Eğitim ve Bilim

Orijinal Makale

2025 Ek Sayı 1 111-127

# Eğitimde Madde Takımlarının Kullanımı: eTIMSS 2019 Örneği *

Kübra Atalay Kabasakal [1], Sebahat Gören [2]

## Öz

Madde takımları; metin, grafik, tablo gibi ortak bir uyarana bağlı birden çok maddenin yer aldığı madde grupları/kümeleridir. Bu tür maddeler ortak bir uyaran paylaşmalarından dolayı madde yanıtlarının birbiriyle ilişkili olma ihtimali oldukça yüksektir ve bu durum Madde Tepki Kuramı'nın (MTK) yerel bağımsızlık varsayımını ihlal ederek madde takımlarında yer alan maddeler arası yerel bağımlılığa neden olur. Bu nedenle bu çalışmada madde takımlarından kaynaklı yerel bağımlılığın madde ve yetenek parametre kestirimi, sınıflama doğruluğu ve Değişen Madde/Madde Grubu Fonksiyonu (DMF/DMGF) üzerindeki etkilerini değerlendirmek için MTK ve Madde Takımı Tepki Kuramı (MTTK) modelleri kullanılmış ve elde edilen sonuçlar karşılaştırmalı olarak incelenmiştir. eTIMSS 2019 uygulamasında matematik alt testi 13. ve 14. kitapçıkta ortak olarak yer alan üç madde takımına verilen cevaplar R yazılımındaki *mirt* paketi kullanılarak analiz edilmiştir. Madde takımlarında genel olarak orta düzeyde yerel bağımlılık derecesi bulunmuş olup her iki modele ait madde ve yetenek parametre kestirimleri arasında çok yüksek bir ilişki bulunmuştur. Sınıflama doğruluğu ele alındığında ise MTTK ile MTK modelleri eşdeğer bir performans göstermiştir. Maddeler bağımsız ve madde takımları olarak iki ayrı şekilde ele alındığında cinsiyete göre DMF/DMGF gösteren hiçbir maddeye rastlanmamıştır. Araştırma bulguları düşük ve orta düzey yerel bağımlılık olduğu MTK'nin madde takımı etkisini tolere edebildiğini göstermektedir.

## Anahtar Kelimeler

Madde takımı
Madde takımı tepki kuramı
Madde parametre kestirimi
Yetenek kestirimi
Sınıflama doğruluğu
Değişen madde/madde grubu fonksiyonu

## Makale Hakkında

Gönderim Tarihi: 05.10.2024
Kabul Tarihi: 30.12.2024
Elektronik Yayın Tarihi: 03.03.2025

DOI: 10.15390/EB.2025.14104

## Giriş

Öğrenci performansının değerlendirilmesi, eğitim araştırmalarında uzun zamandır odak noktası olmuştur. Madde güçlüğü ve öğrenci yeteneğinin doğru tahmin edilmesi, etkili öğretim ve değerlendirme uygulamaları için oldukça önemlidir. Ancak madde güçlük ve yetenek parametrelerinin kestirimi uygulanan testin türüne, testin içerdiği maddelerin türüne, testin yapısına ve uygulandığı gruba göre değişebilmektedir. Okuduğunu anlama metni, şekil veya grafik gibi ortak bir uyarıcıyı paylaşan maddelerden oluşan madde grupları ya da kümeleri madde takımı olarak tanımlanır (Wainer ve Kiely, 1987). Madde takımları, bağlamsal bilgi ve bilişsel süreçler gibi öğrenci yanıtlarını etkileyen faktörlerin daha detaylı bir şekilde modellenmesine olanak tanır. Koziol (2016) bir testte madde takımı kullanmanın amacının, ilgilenilen gizil yapı tarafından açıklananın ötesindeki performansı yakalamak

* Bu çalışmanın bir bölümü 4-6 Ekim 2024 tarihleri arasında düzenlenen Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu'nda sözlü bildiri olarak sunulmuştur.

[1] Hacettepe Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Türkiye, kkatalay@gmail.com
[2] Kütahya Dumlupınar Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Türkiye, sebahatgoren@gmail.com

olduğunu belirtmiştir. Ayrıca madde takımlarının kullanılması ile üst düzey becerilerin daha iyi ölçülebilmesi hedeflenir (DeMars, 2010; Wainer ve Wang, 2000).

Test geliştirme açısından bakıldığında, madde takımlarının daha karmaşık ve birbirleriyle ilişkili maddeleri bir araya getirmenin yanı sıra test verimliliğinin artırılmasına da yardımcı olduğu söylenebilir (Thissen, Steinberg ve Gerrard, 1986). Şöyle ki bir madde takımına yerleştirilmiş madde grupları, test katılımcılarının ortak uyarana bağlı birden fazla madde cevaplandırmasını sağlayarak zaman ve emek açısından ekonomiklik sağlanır (Ho ve Dodd, 2012; Wainer ve Wang, 2000). Ayrıca, madde takımları madde yanıtlarını ortak bir uyarıcıya sabitleyerek yapı ile ilgisiz varyans sorunlarını azaltmaya yardımcı olabilir ve potansiyel olarak sınava girenlerin yetenekleri hakkında yapılan çıkarımların geçerliğini artırabilir. Çünkü bireyler ortak uyarana sahip içeriği yalnızca bir kez değerlendirmek zorunda kalır ve daha sonra bu içerikteki bilgileri madde takımında yer alan tüm maddelerde kullanabilir. Belirtilen nedenlerle özellikle yabancı dil becerilerinin ölçülmesinde bir veya birden fazla metne atıfta bulunan bir dizi madde takımlarının yer aldığı çoktan seçmeli test yapısı kullanımı oldukça yaygındır. Türkiye'de ulusal merkezi sınavlar Ölçme, Seçme ve Yerleştirme Merkezi (ÖSYM) tarafından yapılmaktadır. Hem ÖSYM tarafından gerçekleştirilen sınavlarda (ALES, KPSS YDS, e-YDS, YKS, YÖKDİL) hem de uluslararası sınavlarda (GRE, IELTS, SAT, TOEFL,) madde takımları sağladıkları avantajlar sebebiyle sıklıkla tercih edilmektedir. Bu sınavlardaki madde takımlarına benzer uygulamaların Türkiye'de K-12 düzeyinde test geliştirme süreçlerine entegre edilmesi, ölçme ve değerlendirme süreçlerinin de hem geçerliğini hem de verimliliğini artırabilir. Özellikle ilkokul ve ortaokul öğrencilerinin dikkat sürelerinin sınırlı olduğu düşünüldüğünde, ortak bir uyarana dayalı madde takımları sayesinde zamandan tasarruf sağlanabilir ve öğrencilerin test sırasında daha odaklı bir şekilde ilerlemelerine imkan verebilir.

Madde takımları sağladıkları avantajlar sayesinde sıklıkla kullanılmasından dolayı test puanlarının geçerlik ve güvenirliğinin incelenmesi açısından Klasik Test Kuramı (KTK) dışında Madde Tepki Kuramına (MTK) dayalı çeşitli yöntemleri kullanmayı da gerekli kılmıştır. MTK analizleri, testlerin geçerlik ve güvenirliğini artırmaya yönelik güçlü bir araç olarak hem ulusal hem de uluslararası alanda giderek daha fazla önem kazanmaktadır Çünkü MTK, sadece test puanlarını değil, aynı zamanda her bir maddeye verilen yanıtların arkasındaki temel yetenek düzeylerini modelleyerek, öğrencilerin yeteneklerini daha doğru tahmin etmeye yardımcı olur (Embretson, 2010; Hambleton ve Rogers, 1989). Weiss (1982) ise, test maddelerinin zorluk derecesini bireyin yetenek seviyesine uyarlayarak, bireylerin yetenek düzeylerini daha doğru kestirebilen Bilgisayarda Bireyselleştirilmiş Test (BBT) desenlerinin kullanımının yaygınlaşmasıyla farklı MTK yöntemlerinin kullanımının da artığını vurgulamaktadır. MTK, bireyin yeteneği ile test maddelerine verdiği yanıtlar arasındaki ilişkiyi modelleyen güçlü bir istatistiksel çerçeve sağlamasına rağmen uygulanabilirliğinin altında yatan birkaç temel varsayım vardır (Embretson, 2010). MTK modelleri temelde tek boyutlu modeller ve çok boyutlu modeller olarak ayrılmaktadır. Tek boyutlu MTK'nin ilk varsayımı, ölçülen gizil özelliğin tek boyutlu olduğudur. İkinci olarak tek boyutlu MTK, bir maddeye doğru yanıt verme olasılığının bireyin yetenek düzeyinin monoton olarak artan bir fonksiyonu olduğunu varsayar. Madde karakteristik eğrisi olarak bilinen bu fonksiyon, gizil özellik ile belirli bir yanıtın olasılığı arasındaki ilişkiyi tanımlar (Hambleton ve Rogers, 1989). Tek boyutlu MTK'nin üçüncü varsayımı, farklı test maddelerine verilen yanıtların yerel olarak bağımsız olduğudur; yani bir maddeye verilen yanıtın olasılığı, bireyin yetenek düzeyi göz önüne alındığında, diğer maddelere verilen yanıtlardan etkilenmez. Bu varsayımlar karşılandığı takdirde tek boyutlu MTK, madde parametrelerinin (güçlük ve ayırt edicilik gibi) ve birey yetenek parametrelerinin tahmin edilmesine olanak tanıyarak daha doğru ve güvenilir test puanları sağlamanın yanı sıra uyarlanabilir testler geliştirmek için de kullanılabilir (Hambleton ve Rogers, 1989). Tüm bu cazip özelliklerine rağmen, madde takımları, parametre kestirimlerinde tek boyutlu MTK'nin yerel bağımsızlık varsayımının ihlaline yol açabilmektedir. Yerel bağımsızlık, tek boyutlu MTK modellerinde kritik bir varsayımdır, ancak pratikte, bir madde takımı içindeki maddeler, gizil yetenek kontrol edildikten sonra bile genellikle ilişkili yanıtlar sergileyebilir (Koziol, 2016). MTK, madde takımlarından elde edilen yetenek kestiriminin kesinliğini ve test güvenliğini abartma eğilimindedir ve bu durum da madde güçlüğü ve ayırt edicilik parametreleri için yanlı tahminler verir (Eckes ve Baghaei, 2015; Sireci, Thissen ve Wainer, 1991). Ek olarak madde takımlarından oluşan testlerde yerel bağımsızlık

varsayımı kontrol altına alınmadığı zaman test eşitleme, bağlama ve sınıflama doğruluğu hataları ile karşılaşılabilmektedir (Keller, Swaminathan ve Sireci, 2003; Lee, Kolen, Frisbie ve Ankenmann, 2001; Li, Bolt ve Fu, 2006). Ayrıca BBT uygulamalarında madde takımının kullanılmasıyla bağlam ve sıra etkileri de kontrol edilebilmektedir (Wainer, Bradlow ve Wang, 2007). Fakat madde takımlarından oluşan BBT uygulamalarında tek boyutlu MTK modellerinin kullanılması da yerel bağımlılık varsayımının ihlalinden dolayı madde/madde takımı bilgi fonksiyonlarının yüksek hesaplanmasına sebep olmaktadır (Thissen, Steinberg ve Mooney, 1989). Bu nedenle madde takımlarından oluşan BBT uygulamalarında Madde Takımı Tepki Kuramı Modellerinin (MTTK) tercih edilmesi yetenek kestiriminin doğruluğu ve ölçme kesinliği açısından daha uygun olacaktır.

Madde takımı içeren testlerde, yerel bağımsızlık varsayımı ihlali ile başa çıkmak için iki yöntem önerilmiştir. Yöntemlerden biri, bir madde takımı içindeki maddeleri tek bir çoklu puanlanan madde (süper madde) olarak ele almak ve tek boyutlu çok kategorili maddeler için uygun bir model ile kestirim yapmaktır (Cook, Dodd ve Fitzpatrick, 1999; Sireci vd., 1991; Yen, 1993; Wainer, 1995). Bu yöntem, bir madde takımı içindeki maddeler arasındaki yerel bağımlılığın orta düzeyde olduğu ve testin büyük oranda bağımsız madde içerdiği durumlarda uygundur (Wainer, 1995). Ancak olası yanıt örüntülerinin sayısı bir madde takımındaki madde sayısı ile geometrik olarak arttığından pratik değildir ve bu nedenle sık kullanılmaz (Thissen vd., 1989). Ayrıca madde takımındaki maddelerin toplamı dikkate alındığından bilgi kaybına yol açabilmektedir (Wainer ve Lewis, 1990). Alternatif bir yöntem ise, MTK modellerine genel boyutun yanı sıra belirli boyutların da dahil edilmesiyle madde takımı etkilerinin dikkate alınmasıdır. Bu tür çok boyutlu MTK modelleri araştırmacılar tarafından sıklıkla kullanılmaktadır. Bunlar, iki faktörlü modeller (Gibbons ve Hedeker, 1992) ve rastgele etkili madde takımı tepki modelleridir (Bradlow, Wainer ve Wang, 1999; Wainer vd., 2007). Li ve diğerleri (2006), Rijmen (2010) ve Min ve He (2014) rastgele etkili madde takımı modellerinin iki faktörlü modellerin özel bir durumu olarak kullanılabileceğini belirtmiştir. Özel boyut üzerindeki yüklerin, her bir madde takımı içindeki genel boyut üzerindeki yüklerle orantılı olacak şekilde kısıtlanmasıyla elde edilir. Özet olarak madde takımlarından oluşan testlerde, yerel bağımlılık miktarlarını dikkate alan ve her madde takımı boyunca her bir bireye özgü yerel bağımlılık miktarını belirten ek bir parametrenin yer aldığı Madde Takımı Tepki Kuramı (MTTK) gibi karmaşık modellere ihtiyaç vardır (Wainer, Bradlow ve Du, 2000).

Araştırmacılar genellikle madde takımlarından oluşan testlerde tek boyutlu MTK, MTTK ve bifaktör modellerin parametre ve yetenek kestirimi üzerine yoğunlaşmıştır (Baghaei ve Ravand, 2016; DeMars, 2006; Soysal ve Yılmaz Koğar, 2022; Yılmaz Koğar, 2021). Madde takımlarında DMF analizinin yapıldığı çalışma sayısı oldukça azdır (Paek ve Fukuhara, 2015; Tasdelen Teker ve Dogan, 2015; Wainer, 1995). Bu çalışmada ise MTK ve MTTK modellerine göre parametre kestirimi ile sınav katılımcılarının doğru bir şekilde sınıflandırma performansları ve madde/madde takımlarının DMF/DMGF içerme durumları incelenmiştir. Özellikle eğitim alanında öğrencilerin doğru şekilde sınıflandırılması oldukça önemlidir. Bu nedenle, bu çalışmada madde takımlarından oluşan testlerde yerel bağımlılığın sınıflandırma doğruluğunu nasıl etkilediği ve bu doğruluğun iki ve çok kategorili sınıflandırmalarda nasıl değiştiği incelenmiştir. Ayrıca bu konuda göz önünde bulundurulması gereken önemli bir husus, sınav katılımcılarının belirli alt gruplarının madde güçlüğünün ötesindeki faktörler nedeniyle madde takımları üzerinde farklı performans örüntüleri sergileyebildiği değişen madde fonksiyonu (DMF) etkisidir. Bu tür grup farklılıklarının hesaba katılmaması, yanlı parametre tahminlerine ve öğrenci yeterliğinin yanlış sınıflandırılmasına yol açabilir. Bu nedenle madde takımlarından oluşan testlerde madde grubu dikkate alınarak DMF çalışmalarının yapılarak karşılaştırmalı olarak bu konunun incelenmesinin alan yazına katkı sunacağı düşünülmektedir. Ek olarak madde takımlarından oluşan testlerin nasıl modelleneceği alan yazında genellikle PISA, SAT veya simülasyon verileri ile incelenmiştir (Chang ve Yang, 2010; Koziol, 2016; Yılmaz Koğar, 2021). Fakat bu modellerin farklı gerçek veri setleri ile çalışılması araştırmacılara daha detaylı bilgi ve daha çok karşılaştırma yapma imkânı sunacaktır. Bu nedenle bu çalışmada TIMSS-2019 veri seti ele alınarak geleneksel MTK ve MTTK modelleri kullanılarak parametre kestirimi yapılmış, sınıflama doğruluğu ve DMF'ye ilişkin sonuçlar karşılaştırmalı olarak incelenmiştir. Sınıflama doğruluğu ile ilgili geçti-kaldı gibi iki kategorili sınıflama doğrulukları ile ilgili çalışmalar alan yazında yer almakta olup bu çalışmaların sayısı oldukça azdır (Koziol, 2016; Zhang, 2010). TIMSS veri seti kullanılarak farklı modellere (MTK-MTTK) göre çok

kategorili sınıflandırma ve madde takımlarında DMF sonuçlarının karşılaştırılmasının alanyazına özgün bir katkısı olacağı düşünülmektedir. Ayrıca elde edilen bulgu ve yorumların alan yazındaki diğer çalışma bulgularıyla karşılaştırılarak tartışılması madde takımlarından oluşan testlerde sınıflama doğruluğu ve DMF incelemelerine ilişkin de genel bir perspektif oluşturacaktır. Bu amaçla bu çalışmada eTIMSS 2019 uygulamasına ait Türkiye örnekleminde 13. ve 14. kitapçıkta ortak olarak yer alan yer alan üç madde takımı kullanılarak aşağıdaki araştırma sorularına cevap aranmıştır:

1. Matematik alt testinde yer alan üç madde takımının yerel bağımlılıkları ne düzeydedir?
2. Madde takımlarından oluşan matematik alt testinde 2PL-MTK ve 2PL-MTTK modellerinden elde edilen madde ve yetenek parametreleri arasındaki ilişki nasıldır?
3. Madde takımlarından oluşan matematik alt testinde 2PL-MTK ve 2PL-MTTK modellerinden elde edilen sınıflama doğruluğu nasıldır?
4. Madde takımlarından oluşan matematik alt testinde cinsiyete göre DMF/DMGF içeren maddeler var mıdır?

## Yöntem

### *Araştırma Türü*

Bu çalışmada, madde takımlarından oluşan bir test farklı MTK modelleri kullanılarak parametre kestirimi, sınıflama doğruluğu ve DMF/DMGF açısından incelenmiştir. Farklı yöntemlerden elde edilen sonuçların derinlemesine karşılaştırıldığı var olan durum hakkında daha çok bilgi sağlayan bu çalışma betimsel bir araştırmadır (Creswell, 2014; Karasar, 2016).

### *Çalışma Grubu*

Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS), Uluslararası Eğitim Başarılarını Değerlendirme Kuruluşu (IEA) tarafından dört yılda bir düzenlenen bir başarı izleme çalışmasıdır. İlk kez 1995 yılında yapılan TIMSS, dört yıllık periyotlarla uygulanmakta olup uluslararası gerçekleştirilen önemli bir araştırmadır. TIMSS, dördüncü ve sekizinci sınıf düzeyindeki öğrencilerin matematik ve fen bilimlerindeki başarılarını değerlendirmek amacıyla yapılmaktadır. Gerçekleştirilen 2019 döngüsünde sekizinci sınıf düzeyinde 39 katılımcı ülkeden biri olan Türkiye örneklemini 181 okuldaki 4.077 öğrenci oluşturmaktadır. TIMSS, 2019 uygulamasında bilgisayar tabanlı değerlendirmeye (eTIMSS) geçiş yapmıştır. Bu çalışmada ise eTIMSS 2019 uygulamasına ait Türkiye örnekleminde 13. ve 14. kitapçıkta ortak olarak yer alan yer alan üç madde takımı kullanılmıştır. Bu madde takımları sırasıyla iki, dört ve altı maddeden oluşmaktadır. Analizlerde kayıp veri içeren 89 katılımcı için liste bazında silme yöntemi kullanılmıştır. Kayıp verilerin silinmesiyle toplam 503 öğrencinin yanıtları kullanılmış olup bu öğrencilerin %47.9 'u kız % 52.1'i ise erkektir.

### *Verilerin Analizi*

Model veri uyumu hem MTK hem de MTTK modelleri kapsamında incelenmiş ve en iyi uyumun 2PL modelde sağlandığı tespit edilmiştir. Sonuçlar incelendiğinde, 3PL MTK ve 3PL MTTK modellerinde madde ayırt edicilik ve kesişim parametrelerinin olağan dışı değerler aldığı gözlemlenmiştir. Bu nedenle analizlere 2PL MTK ve 2PL MTTK modelleri ile devam edilmiştir. eTIMSS 2019 uygulamasında 13. ve 14. kitapçıkta ortak olarak yer alan üç madde takımının 2PL-MTTK (Wainer vd., 2007) ve 2PL-MTK (Birnbaum, 1968) modellerine göre madde ve yetenek parametre değerleri kestirilmiş ve karşılaştırmalı olarak incelenmiştir. Bu çalışmada madde parametreleri olarak eğim ve kesişim katsayıları ele alınmıştır. Yetenek kestirimleri beklenen sonsal dağılım (EAP) yöntemi kullanılarak gerçekleştirilmiştir. R yazılımında gerçekleştirilen madde ve yetenek parametre kestirimlerinde ve DMF analizlerinde *mirt* (Chalmers, 2012), sınıflama doğruluğunda *cacIRT* (Lathrop, 2015) paketleri kullanılmıştır.

Bu çalışmada, farklı MTK modellerinden elde edilen madde ve yetenek parametre kestirimlerine karşılık gelen standart hatalar hesaplanmış ve bu modeller karşılaştırılmıştır. Ayrıca madde ve yetenek parametre değerleri arasındaki ilişkiyi incelemek amacıyla Spearman Sıra Farkları Korelasyon Katsayısı kullanılmıştır. Sınıflama doğruluğu hesaplamalarında MTK'ya dayalı

yaklaşımlardan Rudner yöntemi kullanılmıştır. DMF/DMGF belirlenmesinde hem maddeleri bağımsız hem de madde grubu olarak incelenmesine izin veren SIBTEST yöntemi kullanılmıştır.

***Madde Takımı Tepki Kuramı***

Madde Takımı Tepki Kuramı, madde takımlarından oluşan maddelerdeki yerel bağımlılığı ele almaktadır. Yerel bağımlılık uygun şekilde ele alınmazsa, teste bağlı psikometrik sonuçlar olumsuz etkilenebilir. Son yirmi yıldır yerel madde bağımlılığını farklı perspektiflerden yakalamak için madde takımı yapısı modellemesi üzerine çeşitli yöntemler önerilmiştir. Bradlow ve diğerleri (1999) ve Wainer ve diğerleri (2000), madde takımları ve kişiler arasındaki etkileşimi açıklamak için rastgele etki parametresi içeren geleneksel MTK modellerini genişletmiştir. Bu durumda bir madde takımında yer alan maddeler arasındaki yerel bağımlılık düzeylerini hesaba katan $\gamma$ rastgele madde takımı etki parametresi, b madde güçlüğü, a madde ayırt ediciliği parametrelerinden oluşan 2PL-MTTK modeli şu şekildedir:

$$P(\theta_i, \alpha_i, b_i) = \frac{exp\ (\alpha_i(\theta_j - b_i - \gamma_{jd(i)}))}{1 + \ exp\ (\alpha_i(\theta_j - b_i - \gamma_{jd(i)}))}$$

Madde takımı etki parametresi ($\gamma_{jd(i)}$), bireye ve madde takımına özgü bir parametredir. Yerel bağımsızlık varsayımı sağlandığında, bu parametre değeri sıfır yani tüm bireyler için $\gamma_{jd(i)} = 0$ olur ve bu durumda MTTK modeli tek boyutlu MTK modeline dönüşür. $\gamma_{jd(i)}$'nin varyansı tipik olarak her madde takımı için tahmin edilir ve her madde takımı içindeki maddelerin yerel bağımlılık derecesinin bir göstergesi olarak kullanılır. Madde takımı etkisinin varyansları madde takımları boyunca değişmektedir. Ayrıca şans parametresinin ($c_i$) eklenmesi durumunda 3PL-MTTK, 2PL-MTTK'nin özel bir hâli olur. Fakat 3PL MTTK modeli diğer MTTK modellerine göre daha fazla parametre içerdiğinden hesaplama algoritmaları daha karmaşıktır.

Li ve diğerleri (2006) çok boyutluluk perspektifinden genel bir iki parametreli normal ogive madde takımı tepki kuramı modeli (2PNOTRT) önermiştir. Çok boyutlu modeldeki her bir madde yanıtı hem birincil boyuta hem de ikincil madde takımı boyutlarına bağlıdır. Her iki MTTK modeli probit bağlantı fonksiyonu çerçevesinde oluşturulmuştur. Bu temelde, Zhan, Li, Wang ve Bian (2015) madde içi çok boyutlu madde takımı etkisi kavramını önermektedir. Lu, Zhang, Zhang, Xu ve Tao (2021), ikili puanlanan maddeler için logit bağlantı fonksiyonuna dayalı yeni bir madde takımı ayırt edicilik parametresi önermiştir. Bu parametre, büyük ölçekli dil değerlendirmeleri (Eckes, 2014; Rijmen, 2010; Zhang, 2010), hiyerarşik veri analizleri (Jiao, Kamata, Wang ve Jin, 2012) ve bilişsel tanı değerlendirmeleri (Zhan vd., 2018) gibi eğitim ve psikoloji alanlarında da uygulanmıştır.

MTTK modelleri için en yaygın kullanılan tahmin yöntemlerinden biri, beklenti maksimizasyonu (EM; Dempster, Laird ve Rubin, 1977) algoritması aracılığıyla marjinal maksimum olabilirlik yöntemidir (Bock ve Aitkin, 1981; Glas, Wainer ve Bradlow, 2000; Mislevy, 1986; Wang ve Wilson, 2005). Yetenek parametreleri ve madde takımı etkileri, gözlenemeyen gizil değişkenler olarak görülür ve daha sonra gözlenemeyen veriler üzerinden marjinalize edilmiş tam bir veri olasılığının (yanıtlar ve gözlenemeyen veriler) maksimumu hesaplanabilir. Bununla birlikte, MTTK modellerinin marjinal maksimum olabilirlik tahmini, hesaplamaların genellikle analitik olarak zorlayıcı yüksek boyutlu integral içermesi ve dolayısıyla parametrelerin maksimum olabilirlik tahminini bulmanın zor olması nedeniyle engellenmiştir. Daha spesifik olarak, gizil değişken dağılımları üzerindeki integraller Gauss quadrate kullanılarak değerlendirildiğinde (Bock ve Aitkin, 1981), ilgili hesaplamaların sayısı gizil değişken boyutlarının sayısı ile üstel olarak artmaktadır. Uyarlanabilir Gauss quadrate kullanıldığında boyut başına kareleme noktası sayısı azaltılabilse de (Pinheiro ve Bates, 1995), toplam nokta sayısı yine boyut sayısı ile üstel olarak artmaktadır. Tabii ki tüm bu MTTK analiz süreci daha karmaşık ve uzun bir süreci kapsamaktadır.

***Sınıflama Doğruluğu***

Sınıflama doğruluğu, test puanlarına dayalı olarak verilen kararların, puanların herhangi bir ölçme hatası içermemesi durumunda verilecek kararlarla ne ölçüde eşleştiğini ifade eder (Hambleton ve Novick, 1973). Eğitim ve psikolojide tüm ölçmelere ölçme hatası karıştığı için sınıflama doğruluğunun tespit edilmesi gerekmektedir. Bir sınav katılımcısının yanlış sınıflandırılması bir sınıflandırma hatasına işaret eder. Bir sınav katılımcısı gerçek yeterlik kategorisinden daha yüksek bir yeterlilik kategorisinde sınıflandırıldığında ya da gerçek yeteneğinden daha düşük bir kategoriye yerleştirildiğinde sınıflama hatası ortaya çıkar. Sınıflama doğruluğu, yüksek riskli sınavlarda öğrencilerin geleceği için önemli iken, bir değerlendirmenin güçlü ve zayıf yönleri hakkında değerli bilgiler sağlaması da eğitimci ve politika yapıcıların veriye dayalı bilinçli kararlar almasına yardımcı olabilir. Sınıflamanın doğruluğu, öğretim kararlarına rehberlik etmek, program etkinliğini değerlendirmek ve öğrencilerin başarılı olmak için ihtiyaç duydukları desteği almalarını sağlamak gibi kritik kararlar açısından çok önemlidir (Cizek ve Bunch, 2007).

Ölçüt ve norm referanslı yapılan değerlendirmelerin her ikisinde de sınıflandırma yapılmaktadır. İki kategorili sınıflandırmalara örnek olarak "başarılı" ve "başarısız", çoklu sınıflandırma kategorilerine örnek olarak ise "temel", "yeterli" ve "ileri" gibi seviyeler örnek gösterilebilir. Ölçüt referanslı testlerde kullanılan bir sınıflamaya Ölçme Seçme ve Yerleştirme Merkezi'nin (ÖSYM) Yabancı Dil Sınavına (YDS) göre yapılan dil puanı değerlendirmesindeki A, B, C, D gibi kategoriler örnek verilebilir. Norm referanslı testlerdeki sınıflandırmalara örnek ise Yükseköğretim Kurumları Sınavı (YKS) verilebilir. YKS sonuçları, öğrencilerin performansını diğer adaylarla karşılaştırarak değerlendirme yapma imkanı sunar. Sınıflama doğruluğunun hesaplanmasında geliştirilen ilk yöntemler iki uygulamaya dayalı geliştirilmiştir. Ancak pratikte iki uygulamanın getirdiği zorluklar nedeniyle tek uygulamaya dayalı sınıflama doğruluğu indeksleri geliştirme çalışmaları artmıştır. Yöntemlerde temelde geliştirildikleri kurama göre KTK'ye dayalı yöntemler (Hanson ve Brennan, 1990; Huynh, 1976; Lee ve Song, 2004; Livingston ve Lewis, 1995; Subkoviak, 1976) ve MTK'ya dayalı yöntemler (Lee, 2010; Rudner, 2001, 2005) olarak ikiye ayrılmaktadır. MTK çerçevesi altında, yeteneğin nokta tahmini, gizil özellikteki gerçek puan olarak ele alınabilir. Rudner (2001, 2005), sınıflandırmaların beklenen olasılığının hesaplanması yoluyla karar doğruluğunu değerlendirmek için bir yöntem sunmuştur. Rudner (2001, 2005) yetenek (θ) ölçeğine dayanan indeksler üretmiştir. Bu yöntemde, sınıflamaların beklenen olabilirliği hesaplanarak sınıflama doğruluğu elde edilmektedir. Geçme/kesme puanı $\theta_c$, A yanıtlayıcısına ait gerçek yetenek $\theta_n$, B yanıtlayıcısına ait gerçek yetenek $\theta_m$ olsun. $\theta_m > \theta_c > \theta_n$ olduğu için A yanıtlayıcısının, bütün kestirimlerde "kaldı" sınıfında, B yanıtlayıcısı ise bütün kestirimlerde "geçti" şeklinde sınıflandırılmalıdır. Fakat yetenek kestirimindeki hata nedeniyle, her gerçek theta (θ)'ya koşullu bir dağılım eşlik eder. Ancak, aday A şans yoluyla geçti/başarılı olarak sınıflandırılabilir. Bu şans, theta tahminlerinin kesme puanı $\theta_c$'den daha büyük olduğu durumda gerçekleşir. Sınıflandırma terminolojisinde bu şans, başarısız ya da uzman olmayan birinin başarılı veya uzman olarak tanımlandığı yanlış pozitif bir hata yapılması olasılığıdır.

Bu çalışmada öğrencilerin olası puanları, yeterlik düzeyi sınıflandırmalarında kullanılan kesme puanlarına dayalı olarak orta düzey altı ve üstü şeklinde iki kategorili sınıflandırılmıştır. Ardından dört yeterlik düzeyi (alt düzey-orta düzey-üst düzey-ileri düzey) için çok kategorili bir sınıflama yapılmıştır. Farklı MTK modelleri arasında karşılaştırılabilir bu sınıflama için öncelikle kesme puanı ve bu kesme puanına karşılık gelen yetenek düzeyleri belirlenmiştir.

***Değişen Madde ve Madde Grubu Fonksiyonu***

Değişen madde fonksiyonu (DMF), ölçülmek istenilen özellik bakımından aynı yetenek düzeyindeki farklı gruplardaki bireylerin bir maddeyi doğru cevaplama olasılıklarının birbirinden farklı olması durumunda ortaya çıkmaktadır (Clauser ve Mazor, 1998). DMF'nin anlaşılması ve ele alınması, değerlendirme uygulamalarında adalet ve geçerliğin sağlanması için çok önemlidir. DMF'nin tespiti, adilliğin değerlendirilmesinde kritik bir adımdır. DMF belirleme yöntemleri arasında Mantel-Haenszel yöntemi (Holland ve Thayer, 1988), standartlaştırma yöntemi (Dorans ve Kulick, 1986), lojistik regresyon yöntemi (Swaminathan ve Rogers, 1990), SIBTEST yöntemi (Shealy ve Stout, 1993), Lord'un

(1980) ki- kare testi (Wright ve Stone, 1979); olabilirlik oranı testi (Thissen, Steinberg ve Wainer, 1988; Wang ve Yeh, 2003), çoklu göstergeler çoklu nedenler modeli (Finch, 2005; Oort, 1998) sayılabilir. Ayrıca son yıllarda Bilişsel Tanı Modellerine dayalı yöntemlerle DMF tespit edildiği çalışmalar da mevcuttur (Eren, Gündüz ve Tan, 2023; Ma, Terzi ve de la Torre, 2021).

DMF'nin, ölçülmek istenen özellikle ilgili olmayan test maddesinin bazı özelliklerinden dolayı ortaya çıkmasına dayanarak, DMF çok boyutluluk çerçevesinde de tanımlanmıştır. Bu çerçeve, tüm testlerin bir dereceye kadar çok boyutlu olduğu varsayımını temel almaktadır. Bir testte ölçülmek istenen yapı ile ilgili ana bir boyut ve yapı ile alakasız varyans üreten başka boyutlar da bulunabilir. Örneğin, probleme dayalı bir matematik testinde, test matematik yeteneğini yansıtan birincil boyutların yanı sıra okuduğunu anlama veya sözel yetenek gibi diğer ikincil yetenekleri yansıtan boyutlardan oluşacaktır. Bu diğer boyutlar genellikle birincil boyut ile korelasyon göstermektedir. Bu mantıkta DMF'in testteki birincil boyuttan farklı boyutlardan kaynaklandığı düşünülmektedir. Ackerman (1992), çok boyutlu çerçevenin temelini kapsamlı bir şekilde tartışmıştır. Shealy ve Stout (1993) bu çerçeve kapsamında SIBTEST adında bir DMF istatistiği geliştirmişlerdir. SIBTEST, DMF kaynağı olarak çok sayıda boyutun testine izin vermektedir. Bu yöntem, bir tür faktör analizi içerdiğinden, analizde tek tek maddeler yerine madde gruplarının incelemesine de izin vermektedir. Maddelerin gruplanarak DMF analizine alınmasına imkan vermesi nedeniyle DMF kaynaklarına dair genellemelerin daha sağlıklı yapılmasında SIBTEST yöntemi kullanılabilmektedir (Gierl, Bisanz., Bisanz ve Boughton, 2003; Mendes-Barnett ve Ercikan, 2010).

Lojistik regresyon, Mantel Haenszel gibi geleneksel DMF belirleme yöntemlerinin aksine SIBTEST, DMF'yi hem tek tek madde hem de demet/madde takımı düzeyinde inceleyebilme avantajına sahiptir. Açılımı eşzamanlı madde yanlılığı testi olan SIBTEST, ölçülen temel yetenek kontrol edildikten sonra, bir madde veya madde grubunun iki veya daha fazla sınav alt grubu için farklı fonksiyon gösterme derecesini değerlendiren regresyona dayalı bir yöntemdir. Yöntem, öncelikle odak ve referans gruplar arasındaki toplam test puanı ilişkisinin tahmin edilmesini ve ardından madde veya madde grubu için bu genel ilişkiden sapma durumunun test edilmesini içerir. Bu çalışmada hem madde hem de madde grubu bazında çalışabilen SIBTEST yöntemi, DMF ve Değişen Madde Grubu Fonksiyonu (DMGF) belirlemek için kullanılmıştır. SIBTEST yöntemine dayalı DMF belirlemede Roussos ve Stout (1996) tarafından $\beta$ indeksini yorumlayabilmek için bir sınıflama önerilmiştir. $\beta$ indeksinin 0.059'dan küçük olması DMF'nin olmadığının, 0.088'den küçük olması DMF'nin orta düzeyde olduğunun ve 0.088'e büyük ve eşit olması ise DMF'nin yüksek düzeyde olduğunun göstergesidir. DMF için bir etki büyüklüğü sınıflaması mümkün iken DMGF için bu şekilde bir etki büyüklüğü sınıflaması yoktur.

## Bulgular

"Matematik alt testinde yer alan üç madde takımının yerel bağımlılık düzeyleri nedir?" sorusuna yönelik yapılan analizler sonucunda, iki maddeden oluşan madde takımı 1, dört maddeden oluşan madde takımı 2 ve altı maddeden oluşan madde takımı 3'ün yerel bağımlılık düzeyleri orta düzeyde ($\sigma$> 0.5) bulunmuştur. Madde takımı kaynaklı yerel madde bağımlılığına "madde takımı etkisi" adı verilmektedir (Wainer ve Kiely, 1987). Varyans arttıkça, madde takımlarının oluşturduğu etki de artar (Wainer ve Wang, 2000). Varyans değerleri sırasıyla 0 için "madde takımı etkisi yok," 0.5 için "orta" ve 1 için ise "büyük düzeyde madde takımı etkisi" şeklinde yorumlanmaktadır (Wang, Bradlow ve Wainer, 2002; Wang ve Wilson, 2005). Bu çalışmada, 2PL-MTTK kullanılarak üç madde takımına ilişkin madde takımı etkileri sırasıyla 0.575, 0.505 ve 0.615 olarak bulunmuştur. En yüksek yerel bağımlılık üçüncü madde takımında tespit edilse de genel olarak üç madde takımında da yerel bağımlılıklar orta düzeydedir. Sonuç olarak, tüm madde takımları önemli madde takımı etkisine sahip değildir.

İkinci araştırma sorusu olan "Madde takımlarından oluşan matematik alt testinde 2PL-MTK ve 2PL-MTTK modellerinden elde edilen madde ve yetenek parametreleri arasındaki ilişki nasıldır?" için öncelikle 2PL-MTK ve 2PL-MTTK modellerine göre madde ve yetenek parametreleri hesaplanmıştır. Madde parametreleri ve standart hata değerleri Tablo 1'de verilmiştir. Tablo 1'de $\alpha$ eğim, $\delta$ ise kesişim parametresi için kullanılmıştır.

**Tablo 1.** Madde Parametre ve Standart Hata Değerleri

| | | 2PL-MTK | | | | 2PL-MTTK | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Madde Takımı** | **Maddeler** | **$\alpha$** | **$\alpha_{sh}$** | **$\delta$** | **$\delta_{sh}$** | **$\alpha$** | **$\alpha_{sh}$** | **$\delta$** | **$\delta_{sh}$** |
| Madde Takımı I | ME72041A | 4.06 | 0.77 | -0.40 | 0.06 | 4.44 | 1.63 | -0.62 | 0.39 |
| | ME72041B | 4.40 | 0.87 | -1.06 | 0.06 | 5.91 | 2.89 | -1.84 | 0.93 |
| Madde Takımı II | ME72081A | 1.03 | 0.16 | 1.24 | 0.18 | 1.57 | 0.38 | 1.63 | 0.31 |
| | ME72081B | 0.68 | 0.13 | 0.52 | 0.19 | 0.98 | 0.19 | 0.61 | 0.13 |
| | ME72081C | 0.73 | 0.13 | -0.48 | 0.17 | 0.75 | 0.18 | -0.49 | 0.11 |
| | ME72081D | 0.72 | 0.14 | 1.12 | 0.29 | 0.79 | 0.20 | 1.20 | 0.14 |
| Madde Takımı III | ME72140A | 1.76 | 0.26 | 1.89 | 0.12 | 1.93 | 0.35 | 2.33 | 0.31 |
| | ME72140B | 1.84 | 0.34 | 3.21 | 0.19 | 2.13 | 0.47 | 4.03 | 0.61 |
| | ME72140C | 1.64 | 0.25 | 2.20 | 0.15 | 1.98 | 0.37 | 2.85 | 0.39 |
| | ME72140D | 1.02 | 0.21 | 2.35 | 0.38 | 1.32 | 0.25 | 2.83 | 0.29 |
| | ME72140E | 0.75 | 0.14 | 0.91 | 0.23 | 0.68 | 0.14 | 0.94 | 0.12 |
| | ME72140F | 1.48 | 0.25 | 2.48 | 0.20 | 1.44 | 0.31 | 2.74 | 0.29 |

Madde eğim parametresi ($\alpha$), madde ayırt edicilik parametresi olarak yorumlanır. Yüksek değerler, maddenin daha ayırt edici olduğunu gösterir (Baker, 2001). Her iki modele ait parametre değerleri incelendiğinde MTK modeline göre ayırt edicilik parametrelerinin 0.68 ile 4.40 standart hatalarının 0.13 ile 0.87 arasında değiştiği; MTTK modeline göre ise ayırt edicilik değerlerinin 0.68 ile 5.91; bu parametreye ait standart hataların ise 0.14 ile 2.89 arasında değiştiği gözlenmiştir. Madde kesişim parametresi ($\delta$) madde kolaylığı olarak yorumlanır ve madde güçlük parametresinin tersidir. Yüksek değer, maddenin kolay olduğu anlamına gelir (Reckase, 2009). Kesişim parametresinin ise MTK modelinde -1.06 ile 3.21 standart hatalarının 0.06 ile 0.38 arasında değiştiği; MTTK modeline göre ise bu değerlerin -1.84 ile 4.03; bu parametreye ait standart hataların ise 0.11 ile 0.93 arasında değiştiği gözlenmiştir. Tablo 2'de ise bu madde parametre kestirimlerinin ortalama, minimum ve maksimum değerleri verilmiştir.

**Tablo 2.** Kestirilen Madde Parametre Değerlerinin Betimleyici İstatistikleri

| **Parametreler** | **2PL-MTK** | | | **2PL-MTTK** | | |
|---|---|---|---|---|---|---|
| | **Ortalama** | **Min** | **Max** | **Ortalama** | **Min** | **Max** |
| Eğim ($\alpha$) | 1.68 | 0.68 | 4.40 | 1.99 | 0.67 | 5.91 |
| $\alpha_{sh}$ | 0.30 | 0.13 | 0.87 | 0.61 | 0.14 | 2.89 |
| Kesişim ($\delta$) | 1.17 | -1.06 | 3.21 | 1.35 | -1.84 | 4.03 |
| $\delta_{sh}$ | 0.19 | 0.06 | 0.38 | 0.34 | 0.11 | 0.93 |

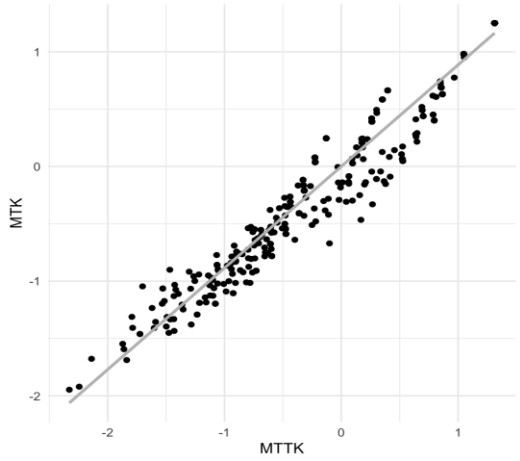
Tablo 2'de yer alan eğim parametreleri ve standart hataları her iki modele göre incelendiğinde. 2PL-MTK modelinin değerleri daha düşük çıkmıştır. Bu fark standart hatalarda daha düşük bulunmuştur. Benzer şekilde madde kesişim parametreleri incelendiğinde de benzer durum elde edilmiş 2PL-MTK modeli ile daha düşük parametre ve standart hata değerleri elde edilmiştir. Ayrıca madde kesişim parametrelerinin standart hataları ayırt ediciliğe göre daha düşük bulunmuştur. Yetenek parametre ve standart hatalarına ilişkin değerler ise Tablo 3'te verilmiştir.

**Tablo 3.** Yetenek Parametresi ve Standart Hatalarına İlişkin Betimsel İstatistikler

| **Model** | **Yetenek Parametre Değeri ($\theta$)** | | | **Standart Hata (sh)** | | |
|---|---|---|---|---|---|---|
| | **Ortalama** | **Min** | **Max** | **Ortalama** | **Min** | **Max** |
| 2PL-MTK | 0.00 | -2.33 | 1.31 | 0.45 | 0.34 | 0.63 |
| 2PL-MTTK | 0.00 | -1.96 | 1.25 | 0.59 | 0.45 | 0.70 |

Tablo 3 incelendiğinde her iki modele ait yetenek parametre ve standart hata değerlerinin benzer olduğu ancak 2PL-MTK parametre değerlerinin ranjının daha geniş, standart hatalarının ise

daha düşük olduğu görülmektedir. Her iki modelden elde edilen yetenek parametreleri arasındaki ilişki saçılım grafiği ile Şekil 1'de verilmiştir.




**Şekil 1.** Yetenek Parametre Kestirimine ilişkin Saçılım Grafiği

Şekil 1'de MTK ve MTTK modellerine göre elde edilen yetenek kestirimlerinin saçılım grafiği gösterilmektedir. Saçılım grafiğine göre her iki modelden elde edilen yetenek parametre kestirimlerinin çok benzer olduğu söylenebilir. Eğim ve kesişim parametreleri arasında da yüksek ilişki olmasına rağmen her iki modele göre kestirilen bu değerler aynı ölçek üzerinde yer almadığından grafiksel gösterim ile karşılaştırmaya yer verilmemiştir. MTK ve MTTK modellerinden elde edilen parametreler arasındaki korelasyon değerleri Tablo 4'te verilmiştir.

**Tablo 4.** MTK ve MTTK Modelleriyle Kestirilen Parametrelere İlişkin Korelasyon Değerleri

| **Madde Takımı** | **Madde sayısı** | **$\alpha$** | **δ** | **θ** |
|---|---|---|---|---|
| Madde Takımı I | 2 | | | |
| Madde Takımı II | 4 | 0.983 | 0.997 | 0.976 |
| Madde Takımı III | 6 | | | |

Tablo 4 incelendiğinde, elde edilen tüm korelasyon değerleri oldukça yüksek bulunmuştur (r>.95). Bu korelasyon değerleri eğim parametresi ($\alpha$) için 0.983. kesişim parametresi (δ) için 0.997 ve yetenek kestirimleri (θ) için 0.976 bulunmuştur.

Üçüncü araştırma sorusunu "Madde takımlarından oluşan matematik alt testinde 2PL-MTK ve 2PL-MTTK modellerinden elde edilen sınıflama doğruluğu nasıldır?" cevaplamak için öncelikle kesme puan değerleri bulunmuştur. Kesme puanın hesaplanmasında öğrencilerin matematik testi olası değerleri ve TIMSS uygulamasında öğrencilerin başarılarının davranış göstergelerini oluşturmak üzere tanımlanan dört farklı yeterlilik düzeyi (alt, orta, yüksek, ileri) temel alınmıştır. İki MTK modeli arasında karşılaştırılabilir sınıflama için belirlenen kesme puanına karşılık gelen yetenek düzeyleri belirlenmiştir. Belirlenen yetenek düzeyinin altında kalanlar "düzey altı", üzerindekiler ise "düzey üstü" olarak sınıflandırılmıştır. Bu sınıflama alt düzey altı-üstü, orta düzey altı-üstü, yüksek düzey altı-üstü ve ileri düzey altı-üstü olmak üzere TIMSS'te yer alan dört farklı yeterlilik düzeyine ilişkin yapılmıştır. Sınıflama doğruluğuna ilişkin elde edilen bulgular Tablo 5'te yer almaktadır.

**Tablo 5.** İki Kategorili Sınıflama Doğruluğuna İlişkin Değerler

| **Model** | **2PL-MTK** | | | | **2PL-MTTK** | | | |
|---|---|---|---|---|---|---|---|---|
| **Düzey** | **Alt** | **Orta** | **Yüksek** | **İleri** | **Alt** | **Orta** | **Yüksek** | **İleri** |
| Kesme değer (θ) | -0.74 | -0.32 | 0.40 | 1.31 | -0.64 | -0.21 | 0.39 | 1.25 |
| Sınıflama doğruluğu | 0.94 | 0.90 | 0.93 | 0.95 | 0.90 | 0.93 | 0.94 | 0.91 |
| Sınıflama tutarlılığı | 0.91 | 0.86 | 0.91 | 0.94 | 0.86 | 0.90 | 0.93 | 0.86 |

Tablo 5 incelendiğinde, 2PL-MTK için alt düzey kesme noktası -0.74, orta düzey kesme noktası -0.32, yüksek düzey kesme noktası 0.40, ileri düzey kesme noktası ise 1.31'dir. 2PL-MTTK için alt düzey kesme noktası -0.64, orta düzey kesme noktası -0.21, yüksek düzey kesme noktası 0.39, ileri düzey kesme noktası ise 1.25'tir. İki kategorili sınıflandırma için belirlenen kesme değerler her bir modelde düzey arttıkça beklenen bir şekilde artmaktadır. Çünkü düzeylerin tanımları gereği alt düzeyden ileri düzeye gidildikçe öğrencilerin bilgileri artmakta ve bu bilgiyi kullanma şekilleri karmaşıklaşmaktadır. Örneğin alt düzeydeki bir öğrencinin matematiğe ilişkin temel bilgiye sahip olması yeterli iken ileri düzeydeki öğrencinin matematiğe ilişkin bilgilerini karmaşık durumlara uygulayabilir ve gerekçelerini açıklayabilir olması gerekmektedir. Bu nedenle ileri düzeyde bir öğrencinin daha yüksek bir yeteneğe sahip olması beklenir. İki kategorili orta ve yüksek düzeyde sınıflama doğruluğu ve tutarlılığı 2PL-MTTK modelinde az da olsa daha yüksek bulunmuşken alt düzey ve ileri düzeyde 2PL-MTK modelinde daha yüksek bulunmuştur. Sonuç olarak elde edilen değerlerin tüm düzeylerde oldukça yüksek ve benzer olduğu söylenebilir.

**Tablo 6.** Çok Kategorili Sınıflama Doğruluğuna İlişkin Değerler

| **Model** | **2PL-MTK** | **2PL-MTTK** |
|---|---|---|
| Sınıflama doğruluğu | 0.74 | 0.73 |
| Sınıflama tutarlılığı | 0.68 | 0.64 |

Tablo 6 incelendiğinde çok kategorili sınıflandırma için alt, orta, yüksek ve ileri olmak üzere dört düzey birlikte değerlendirildiğinde az da olsa 2PL-MTK modelinin sınıflama doğruluğu ve tutarlılığı 2PL-MTTK modelinden daha yüksek bulunmuştur.

"Madde takımlarından oluşan matematik alt testinde cinsiyete göre DMF/DMGF gösteren maddeler var mıdır?" olan dördüncü araştırma sorusunu cevaplandırmak için R yazılımından *mirt* paketinin SIBTEST fonksiyonu kullanılmıştır. Tablo 7'de her bir madde ve her bir madde grubu için sonuçlara yer verilmiştir.

**Tablo 7.** Değişen Madde ve Değişen Madde Grubu Fonksiyonu Sonuçları

| | | **DMF** | | **DMGF** | |
|---|---|---|---|---|---|
| | **Maddeler** | **β** | **p** | **β** | **p** |
| **Madde Takımı I** | ME72041A | -0.012 | 0.797 | -0.164 | 0.056 |
| | ME72041B | -0.065 | 0.136 | | |
| **Madde Takımı II** | ME72081A | -0.062 | 0.171 | -0.037 | 0.741 |
| | ME72081B | -0.009 | 0.855 | | |
| | ME72081C | 0.064 | 0.175 | | |
| | ME72081D | 0.033 | 0.468 | | |
| **Madde Takımı III** | ME72140A | 0.016 | 0.719 | 0.137 | 0.38 |
| | ME72140B | -0.007 | 0.872 | | |
| | ME72140C | 0.056 | 0.269 | | |
| | ME72140D | -0.052 | 0.156 | | |
| | ME72140E | -0.055 | 0.244 | | |
| | ME72140F | 0.074 | 0.107 | | |

Tablo 7 incelendiğinde SIBTEST yöntemine göre madde takımlarındaki hiçbir maddenin DMF göstermediği ($p > .05$), ayrıca hiçbir madde madde grubunun da DMGF göstermediği belirlenmiştir.

## Tartışma, Sonuç ve Öneriler

Bu araştırmada, eTIMSS 2019 uygulamasına ait Türkiye örnekleminde 13. ve 14. kitapçıkta ortak olarak yer alan üç madde takımına ait yerel bağımlılık, MTTK (Wainer vd., 2007) aracılığıyla modellenmiştir. Madde takımlarında birbiriyle ilişkili maddelerin gruplandırılması, test sırasında bilişsel yükü azaltarak özellikle küçük yaş grupları için yararlı olabilir (Yen, 1993). Madde takımları ayrıca belirli alanlarda zorluk yaşayan öğrenciler hakkında daha ayrıntılı bilgi sunarak hedefe yönelik eğitim olanaklarını artırabilir. Hem ulusal hem de uluslararası alanda kullanımı yaygınlaşan bu tür maddelerde öncelikle incelenmesi gereken nokta, maddelerin birbirleriyle ilişkisini ortaya koyan yerel bağımlılık derecesidir. Bu nedenle ilk araştırma sorusu, her bir madde takımında var olabilecek yerel bağımlılık derecesiyle ilgilidir. Madde takımı etki büyüklükleri sırası ile 0.575, 0.505, 0.612 olarak hesaplanmıştır. Elde edilen değerler literatürde orta düzeyde yerel bağımlılığın göstergesi olarak belirlenen kritik değerler aralığındadır (Li vd., 2006; Wang ve Wilson, 2005). Murphy, Dodd ve Vaughn (2010), madde takımı tabanlı bilgisayarda bireyselleştirilmiş testlerlerde MTK ile MTTK modellerinin düşük ve orta düzeyde madde takımı etkisine sahip durumlarda benzer performans gösterdiğini bulmuştur. Yapılan bu çalışma, özellikle madde etkisinin orta düzeyde olduğu durumlarda, geleneksel MTK yöntemlerinin kullanılabilirliğini ve bu yöntemlerin zaman ve emek açısından daha pratik bir alternatif olabileceğini göstermiştir.

İkinci araştırma sorusu ile 2PL-MTK ve 2PL-MTTK modellerinden elde edilen parametre kestirimleri (eğim, kesişim ve yetenek parametreleri) karşılaştırmalı olarak incelenmiştir. MTTK modeli ve MTK modeline dayalı yetenek parametre kestirimleri arasındaki korelasyon oldukça yüksek olup bu değer 0.976 bulunmuştur. Ancak yetenek kestirimlerinin standart hataları incelendiğinde MTK modelinin hata değerlerinin daha düşük olduğu gözlemlenmiştir. Yerel bağımlılık ihlal edildiği MTK modellerinde yetenek, daha düşük standart hatalar ile kestirilir (Chang ve Wang, 2010; Eckes, 2014; Koziol, 2016; Wainer ve Wang, 2000). Madde parametre kestirimleri incelendiğinde ise eğim ve kesişim parametrelerinin ilişkisi de oldukça yüksek bulunmuştur ($r>0.98$). Madde takımlarındaki maddeler arasında bulunan orta düzeyde yerel bağımlılık, modellere göre yapılan parametre kestirimleri arasındaki yüksek uyuma katkı sağlamış olabilir. Sonuç olarak, MTK ve MTTK modelleri ile yapılan analizlerin karşılaştırmalı bir şekilde incelenmesi madde takımlarının özellikle K-12 değerlendirmelerinde daha etkili tasarlanmasına olanak tanır.

Üçüncü araştırma sorusunda sınıflama doğruluğu sonuçları MTK ve MTTK modelleri ile incelenmiştir. Bu çalışmada iki kategorili sınıflandırmalar için "ileri düzey altı" ve "ileri düzey üstü" gibi bir sınıflandırma her düzey için gerçekleştirilmiştir. Sınıflandırma kararının alt, orta, yüksek ve ileri olmak üzere dört düzey için gerektirdiği durumlarda her bir düzey için kesme puanı tanımlanmıştır. İki kategorili sınıflandırma için belirlenen kesme değerler her bir modelde düzey arttıkça beklenen bir şekilde artmıştır. Alan yazında farklı MTK modelleri ile yapılan sınıflama doğrulukları çalışmalarında kesme değerler birbirlerine oldukça benzer bulunmuş iken (Lee, 2010; Zhang, 2010) bu çalışmada her iki modelin de kesme puanları yüksek düzey hariç birbirlerinden farklılaşmaktadır. Genel olarak iki ve çok kategorili sınıflandırmalardan elde edilen sınıflama doğruluğu ve tutarlılığı için oldukça yüksek değerler bulunmuş olup her iki model için de oldukça tutarlı sonuçlar elde edilmiştir. İki kategorili sınıflandırmada orta ve yüksek düzey için belirlenen kesme puanlarla yapılan analizlerde sınıflama doğruluğu 2PL-MTTK modelinde daha iyi çıkmıştır. Çok kategorili sınıflama doğruluk ve tutarlılıkları ise iki kategorili sınıflandırma değerlerinden daha düşük çıkmıştır. Bunun nedeni belirlenen düzey sayısının sınıflama doğruluğu ve tutarlılığının hesaplanmasında önemli bir ölçüt olmasıdır (Lathrop ve Cheng, 2014). Bu alanda yapılan çalışma bulguları MTTK modelinin, özellikle verilerde güçlü madde takımı etkisi mevcut olduğunda, sınıflandırma doğruluğu açısından diğer yaklaşımlardan daha iyi performans gösterdiğini veya eşdeğer performans sergilediğini göstermektedir (Keller vd., 2003; Zhang, 2010). Fakat Koziol (2016) çalışmasında küçük madde takımı etkisi koşulları altında MTK ve MTTK modelleri benzer performanslar elde etse de büyük madde takımı etkisi koşulları altında sınıflama doğruluğu yüzdesini daha düşük bulmuştur. Bu nedenle bu çalışmada madde takımlarının önemli düzeyde yerel bağımlılık içermemesinden kaynaklı benzer ve yüksek sınıflama doğruluğu yüzdeleri elde edilmiş olabilir.

Dördüncü araştırma sorusunda ise madde takımlarının cinsiyete göre DMF gösterip göstermediği hem madde hem de madde grubu bazında ele alınmıştır ve SIBTEST yöntemi uygulanmıştır. Bu yöntemin seçilmesinin nedeni MTK tabanlı yöntemlerin KTK tabanlı yaklaşımlardan daha güçlü olduğunun öne sürülmesidir (Hambleton ve Swaminathan, 2013). Fakat MTK temelli yaklaşımlar, modern test oluşturmanın giderek artan bir şekilde madde takımları kullanımına yönelmesi nedeniyle karşılanması zor olan yerel bağımsızlık varsayımını gerektirir (Ferne ve Rupp, 2007; Wainer ve Lukhele, 1997). Bu gibi durumlarda, uygulanabilecek yöntemlerden biri olan SIBTEST yönteminde DMGF analizinin uygulanmasına tabi tutulmuştur. Bu çalışmanın amacı madde takımlarının hem bağımsız maddeler hem de madde grupları olarak ele alındığındaki sonuçları karşılaştırmak olduğundan SIBTEST yöntemi ile bu karşılaştırmanın yapılabilmesidir. Maddeler bağımsız olarak ve madde grupları olarak ele alındığında DMF ve DMGF içeren madde veya madde takımına rastlanmamıştır. Bu durum madde takımlarında düşük ve orta düzeyde madde takımı etkisi bulunmasından kaynaklanabilir.

Sonuç olarak bu çalışmada parametre kestirimlerinin korelasyonu, sınıflama doğruluğu ve DMF sonuçları standart MTK ve madde takımı tabanlı MTK modellerinde oldukça benzer olmakla birlikte MTK modellerinde standart hataları daha düşük ve sınıflama doğruluğu yüzdesi daha yüksek bulunmuştur. Bu nedenle, madde takımlarının kavramsal avantajlarının istatistiksel dezavantajlardan daha ağır basıp basmadığı belirlenerek analizlere devam edilmelidir. Genel olarak performanstaki düşüş ihmal edilebilir düzeyde ise bu testlerin avantajları dezavantajlarından daha ağır basabilir. Alanyazında tek boyutlu MTK ile MTTK modellerinin genellikle parametre kestirimi açısından karşılaştıran birçok çalışma vardır (Bradlow vd., 1999; Glas vd., 2000; Wainer vd., 2000; Wainer ve Wang, 2000). Fakat bu modeller arası sınıflama doğruluğu ve DMF gibi konuların araştırıldığı çalışma sayısı yeterli değildir. Bu nedenle gelecekteki çalışmalar yüksek madde takımı etkisi olduğunda sınıflama doğruluğu ve farklı yöntemlere göre madde takımlarında DMF belirleme yöntemlerine yönelik yapılmalıdır. Özellikle madde takımı etkisinin yüksek olduğu ve DMF içeren maddelerin olduğu durumlar için sınıflama doğruluğu araştırmaları yapılabilir. Ayrıca MTK ve MTTK modelleri arasındaki sınıflandırma doğruluğundaki gözlemlenen benzerlikler, kullanılan belirli kesme noktasıyla sınırlı olabilir. Kesme noktası gizil yetenek dağılımının uç noktasına yakınlaştıkça sınıflama doğruluğu yüzdesi de düşebilir. Bu yüzden kesme noktasının modellerin göreceli performansı üzerindeki etkisini belirlemek için de çalışmalar yapılabilir. K-12 sınıf içi ölçmelerde, öğrencilerin beceri düzeylerinin doğru bir şekilde belirlenmesi, öğretmenlerin öğrencilere yönelik bireyselleştirilmiş öğretim stratejileri geliştirebilmesi için kritik öneme sahiptir. Madde takımlarının kullanılması, sınıf içindeki çeşitli beceri düzeylerini daha ayrıntılı şekilde değerlendirebilir. Örneğin, bir öğrencinin "ileri düzey altı" veya "ileri düzey üstü" olarak sınıflandırılması, öğretmenin hangi alanlarda ek destek sağlaması gerektiğini belirlemesine yardımcı olabilir. Ayrıca K-12 sınıf içi değerlendirmelerde, çok kategorili sınıflamaların kullanımı, öğrenci gelişimini daha ayrıntılı bir şekilde izlemeye olanak tanıyabilir, ancak sınıflama doğruluğu açısından daha dikkatli bir analiz gerektirebilir.

K-12 düzeyinde madde takımları; ölçüm hassasiyetini artırma, test süresini kısaltma, özgünlük sağlama, öğrenme hedefleriyle uyum geliştirme ve güvenirliği artırma gibi önemli avantajlar sunmaktadır. Bu faydalar, madde takımlarının hem öğrenmeyi hem de değerlendirmeyi destekleme potansiyelini vurgulayan çeşitli çalışmalar tarafından desteklenmektedir. Bu madde türünün avantajlarından yararlanırken, psikometrik analizlere olan etkisini dikkate almak gerekmektedir. Bu nedenle yapılan bu çalışmanın benzer konudaki ileriki araştırmalara yol gösterici olacağı düşünülmektedir.

## 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.

Baghaei, P. ve Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicológica, 37*(1), 85-104.

Baker, F. B. (2001). *The basics of item response theory.* https://files.eric.ed.gov/fulltext/ED458219.pdf adresinden erişildi.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. F. M. Lord ve M. R. Novick (Ed.), *Statistical theories of mental test scores* içinde (s. 397-472). Reading, MA: Addison- Wesley.

Bock, R. D ve Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459. doi:10.1007/BF02293801

Bradlow, E. T., Wainer, H. ve Wang, X. (1999). Bayesian random effects model for testlets. *ETS Research Report Series, 1998*(1). doi:10.1002/j.2333-8504.1998.tb01752.x

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, *48*(6). doi:10.18637/jss.v048.i06

Chang, Y ve Wang, J. (2010). *Examining testlet effects on the PIRLS 2006 assessment.* 4th IEA International Research Conference sunulan bildiri, Gothenburg, Sweden.

Cizek, G. J ve Bunch, M. (2007). *Standard setting: A practitioner's guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE.

Clauser, B. E. ve Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Cook, K. F., Dodd, B. G. ve Fitzpatrick, S. J. (1999). A comparison of three polytomous item response theory models in the context of testlet scoring. *Journal of Outcome Measurement, 3*(1), 1-20.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative and mixed methods approaches* (4. bs.). Thousand Oaks, CA: Sage.

DeMars, C. (2010). *Item response theory.* Oxford: Oxford University Press.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145-168. doi:10.1111/j.1745-3984.2006.00010.x

Dempster, A. P., Laird, N. M. ve Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1-22.

Dorans, N. J. ve Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement, 23*(4), 355-368.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing, 31*(1), 39-61. doi:10.1177/0265532213492969

Eckes, T. ve Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-Tests. *Applied Measurement in Education, 28*(2), 85-98. doi:10.1080/08957347.2014.1002919

Embretson, S. E. (2010). *Measuring psychological constructs: Advances in model-based approaches*. Washington: American Psychological Association. doi:10.1037/12074-000

Eren, B., Gündüz, T. ve Tan, Ş. (2023). Comparison of methods used in detection of DIF in cognitive diagnostic models with traditional methods: Applications in TIMSS 2011. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(1), 76-94. doi:10.21031/epod.1218144

Ferne, T. ve Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*(2), 113-148. doi:10.1080/15434300701375923

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement, 29*(4), 278-295.

Gibbons, R. D. ve Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423-436. doi:10.1007/BF02295430

Gierl, M. J., Bisanz, J., Bisanz, G. L. ve Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement*, *40*(4), 281-306

Glas, C. A. W., Wainer, H. ve Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. W. J. van der Linden ve C. A. W. Glas (Ed.), *Computerized adaptive testing: Theory and practice* içinde (s. 271-287). Kluwer-Nijhoff.

Hambleton, R. ve Novick, M. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, *10*(3), 159-170.

Hambleton, R. K. ve Rogers, H. (1989). Solving criterion-referenced measurement problems with item response models. *International Journal of Educational Research, 13*(2), 145-160. doi:10.1016/0883-0355(89)90003-7

Hambleton, R. K. ve Swaminathan, H. (2013). *Item response theory: Principles and applications.* Berlin: Springer Science & Business Media.

Hanson, B. A. ve Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, *27*(4), 345-359.

Ho, T.-H. ve Dodd, B. G. (2012). Item selection and ability estimation procedures for a mixed-format adaptive test. *Applied Measurement in Education, 25*(4), 305-326. doi:10.1080/08957347.2012.714686

Holland, P. W. ve Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. H. Wainer ve H. I. Braun (Ed.), *Test validity* içinde (s. 129-145). Mahwah, NJ: Lawrence Erlbaum Associates.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*(4), 253-264.

Jiao, H., Kamata, A., Wang, S. ve Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*(1), 82-100. doi:10.1111/j.1745-3984.2011.00161.x

Karasar, N. (2016). *Bilimsel araştırma yöntemi*. Ankara: Nobel Yayıncılık.

Keller, L. A., Swaminathan, H. ve Sireci, S. G. (2003). Evaluating scoring procedures for context-dependent item sets. *Applied Measurement in Education, 16*(3), 207-222. doi:10.1207/S15324818AME1603_3

Koziol, N. A. (2016). Parameter recovery and classification accuracy under conditions of testlet dependency: A comparison of the traditional 2PL, Testlet, and Bi-Factor models. *Applied Measurement in Education, 29*(3), 184-195. doi:10.1080/08957347.2016.1171767

Lathrop, Q. N. (2015). *cacIRT: Classification accuracy and consistency under item response theory.* https://CRAN.R-project.org/package=cacIRT adresinden erişildi.

Lathrop, Q. N. ve Cheng, Y. (2014). A nonparametric approach to estimate classification accuracy and consistency. *Journal of Educational Measurement, 51*(3), 318-334. doi:10.1111/jedm.12048

Lee, G., Kolen, M. J., Frisbie, D. A. ve Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25*(4), 357-372. doi:10.1177/01466210122032226

Lee, S. Y. ve Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39*(4), 653-686. doi:10.1207/s15327906mbr3904_4

Lee, W.-C. (2010). Classification consistency and accuracy for complex assessment using item response theory. *Journal of Educational Measurement*, *47*(1), 1-17.

Li, Y., Bolt, D. M. ve Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*(1), 3-21. doi:10.1177/0146621605275414

Livingston, S. A. ve Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179-197.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge.

Lu, J., Zhang, J., Zhang, Z., Xu, B. ve Tao, J. (2021). A novel and highly effective Bayesian sampling algorithm based on the auxiliary variables to estimate the testlet effect models. *Frontiers in Psychology,* 12. doi:10.3389/fpsyg.2021.509575

Ma, W., Terzi, R. ve de la Torre, J. (2021). Detecting differential item functioning using multiple-group cognitive diagnosis models. *Applied Psychological Measurement*, *45*(1), 37-53. doi:10.1177/0146621620965745

Mendes-Barnett, S. ve Ercikan, K. (2010). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education*, *19*(4), 289-304.

Min, S. ve He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing, 31*(4), 453-477. doi:10.1177/0265532214527277

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, 51, 177-195. doi:10.1007/BF02293979

Murphy, D. L., Dodd, B. G. ve Vaughn, B. K. (2010). A comparison of item selection techniques for testlets. *Applied Psychological Measurement*, *34*(6), 424-437. doi:10.1177/0146621609349804

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5*(2), 107-124.

Paek, I. ve Fukuhara, H. (2015). An investigation of DIF mechanisms in the context of differential testlet effects. *British Journal of Mathematical and Statistical Psychology*, *68*(1), 142-157. doi:10.1111/bmsp.12039

Pinheiro, J. C. ve Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*(1), 12-35. doi:10.1080/10618600.1995.10474663

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Rijmen, F. (2010). Formal relations and an empirical comparison among the Bi-Factor, the Testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*(3), 361-372. doi:10.1111/j.1745-3984.2010.00118.x

Roussos, L. ve Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355-371. doi:10.1177/014662169602000404

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation, 7*(1), 14. doi:10.7275/an9m-2035

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation, 10*(1), 13. doi:10.7275/56a5-6b14

Shealy, R. ve Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIBF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Sireci, S. G., Thissen, D. ve Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237-247. doi:10.1111/j.1745-3984.1991.tb00356.x

Soysal, S. ve Yılmaz Koğar, E. (2022). Item parameter recovery via traditional 2PL, Testlet and Bi-factor models for Testlet-Based tests. *International Journal of Assessment Tools in Education, 9*(1), 254-266. doi:10.21449/ijate.948182

Subkoviak, M. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, *13*(4), 265-275.

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

Tasdelen Teker, G. ve Dogan, N. (2015). The effects of testlets on reliability and differential item functioning. *Educational Sciences: Theory and Practice*, *15*(4), 969-980. doi:10.12738/estp.2015.4.2577

Thissen, D., Steinberg, L. ve Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*(1), 118-128.

Thissen, D., Steinberg, L. ve Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*(3), 247-260. doi:10.1111/j.1745-3984.1989.tb00331.x

Thissen, D., Steinberg, L. ve Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. H. Wainer ve H. I. Braun (Ed.), *Test validity* içinde (s. 147-169). Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8*(2), 157-186. doi:10.1207/s15324818ame0802_4

Wainer, H., Bradlow, E. T. ve Du, Z. (2000). Testlet response theory: An analog for the 3-PL useful in testlet-based adaptive testing. W. J. van der Linden ve C. A. W. Glas (Ed.), *Computerized adaptive testing, theory and practice* içinde (s. 245-270). Kluwer-Nijhoff.

Wainer, H., Bradlow, E. T. ve Wang, X. (2007). *Testlet response theory and its applications.* Cambridge: Cambridge University Press.

Wainer, H. ve Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201. doi:10.1111/j.1745-3984.1987.tb00274.x

Wainer, H. ve Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*(1), 1-14. doi:10.1111/j.1745-3984.1990.tb00730.x

Wainer, H. ve Lukhele, R. (1997). Managing the influence of DIF from big items: The 1988 advanced placement history test as an example. *Applied Measurement in Education*, *10*(3), 201-215. doi.org/10.1207/s15324818ame1003_1

Wainer, H. ve Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203-220. doi:10.1111/j.1745-3984.2000.tb01083.x

Wang, X., Bradlow, E. T. ve Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, *26*(1), 109-128. doi: 10.1177/014662160202600100

Wang, W. C. ve Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, *29*(4), 296-318. doi:10.1177/0146621605276281

Wang, W. C. ve Yeh, L. Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, 27, 479-498. doi:10.1177/0146621603259902

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473-492. doi:10.1177/014662168200600408

Wright, B.D. ve Stone, M.H. (1979). *Best test design*. Chicago: MESA

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213. doi:10.1111/j.1745-3984.1993.tb00423.x

Yılmaz Koğar, E. (2021). Comparison of testlet effect on parameter estimates using different item response theory models. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(3), 254-266. doi:10.21031/epod.948227

Zhan, P., Li, X., Wang, W.-C. ve Bian, Y. (2015). *The logistic testlet framework for within-item multidimensional testlet-effect.* 2015 International Meeting of the Psychometric Society (IMPS), Beijing Normal University, Beijing, China.

Zhan, P., Liao, M., & Bian, Y. (2018). Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Frontiers in Psychology*, 9, 607. doi:10.3389/fpsyg.2018.00607

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, *27*(1), 119-140. doi:10.1177/0265532209347