çalışmada kullanılan 20 maddeye ait tepki kategori sayısı dört olarak belirlenmiştir. Bu kategori sayısı belirlenirken, gerçek test koşullarında bilişsel ve duyuşsal özelliklerin ölçülmesine uygun olması dikkate alınmıştır. Örneğin TIMSS ve PISA gibi madde ve yetenek parametrelerinin MTK ölçeklemesine göre yapıldığı uygulamalarda başarı, tutum ve öz yeterlik algısı gibi değişkenler için dört kategorili tepkiler kullanılmıştır. Ayrıca, Dodeen ve Johanson (2003), Dodeen (2004), Artar (2007) ve Garrett (2009) yaptıkları çalışmada maddelere ait tepki kategori sayısını dört olarak belirlemiştir.

**MİF Türü.** Kısmı Puan Modeli'ne uygun olarak üretilen veriler sadece adım güçlüğü parametresi ($\delta$) bakımından farklılaştığı için referans ve odak grup arasında tek biçimli (uniform) MİF gözlenmektedir. Tek biçimli olmayan MİF gözlenebilmesi için grupların kategori ayırıcılık parametreleri (α) bakımından farklılık göstermesi gerekmektedir.

## Değişimleme Yapılan Faktörler

**Referans (R) ve Odak (O) Gruba Ait Örneklem Büyüklükleri.** Bu çalışmada kullanılan MİF belirleme testleri hipotez testlerine dayandığı için grupların örneklem büyüklüğü bu testlerin performansını etkileyen önemli bir değişkendir. MTK-OOT'i ve Mantel Test'in kullanıldığı simülasyon çalışmalarında örneklem büyüklüğünün 500 ile 2000 arasında değiştiği görülmektedir (örn. Artar, 2007; Su ve Wang, 2005; Thurman, 2009).

Bu çalışmada referans ve odak grup için farklı örneklem büyüklüklerini (500, 1250 ve 2000) içeren koşullar oluşturulmuştur. Böylece küçük, orta ve büyük örneklem büyüklüklerin temsil edilmesi sağlanmıştır (Atar ve Kamata, 2011; Narayanan ve Swaminathan, 1996; Rogers ve Swaminathan, 1993). Su ve Wang (2005) yaptıkları çalışmada referans ve odak grup büyüklüklerinin denk olduğu koşullar oluşturmalarına karşın, Kristjansson ve diğerleri (2005) referans ve odak grup büyüklüklerinin denk olmadığı durumların MİF belirleme tekniklerinin performansını etkileyebileceğine dikkat çekmiştir. Genellikle odak

gruptaki birey sayısının referans gruba göre daha az olduğu (Ankemann ve diğerleri, 1997; Garrett, 2009; Wood, 2011) dikkate alınarak, bu çalışma için referans ve odak grup büyüklük oranları eşit olduğu (250:250 ve 1000:1000) durumlara ek olarak, referans ve odak grup büyüklük oranının 4:1 (1000:250) olduğu koşul da elde edilmiştir. Bu çalışmada MULTİLOG'da parametre kestirim hatasını en aza indirmek için parametre kestirimlerinin doğruluğu da dikkate alınarak odak gruptaki birey sayısının en az 250 olduğu koşullar elde edilmiştir.

**Referans (R) ve Odak (O) Gruba Ait Dağılım Özellikleri.** MİF çalışmalarında, grupların aynı ve farklı yetenek dağılımlarına sahip olma durumları sıklıkla ele alınan bir değişkendir. Çünkü grupların yetenek dağılımları farklılaştıkça I. Tip hata ve güç oranının da değiştiği bilinmektedir (Kristjansson ve diğerleri, 2005; Narayanan ve Swaminathan, 1996; Penfield, 2001). İlgili çalışmaların büyük kısmında referans ve odak grubun yetenek dağılımları arasında -0.25, -0.50, -1 ve -1.5 biçiminde farklılaşan sapma değerleri kullanılmıştır (Örn. Bolt, 2002; Garrett, 2009, Wang ve diğerleri, 2013; Wood, 2011).

Bu çalışmada grup yetenek dağılımı için üç farklı koşul oluşturulmuştur. İlk koşulda, odak ve referans grup yetenek ölçülerinin dağılımlarının ortalaması 0 standart sapması 1 olan birim normal dağılım özelliği taşımaktadır (R~*N*(0,1), O~*N*(0,1)). Roussos ve Stout (1996) ve Tian (1999), gerçek test uygulamaları ve uzman görüşüne dayanarak grupların yetenek dağılım ortalaması arasında 0.5 ve 1 standart sapmalık farkların gerçek test uygulamalarına yakın olduğunu vurgulamıştır. Bu durum dikkate alınarak ikinci koşulda, referans grup dağılım özellikleri aynı kalmış, odak grubun yetenek dağılım ortalaması -0.5 standart sapma ((R~*N*(0,1), O~*N*(-0.5,1)) ve son koşulda yine referans grup birim normal dağılım özelliği gösterirken, odak gruba ait yetenek dağılımının ortalaması -1standart sapma (R~*N*(0,1), O~*N*(-1,1)) fark gösterecek biçimde değiştirilmiştir.

**MİF Miktarı.** Bu çalışmada yorumlanabilir sonuçlar elde etmek amacıyla MİF miktarı 0.43 ve 0.64 lojit birim olarak belirlenmiştir. Örneğin Thurman (2009) çalışmasında MİF miktarı olarak 0.10 lojit birimi kullanmış; ancak yorumlanabilir değerler elde edememiştir.

Çalışmada esas alınan lojit birimler maddelerin farklı düzeylerde sahip olduğu MİF'i göstermek üzere Educational Testing Service (ETS) tarafından yapılan sınıflamaya (Dorans ve Holland, 1993) göre orta düzey (moderate) ve büyük (large) MİF miktarlarını temsil etmektedir. MİF miktarının 0.43 olduğu koşul için, referans gruba ait maddelerin adım güçlüğü parametreleri aynı kalırken, odak gruba ait ilgili maddelerin adım güçlüğü parametresi 0.43 lojit birim arttırılmıştır ($\delta_{i1o} = \delta_{i1r} + 0.43$).

**MİF Örüntüsü.** MİF örüntüsü, adım güçlüğü parametresi bakımından referans ve odak grup arasındaki farklılaşmayı ifade etmektedir. MİF örüntüsü bakımından ilgili araştırmalar incelendiğinde, çalışmalarda ağırlıklı olarak sabit veya dengeli MİF örüntüsünün yer aldığı görülmektedir (örn. Ankemann ve diğerleri, 1999; Kristjansson ve diğerleri, 2005; Tian, 1999; Wang ve Su, 2005; Wang ve diğerleri, 2013). Ancak sayıları fazla olmasa da son yapılan çalışmalarda çok kategorili puanlanan maddeler için çok farklı kombinasyonlarda MİF örüntüsünün ortaya çıkabileceği vurgulanarak sabit (constant) ve dengeli (balanced) örüntüler dışında düşük (low), yüksek (high), ıraksak (convergent) ya da yakınsak (divergent) MİF örüntülerinin de incelendiği görülmektedir (Artar, 2007; Fidalgo ve Bartram, 2010; Thurman, 2009; Wood, 2011). Bu çalışmada yorumlama kolaylığı ve gerçek test koşullarına uygun olduğu düşünülerek, güç çalışmaları kapsamında MİF içeren üç madde için iki farklı MİF örüntüsü elde edilmiştir.

İlk MİF örüntüsü koşulunda, birinci puan kategorisi ile ikinci puan kategorisi arasında yer alan ilk adım güçlüğü parametresi, belirlenen miktarlarda odak grup parametreleri için daha fazladır. Diğer adım güçlüğü parametreleri referans ve odak grup için aynı kalmıştır ($\delta_{i1o} = \delta_{i1r} + 0.43$, $\delta_{i2o} = \delta_{i2r}$, $\delta_{i3o} = \delta_{i3r}$). Bu koşul düşük MİF örüntüsüdür (low shift DIF pattern). İkinci koşulda, üçüncü ve dördüncü puan kategorisi arasında yer alan üçüncü

adım güçlüğü parametresi, belirlenen miktarlarda odak grup parametreleri için daha fazladır. Diğer adım güçlüğü parametreleri her iki grup için aynı kalmıştır ($\delta_{i1o}= \delta_{i1r}$, $\delta_{i2o}= \delta_{i2r}$, $\delta_{i3o}= \delta_{i3r}$ +0.43). Bu koşul yüksek MİF örüntüsüdür (high shift DIF pattern).

Her bir koşul için Kısmi Puan Modeli'ne uygun olarak maddeleri dört kategoride puanlanan bir testin iki farklı formu kullanılmıştır. İlk test formu, I. Tip hata çalışmaları için tüm maddeler MİF içermeyecek biçimde modellenmiştir. Güç çalışmaları kapsamında kullanılan ikinci test formunda ise 17 madde ilk test formunda olduğu gibi MİF göstermemiş, sadece üç madde (4,13 ve 17) ise farklı MİF örüntüsü ve MİF miktarına uygun olarak MİF gösterecek biçimde modellenmiştir. Bu maddeler seçilirken adım güçlüğü parametresi bakımından herhangi bir değişkenlik oluşturmaması için madde parametrelerinin birbirine yakın olmasına dikkat edilmiştir. Bu maddeler dışındaki diğer maddeler temel (core) madde olarak adlandırılmaktadır ve bunlar tüm koşullar için sabit kalmıştır. MTK-OOT için bu maddeler referans ve odak grup ortak bağ maddeleri (common anchor item) olarak kullanılmıştır; çünkü bu maddeler MİF içermeyecek biçimde modellenmiştir.

Çalışma kapsamında elde edilen koşullarla ilgili olarak güvenilir sonuçlar elde etmek amacıyla her koşul için belirli sayıda veri yeniden üretilmiş ve bu veriler üzerinde MİF analizleri tekrar edilmiştir. National Council on Measurement in Education (NCME) tarafından basılan 2009 yayınlarında MİF belirleme tekniklerinin karşılaştırıldığı yapay veri içeren çalışmalarda ortak kullanılan tekrarlama sayısının 100 olduğu görülmektedir (Akt. Kim,2010). Ayrıca, Bolt (2002), Fortman-Johnson (2007) ve Kim (2000) tarafından yapılan çalışmalarda da tekrarlama sayısı aynı şekildedir. Bu çalışma için de tekrarlama sayısı 100 olarak belirlenmiştir.

Son durumda I. Tip hata çalışmaları için elde edilen simülasyon Çizelge 2'de verilmektedir.

Çizelge 2. I. Tip Hata Çalışmaları İçin Simülasyon Deseni

| **Yetenek Dağılımı** | | **Örneklem Büyüklüğü** | |
|---|---|---|---|
| **Referans (R)** | **Odak (O)** | **Referans (R)** | **Odak (O)** |
| R~*N*(0,1), O~*N*(0,1) | | 250 | 250 |
| | | 1000 | 250 |
| | | 1000 | 1000 |
| R~*N*(0,1), O~*N*(-0.5,1) | | 250 | 250 |
| | | 1000 | 250 |
| | | 1000 | 1000 |
| R~*N*(0,1), O~*N*(-1,1) | | 250 | 250 |
| | | 1000 | 250 |
| | | 1000 | 1000 |

Çizelge 2 incelendiğinde, son durumda I. Tip hata çalışması için 18 [3 (yetenek dağılımı) x 3 (örneklem büyüklüğü) x 2 (MİF belirleme testi)] simülasyon koşulu ortaya çıkmıştır. Her bir koşul için 100 tekrar yapıldığından 1800 veri dosyası elde edilmiştir.

Son durumda güç çalışmaları için elde edilen simülasyon Çizelge 3'te verilmektedir.

Çizelge 3. Güç Çalışmaları İçin Simülasyon Deseni

| Yetenek Dağılımı | | Örneklem Büyüklüğü | | MİF Miktarı | MİF Örüntüsü |
|---|---|---|---|---|---|
| **R** | **O** | **Referans** | **Odak** | | |
| R~*N*(0,1), O~*N*(0,1) | | 250 | 250 | 0.42 | D |
| | | | | | Y |
| | | | | 0.64 | D |
| | | | | | Y |
| | | 1000 | 250 | 0.42 | D |
| | | | | | Y |
| | | | | 0.64 | D |
| | | | | | Y |
| | | 1000 | 1000 | 0.42 | D |
| | | | | | Y |
| | | | | 0.64 | D |
| | | | | | Y |
| R~*N*(0,1), O~*N*(-0.5,1) | | 250 | 250 | 0.42 | D |
| | | | | | Y |
| | | | | 0.64 | D |
| | | | | | Y |
| | | 1000 | 250 | 0.42 | D |
| | | | | | Y |
| | | | | 0.64 | D |
| | | | | | Y |
| | | 1000 | 1000 | 0.42 | D |
| | | | | | Y |
| | | | | 0.64 | D |
| | | | | | Y |
| R~*N*(0,1), O~*N*(-1,1) | | 250 | 250 | 0.42 | D |
| | | | | | Y |
| | | | | 0.64 | D |
| | | | | | Y |
| | | 1000 | 250 | 0.42 | D |
| | | | | | Y |
| | | | | 0.64 | D |
| | | | | | Y |
| | | 1000 | 1000 | 0.42 | D |
| | | | | | Y |
| | | | | 0.64 | D |
| | | | | | Y |

D: Düşük Y: Yüksek

Çizelge 3'e göre, istatistiksel güç çalışmaları için 76 [3 (yetenek dağılımı) x 3 (örneklem büyüklüğü) x 2 (MİF miktarı) x 2 (MİF örüntüsü) x 2 (MİF belirleme tekniği)] simülasyon koşulu ortaya çıkmıştır. Böylece güç çalışmaları için 7600 veri dosyası elde edilmiştir.

## Verilerin Üretilmesi

Belirlenen koşullara uygun veri üretmek için WinGen3 yazılım programı kullanılmıştır. WinGen, Madde Tepki Kuramı modellerine uygun ve farklı yetenek ve madde parametre dağılımlarına göre madde ve yetenek parametrelerinin üretilmesine izin vermektedir. Program ilgili yetenek ve madde parametrelerine uygun olarak 1.000.000 tekrara kadar yanıt örüntüleri içeren veri dosyası oluşturabilmektedir (Han ve Hambleton, 2007). Programın, araştırma amacına uygun veri setlerinin oluşturulmasına olanak sağlaması ve kullanım kolaylığı dikkate alınarak araştırma verileri WinGen3 programı kullanılarak üretilmiştir.

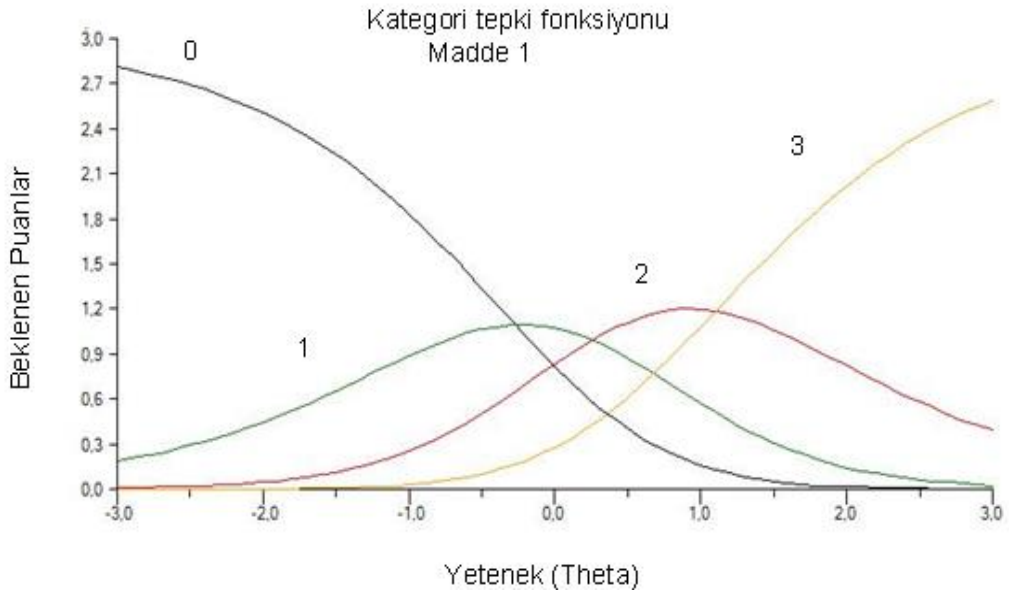
Belirlenen koşullara Kısmı Puan Modeli'ne uygun olarak veriler referans ve odak grup için ayrı ayrı üretilmiş ve daha sonra bu veriler analiz için tek bir veri dosyasında birleştirilmiştir. WinGen3 programında ilk olarak birey sayısı ve yetenek dağılım özellikleri tanımlanarak bireylerin yetenek kestirimleri yapılmıştır. Daha sonra madde sayısı, tepki kategori sayısı ve ilgili MTK modeli tanımlanarak madde parametreleri elde edilmiştir. I. Tip hata çalışmaları kapsamında kullanılan ve MİF içermeyen maddelere ait adım güçlüğü parametreleri Çizelge 4'te verilmektedir.

Çizelge 4. MİF İçermeyen Maddelere Ait Adım Güçlüğü Parametreleri

| Maddeler | $\delta_1$ | $\delta_2$ | $\delta_3$ | Maddeler | $\delta_1$ | $\delta_2$ | $\delta_3$ |
|---|---|---|---|---|---|---|---|
| **1** | -0.274 | 0.257 | 1.100 | **11** | -1.372 | -0.274 | -0.139 |
| **2** | -1.326 | -1.156 | 0.773 | **12** | -1.212 | 0.656 | 0.871 |
| **3** | -0.963 | 0.433 | 0.450 | **13** | 0.121 | 1.309 | 1.502 |
| **4** | 0.422 | 0.709 | 1.109 | **14** | -0.957 | 1.253 | 2.159 |
| **5** | -0.082 | 0.510 | 1.637 | **15** | -0.185 | 0.271 | 1.065 |
| **6** | -1.071 | -0.152 | 2.138 | **16** | -0.611 | 0.703 | 1.875 |
| **7** | -0.781 | -0.108 | 0.045 | **17** | 0.503 | 1.112 | 1.952 |
| **8** | -1.122 | -0.518 | 1.239 | **18** | -1.246 | 0.766 | 1.797 |
| **9** | -1.209 | -0.854 | 0.766 | **19** | -2.192 | -0.701 | 1.15 |
| **10** | -0.996 | 0.016 | 0.814 | **20** | -1.592 | -0.75 | -0.553 |

$\delta$: adım güçlüğü parametresi

Bu maddeler için kategori tepki fonksiyonlarını gösteren grafikler de elde edilmiştir. Örnek olarak bir numaralı madde için grafiksel gösterim Şekil 7'de verilmektedir. Diğer maddelere ilişkin kategori tepki fonksiyonları Ek 1'de ayrıca sunulmaktadır.




Şekil 7. Bir Numaralı Madde İçin Kategori Tepki Fonksiyonu

$(\delta_1 : -0.274, \delta_2 : 0.257, \delta_3 : 1.100)$

Güç çalışmaları kapsamında MİF içeren maddeleri üretmek için, odak grupta belirlenen üç maddeye ait adım güçlüğü parametreleri ilgili koşulları (MİF miktarı ve MİF örüntüsü) sağlayacak biçimde yeniden düzenlenmiştir. Referans gruba ait maddelerin adım güçlüğü parametreleri ise değişmeden kalmıştır. Odak grup için üç madde MİF içerecek biçimde düzenlendikten sonra Kısmı Puan Modeli'ne uygun olarak tepki örüntüleri ilgili koşullarda tekrar üretilmiştir. Güç çalışmalarında kullanılan MİF'li maddelerin ilgili MİF miktarı ve MİF örüntüsü koşullarında elde edilen adım güçlüğü parametreleri Çizelge 5'te verilmektedir.

Çizelge 5. MİF İçeren Maddelerin Adım Güçlüğü Parametreleri

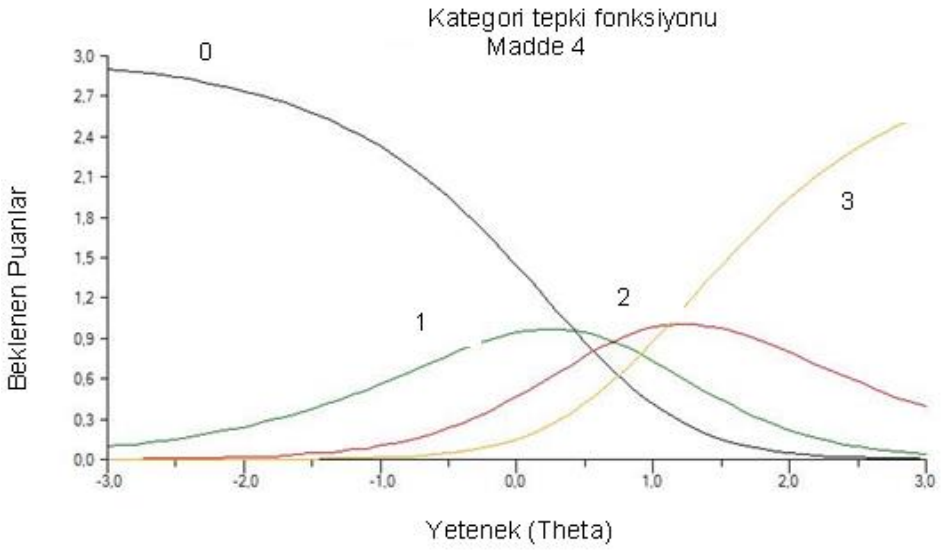
a) (D- MİF Örüntüsü Koşulunda)

| | **Referans Grup** | | | **Odak Grup (MİF Miktarı 0.43)** | | |
|---|---|---|---|---|---|---|
| **Maddeler** | **1** | **2** | **3** | **1** | **2** | **3** |
| 4 | 0.422 | 0.709 | 1.109 | 0.852 | 0.709 | 1.109 |
| 13 | 0.121 | 1.309 | 1.502 | 1.502 | 1.309 | 1.502 |
| 17 | 0.503 | 1.112 | 1.952 | 1.952 | 1.112 | 1.952 |
| | **Referans Grup** | | | **Odak Grup (MİF Miktarı 0.64)** | | |
| **Maddeler** | **1** | **2** | **3** | **1** | **2** | **3** |
| 4 | 0.422 | 0.709 | 1.109 | 1.062 | 0.709 | 1.109 |
| 13 | 0.121 | 1.309 | 1.502 | 0.761 | 1.309 | 1.502 |
| 17 | 0.503 | 1.112 | 1.952 | 1.143 | 1.112 | 1.952 |

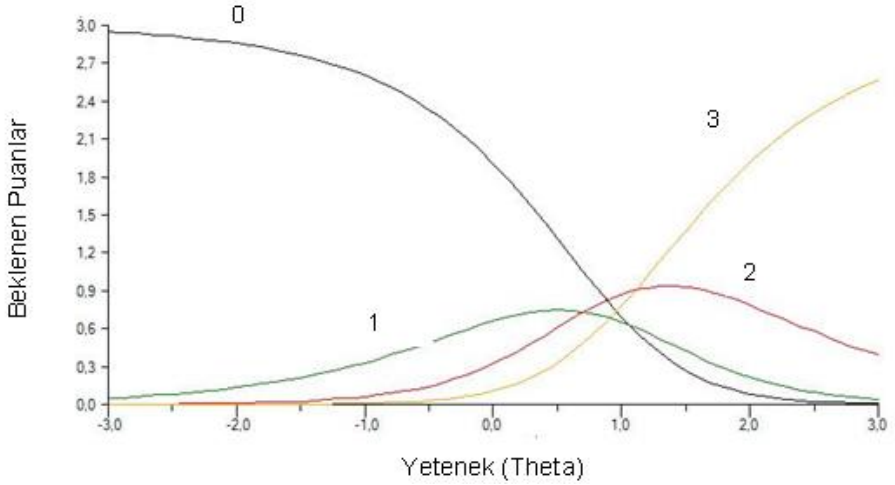
b) (Y- MİF Örüntüsü Koşulunda)

| | **Referans Grup** | | | **Odak Grup (MİF Miktarı 0.43)** | | |
|---|---|---|---|---|---|---|
| **Maddeler** | **1** | **2** | **3** | **1** | **2** | **3** |
| **4** | 0.422 | 0.709 | 1.109 | 0.422 | 0.709 | 1.539 |
| **13** | 0.121 | 1.309 | 1.502 | 0.121 | 1.309 | 1.932 |
| **17** | 0.503 | 1.112 | 1.952 | 0.503 | 1.112 | 2.382 |
| | **Referans Grup** | | | **Odak Grup (MİF Miktarı 0.64)** | | |
| **Maddeler** | **1** | **2** | **3** | **1** | **2** | **3** |
| **4** | 0.422 | 0.709 | 1.109 | 0.422 | 0.709 | 1.749 |
| **13** | 0.121 | 1.309 | 1.502 | 0.121 | 1.309 | 2.142 |
| **17** | 0.503 | 1.112 | 1.952 | 0.503 | 1.112 | 2.592 |

MİF içeren maddeler için referans ve odak gruptan kategori tepki fonksiyonları elde edilmiştir. Örnek olarak Şekil 8 ve 9'da düşük MİF örüntüsü ve MİF miktarının 0.64 lojit birim olduğu koşulda referans ve odak gruptan dört numaralı madde için elde edilen kategori tepki fonksiyonları verilmektedir.




Şekil 8. Referans Gruptan Elde Edilen Kategori Tepki Fonksiyonu




Şekil 9. Odak Gruptan Elde Edilen Kategori Tepki Fonksiyonu

En son aşamada ise tekrar sayısı (100) girilerek her bir koşula ilişkin ilgili sayıda tepki örüntüleri elde edilmiştir. Referans ve odak grup için veri dosyaları ayrı ayrı üretilmiş ve daha sonra tek bir dosyada analiz etmek üzere veri dosyaları birleştirilmiştir.

WinGen3'te üretilen tepki örüntülerine ait dosyalar "wgr" uzantılıdır. Bu dosyaların MULTILOG'da çalıştırılabilmesi için "dat" uzantılı dosyaya dönüştürülmesi gerekmektedir. Bu amaçla "wgr" uzantılı referans ve odak gruba ait tepki örüntüleri EXCEL yardımı ile önce "prn" ve daha sonra "dat" uzantılı dosyalar haline getirilmiştir. Ayrıca, referans ve odak grup veri dosyaları birleştirilirken, WİNGEN3'te elde edilen 0 kodları için yeniden kodlama yapılmıştır. Çünkü MULTILOG 0 kodunu kayıp veri olarak işleme almaktadır. Son durumda, EXCEL dosyalarında 0 kodu 1'e, 1 kodu 2'ye, 2 kodu 3'e ve 3 kodu 4'e dönüştürülerek "dat" uzantılı veri dosyaları elde edilmiştir. Mantel Test analizleri için, "wgr" uzantılı tepki örüntüleri EXCEL yardımı ile "txt" uzantılı hale getirilmiştir. Belirtilen işlemler araştırma koşullarına ait her bir tekrarlama (replikasyon) veri dosyası için tekrarlanmıştır.

## Verilerin Analizi ve Yorumlanması

### MTK- Olabilirlik Oran Testi Analizleri

Model karşılaştırmaya dayanan MTK-OOT'ye ilişkin analizleri için, geniş (augmented model) ve dar (compact model) modeller kurulmuştur. I. Tip hata çalışmaları için kurulan dar modelde, testteki tüm maddelere ait parametreler referans ve odak gruplarda eşitlenmiştir. Daha sonra her bir madde için, ilgili maddeye ait parametrelerin gruplarda serbest bırakıldığı geniş modeller oluşturulmuştur. Güç çalışmalarında bağ maddeleri dışında kalan her bir incelenen (studied) madde için, bağ maddeleri ile birlikte gruplarda parametrelerin eşitlendiği dar modeller ve incelenen maddelere ait parametrelerin her iki grupta serbest bırakıldığı geniş model kurulmuştur.

Dört numaralı için kurulan dar modelde, bağ maddeleri ve dört numaralı madde parametreleri referans ve odak grupta eşitlenmiş diğer iki maddenin (13 ve 17 numaralı maddeler) parametreleri serbest bırakılmıştır. Her bir "incelenen madde" için bir dar model kurulmuştur. Dar ve geniş modeller için komut dosyaları ayrı ayrı oluşturulmuştur (Ek 2'de örnek komut dosyası verilmektedir). Geniş ve dar modellerden elde edilen "-2 log Likelihood" değerleri arasındaki fark ilgili serbestlik derecesinde karşılaştırılarak ilgili maddenin MİF gösterip göstermediğine karar verilmiştir.

**Mantel Test Analizleri**

Mantel Test analizlerinde ilgili maddelerin MİF gösterip göstermediğine karar vermek üzere için Mantel ki-kare istatistiği kullanılmıştır. Referans ve odak grup bireylerini eşleştirmek üzere toplam puanlar kullanılmıştır. Analizin yapıldığı program bireyleri eşleştirmek üzere 10 aralık düzeyi belirlemektedir. Analiz sonucu elde edilen ki-kare istatistiği bir serbestlik derecesinde ki-kare dağılımı göstermektedir (Mantel, 1963; Zwick ve diğerleri,1993; Zwick ve diğerleri, 1997) ve bu istatistik için 0.05 anlamlılık düzeyinde kritik değer 3.84'tür.

Yukarıdaki analizler her bir koşulda 100 defa tekrarlanmıştır. Tanımlanan farklı koşullarda MİF analizleri yapıldıktan sonra, her bir testin belirlenen koşullardaki I. Tip hata ve güç oranları hesaplanmıştır. I. Tip hata oranı, 0.05 manidarlık düzeyi için tekrarlanan 100 analizde MİF'in yanlış belirlendiği analiz oranı ve istatistiksel güç ise doğru belirlenen MİF oranıdır. I. Tip hata oranları MİF içermeyen 20 madde için hesaplanırken güç oranları MİF içerecek biçimde modellenen üç madde üzerinden hesaplanmıştır.

I. Tip hatanın değerlendirilmesi için Bradley'in esnek referansı (Bradley's liberal criterion) kullanılmıştır. Buna göre 0.05 α düzeyinde $0.025 \leq$ I. Tip hata oranı $\leq 0.075$ olduğu durumlarda I.Tip hatanın iyi kontrol edildiği sonucuna varılmıştır. Güç çalışmalarında tekrar üretilen 100 veri seti için MİF'in doğru belirlendiği analiz oranın 0.80'denk ya da büyük olduğu durumlarda ilgili tekniğin MİF'i belirlemede yeterli; oranın 0.80'den küçük

olduğu durumlarda ise bu tekniğin MİF'i belirlemede yetersiz olduğu değerlendirmesi yapılmıştır.

Çalışmada veri üretmek için WinGen3, MTK-OOT karşılaştırmaları için MULTİLOG (Thissen, Chen ve Bock, 2002) ve Mantel Test analizleri için DIFAS (Penfield, 2005) programı kullanılmıştır.

# BÖLÜM 4

# BULGULAR VE YORUMLAR

Bu bölümde araştırmanın alt amaçlarına uygun olarak elde edilen bulgulara ve bu bulgulara ilişkin yorumlara yer verilmiştir.

## Mantel Test ve MTK-OOT İçin I. Tip Hata Oranları

I. Tip hata çalışması kapsamında grupların yetenek dağılımı ortalaması ve örneklem büyüklüklerine bağlı olarak Mantel Test ve MTK-OOT'nin I. Tip hata oranları incelenmiştir. Bu kapsamda grupların yetenek dağılımı ortalamasının aynı ve farklı olduğu koşullara ilişkin bulgu ve yorumlara yer verilmiştir.

### R~N(0,1), O~N(0,1) dağılım koşulunda farklı örneklem büyüklükleri için Mantel Test ve MTK-OOT'nin I. Tip Hata Oranları

Bu kısımda I. Tip hata çalışmaları kapsamında araştırmanın ilk alt amacı olan *"Mantel Test ve MTK-OOT için referans ve odak gruba ait yetenek dağılımlarının birim normal dağılım gösterdiği koşulda, grupların örneklem büyüklüklerindeki değişime bağlı olarak Mantel Test ve MTK-OOT için I.Tip hata oranları nasıl değişmektedir?"* sorusuna ilişkin bulgu ve yorumlara yer verilmiştir.

İlgili koşulda değişen örneklem büyüklük oranlarına bağlı olarak Mantel Test ve MTK-OOT için I. Tip hata oranları Çizelge 6'da verilmektedir.

Çizelge 6. R~N(0,1), O~N(0,1) Dağılım Koşulunda Değişen Örneklem Büyüklükleri için Mantel Test ve MTK-OOT'nin I. Tip Hata Oranları

| **Yetenek Dağılımı** | | **Örneklem Büyüklüğü** | | **MİF Belirleme Testi** | |
|---|---|---|---|---|---|
| **Referans** | **Odak** | **Referans** | **Odak** | **Mantel Test** | **MTK-OOT** |
| R~*N*(0,1), O~*N*(0,1) | | 250 | 250 | 0.045 | 0.052 |
| | | 1000 | 250 | 0.047 | 0.038 |
| | | 1000 | 1000 | 0.053 | 0.041 |

R: Referans grup O: Odak grup

Çizelge 6 incelendiğinde, R~N(0,1), O~N(0,1) dağılım koşulunda, I. Tip hata oranları Mantel Test için 0.045 - 0.053 ve MTK-OOT için 0.038 - 0.052 arasında değişmiştir. Her iki MİF belirleme testin de I. Tip hata değerlerinin 0.05 α düzeyine yakın veya altında olduğu görülmektedir. Bradley'in esnek referansı dikkate alındığında ilgili koşulda her iki MİF belirleme testinin de I. Tip hata oranları 0.075'in altında kalmış ve ilgili testler I. Tip hata kontrolünü sağlamıştır. İlgili literatürde Kısmi Puan Modeli ya da diğer çok kategorili puanlanan MTK modellerinin (Dereceli Tepki Modeli, Genelleştirilmiş Kısmi Puan Modeli vb.) kullanılarak Mantel Test ve MTK-OOT'nin farklı test koşullarında karşılaştırıldığı birçok araştırma bulunmaktadır. Bu araştırmalarda da referans ve odak grup yetenek dağılımları birim normal dağılım gösterdiği koşullarda Mantel Test ve MTK-OOT'ye ilişkin I. Tip hata oranlarının 0.05 α düzeyine yakın veya altında olduğu ve I. Tip hatayı kontrol ettiği sonucuna varıldığı görülmektedir (Ankenmann ve diğerleri, 1999; Artar, 2007; Bolt, 2002; Chang ve diğerleri,1996; Garrett, 2009; Johnson- Frotman, 2007; Kim ve Cohen, 1997; Kristanjansonn, 2001;Kristanjansonn ve diğerleri, 2005; Thurman, 2009; Tian, 1999).

Grup yetenek dağılımlarının benzer olduğu koşul için artan örneklem büyüklüğüne bağlı olarak I. Tip hata oranları Mantel Test için yükselirken MTK-OOT için düşme eğilimi göstermiştir. Küçük örneklemde (250:250) Mantel Test ve büyük örneklemde (1000:1000) MTK-OOT daha düşük I. Tip hata oranı vermiştir Yani küçük örneklem grubunda, Mantel Test, büyük örneklem grubunda ise MTK-OOT daha iyi I. Tip hata sonuçları vermiştir. Bu durum Mantel Test ve MTK-OOT'nin MİF istatistiğini elde etme yolu dikkate alınarak açıklanabilir.

Bilindiği gibi grupların gözlenen puanlar yoluyla eşleştirildiği Mantel Test gruplar arası madde ortalama farklarını Ki-Kare istatistiğine dayalı olarak değerlendirmektedir. Bu istatistik artan örneklem büyüklüğüne duyarlı olduğu için gerçekte MİF içermeyen ya da çok küçük miktarlarda MİF içeren maddelerin artan örneklem büyüklüğüyle birlikte MİF'li olarak belirlenme olasılığı da artmaktadır. Bu durumda artan örneklem büyüklüğüne bağlı olarak Mantel Test için I. Tip hata oranları yükselme eğilimi göstermektedir.

MTK OOT ise grupları örtük bir değişken üzerinden eşleştirmekte ve model karşılaştırma yoluyla madde parametrelerinin farklılığını değerlendirmektedir. MTK OOT analizleri için kullanılan MULTİLOG programı madde parametrelerinin kestirimi için marginal olabilirlik ve yetenek parametrelerinin kestirimi için maksimum olabilirlik kestirim yöntemini kullanmaktadır. Gruplardan elde edilen madde ve yetenek parametre kestirimlerinin doğruluğunda gruplardaki birey sayısı ve yetenek dağılımların önemli rol oynayabilmektedir (Bahry, 2012; de Ayala, 2009; Reise ve Yu, 1990). Bu bakımdan artan örneklem büyüklüğüne bağlı olarak daha doğru madde ve yetenek parametre kestirimi, MTK-OOT'nin daha düşük I. Tip hata oranları vermesine katkı sağlamış olabilir. Benzer şekilde Mantel Test ve MTK-OOT'yi karşılaştıran araştırma sonuçları da (örn., Ankenmann ve diğerleri, 1999; Garrett, 2009; Johnson- Frotman, 2007) küçük örneklem gruplarında Mantel Test'in MTK-OOT'ye göre daha düşük I. Tip hata oranları verdiğini ancak artan örneklem büyüklüğüne bağlı olarak MTK-OOT'nin daha iyi performans gösterdiğini ortaya koymaktadır.

**Referans ve odak grup yetenek dağılımlarının farklılaştığı koşulda farklı örneklem büyüklükleri için Mantel Test ve MTK-OOT'nin I. Tip Hata Oranları**

Bu kısımda *"Referans ve odak gruba ait yetenek dağılımlarının farklılaştığı koşulda grupların örneklem büyüklüklerindeki değişime bağlı olarak Mantel Test ve MTK-OOT için I.Tip hata oranları nasıl değişmektedir?"* sorusuna yanıt aranmıştır. Bu kapsamda, referans gruba ait yetenek dağılımı aynı kalırken, odak gruba ait yetenek dağılımlarının O~N(-0.5,1) ve O~N(-1,1) olduğu iki farklı koşul incelenmiştir.

**R~N(0,1), O~N(-0.5,1) dağılım koşulunda farklı örneklem büyüklükleri için Mantel Test ve MTK-OOT'nin I. Tip Hata Oranları.** *"Referans gruba ait yetenek dağılımının birim normal dağılım ve odak gruba ait yetenek dağılım ortalamasının -0.5 koşulda grupların örneklem büyüklüklerindeki değişime bağlı olarak Mantel Test ve MTK-OOT için I.Tip hata oranları nasıl değişmektedir?"* sorusuna ilişkin bulgular Çizelge 7'de verilmektedir.

Çizelge 7. R~N(0,1), O~N(-0.5,1) Dağılım Koşulunda Değişen Örneklem Büyüklükleri için Mantel Test ve MTK-OOT'nin I. Tip Hata Oranları

| **Yetenek Dağılımı** | | **Örneklem Büyüklüğü** | | **MİF Belirleme Testi** | |
|---|---|---|---|---|---|
| **Referans** | **Odak** | **Referans** | **Odak** | **Mantel Test** | **MTK-OOT** |
| R~*N*(0,1), O~*N*(-0.5,1) | | 250 | 250 | 0.043 | 0.069 |
| | | 1000 | 250 | 0.048 | 0.071 |
| | | 1000 | 1000 | 0.045 | 0.091 |

Çizelge 7'ye göre, R~N(0,1), O~N(-0.5,1) dağılım koşulunda, I. Tip hata oranları Mantel Test için 0.054 - 0.063 ve MTK-OOT için 0.069 - 0.091 arasında değişmiştir. Bradley'in esnek referansı dikkate alındığında ilgili

koşulda Mantel Test tüm örneklem büyüklüğü koşullarında, MTK-OOT ise büyük örneklem koşulu (1000:1000) hariç I. Tip hata kontrolü sağlamıştır. Tüm örneklem büyüklüğü koşullarında Mantel Test MTK-OOT'ye göre daha düşük I. Tip hata oranı vermiştir. Bu koşulda artan örneklem büyüklüğüne bağlı olarak, Mantel Test'in I.Tip hata oranları azalma eğilimi gösterirken, MTK-OOT için artma eğilimi göstermiştir.

R~N(0,1), O~N(0,1) dağılım koşullarıyla karşılaştırıldığında, I. Tip hata oranları Mantel Test için bir miktar azalırken, MTK-OOT için yükselmiştir. Mantel Test, odak grubun yetenek dağılım ortalamasındaki sapmaya bağlı olarak MTK-OOT'ye göre daha düşük I. Tip hata oranları vermiştir.

**R~N(0,1), O~N(-1,1) dağılım koşulunda farklı örneklem büyüklükleri koşulda Mantel Test ve MTK-OOT İçin I. Tip Hata Oranları.** *"Referans gruba ait yetenek dağılımının birim normal dağılım ve odak gruba ait yetenek dağılım ortalamasının -1 olduğu koşulda grupların örneklem büyüklüklerindeki değişime bağlı olarak Mantel Test ve MTK-OOT için I.Tip hata oranları nasıl değişmektedir?"* sorusuna ilişkin bulgular Çizelge 8'de verilmektedir.

Çizelge 8. R~N(0,1), O~N(-1,1) Dağılım Koşulunda Değişen Örneklem Büyüklükleri için Mantel Test ve MTK-OOT'nin I. Tip Hata Oranları

| **Yetenek Dağılımı** | | **Örneklem Büyüklüğü** | | **MİF Belirleme Testi** | |
|---|---|---|---|---|---|
| **Referans** | **Odak** | **Referans** | **Odak** | **Mantel Test** | **MTK-OOT** |
| R~*N*(0,1), O~*N*(-1,1) | | 250 | 250 | 0.053 | 0.069 |
| | | 1000 | 250 | 0.058 | 0.083 |
| | | 1000 | 1000 | 0.063 | 0.122 |

Çizelge 8'e göre, R~N(0,1), O~N(-1,1) dağılım koşulunda, I.Tip hata oranları Mantel Test için 0.053 - 0.063 ve MTK-OOT için 0.069 – 0.122

arasında değişmiştir. Bradley'in esnek referansı dikkate alındığında ilgili koşulda Mantel Test tüm örneklem büyüklüğü koşullarında, MTK-OOT ise küçük örneklem koşulu (250:250) hariç I. Tip hatayı kontrol etmiştir. Bu koşulda, her iki MİF belirleme testi için de artan örneklem büyüklüğüne bağlı olarak I. Tip hata oranları da artmıştır ve en yüksek I. Tip hata oranları sapma miktarının -1 olduğu büyük örneklem (1000:1000) koşulunda elde edilmiştir. Tüm örneklem büyüklüğü koşullarında Mantel Test MTK-OOT'ye göre daha düşük I. Tip hata oranı vermiştir Grupların yetenek ortalamasındaki sapma miktarlarındaki artış en çok MTK-OOT'nin I. Tip hata oranlarını yükseltmiştir.

Odak grup yetenek dağılım ortalamasındaki sapma miktarları birlikte değerlendirildiğinde, her iki MİF belirleme testi için de artan örneklem büyüklüğü ve odak grup yetenek dağılım ortalamasındaki sapmalara bağlı olarak I. Tip hata oranları artma eğilimi göstermiştir. Pommerivh, Spray ve Parshall (1996), Sweeney (1996) ve Zwick (1990) yetenek dağılımındaki bozulmaların MİF belirlemede tutarsızlıklar oluşturduğuna dikkat çekmektedir. Genel olarak bakıldığında odak grubun yetenek dağılım ortalamasındaki sapmaya bağlı olarak her iki MİF belirleme testinin de I. Tip hata oranları yükselmiştir.

Mantel Test için I. Tip hata oranlarındaki artış, grupların yetenek ortalamalarındaki değişime bağlı olarak bazı ham puan aralıklarında beklenen yetenek ($\theta$) değerlerinin de farklılaşmasından kaynaklanabilir. Böylece gözlenen madde puan ortama farkları belirli puan aralıkları için daha belirgin hale gelecektir. Bu araştırma bulgularına paralel olarak yapay verilere yapılan diğer birçok araştırmada grupların yetenek dağılım ortalamasındaki sapmaya bağlı olarak Mantel Test için I. Tip hata oranları yükselmiş ancak bu değerler Bradley'in ölçütleri içinde yer almış ve Mantel Test'in artan sapma miktarına bağlı olarak I. Tip hatayı iyi kontrol ettiği sonucuna ulaşılmıştır (örn., Chang ve diğerleri,1996; Demars, 2007; Garrett, 2009; Kristanjansonn ve diğerleri, 2005; Thurman, 2009; Tian, 1999).

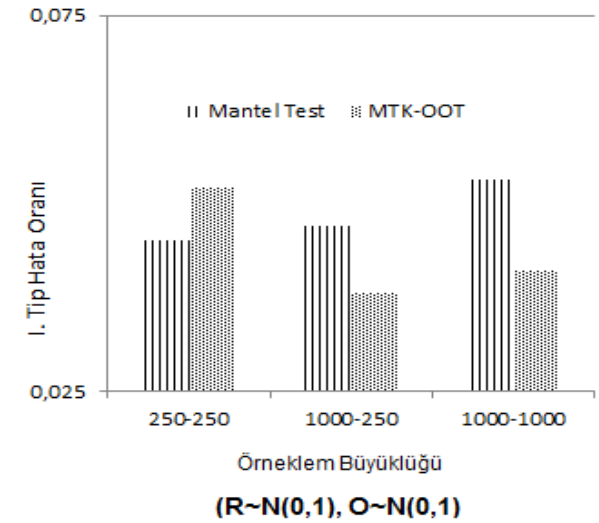
Her iki test için de en yüksek I. Tip hata değerleri büyük örneklem büyüklüğü ve odak grubun yetenek ortalamasındaki artan sapma miktarı koşulunda elde edilmiştir. Mantel Test ile karşılaştırıldığında, artan örneklem büyüklüğü ve grupların yetenek ortalamasındaki sapmaya bağlı olarak MTK-OOT için I. Tip hata değerlerindeki artış daha yüksek olmuştur. Grupların yetenek dağılımlarındaki sapmalar karşısında MTK-OOT I. Tip hatayı kontrol etmekte zayıf kalmış ve MTK-OOT'nin gerçekte MİF içermeyen bir maddeyi MİF'li olarak belirleme olasılığını yükseltmiştir. Bu durum θ dağılımlarındaki kaymaların madde parametrelerinin kestirimi üzerindeki etkisiyle açıklanabilir.

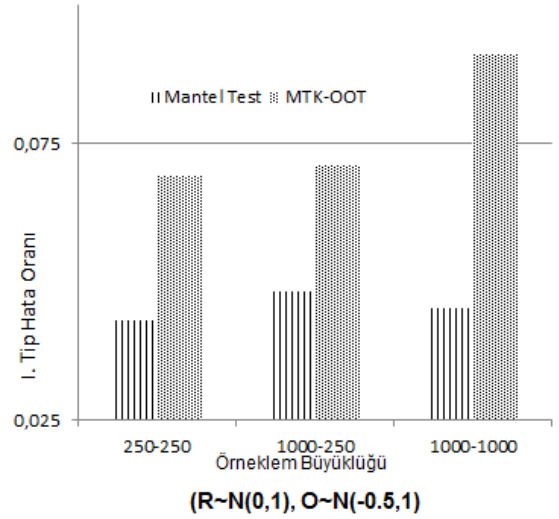
Odak grup yetenek dağılım ortalamasının referans gruba göre daha düşük olduğu durumlarda, madde parametreleri (adım güçlüğü parametresi) aşağı doğru çekilen yetenek dağılımıyla bağlantılı olarak daha yüksek değerler alabilecektir. Diğer bir deyişle, maddeye ait adım güçlüğü parametresi daha düşük θ değerleri için daha yüksek değerler alabilir. Buna bağlı olarak madde parametrelerinin her iki grupta da eşitlendiği dar modelden elde edilen parametre kestirimleri için hata değerleri yükselecek ve bu modelden edilen model-veri uyum istatistiği geniş modele göre daha kötü olacaktır. Çünkü geniş modelde ilgili maddeye ait parametre kestirimleri her bir grup için ayrı ayrı yapıldığı için daha doğru kestirimler elde edilebilecektir. Böylece daha iyi uyum gösteren geniş model ilgili maddenin MİF gösteriyor olarak işaretlenmesine neden olabilir.

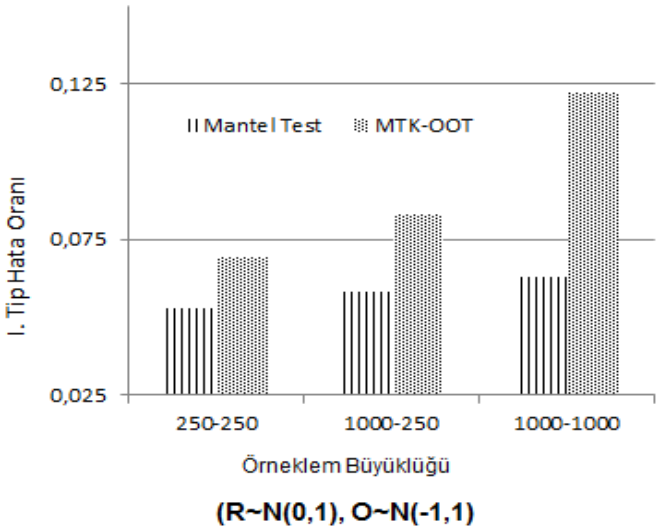
Grupların yetenek dağılımlarındaki sapmalar karşısında MTK-OOT I. Tip hatayı kontrol etmekte zayıf kalması madde parametreleri ile θ parametreleri arasındaki uyumla da açıklanabilir. Bilindiği gibi madde parametre dağılımı θ parametreleri dağılımına yaklaştıkça daha doğru kestirimler yapılmaktadır (Bahry, 2012). Araştırmada elde edilen adım güçlüğü parametreleri, ortalaması 0 standart sapması 1 olan birim normal dağılımdan elde edilmiştir. Odak grubun θ dağılımındaki kaymalar adım güçlüğü parametreleriyle olan uyumu azaltmakta ve buna bağlı olarak en çok olabilirlik ve marjinal olasılık yöntemine dayalı parametre kestirimindeki hatalar artmaktadır. Bu durum gruplar arasında gözlenen farklılığın kaynağı olarak açıklanabilir.

Bu araştırma bulgularına paralel olarak yapay verilerle yapılan diğer araştırmalarda grupların yetenek dağılım ortalamasındaki sapmaya bağlı olarak MTK-OOT için I. Tip hata oranları yükselmiş ve MTK-OOT I. Tip hatayı kontrol etmekte zorlanmıştır (Bolt, 2002; Garrett, 2009; Stark ve diğerleri, 2006). Ancak Ankenmann ve diğerleri (1999) ve Kim ve Cohen (1997) yaptıkları çalışmada MTK-OOT için farklılaşan yetenek dağılım koşullarında normal dağılım koşullarına benzer I. Tip hata değerleri elde etmiştir. Bu çalışmalarda kullanılan bağımsız değişkenlerin ve bu değişkenler arasındaki etkileşime bağlı olarak MTK-OOT için farklı I. Tip hata performansı elde edilmiş olabilir. Örneğin Kim ve Cohen (1997) gruplara ait yetenek dağılımlarının benzer ve farklı olduğu durumları bu araştırmadan farklı olarak (R~N(1,1), O~N(1,1), (R~N(1,1), O~N(0,1) biçiminde oluşturmuştur. Ayrıca, her iki çalışmada da Dereceli Tepki Modeline göre simule edilmiş veriler kullanılmıştır. Kim ve Cohen (1997) sadece çok kategorili puanlanan maddelerden oluşan bir test formu kullanırken, Ankenmann ve diğerleri (1999) tarafından kullanılan test ayrıca ikili puanlanan maddeler de içermektedir.

Son durumda, Mantel Test ve MTK-OOT için grupların değişen örneklem büyüklükleri ve yetenek dağılım ortalamalarına bağlı olarak I. Tip hata bulgularını özetleyen grafikler aşağıda verilmektedir.




Şekil 10. Farklı Yetenek Dağılımları ve Örneklem Büyüklüklerinde Mantel Test ve MTK-OOT'nin I. Tip Hata Oranları

Şekil 10 incelendiğinde, ilgili MİF belirleme testine ilişkin I. Tip hata oranları ile grupların örneklem büyüklüğü ve yetenek dağılım özellikleri arasında bir örüntü ortaya çıktığı görülmektedir. Bu örüntüye göre, artan örneklem büyüklüğü ve grupların yetenek dağılımlarındaki farklılaşmaya bağlı olarak I. Tip hata oranları; yani gerçekte MİF içermeyen bir maddenin MİF'li olarak belirlenme olasılığı yükselmiştir. En yüksek I. Tip hata oranları büyük örneklem (1000:1000) ve sapma koşullarında (O~*N*(-1,1)) elde edilmiştir. Tüm yetenek dağılımları koşulları için Mantel Test küçük örneklemlerde I. Tip hatayı MTK-OOT'ye göre daha düşük I. Tip hata oranları vermiştir. Odak grubun yetenek dağılımındaki sapma MTK-OOT'nin I. Tip hata oranlarını Mantel Test'e göre daha fazla yükseltmiştir. Bu durumda artan örneklem büyüklüğü ve odak grubun dağılımındaki sapma, Mantel Test'e göre MTK-OOT'nin gerçekte MİF içermeyen bir maddeyi MİF'li olarak belirleme olasılığını daha fazla yükseltmiştir.

## Mantel Test ve MTK-OOT İçin İstatistiksel Güç Oranları

Araştırmanın bu kısmında, grupların değişen yetenek dağılım ortalamaları ve örneklem büyüklükleri, MİF örüntüsü ve MİF miktarı için ilgili MİF belirleme testlerinin istatistiksel güç oranlarına ilişkin bulgu ve yorumlara yer verilmiştir.

### R~N(0,1), O~N(0,1) dağılım koşulunda Mantel Test ve MTK-OOT'nin İstatistiksel Güç Oranları

Güç çalışmaları kapsamında ilk alt amaç olarak Mantel Test ve MTK-OOT için, referans ve odak gruba ait yetenek dağılımlarının birim normal dağılım özelliği gösterdiği koşulda, değişen grup örneklem büyüklüklerine MİF örüntüsü ve MİF miktarına bağlı olarak güç oranları incelenmiştir. Bu amaçla ilgili koşullar düşük ve yüksek MİF örüntüsü bağlamında incelenmiştir.

**Düşük MİF Örüntüsü Koşulu.** "*Referans ve odak gruba ait yetenek dağılımlarının birim normal dağılım özelliği gösterdiği düşük MİF örüntüsü için, değişen grup örneklem büyüklüklerine ve MİF miktarına bağlı olarak Mantel Test ve MTK-OOT'nin güç oranları nasıl değişmektedir?"* sorusuna ilişkin bulgu ve yorumlara yer verilmiştir.

İlgili koşul için değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak güç oranları Çizelge 9'da verilmektedir.

Çizelge 9. R~N(0,1), O~N(0,1) Dağılım Koşulu ve Düşük MİF Örüntüsünde, Değişen Örneklem Büyüklükleri ve MİF Miktarına göre Mantel Test ve MTK-OOT'nin İstatistiksel Güç Oranları

| **Yetenek Dağılımı** | | **Örneklem Büyüklüğü** | | **MİF belirleme Testi** | | | |
|---|---|---|---|---|---|---|---|
| | | | | **Mantel Test** | | **MTK-OOT** | |
| | | | | **MİF Miktarı** | | | |
| **Referans** | **Odak** | **Referans** | **Odak** | **0.43** | **0.64** | **0.43** | **0.64** |
| R~*N*(0,1), O~*N*(0,1) | | 250 | 250 | 0.292 | 0.543 | 0.313 | 0.700 |
| | | 1000 | 250 | 0.453 | 0.773 | 0.570 | 0.890 |
| | | 1000 | 1000 | 0.773 | 0.986 | 0.933 | 1 |

Çizelge 9 incelendiğinde, R~N(0,1), O~N(0,1) dağılım ve düşük MİF örüntüsü için, MİF miktarının 0.42 olduğu koşulda istatistiksel güç oranları Mantel Test için 0.292 ile 0.773 ve MTK-OOT için 0.313 ile 0.933 arasında değişmiştir. Bu MİF miktarı için tüm örneklem büyüklüklerinde Mantel Test'in istatistiksel güç oranları 0.80'in altında kalmıştır; yani Mantel Test gerçekte MİF içeren bir maddeyi MİF'li olarak belirlemede yetersiz kalmıştır. MTK-OOT sadece büyük örneklem (1000:1000) koşulunda 0.80'den büyük istatistiksel güç değeri vermiştir ve MİF'i belirlemede yeterli olmuştur.

MİF miktarının 0.64 olduğu koşulda ise istatistiksel güç oranları Mantel Test için 0.543 ile 0.986 MTK-OOT için 0.700 ile 1 arasında değişmiştir. Bu MİF miktarı için Mantel Test büyük örneklemde (1000:1000), MTK-OOT ise küçük örneklem (250:250) dışındaki koşullarda 0.80'in üzerinde istatistiksel güç değerleri vermiş ve MİF'i belirlemede yeterli olmuştur. Her iki MİF belirleme testi için de, düşük MİF örüntüsünde artan MİF miktarı ve örneklem büyüklüğüne bağlı olarak ilgili testlerinin istatistiksel güç oranlarını yükselmiştir.

İlgili literatür incelendiğinde, Mantel Test ve MTK-OOT için istatistiksel güç çalışmalarında ağırlıklı olarak tüm adım güçlüğü parametrenin sabit miktarda odak grup için yüksek olduğu *sabit* (constant) MİF örüntüsünün çalışıldığı dikkat çekmektedir (örn., Ankenmann ve diğerleri, 1999; Bolt, 2002; Chang ve diğerleri,1996; Garrett, 2009; Johnson- Frotman, 2007; Kristanjansonn ve diğerleri, 2005). Bu ilgili araştırma sonuçları da MİF belirleme testlerinin performansının öncelikli olarak örneklem büyüklüğünden etkilendiğini göstermiştir.

Artar (2007), Fidalgo ve Bartram (2010), Stark ve diğerleri (2006) ve Thurman (2009) tarafından yapılan çalışmalar düşük MİF örüntüsü koşulunu içermektedir. Artar (2007) ve Stark ve diğerleri (2006), düşük MİF örüntüsü koşulunda MTK-OOT için artan MİF miktarı ve örneklem büyüklüğüne bağlı olarak istatistiksel güç oranlarının yükseldiğini belirtmektedir. Stark ve diğerleri (2006) tarafından yapılan çalışmada da düşük MİF örüntüsünde MTK-OOT için bu çalışmadakine benzer istatistiksel güç değerleri edilmiştir. Fidalgo ve Bartram (2010) ve Thurman (2009) tarafından yapılan çalışmalarda ise Dereceli tepki Modeli ve Genelleştirilmiş Kısmi Puan Modeli kullanılarak; 500:500 örneklem büyüklüğünde, 0.40 MİF miktarı ve düşük MİF örüntüsünde Mantel Test için yeterli istatistiksel güç oranları elde edilmemiştir.

Bu araştırma bulguları ayrıca tüm örneklem büyüklüğü ve MİF miktarı koşullarında MTK-OOT'in istatistiksel güç oranlarının, Mantel Test'e göre daha yüksek olduğunu göstermektedir. Bu durum ilgili testlerin MİF belirleme yaklaşımıyla açıklanabilir. Mantel Test'in MİF belirleme süreci grupları gözlenen toplam puanlar yoluyla eşleştirerek gruplar arası beklenen puanları karşılaştırmasını içermektedir. MTK-OOT ise bir örtük değişken kullanarak

grupları eşleştirmekte ve tek tek gruplar arası adım güçlüğü parametrelerinin farklılığını test etmektedir. Bu durum MTK-OOT'nin Mantel Test'e göre daha hassas sonuçlar vermesine katkıda bulunabilir.

Bu kapsamda ayrıca, Mantel Test ve MTK-OOT için referans ve odak gruba ait yetenek dağılımlarının birim normal dağılım özelliği gösterdiği yüksek MİF örüntüsü için güç oranları incelenmiştir.

**Yüksek MİF Örüntüsü Koşulu.** "*Referans ve odak gruba ait yetenek dağılımlarının birim normal dağılım özelliği gösterdiği yüksek MİF örüntüsü için, değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak Mantel Test ve MTK-OOT'nin güç oranları nasıl değişmektedir?"* sorusuna ilişkin bulgu ve yorumlara yer verilmiştir.

İlgili koşul için değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak Mantel Test ve MTK-OOT için güç oranları Çizelge 10'da verilmektedir.

Çizelge 10. R~N(0,1), O~N(0,1) Dağılım Koşulu ve Yüksek MİF Örüntüsünde, Değişen Örneklem Büyüklükleri ve MİF Miktarına göre Mantel Test ve MTK-OOT'nin İstatistiksel Güç Oranları

| **Yetenek Dağılımı** | | **Örneklem Büyüklüğü** | | **MİF belirleme Testi** | | | |
|---|---|---|---|---|---|---|---|
| | | | | **Mantel Test** | | **MTK-OOT** | |
| | | | | **MİF Miktarı** | | | |
| **Referans** | **Odak** | **Referans** | **Odak** | **0.43** | **0.64** | **0.43** | **0.64** |
| R~*N*(0,1), O~*N*(0,1) | | 250 | 250 | 0.111 | 0.142 | 0.203 | 0.283 |
| | | 1000 | 250 | 0.143 | 0.170 | 0.236 | 0.333 |
| | | 1000 | 1000 | 0.216 | 0.383 | 0.446 | 0.751 |

Çizelge 10 incelendiğinde, R~N(0,1), O~N(0,1) dağılım ve yüksek MİF örüntüsü için, MİF miktarının 0.42 olduğu koşulda istatistiksel güç oranları Mantel Test için 0.111- 0.216 ve MTK-OOT için 0.151 - 0.483; MİF miktarının 0.64 olduğu koşulda ise istatistiksel güç oranları Mantel Test için 0.142 ile 0.383 MTK-OOT için 0.283 ile 0.751 arasında değişmiştir. Yüksek MİF örüntüsü için, değişen grup örneklem büyüklüğü ve MİF miktarına bağlı olarak her iki MİF belirleme testinin istatistiksel güç oranları artış göstermiştir. Ancak her iki MİF miktarı ve örneklem büyüklüğü için elde edilen istatistiksel güç oranları Mantel Test ve MTK-OOT için 0.80'in çok altında kalmıştır.

Fidalgo ve Bartram (2010) ve Thurman (2009) da tarafından yapılan çalışmalarda 500:500 örneklem büyüklüğünde, 0.40 MİF miktarı ve yüksek MİF örüntüsünde farklı güçlükteki adım güçlüğü parametrelerinin olduğu koşullarda Mantel Test için yeterli istatistiksel güç oranları elde edilmemiştir. Ancak Artar (2007), farklı örneklem büyüklüklerinde (600, 1200 ve 2400) her iki MİF miktarı (0.43 ve 0.53) koşulunda MTK-OOT için yüksek istatistiksel güç oranları elde etmiştir. Bu durum Artar'ın (2007) kullandığı adım güçlüğü parametrelerinin özellikleriyle açıklanabilir.

Artar'ın (2007) ilgili koşulda adım güçlüğü parametreleri (-1,25, -0.50 ve 0.75), bu çalışmada kullanılan parametrelere göre daha kolaydır. Bu durumda zaten kolay bir maddenin en son kategorisine MİF ekleyerek onu daha zor hale getirmek tüm yetenek aralığındaki bireyleri etkilediği için bu durumda grupların madde parametreleri arasında fark belirlemek daha kolay hale gelmektedir. Sonuç olarak yüksek MİF örüntüsünde kolay bir madde için MİF belirlemek daha olası olduğu için Artar (2007) çalışmasında bu çalışmaya göre daha yüksek istatistiksel güç oranları elde etmiş olabilir. Yüksek MİF örüntüsünde yeterli istatistiksel güç oranları elde edemeyen Thurman (2009) da çalışmasında daha zor adım güçlüğü parametreleri (0,1,2) kullanmıştır.

Referans ve odak gruba ait yetenek dağılımlarının birim normal dağılım özelliği gösterdiği koşullar için, düşük MİF örüntüsüyle karşılaştırıldığında, yüksek MİF örüntüsü için istatistiksel güç oranları tüm araştırma koşullarına çok düşüktür. Bu durum bu çalışmada kullanılan adım güçlüğü parametrelerinin özellikleriyle açıklanabilir. Zor bir maddenin en son kategorisindeki MİF (yüksek MİF örüntüsü) yetenek aralıklarının en üstünde

yer alan az sayıdaki bireyi etkilemektedir. Buna bağlı olarak gruplar arasında MİF bulmak zorlaştığı için daha düşük istatistiksel güç oranları elde edilmektedir. Ancak zor bir maddeyi ilk adımda daha zor hale getirmek (düşük MİF örüntüsü) tüm yetenek aralığındaki bireyleri etkilediği için daha yüksek istatistiksel güç oranları elde edilmesi daha olasıdır. Benzer durum zor adım güçlüğü parametreleri kullanan Fidalgo ve Bartram (2010) ve Thurman'ın (2009) çalışmasında da ortaya çıkmıştır.

Güç çalışmaları kapsamında, referans ve odak grubun yetenek dağılım ortalamalarının farklılaştığı koşul için, referans gruba ait yetenek dağılımının birim normal dağılım ve odak gruba ait yetenek dağılım ortalamasının -0.5 ve -1 olduğu iki farklı koşul incelenmiştir.

### R~N(0,1), O~N(-0.5,1) dağılım koşulunda Mantel Test ve MTK-OOT'nin İstatistiksel Güç Oranları

Odak gruba ait yetenek dağılım ortalamasının -0.5 olduğu koşul düşük ve yüksek MİF örüntüsü bağlamında incelenmiştir.

**Düşük MİF Örüntüsü Koşulu.** "*Referans gruba ait yetenek dağılımının birim normal dağılım ve odak gruba ait yetenek dağılım ortalamasının -0.5 olduğu düşük MİF örüntüsü için, değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak Mantel Test ve MTK-OOT'nin güç oranları nasıl değişmektedir?"* sorusuna ilişkin bulgu ve yorumlara yer verilmiştir.

İlgili koşulda, değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak Mantel Test ve MTK-OOT için güç oranları Çizelge 11'de verilmektedir.

Çizelge 11. R~N(0,1), O~N(-0.5,1) Dağılım Koşulu ve Düşük MİF Örüntüsünde, Değişen Grup Örneklem Büyüklükleri ve MİF Miktarına göre Mantel Test ve MTK-OOT'nin İstatistiksel Güç Oranları

| Yetenek Dağılımı | | Örneklem Büyüklüğü | | MİF belirleme Testi | | | |
|---|---|---|---|---|---|---|---|
| | | | | Mantel Test | | MTK-OOT | |
| | | | | MİF Miktarı | | | |
| **Referans** | **Odak** | **Referans** | **Odak** | **0.43** | **0.64** | **0.43** | **0.64** |
| R~*N*(0,1), O~*N*(-0.5,1) | | 250 | 250 | 0.315 | 0.563 | 0.406 | 0.753 |
| | | 1000 | 250 | 0.466 | 0.840 | 0.636 | 0.954 |
| | | 1000 | 1000 | 0.820 | 0.986 | 0.845 | 0.993 |

Çizelge 11'e göre, R~N(0,1), O~N(-0.5,1) dağılım ve düşük MİF örüntüsü için, MİF miktarının 0.42 olduğu koşulda istatistiksel güç oranları Mantel Test için 0.315 ile 0.820 ve MTK-OOT için 0.406 ve 0.845 arasında değişmiştir. MİF miktarının 0.42 olduğu durumda, her iki MİF belirleme testi sadece büyük örneklem (1000:1000) koşulunda 0.80'den büyük istatistiksel güç oranları vermiş ve MİF'i belirlemede yeterli olmuştur.

MİF miktarının 0.64 olduğu koşulda ise istatistiksel güç oranları Mantel Test için 0.563 ile 0.986 ve MTK-OOT için 0.753 ve 0.993 arasında değişmiştir. Bu durumda, küçük örneklem (250:250) koşulu hariç Mantel Test ve MTK-OOT MİF'i belirlemede yeterli olmuştur.

Her iki MİF belirleme testi için de, artan MİF miktarı ve örneklem büyüklüğü ilgili testlerin istatistiksel güç oranlarını yükseltmiştir. Buna bağlı olarak ilgili MİF belirleme testleri için en yüksek istatistiksel güç oranları 0.64 MİF miktarı ve 1000:1000 örneklem koşulunda elde edilmiştir.

**Yüksek MİF Örüntüsü Koşulu.** "*Referans gruba ait yetenek dağılımının birim normal dağılım ve odak gruba ait yetenek dağılım ortalamasının -0.5 olduğu yüksek MİF örüntüsü için, değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak Mantel Test ve MTK-OOT'nin güç oranları nasıl değişmektedir?"* sorusuna ilişkin bulgu ve yorumlara yer verilmiştir.

İlgili koşulda, değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak Mantel Test ve MTK-OOT için güç oranları Çizelge 12'de verilmektedir.

Çizelge 12. R~N(0,1), O~N(-0.5,1) Dağılım Koşulu ve Yüksek MİF Örüntüsünde, Değişen Grup Örneklem Büyüklükleri ve MİF Miktarına göre Mantel Test ve MTK-OOT'nin İstatistiksel Güç Oranları

| **Yetenek Dağılımı** | | **Örneklem Büyüklüğü** | | **MİF belirleme Testi** | | | |
|---|---|---|---|---|---|---|---|
| | | | | **Mantel Test** | | **MTK-OOT** | |
| | | | | **MİF Miktarı** | | | |
| **Referans** | **Odak** | **Referans** | **Odak** | **0.42** | **0.64** | **0.42** | **0.64** |
| R~*N*(0,1), O~*N*(-0.5,1) | | 250 | 250 | 0.100 | 0.111 | 0.151 | 0.283 |
| | | 1000 | 250 | 0.130 | 0.135 | 0.220 | 0.333 |
| | | 1000 | 1000 | 0.166 | 0.250 | 0.483 | 0.751 |

Çizelge 12'ye göre, R~N(0,1), O~N(-0.5,1) dağılım ve yüksek MİF örüntüsü için, MİF miktarının 0.42 olduğu koşulda istatistiksel güç oranları Mantel Test için 0.100 ile 0.166 ve MTK-OOT için 0.151 ve 0.483 arasında değişmiştir. MİF miktarının 0.64 olduğu koşulda ise istatistiksel güç oranları Mantel test için 0.111 ile 0.250 ve MTK-OOT için 0.283 ve 0.751 arasında değişmiştir. Her iki MİF miktarı için de tüm örneklem büyüklüğü koşullarında

Mantel Test ve MTK-OOT'nin istatistiksel güç oranları 0.80'nin altında kalmıştır; yani her iki MİF belirleme testi de gerçekte MİF içeren bir maddeyi MİF'li belirlemede yetersiz kalmıştır.

.

**R~N(0,1), O~N(-1,1) dağılım koşulunda Mantel Test ve MTK-OOT'nin İstatistiksel Güç Oranları**

Odak gruba ait yetenek dağılım ortalamasının -1 olduğu koşul düşük ve yüksek MİF örüntüsü bağlamında incelenmiştir.

**Düşük MİF örüntüsü Koşulu.** "*Referans gruba ait yetenek dağılımının birim normal dağılım ve odak gruba ait yetenek dağılım ortalamasının -1 olduğu düşük MİF örüntüsü için, değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak Mantel Test ve MTK-OOT'nin güç oranları nasıl değişmektedir?"* sorusuna ilişkin bulgu ve yorumlara yer verilmiştir.

İlgili koşulda değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak Mantel Test ve MTK-OOT için güç oranları Çizelge 13'te verilmektedir.

Çizelge 13. R~N(0,1), O~N(-1,1) Dağılım Koşulu ve Düşük MİF Örüntüsünde, Değişen Grup Örneklem Büyüklükleri ve MİF Miktarına göre Mantel Test ve MTK-OOT'nin İstatistiksel Güç Oranları

| **Yetenek Dağılımı** | | **Örneklem Büyüklüğü** | | **MİF belirleme Testi** | | | |
|---|---|---|---|---|---|---|---|
| | | | | **Mantel Test** | | **MTK-OOT** | |
| | | | | **MİF Miktarı** | | | |
| **Referans** | **Odak** | **Referans** | **Odak** | **0.42** | **0.64** | **0.42** | **0.64** |
| R~*N*(0,1), O~*N*(-1,1) | | 250 | 250 | 0.316 | 0.513 | 0.516 | 0.836 |
| | | 1000 | 250 | 0.476 | 0.763 | 0.740 | 0.953 |
| | | 1000 | 1000 | 0.833 | 0.986 | 0.950 | 1 |

Çizelge 13'e göre, R~N(0,1), O~N(-1,1) dağılım ve düşük MİF örüntüsü için, MİF miktarının 0.42 olduğu koşulda Mantel Test için istatistiksel güç oranları 0.316 ile 0.833 ve MTK-OOT için 0.516 ve 0.950 arasında değişmiştir. MİF miktarının 0.42 olduğu durumda, her iki MİF belirleme testi için de sadece büyük örneklem koşulunda 0.80'nin üzerinde istatistiksel güç oranları elde edilmiştir ve bu koşulda testler MİF'i belirlemede yeterli olmuştur. MİF miktarının 0.64 olduğu koşulda ise istatistiksel güç oranları Mantel test için 0.513 ile 0.986 ve MTK-OOT için 0.836 ve 1 arasında değişmiştir. Bu durumda, Mantel Test küçük örneklem koşulu dışında, MTK-OOT ise tüm örneklem büyüklüğü koşullarında MİF'i belirlemede yeterli olmuştur.

Düşük MİF örüntüsü ve odak grup yetenek dağılım ortalamasının, referans grup yetenek dağılımına göre daha düşük ortalama yönünde sapma gösterdiği koşullar genel olarak incelendiğinde, her iki MİF miktarı için de ilgili MİF belirleme testinin istatistiksel gücü birim normal dağılım koşullarına göre bir miktar artmıştır. Tüm koşullarda MTK-OOT'nin istatistiksel güç değerleri Mantel Test'ten daha yüksek olmuştur.

MİF miktarının 0.42 lojit birim olduğu koşullar için artan sapma miktarına bağlı olarak Mantel Test için istatistiksel güç oranları birbirine yakın değerler alırken, MTK-OOT için bu değerler bir miktar artmıştır. MİF miktarının 0.64 lojit birim olduğu koşullarda ise, ilgili MİF belirleme testleri için grup yetenek dağılım ortalamalarındaki sapma ve değişen örneklem büyüklükleri arasında belirgin bir örüntü elde edilmemiştir. Her iki MİF belirleme testi de artan MİF miktarına bağlı olarak 0.64 lojit birim MİF miktarı için daha yüksek istatistiksel güç oranları vermiştir.

**Yüksek MİF Örüntüsü Koşulu.** "*Referans gruba ait yetenek dağılımının birim normal dağılım ve odak gruba ait yetenek dağılım ortalamasının -1 olduğu yüksek MİF örüntüsü için, değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak Mantel Test ve MTK-OOT için güç oranları nasıl değişmektedir?"* sorusuna ilişkin bulgu ve yorumlara yer verilmiştir.

İlgili koşulda değişen grup örneklem büyüklükleri ve MİF miktarına bağlı olarak Mantel Test ve MTK-OOT için güç oranları Çizelge 14'te verilmektedir.

Çizelge 14. R~N(0,1), O~N(-1,1) Dağılım Koşulu ve Yüksek MİF Örüntüsünde, Değişen Grup Örneklem Büyüklükleri ve MİF Miktarına göre Mantel Test ve MTK-OOT'nin İstatistiksel Güç Oranları

| **Yetenek Dağılımı** | | **Örneklem Büyüklüğü** | | **MİF belirleme Testi** | | | |
|---|---|---|---|---|---|---|---|
| | | | | **Mantel Test** | | **MTK-OOT** | |
| | | | | **MİF Miktarı** | | | |
| **Referans** | **Odak** | **Referans** | **Odak** | **0.43** | **0.64** | **0.43** | **0.64** |
| R~*N*(0,1), O~*N*(-1,1) | | 250 | 250 | 0.090 | 0.097 | 0.174 | 0.210 |
| | | 1000 | 250 | 0.116 | 0.129 | 0.182 | 0.283 |
| | | 1000 | 1000 | 0.133 | 0.166 | 0.490 | 0.700 |

Çizelge 14'e göre, R~N(0,1), O~N(-1,1) dağılım ve yüksek MİF örüntüsü için, MİF miktarının 0.42 olduğu koşulda Mantel Test için istatistiksel güç oranları 0.090 ile 0.133 ve MTK-OOT için 0.174 ve 0.490 arasında değişmiştir. MİF miktarının 0.64 olduğu koşulda ise istatistiksel güç oranları Mantel test için 0.097 ile 0.166 ve MTK-OOT için 0.210 ve 0.700 arasında değişmiştir. Her iki MİF miktarı için de, tüm örneklem büyüklüğü koşullarında Mantel Test ve MTK-OOT'nin istatistiksel güç oranları 0.80'nin altında kalmış ve testler bu koşullarda MİF'i belirlemede yetersiz kalmıştır.

Odak grup yetenek dağılım ortalamasının, referans grup yetenek dağılımına göre sapma gösterdiği yüksek MİF örüntüsü koşulları için, her iki MİF miktarı için de ilgili MİF belirleme testlerinin istatistiksel gücü birim normal dağılım koşullarına göre düşüş eğilimi göstermiştir. Bu MİF örüntüsü için değişen grup yetenek dağılımları, örneklem büyüklüğü ve MİF miktarında her iki test de MİF'i belirlemede yetersiz kalmıştır. Bu koşulda, artan grupların dağılım ortalamalarındaki sapma miktarı ile grupların örneklem büyüklüğü ve

MİF miktarı arasında belirgin bir örüntü ortaya çıkmamıştır. Düşük MİF örüntüsüyle karşılaştırıldığında, yüksek MİF örüntüsünün tüm araştırma koşullarına ait istatistiksel güç oranları çok düşüktür.

Mantel Test ve MTK-OOT için grupların yetenek ortalamasındaki sapma miktarı ve değişen MİF örüntüsü birlikte değerlendirildiğinde, her iki MİF örüntüsünde de ilgili testlerin MİF belirlemedeki gücü odak grubun yetenek ortalamasındaki sapmaya bağlı olarak bir miktar yükselmiştir. İstatistiksel güç oranlarındaki bu yükseliş değerlendirilirken istatistiksel güç ile I. Tip hata arasındaki ilişkiye dikkat edilmelidir. Örneğin büyük örneklem koşullarında yüksek istatistiksel güç oranı odak grup dağılım ortalamasında meydana gelen sapmanın MTK-OOT'nin I. Tip hata oranlarını yükseltmesiyle açıklanabilir. Ankenmann ve diğerleri (1999), Jodoin ve Gierl (2001) ve Li ve Stout (1996), MİF belirleme testlerin istatistiksel güçlerinin yorumlanabilmesi için I. Tip hata oranlarının karşılaştırılabilir olması gerektiğini, çünkü şişirilmiş I.Tip hata oranlarının bir testin istatistiksel gücünü yapay olarak artırdığını vurgulamaktadır. Bu yüzden büyük örneklem koşulunda MTK-OOT'nin istatistiksel gücü I. Tip hata oranlarına bağlı olarak olduğundan daha yüksek değerler almış olabilir. Benzer şekilde Mantel Test için kullanılan eşleştirme değişkeni MİF'li maddeleri de ham puanlardan oluşmaktadır. Literatürde MİF içeren eşleştirme değişkeninin MİF belirleme testlerinin I. Tip hata oranlarını şişirdiği vurgulanmaktadır (Stark ve diğerleri, 2006; Wodds, 2011). Bu durumda farklı MİF örüntüsü, MİF miktarı ve artan sapma miktarına bağlı olarak istatistiksel güç oranlarında yapay bir yükseliş meydana geldiği ifade edilebilir. Genel olarak farklı sapma koşulları için ilgili testlerin istatistiksel gücü birbirine yakın değerler almıştır. Benzer durum Fidalgo ve Bartram (2010) ve Stark ve diğerleri (2006) tarafından yapılan çalışmalar için de geçerlidir. Sabit ya da dengeli MİF örüntüsünü değişen yetenek ortalamasına bağlı olarak inceleyen diğer çalışmalar (örn., Bolt, 2002; Johnson- Frotman, 2007; Kristanjansonn, 2001; Kristanjansonn ve diğerleri, 2005), Mantel test ya da MTK-OOT'nin istatistiksel güç oranlarının odak grubun yetenek dağılımındaki farklardan az etkilendiğini ortaya koymaktadır. Benzer şekilde bu ilgili araştırma sonuçlarında da ilgili testlerin istatistiksel güç oranlarının öncelikli olarak örneklem büyüklüğünden etkilendiği vurgulanmaktadır.

Son durumda Mantel Test ve MTK-OOT'nin istatistiksel güç oranları Çizelge 15'de verilmektedir.

Çizelge 15. Mantel Test ve MTK-OOT'nin İstatistiksel Güç Oranları

| Yetenek Dağılımı | | Örneklem Büyüklüğü | | MİF Belirleme Testi | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Mantel Test | | | | | | MTK-OOT | |
| | | | | MİF örüntüsü | | | | | | | |
| | | | | Düşük | | Yüksek | | Düşük | | Yüksek | |
| | | | | MİF Miktarı | | | | | | | |
| R | O | R | O | 42 | 64 | 42 | 64 | 42 | 64 | 42 | 64 |
| R~*N*(0,1), O~*N*(0,1) | | 250 | 250 | 0.292 | 0.543 | 0.111 | 0.142 | 0.313 | 0.700 | 0.203 | 0.283 |
| | | 1000 | 250 | 0.453 | 0.773 | 0.143 | 0.170 | 0.570 | 0.890 | 0.236 | 0.333 |
| | | 1000 | 1000 | 0.773 | 0.986 | 0.216 | 0.383 | 0.933 | 1 | 0.446 | 0.751 |
| R~*N*(0,1), O~*N*(-0.5,1) | | 250 | 250 | 0.315 | 0.563 | 0.100 | 0.111 | 0.406 | 0.753 | 0.151 | 0.283 |
| | | 1000 | 250 | 0.466 | 0.840 | 0.130 | 0.135 | 0.636 | 0.954 | 0.220 | 0.333 |
| | | 1000 | 1000 | 0.820 | 0.986 | 0.166 | 0.250 | 0.845 | 0.993 | 0.483 | 0.751 |
| R~*N*(0,1), O~*N*(-1,1) | | 250 | 250 | 0.316 | 0.513 | 0.090 | 0.097 | 0.516 | 0.836 | 0.174 | 0.210 |
| | | 1000 | 250 | 0.476 | 0.763 | 0.116 | 0.129 | 0.740 | 0.953 | 0.182 | 0.283 |
| | | 1000 | 1000 | 0.833 | 0.986 | 0.133 | 0.166 | 0.950 | 1 | 0.490 | 0.700 |

Çizelge 15 incelendiğinde, her iki MİF belirleme testi için de düşük MİF örüntüsünde elde edilen istatistiksel güç değerlerinin yüksek MİF örüntüsüne göre daha yüksek olduğu görülmektedir. Yüksek MİF örüntüsü için elde edilen istatistiksel güç değerleri çok düşüktür. İlgili testlerin istatistiksel gücü öncelikli olarak grupların örneklem büyüklüğü ve MİF miktarından etkilenmiştir. Grupların yetenek dağılımındaki farklılaşma MİF belirleme testlerinin istatistiksel güç oranlarında bir fark oluşturmamıştır.

# BÖLÜM 5

# SONUÇLAR VE ÖNERİLER

Bu bölümde araştırmadan elde edilen sonuç ve tartışmalara yer verilmiş ve araştırma sonuçlarına uygun olarak öneriler sunulmuştur.

## Sonuçlar

Bu çalışmada, Kısmi Puan Modeline uygun olarak üretilen çok kategorili tepki gerektiren maddeler için örneklem büyüklüğü, odak grubun yetenek dağılım ortalaması, MİF miktarı ve MİF örüntüsü gibi farklı test koşullarında MİF belirleme testlerinden Mantel Test ve MTK-OOT'nin performansı incelenmiştir. Bu kapsamda, farklı test koşullarında bu testlere ilişkin I.Tip Hata ve istatistiksel güç oranları dikkate alınarak karşılaştırma yapılmıştır. Araştırma sonuçları I. Tip Hata ve güç oranları bağlamında aşağıda verilmiştir.

### Mantel Test ve MTK-OOT'nin I.Tip Hata Sonuçları

Grupların yetenek dağılımı aynı ve birim normal dağılım gösterdiği farklı örneklem büyüklüğü koşullarında her iki MİF belirleme testide de I. Tip hatayı kontrol etmiştir. Artan örneklem büyüklüğüne bağlı olarak I. Tip hata oranları Mantel Test için artarken MTK-OOT için azalmıştır. Küçük örneklem koşulunda Mantel Test, büyük örneklem koşulunda ise MTK-OOT daha düşük I. Tip hata sonuçları üretmiştir.

Her iki MİF belirleme testi için de odak grup yetenek dağılım ortalamasının, referans grup yetenek dağılımına göre sapma gösterdiği

koşullarda elde edilen I. Tip Hata değerleri, her iki grubun yetenek dağılım ölçülerinin birim normal dağılım gösterdiği koşullara göre daha yüksek olmuştur. Yani odak grubun yetenek dağılım ortalamasındaki sapmaya bağlı olarak her iki MİF belirleme testinin de I. Tip hata oranları yükselmiştir.

Her iki test için de en yüksek I. Tip hata değerleri büyük örneklem büyüklüğü ve odak grubun yetenek ortalamasındaki artan sapma miktarı koşulunda elde edilmiştir. Mantel Test ile karşılaştırıldığında, artan örneklem büyüklüğü ve grupların yetenek ortalamasındaki sapmaya bağlı olarak MTK-OOT için I. Tip hata değerlerindeki artış daha yüksek olmuştur. Grupların yetenek dağılımlarındaki sapmalar karşısında MTK-OOT I. Tip hatayı kontrol etmekte zayıf kalmıştır.

**Mantel Test ve MTK-OOT'nin İstatistiksel Güç Sonuçları**

Referans ve odak gruba ait yetenek dağılımlarının aynı ve birim normal dağılım özelliği gösterdiği düşük MİF örüntüsü için, değişen grup örneklem büyüklüğü ve MİF miktarına bağlı olarak her iki MİF testinin istatistiksel güç oranları artış göstermiştir. Orta düzey MİF miktarı koşulunda Mantel Test ve MTK-OOT sadece büyük örneklem koşulunda MİF'i belirlemede yeterli olmuştur. Büyük MİF miktarı koşulunda ise Mantel Test büyük örneklem, MTK-OOT orta ve büyük örneklemlerde MİF'i belirlemede yeterli olmuştur. Her iki MİF miktarı ve tüm örneklem büyüklükleri için MTK-OOT'nin istatistiksel güç oranları Mantel Test'ten daha yüksek olmuştur.

Referans ve odak gruba ait yetenek dağılımlarının aynı ve birim normal dağılım özelliği gösterdiği yüksek MİF örüntüsü için, değişen grup örneklem büyüklüğü ve MİF miktarına bağlı olarak her iki MİF belirleme testinin istatistiksel güç oranları artış göstermiştir. Ancak yüksek MİF örüntüsü koşullarında ilgili testlerin MİF'i belirlemedeki performansı kötüdür.

Odak grup yetenek dağılım ortalamasının, referans grup yetenek dağılımına göre sapma gösterdiği düşük MİF örüntüsü için, orta düzey MİF miktarı için ilgili MİF belirleme testlerinin istatistiksel gücü birim normal dağılım koşullarına göre bir miktar artarken; büyük MİF miktarı için istatistiksel güç

oranları birbirine yakın değerler almıştır. Genel olarak grup yetenek dağılım ortalamalarındaki sapma ile istatistiksel güç oranları arasında belirgin bir örüntü elde edilememiştir. Aratan örneklem büyüklüğü ve MİF miktarına bağlı olarak her iki MİF belirleme testinin istatistiksel güç oranları yükselmiştir. Tüm koşullar için MTK-OOT'nin istatistiksel güç değerleri Mantel Test'ten daha yüksek olmuştur.

Odak grup yetenek dağılım ortalamasının, referans grup yetenek dağılımına göre sapma gösterdiği yüksek MİF örüntüsü koşullarında her iki MİF belirleme testinin istatistiksel güç değerleri çok düşüktür. Bu MİF örüntüsü için değişen grup yetenek dağılımları, örneklem büyüklüğü ve MİF miktarında her iki testi de MİF'i belirlemede yetersiz kalmıştır. Bu koşulda, artan sapma miktarı ile grupların örneklem büyüklüğü ve MİF miktarı arasında belirgin bir örüntü ortaya çıkmamıştır. Düşük MİF örüntüsüyle karşılaştırıldığında, yüksek MİF örüntüsünün tüm koşullarına ait istatistiksel güç oranları düşük MİF örüntüsü koşullarına göre çok düşüktür.

## Öneriler

Bu araştırmanın sonuçlarına dayanarak çok kategorili tepki gerektiren maddeler için MİF analizleri içeren pratik test uygulamalarında uygulayıcılar veya araştırmacılar için aşağıdaki öneriler sunulabilir:

1) Araştırma sonuçlarına göre, çok kategorili tepki gerektiren maddelerle yapılacak MİF analizlerinde grupların yetenek dağılımları birim normal dağılım özelliği gösterdiğinde, araştırmacıların referans ve odak gruptaki birey sayılarını değerlendirerek, büyük örneklem durumlarında MTK-OOT, küçük örneklem durumlarında Mantel Test'i kullanmaları önerilebilir.
2) Referans ve odak grubun yetenek dağılımları farklılaştığı durumlarda Mantel Test, MTK-OOT'ye göre I. Tip hatayı kontrol etmekte daha iyi performans göstermiştir. Bu yüzden araştırmacılar veya uygulayıcılar grupların yetenek dağılımlarındaki sapmaları değerlendirerek öncelikle Mantel Test kullanmayı tercih edebilir.

3) MTK-OOT, odak grup için yetenek dağılım ortalaması için sapma miktarı arttığında ve örneklem büyüdüğünde I. Tip hatayı kontrol etme konusunda iyi performans göstermemiştir. Bu tür test koşulları için eğer bu test kullanılacaksa belirtilen test koşullarına daha az duyarlı başka MİF belirleme tekniklerinin (örneğin parametrik olmayan MİF belirleme yaklaşımları) de sonuçlarının beraber değerlendirilerek maddelerin MİF gösterip göstermediğine karar verilmelidir.

4) Mantel Test ve MTK-OOT'nin istatistiksel gücü öncelikli olarak grupların örneklem büyüklüğünden etkilenmiştir. Bu durumda araştırmacılar veya uygulayıcılar bu MİF belirleme testlerini kullanırken daha güçlü istatistiksel sonuçlar elde etmek üzere olanaklarını değerlendirerek örneklem sayılarını artırma yoluna gidebilir.

Çok kategorili tepki gerektiren maddeler için MİF çalışması yapmak isteyen araştırmacılar için aşağıdaki öneriler getirilebilir:

1) Bu araştırmada, çok kategorili tepki gerektiren MTK modellerinden Kısmı Puan Modeli'ne uygun veriler kullanılmıştır. Daha sonraki çalışmalar için özellikle tutum, algı vb. değişkenlerin ölçeklenmesinde sıklıkla kullanılan ve kategori ayırıcılık parametresini de içeren modellere uygun veriler üzerinden MİF belirleme teknikleri karşılaştırılabilir.

2) Bu çalışma kapsamında çok kategorili tepki gerektiren maddeler üzerinden MİF belirleme testleri karşılaştırılmıştır. Bundan sonraki çalışmalar için hem iki (dikotomus) hem de çok kategoride tepki gerektiren madde formatını içeren karışık (mixed) testler üzerinden MİF belirleme tekniklerinin performansları karşılaştırılabilir.

3) Bu çalışma kapsamında ele alınan referans ve odak gruba ait grup büyüklükleri dışında, daha büyük veya küçük örneklem büyüklükleri veya örneklem büyüklük oranları içeren simülasyon desenleri oluşturulabilir.

4) MİF belirleme tekniklerinin performansını etkileyebilecek ancak bu araştırma için sabit değişken olarak ele alınan MİF içeren madde sayısı, adım güçlüğü parametresinin ($\delta$) zorluk düzeyi gibi farklı değişkenleri içeren simülasyon koşulları oluşturularak MİF belirleme tekniklerinin performansları incelenebilir.

5) Bu çalışmada referans ve odak grup yetenek dağılımları arasındaki sapmaların az olduğu durumlarda büyük örneklem grupları için MTK- OOT'nin daha iyi sonuçlar verdiği belirlenmiştir. Özellikle büyük örneklem grupları için MTK ölçeklemesine dayalı diğer MİF belirleme tekniklerinin performansları incelenerek yetenek ortalamasındaki sapmalara daha az duyarlı MTK MİF belirleme teknikleri belirlenebilir.

6) MİF belirleme tekniklerinin performansları karşılaştırılırken örneklem büyüklüğünden kaynaklanan sınırlılığı aşmak üzere I. Tip hata ve istatistiksel güç oranları dışında standardize ortalama farkların etki büyüklüğü (standardized mean difference effect size measure) ve $R^2$ etki büyüklüğü ölçüsü (R-squared effect size measure) gibi ölçütler kullanılarak da ilgili MİF belirleme tekniklerine ilişkin değerlendirmeler yapılabilir.

# KAYNAKÇA

Ackerman, T.A. (1992). A Didactic Explanation of Item Bias, Item Impact, and Item Validity from A Multidimensional Perspective. *Journal of Educational Measurement*, 29(1), 67-91.

Angoff, W. H. (1993). Perspectives on Differential Item Functioning Methodology. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp.3-23). Hillsdale, NJ: Erlbaum.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Ankenmann, R.D., Witt, E. A., and Dunbar, S. B. (1999) An Investigation of the Power of the Likelihood Ratio Goodness of Fit Statistic in Detecting Differential Item Functioning. *Journal of Educational Measurement, 36*(4), 277-300.

Artar, B. (2007).*Differential Item Functioning Analyses For Mixed Response Data Using IRT Likelihood-Ratio Test, Logistic Regression and Gllamm Procedures.* Unpublished doctoral dissertation, Florida State University.

Atar, B., and Kamata, A. (2011). Comparison of IRT Likelihood Ratio Test and Logistic Regression DIF Detection Procedures. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 41*, 36–47.

Bahry, M.L. (2012). *Polytomous Item Response Theory Parameter Recovery: An Investigation of Non-Normal Distributions And Small Sample Size*. Unpublished master's thesis, University of Alberta, Canada.

Bakan Kalaycıoglu, D., and Berberoglu, G. (2010). Differential Item Functioning Analysis of the Science and Mathematics Items in the University Entrance Examinations in Turkey. *Journal of Psychoeducational Assessment, 20*, 1-12.

Barrick, M.R., and Mount, M.K. (1991). The Big Five Personality Dimensions and Job Performance: A meta-analysis. *Personnel Psychology*, *44*, 1–26.

Beretvas, N., S. (2000).*To Meet Or Not Meet Standard: Proficiency Estimation Using Different Polytomous IRT Models*. Unpublished doctoral dissertation,University of Washington.

Bolt, D. M. (2002). A Monte Carlo Comparison of Parametric and Nonparametric Polytomous DIF detection methods. *Applied Measurement in Education, 15,*113-141.

Boomsma, A. (2013) Reporting Monte Carlo Studies in Structural Equation Modeling, Structural Equation Modeling. *A Multidisciplinary Journal*, 20(3), 518-540.

Camilli, G., and Shepard, L.A. (1994). *Methods for Identifying Biased Test Items*.California: Sage Publications.

Carter, N. T. (2011). Applıcatıons Of Dıfferentıal Functıonıng Methods To The Generalızed Graded Unfoldıng Model. Unpublished doctoral dissertation, Bowling Green State University.

Chang, H., Mazzeo, J., and 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., and Mazor, K. M. (1998). Using Statistical Procedures To Identify Differential Item Functioning Test Items. *Educational Measurement: Issues and Practice,* 17, 31-44.

Cohen, A. S., Kim, S.-H., and Baker, F. B. (1993). Detection of Differential Item Functioning in the Graded Response Model. *Applied Psychological Measurement, 17*, 335–350.

Crocker, L., and Algina, J. (1986). *Introduction to Classical and Modern Test Theory.*New York: Holt, Rinehart and Winston.

De Ayala, R. J. (1993). An Introduction To Polytomous Item Response Theory Models. *Measurement & Evaluation in Counseling & Development, 25(4).*

De Ayala, R. J. (2009). *The theory and practice of item response theory.* NY: Guilford Press.

De Leo, J.A.,Van Dam, N. T., Hobkirk, A. L., and Earleywine, M. (2010). Examining bias in the impulsive sensation seeking (ImpSS) Scale using Differential Item Functioning (DIF) – An item response analysis. *Personality and Individual Differences*, 50 (2011) 570–576

DeMars, C. E. (2007). Polytomous Differential Item Functioning and Violations of Ordering of the Expected Latent Trait by the Raw Score. *Educational and Psychological Measurement,* 68( 3), 379-396

Dodd, B. G., De Ayala, R. J., and Koch, W. R. (1995). Computerized Adaptive Testing With Polytomous Items. *Applied Psychological Measurement, 19,* 5-22.

Dodeen, H., and Johanson G. (2003). An Analysis of Sex-Related Differential Item Functioning in Attitude Assessment. *Assessment & Evaluation in Higher Education, 28* (2), 129-134.

Dodeen, H. (2004). Stability of Differential Item Functioning Over a Single Population in Survey Data. *Journal of Experimental Education, 72*, 181-193.

Donoghue, J. R., and Allen, N. L. (1993). Thin Versus Thick Matching İn The Mantel-Haenszel Procedure For Detecting DIF. *Journal of Educational and Behavioral Statistics, 18*, 131-154.

Dorans, N. J., and Holland, P. W. (1993). DIF *detection and description*: Mantel-Haenszel and standardization. In P. W. Holland and H. Wainer (Eds.), *Differential Item Functioning* (p 35-66). Hillsdale, NJ: Erlbaum.

Dorans, N. J., and Potenza, M. T. (1994). *Equity Assessment for Polytomously Scored Items: A Taxonomy of Prodecures for Assessing Differential Item Functioning*. Web: http://www.eric.ed.gov/PDFS/ED380499.pdf adresinden 12 Aralık 2010'da erişilmiştir.

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

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 12 Nisan 2013'de erişilmiştir.

Ercikan, K. (1998). Translation Effects in International Assessments. *International Journal of Educational Research*, 29, 543-553.

Feingold, A. (1994). Gender differences in personality: A metaanalysis. *Psychological Bulletin*, *116*, 429–456

Fidalgo, A. M., and Bartram, D. (2010).A Comparison of the LR and DFIT Frameworks of Differential Functioning Applied to the Generalized Graded Unfolding Model. *Applied Psychological Measurement*, 34(8) 600–606.

Flowers, C. P., Oshima, T. C., and Raju, N. S. (1999). A Description and Demonstration of the Polytomous-DFIT Framework. *Applied Psychological Measurement, 23,* 309–326.

French, B. F., Hand, B., Therrien, W. J., and Valdivia Vazquez, J. A. (2012). Detection of sex differential item functioning in the Cornell Critical Thinking Test. European Journal of Psychological Assessment, 28(3), 201-207.

Garrett, P. (2009). *A Monte Carlo Study Investigating Missing Data, Differential Item Functioning and Effect Size*. Unpublished doctoral dissertation, Georgia State University.

Gierl, M. J. (2005). Using Dimensionality-Based DIF Analyses to Identify and Interpret Constructs That Elicit Group Differences. *Educational Measurement: Issues and Practice, 24*, 3-14.

Haladyna, T. (1999). *Developing and Validity Multiple-choice Test Items*. Mahwah: Lawrence Erlbaum Associates.

Hambleton, R.K., and Swaminathan, H. (1991). *Item Response Theory: Principles and Applications.* Norwell, MA: Kluwer-Nijhoff Publishing.

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

Hambleton, R. K., and H. Swaminathan (1985). *Item Response Theory: Principles and Application.* Boston: Kluwer-Nijhoff Publising.

Han, K. T., and Hambleton, R. K. (2007). *Windows Software that Generates IRT Model Parameters and Item Responses (WinGen).* Web: http://www.umass.edu/remp/software/simcata/wingen/modelsF.html adresinden 10 Aralık 2010'da erişilmiştir.

Harwell, M., Stone, C.A., Hsu, T.-C., and Kirisci, L. (1996). Monte Carlo Studies In Item Response Theory. *Applied Psychological Measurement*, 20, 101-125.

Hong, S., and Roznowski, M. (2001). An Investigation of the Influence of Internal Test Bias On Regression Slope. *Applied Measurement in Education*, 14(4), 351-68.

Horst, P. (1966). *Psychological Measurement and Prediction*. Belmont: Wadsworth Pub. Co.

Jodoin, M. G. and Gierl, M. J. (1999). Evaluating Type I Error and Power Using an Effect Size Measure with the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, *14*(4), 329– 349.

Johnson- Frotman, K., A. (2007). *The Evaluation of New Criteria for Polytomous DIF in the DFIT Framework*. Unpublished doctoral dissertation. Illinois Institute of Technology, Chigago.

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

Kavcar, B. (2004). *Simülasyon Yöntemi Kullanılarak Yapılan Satış Tahminleriyle Satış Bütçesi Hazırlanması*. Yayımlanmamış yüksek lisans tezi, Ankara Üniversitesi Sosyal Bilimleri Enstitüsü, Ankara.

Kim, S-H, Cohen, A.S (1997). *An investigation of the Likelihood ratio test for detection of Differantial Item Funvtioning under the Graded Response* Model. Paper presented at the annual meeting of the American Educational Research Associatio, Chigago. Web: http://eric.ed.gov/?id=ED408304 adresinden 12 Nisan 2011'de erişilmiştir.

Kim, S.-H, Cohen, A.S., DiStefano, C.A., and Kim, S. (1998, April). *An investigation of the likelihood ratio test for detection of differential item functioning under the partial credit model.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA. Web: http://eric.ed.gov/?id=ED443837adresinden 17 Aralık 2010'da erişilmiştir.

Kim, S.-H. (2000, April). *An investigation of the Likelihood Ratio Test, the Mantel Test, and the Generalized Mantel–Haenszel Test of DIF*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Web: http://eric.ed.gov/?id=ED441007adresinden 17 Aralık 2010'da erişilmiştir.

Kim, J. (2010). *Controlling Type I Error Rate in Evaluating Differential Item Functioning for Four DIF Methods: Use of Three Procedures for Adjustment of Multiple Item Testing*. Unpublished doctoral dissertation Georgia State University.

Kim, S., Cohen, A. S., Alagoz, C., and Kim, S. (2007). DIF Detection and Effect Size Measures for Polytomously Scored Items. *Journal of Educational Measurement, 44*(2), 93-116.

Kristanjansonn E. (2001). Detecting DIF In Polytomous Items: An Empirical Comparision of the Ordinal Logistic Regression, Logistic Discriminant Functon Analysis, Mantel and Generalized Mantel Haenszel Procedures. Unpublished doctoral dissertation, University of Ottova, Canada.

Kristanjansonn E., Aylesworth, R., McDowell, I., and Zumbo, B. D. (2005). A Comparison of Four Methods for Detecting Differential Item Functioning in Ordered Response Model. *Educational and Psychological Measurement, 65*(6), 935-953.

Kubiszyn, T., and Borich, G. (2000). *Educational Testing and Measurement*. New York:John Wiley & Sons.

Li, H. and Stout, W. (1996). A new procedure for detection of crossing DIF. Psychometrika, 61, 647–677.

Li, Z., and Zumbo, B. D. (2009). Impact of Differential İtem Functioning On Subsequent Statistical Conclusions Based On Observed Test Score Data. *Psicológica, 30*, 343–370.

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, *47*, 149- 174.

Mendes-Barnett, S., and Ercikan, K. (2006). Examining Sources Of Gender DIF In Mathematics Assessments Using A Confirmatory Multidimensional Model Approach. *Applied Measurement in Education, 19*, 289-304.

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances As Scientific Inquiry Into Score Meaning. *American Psychologist*, 50(9), 741-749.

Meyer, J. P., Huynh, H., and Seaman, M. A. (2004). Exact Small-sample Differential Item Functioning Methods for Polytomous Items with Illustration Based on An Attitude Survey. *Journal of Educational Measurement, 41*(4), 331-344.

Narayanan, P., and Swaminathan H. (1996). Identification Of Items That Show Nonuniform DIF. *Applied Psychological Measurement,* 20(3), 257-274.

Ostini, R., and Nering, M. L. (2006). *Polytomous Item Response Theory Models.*Thousand Oaks, CA:Sage.

Oshima, T.C., Raju, N. S., and Nanda, A. O. (2006). A New Method for Assessing the Statical Significance in the Differential Functioning of Items and Test (DFIT) framework. *Journal of Educational Measurement, 43*, 1-17.

Penfield, D. R. (2001). Assessing Differential Item Functioning Among Multiple Groups: A Comparison Of Three Mantel-Haanszel Procedures. *Applied Measurement in Education*, 14 (3), 253–259.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement*, 29, 150–151.

Popham, W. C. (1999). *Classroom Assessment: What Teachers Need to Know.* 2. Baskı Boston: Allyn and Bacon.

Reise, S. P., and Yu, J. (1990). Parameter Recovery in The Graded Response Model Using MULTILOG. *Journal of Educational Measurement, 27(2),* 133-144

Roever, C. (2005). “That’s Not Fair!” Fairness, Bias, and Differential Item Functioning In Language Testing. Web: http://www2.hawaii.edu/~roever/brownbag.pdf adresinden 18 Kasım 2012’de erişilmiştir.

Rogers, J. H., and Swaminathan, H. (1993). A Comparison Of Logistic Regression And Mantel-Haenszel Procedures For Detecting Differential Item Functioning. *Applied Psychological Measurement, 17* (2), 105-116.

Roussos, L. A., and Stout, W. F. (1996). Simulation Studies Of The Effects Of Small Sample Size And Studied İtem Parameters on SIBTEST and Mantel-Haenszel Type I Error Performance. *Journal of Educational Measurement, 33*, 215-230.

Samejima, F. (1969). Estimation Of Latent Trait Ability Using A Response Pattern Of Graded Scores.*Psychometrika Monograph,* No. 17.

Schaeffer, G. A., Henderson-Montero, D., Julian, M., and Bené, N. H. (2002). A Comparison Of Three Scoring Methods For Tests With Selected Response and Constructed-Response Items. *Educational Asssessment, 8(4),* 317-340.

Shealy, R., and Stout, W. F. (1993). A Model-Based Standardization Approach That Separates True Bias/ DIF From Group Ability Differences And Detects Test Bias/DTF As Well As İtem Bias/DIF. *Psychometrika, 58*, 159–194.

Sweeney, K. P. (1996). *A Monte Carlo Investigation Of The Likelihood-Ratio Procedure In The Detection Of Differential Item Functioning.* Unpublished doctoral dissertation, Fordham University, New York, NY.

Sykes, R. C., and Hou, L. (2003). Weighting Constructed-Response Items In IRT-Based Exams. *Applied Measurement in Education, 16(4),* 257-275.

Su, Y.-H., and Wang, W.-C. (2005). Efficiency of the Mantel, Generalized Mantel-Haenszel, and Logistic Discriminant Function Analysis Methods in Detecting Differential Item Functioning for Polytomous Items. *Applied Measurement in Education, 18*, 313-350.

Stark, S., Chernyshenko, O. S., and Drasgow, F. (2006). Detecting Differential Item Functioning With Confirmatory Factor Analysis and Item Response Theory: Toward a Unified Strategy. *Journal of Applied Psychology*, 91(6), 1292–1306

Stocking, M., and Lord, F. M. (1983). Developing A Common Metric In Item Response Theory. *Applied Psychological Measurement, 7,* 207-210.

Thissen, D., Steinberg, L., and Gerrard, M. (1986). Beyond Group-Mean Differences: The Concept Of Item Bias. *Psychological Bulletin,* 99(1), 118-128.

Thissen, D., and Wainer, H. (2001). *Test Scoring*. New Jersey: Lawrence Erlbaum

Thissen, D., Chen, W.-H., and Bock, R. D. (2002). MULTILOG [Computer program]. Lincolnwood,IL: Scientific Software International.

Thurman, C. J. (2009). *A Monte Carlo Study Investigating the Influence of Item Discrimination, Category Intersection Parameters, and Differential Item Functioning Patterns on Detection of Differential Item Functioning in Polytomous Items.* Unpublished doctoral dissertation, Georgia State University.

Tian F. (1999). *Detecting DIF In Polytomous Item Responses*. Unpublished dissertation.University of Ottowa.

Uiterwijk, H., and Valen, T. (2005). Linguistic Sources of Item Bias for Second Generation Immigrants in Dutch Tests. *Language Testing*, 22(2), 211 234.

Wang, W.-C., and Su, Y.-H. (2004). Factors Influencing The Mantel and Generalized Mantel-Haenszel Methods for the Assessment of Differential Item Functioning in Polytomous Items. *Applied Psychological Measurement, 28*, 450-481.

Wang, H. (2009). *Creatıng a DIF index —A Combination of DIF Measure, DIF Direction, and MIF Impact on People*. Unpublished doctoral dissertation, Michigan State University.

Wang, W., Tay, L., and Drasgow, F. (2013). Detecting Differential Item Functioning of Polytomous Items for an Ideal Point Response Process. *Applied Psychological Measurement* 37(4) 316–335

Wetzel, E. , and Hell, B. (2013). Gender-related differential item functioning in vocational interest measurement: An analysis of the AIST-R. *Journal of Individual Differences*, 34(3), 170-183.

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.

Wood, W. S. (2011a). *Dıfferentıal Item Functıonıng Procedures for Polytomous Items When Examınee Sample Sızes Are Small*. Unpublished doctoral dissertation, Graduate College of The University of Iowa.

Wood, W. S. (2011b). DIF Testing For Ordinal Items With Poly-SIBTEST, The Mantel And GMH Tests, And IRT-LR-DIF When The Latent Distribution Is Nonnormal For Both Groups. *Applied Psychological Measurement*, 35(2) 145–164

Wyse A. E., and Mapuranga, R. (2009). Differential Item Functioning Analysis Using Rasch Item Information Functions. *International Journal of Testing,9*(4), 333 – 357.

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.* Web: http://educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf adresinden 19 Haziran 2010'da erişilmiştir.

Zwick, R. (1990). When Do Item Response Function And Mantel-Haenszel Definitions Of Differential Item Functioning Coincide? *Journal of Educational Statistics*, 15, 185–197.

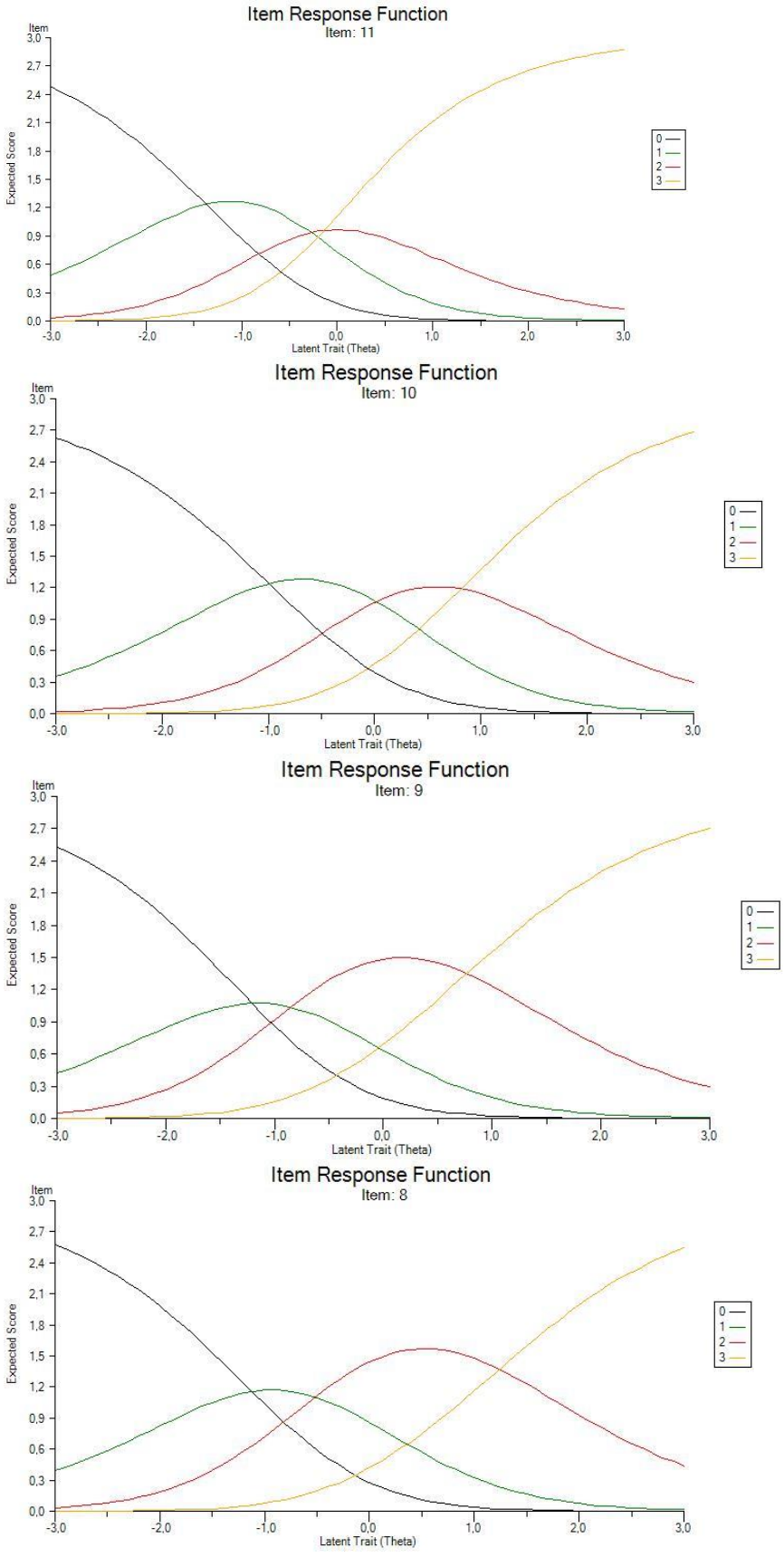
Zwick, R., Donoghue, J. R., and Grima, A. (1993). Assessment of Differential Item Functioning for Performance Tasks. *Journal of Educational Measurement, 30,*233–251.

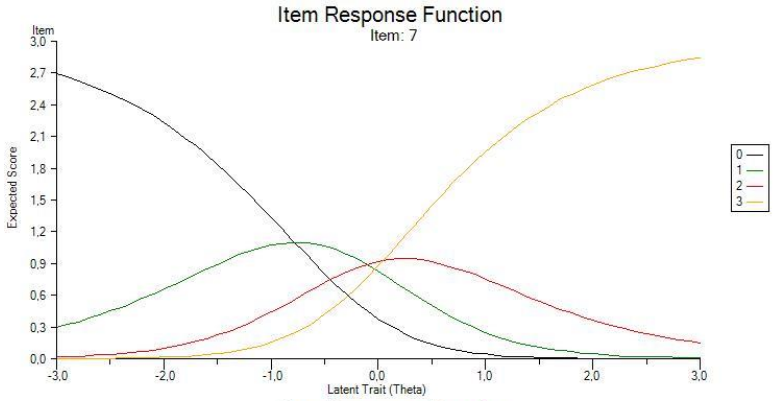
Zwick, R., and Thayer, D. T. (1996). Evaluating the Magnitude of Differential Item Functioning in Polytomous Items. *Journal of Educational and Behavioral Statistics, 21,* 187–201.

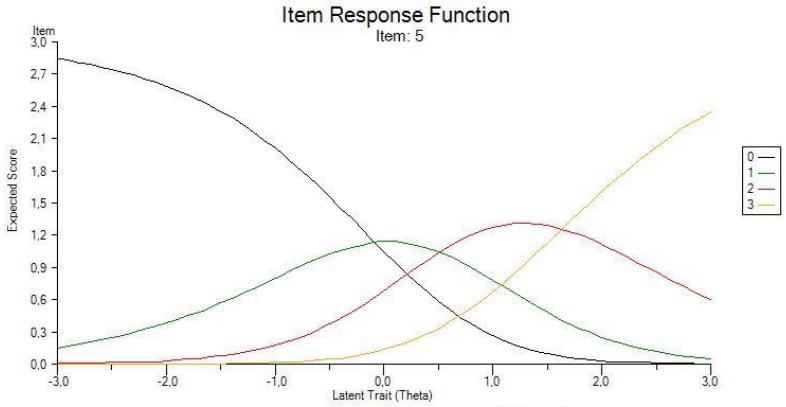
Zwick, R., Thayer, D. T., and Mazzeo, J. (1997). Descriptive and Inferential Procedures for Assessing Differential Item Functioning in Polytomous Items. *Applied Measurement in Education, 10*(4), 321-344.

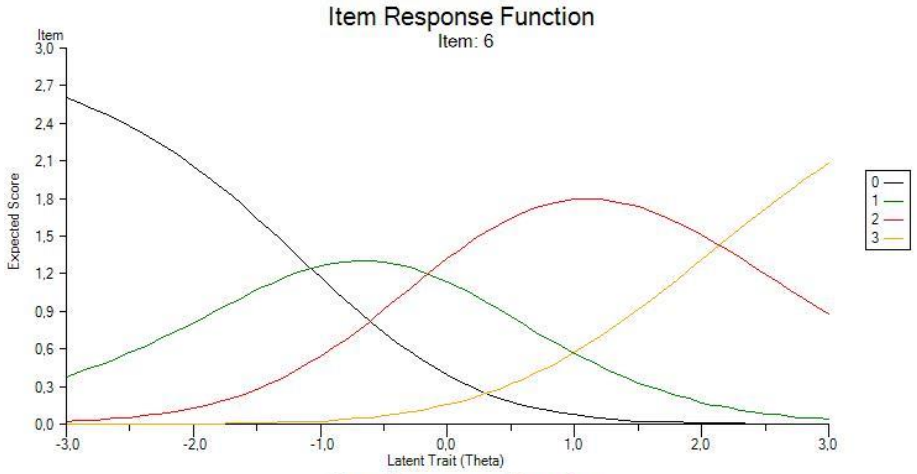
# EKLER

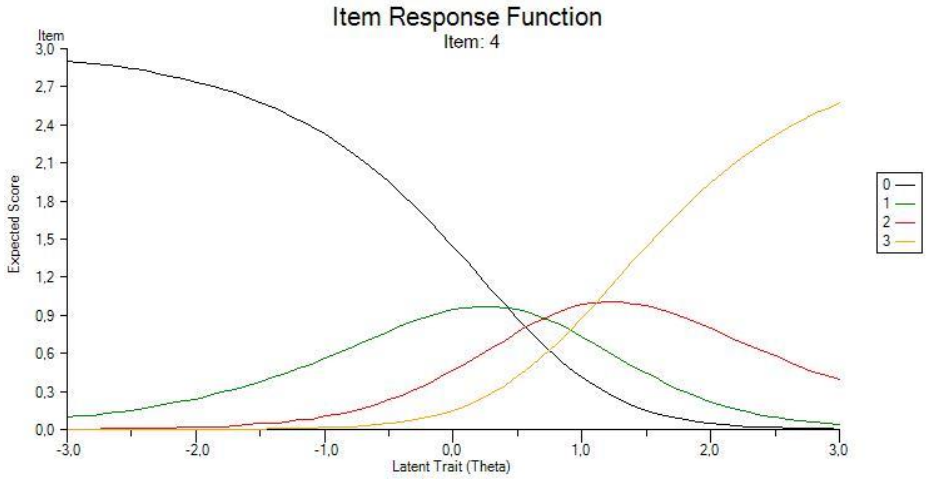
## EK1: Madde Karakteristik Eğrileri

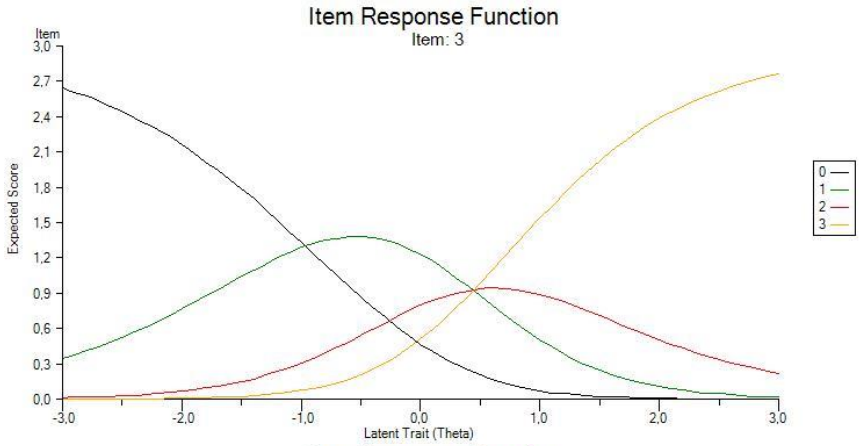


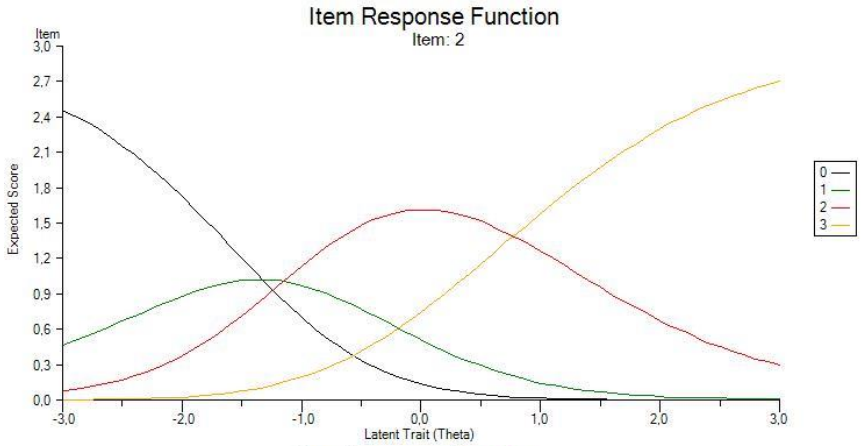

Item Response Function
Item: 7
Item
Expected Score
Latent Trait (Theta)
0
1
2
3

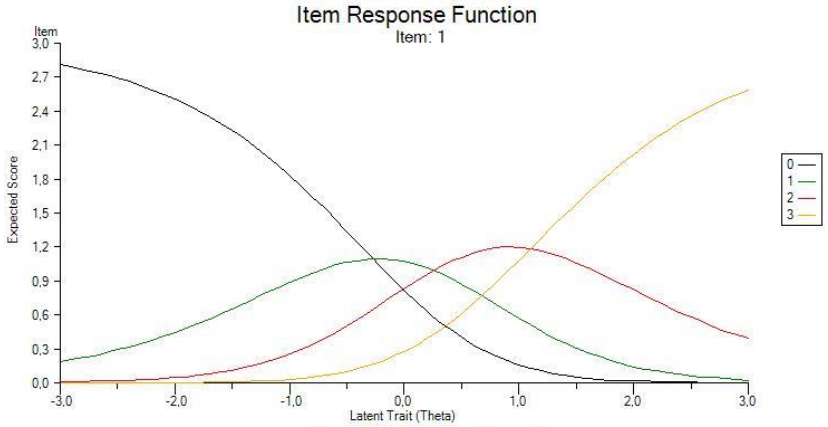

Item Response Function
Item: 5
Item
Expected Score
Latent Trait (Theta)
0
1
2
3

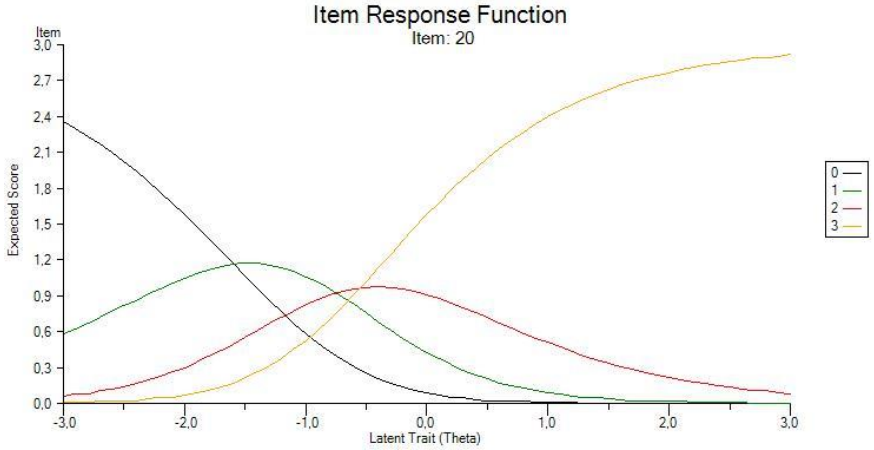

Item Response Function
Item: 6
Item
Expected Score
Latent Trait (Theta)
0
1
2
3

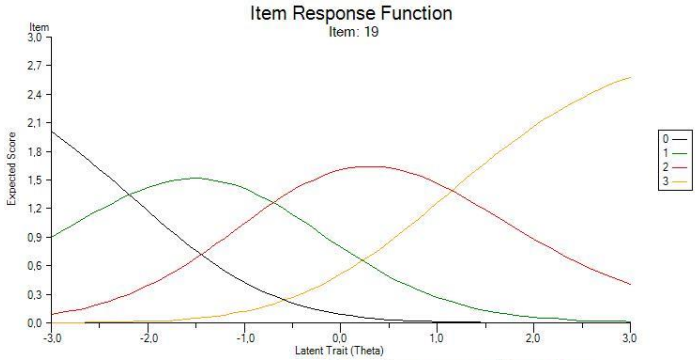

Item Response Function
Item: 4
Item
Expected Score
Latent Trait (Theta)
0
1
2
3

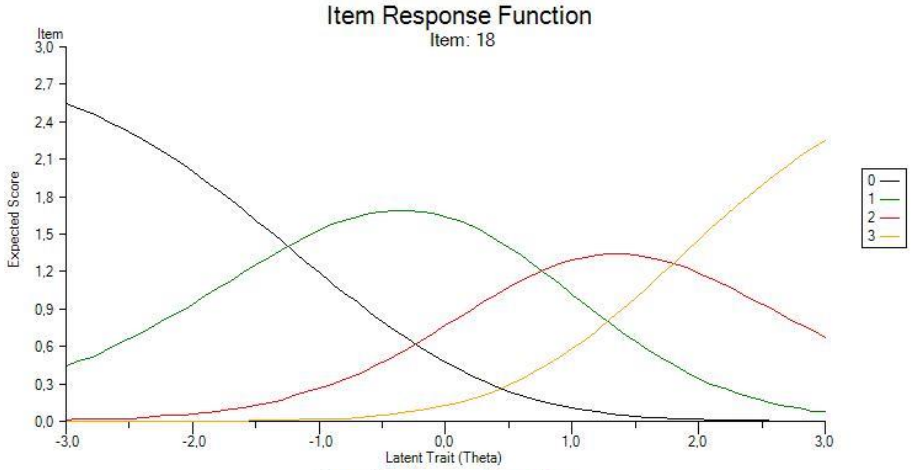

Item Response Function
Item: 3
Item
Expected Score
Latent Trait (Theta)
0
1
2
3

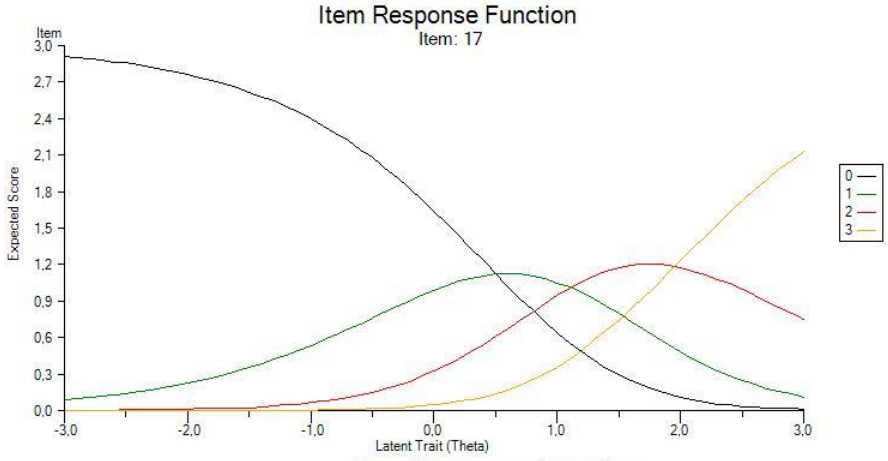

Item Response Function
Item: 2
Item
Expected Score
Latent Trait (Theta)
0
1
2
3

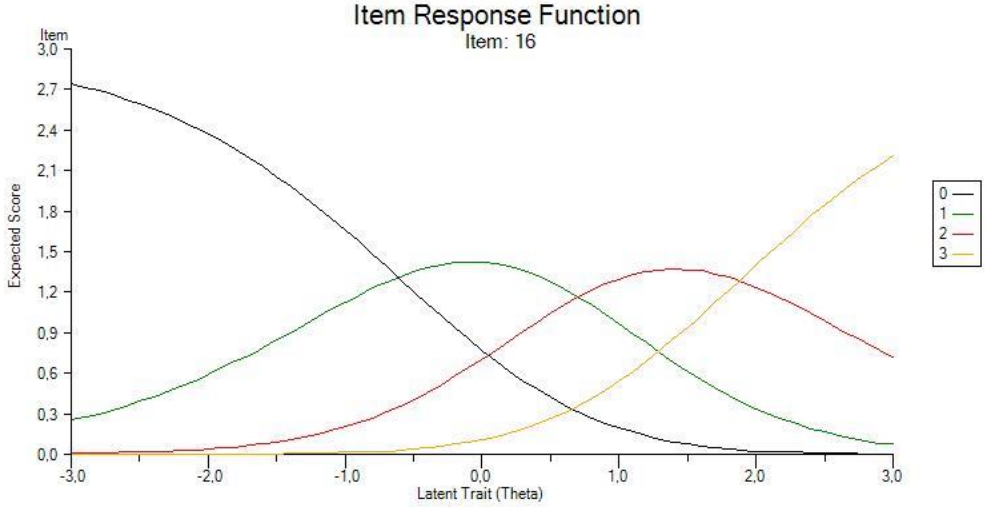

Item Response Function
Item: 1
Item
Expected Score
Latent Trait (Theta)
0
1
2
3

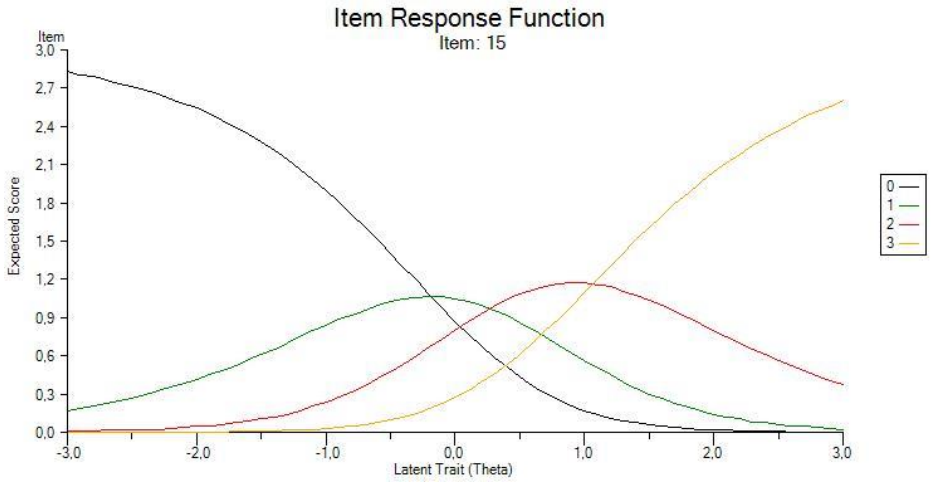

Item Response Function
Item: 20
Item
Expected Score
Latent Trait (Theta)
0
1
2
3

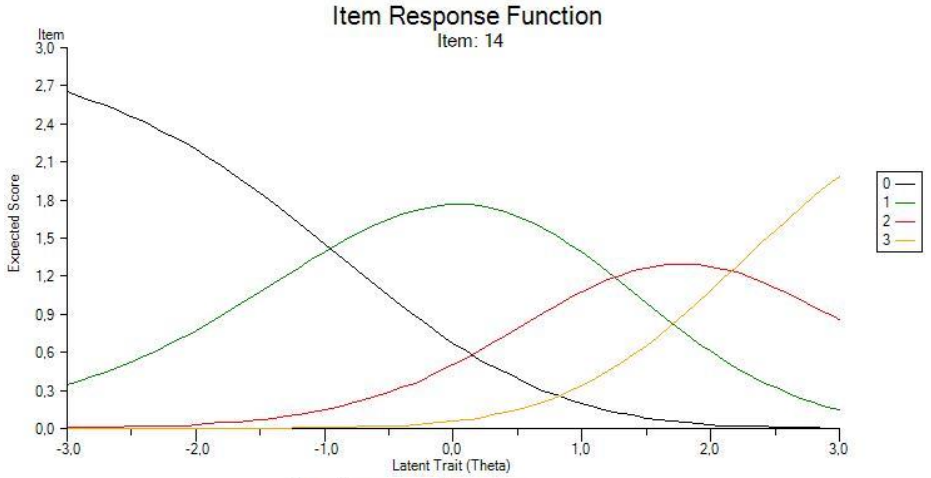

Item Response Function
Item: 19
Item
Expected Score
Latent Trait (Theta)
0
1
2
3

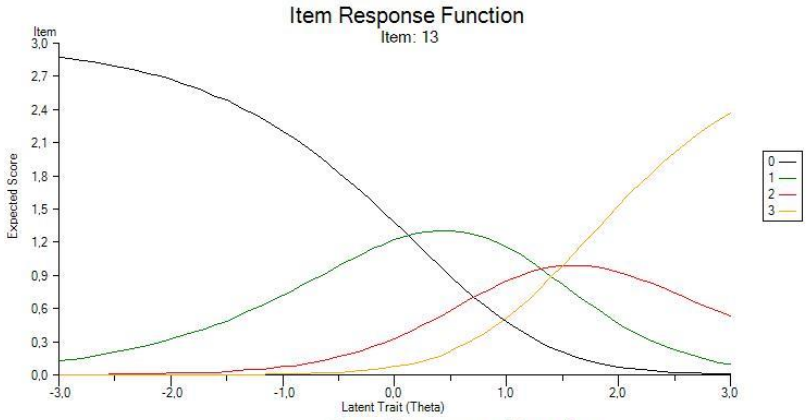

Item Response Function
Item: 18
Item
Expected Score
Latent Trait (Theta)
0
1
2
3

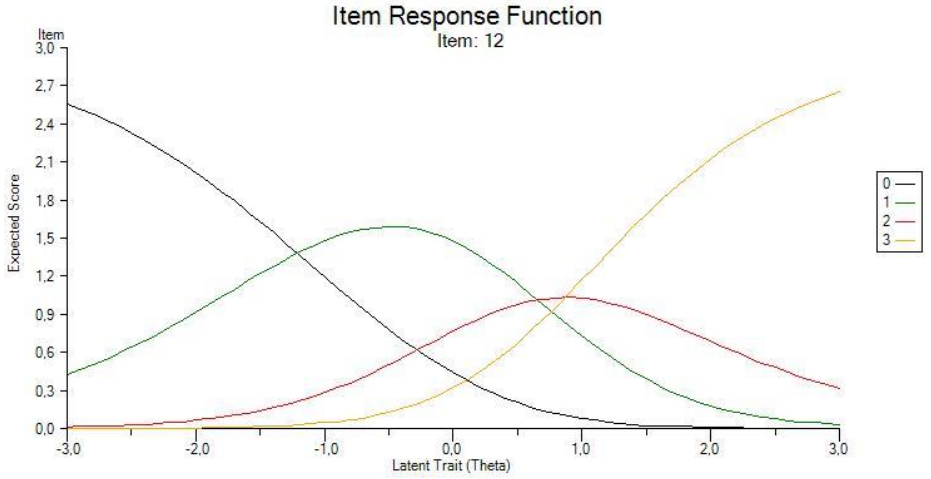

Item Response Function
Item: 17
Item
Expected Score
Latent Trait (Theta)
0
1
2
3


Item Response Function
Item: 16
Item
Expected Score
Latent Trait (Theta)
0
1
2
3


Item Response Function
Item: 15
Item
Expected Score
Latent Trait (Theta)
0
1
2
3


Item Response Function
Item: 14
Item
Expected Score
Latent Trait (Theta)
0
1
2
3


Item Response Function
Item: 13
Item
Expected Score
Latent Trait (Theta)
0
1
2
3


Item Response Function
Item: 12
Item
Expected Score
Latent Trait (Theta)
0
1
2
3

**EK2.** MULTILOG Programında Olabilirlik Oran Testi İçin Yazılan Komut Dosyası Örneği

```
From Safiye Bilican <safiyebilican@yahoo.com> on 6 DEC 2011
      Next, run a DIF model on the expanded data file.
>PROBLEM  RANDOM, INDIVIDUAL, DATA='referans_focal_1.dat',
      NITEMS=40, NGROUPS=2, NEXAMIEES=600, NCHARS=3;
>TEST     ALL, NOMINAL, NC=(4(0)40), HIGH=(4(0)40);
>TMATRIX  ALL, AK, POLYNOMIAL;
>EQUAL    ALL, AK=1;
>FIX      ALL, AK=(2,3), VALUE=0.0;
>TMATRIX  ALL, CK, TRIANGLE;
>EQUAL    AK=(1,2,3,4),
      ITEMS=(21,22,23,25,26,27,28,30,31,32,33,34,35,36,37,38,40),
      WITH= ( 1, 2, 3, 5, 6, 7, 8,10,11,12,13,14,15,16,17,18,20);
>EQUAL    CK= (1,2,3,4),
      ITEMS=(21,22,23,25,26,27,28,30,31,32,33,34,35,36,37,38,40),
      WITH= ( 1, 2, 3, 5, 6, 7, 8,10,11,12,13,14,15,16,17,18,20);
>FIX      MU, GROUPS=1, VALUE=0.0;
>FIX      SD, GROUPS=1, VALUE=1.0;
>FIX      SD, GROUPS=2, VALUE=1.0;
>ESTIMATE NCYCLES=100;
>SAVE;
>END;
4
1234
11111111111111111111111111111111111111
22222222222222222222222222222222222222
33333333333333333333333333333333333333
44444444444444444444444444444444444444
(3A1,1X,I1,40A1)
```