ATATÜRK ÜNİVERSİTESİ YAYINLARI
ATATURK UNIVERSITY PUBLICATIONS

Araştırma Makalesi *Research Article* DOI: 10.5281/zenodo.15036409

Validity of Simulation Studies: A Case Research in the Context of Differential Item Functioning Detection

Simülasyon Çalışmalarının Geçerliği: Değişen Madde Fonksiyonu Belirleme Çalışması Bağlamında Bir Örnek Araştırma

Özkan SAATÇİOĞLU[1]
Selçuk Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Konya, Türkiye

ABSTRACT

The aim of this study is to examine the simulation validity by determining whether the simulation process produces results that are realistically close to expectations, through the generation of artificial data containing Differential Item Functioning (DIF) and assessing whether the data were accurately generated. In the study, which involves one reference group and two focal groups, 2250 different conditions were simulated by considering factors such as the sample size of the reference group, the sample size ratios of the focal groups, the amount of DIF, and the DIF technique. During the data generation process, random data for difficulty and discrimination parameters were generated using the Two-Parameter Logistic Model (2PLM), and it was planned that 20% of the items in the test would contain DIF. To test the validity of the simulation, mean absolute bias and RMSE values for the difficulty and discrimination parameters were calculated both at the item level and by considering the relevant factors. The analysis results revealed that the mean absolute bias and RMSE values calculated for the difficulty and discrimination parameters were low and close to zero. This indicates that estimation errors were minimal and supports the validity of the results. Additionally, it was found that the sample size of the reference group and the sample size ratios of the focal groups had a statistically significant effect on the mean absolute bias and RMSE values for both difficulty and discrimination parameters, and it was observed that as the sample size increased, the mean absolute bias and RMSE values decreased. However, it was concluded that the amount of DIF added to the focal groups did not have a significant effect on the accuracy of parameter estimations. The findings demonstrate that sample size plays a critical role in the accuracy of parameter estimations, while the amount of DIF does not significantly impact this process, and the results of the study are consistent with relevant research in the literature. As a result of this research, it has been recommended that validity evidence for the simulation should be provided not only in DIF investigation studies but also in simulation studies conducted in various subject areas within the field of psychometrics.

Keywords: Simulation, validity of simulation, differential item functioning, item difficulty parameter, item discrimination parameter

Geliş Tarihi/Received 14.02.2025
Kabul Tarihi/Accepted 14.03.2025
Yayın Tarihi/Publication Date 17.03.2025

Sorumlu Yazar/Corresponding author: Özkan SAATÇİOĞLU
E-mail: ozkan.saatcioglu[at]selcuk.edu.tr
Cite this article: Saatçioğlu, Ö. (2025). Validity of simulation studies: A case research in the context of differential item functioning detection. *Journal of Psychometric Research, 3*(1), 24-40.

ÖZ

Bu çalışmanın amacı, simülasyon sürecinin beklenen şekilde gerçeğe yakın sonuçlar ortaya çıkarıp çıkarmadığını belirlemek amacıyla Değişen Madde Fonksiyonu (DMF) içeren yapay verilerin doğru bir şekilde üretilip üretilmediğine yönelik simülasyon geçerliğinin incelenmesidir. Bir referans iki odak olmak üzere üç grubun ele alındığı araştırmada, referans grubun örneklem büyüklüğü, odak grupların örneklem büyüklüğü oranları, DMF miktarı ve DMF tekniği faktörleri dikkate alınarak 2250 farklı koşul simüle edilmiştir. Veri üretim sürecinde, İki Parametreli Lojistik Model (2PLM) ile güçlük ve ayırt edicilik parametreleri için rastgele veriler oluşturulmuş ve testteki maddelerin %20'sinin DMF içermesi planlanmıştır. Simülasyonun geçerliğini test etmek amacıyla, güçlük ve ayırt edicilik parametrelerine ilişkin ortalama mutlak yanlılık ve RMSE değerleri hem madde düzeyinde hem de ilgili faktörler dikkate alınarak hesaplanmıştır. Analizler sonucunda, güçlük ve ayırt edicilik parametreleri için hesaplanan ortalama mutlak yanlılık ve RMSE değerlerinin düşük ve sıfıra yakın olduğu bulunmuştur. Bu durum kestirim hatalarının az olduğunu ve sonuçların geçerliğinin desteklendiğini ortaya koymuştur. Ayrıca referans grubun örneklem büyüklüğünün ve odak grupların örneklem büyüklüğü oranlarının hem güçlük hem de ayırt edicilik parametreleri için ortalama mutlak yanlılık ve RMSE değerleri üzerinde istatistiksel olarak manidar bir etkiye sahip olduğu belirlenmiş ve örneklem büyüklüğü arttıkça ortalama mutlak yanlılık ve RMSE değerlerinin azaldığı tespit edilmiştir. Bununla birlikte, odak gruplara eklenen DMF miktarlarının, parametre kestirimlerinin doğruluğu üzerinde anlamlı bir etki oluşturmadığı sonucuna ulaşılmıştır. Elde edilen bulgular, örneklem büyüklüğünün parametre kestirimlerinin doğruluğu üzerinde kritik bir rol oynadığını ve DMF miktarının bu süreçte anlamlı bir etki yaratmadığını ortaya koymuş ve çalışmanın bulguları alanyazındaki ilgili araştırmalar ile tutarlılık göstermiştir. Yapılan bu araştırma sonucunda DMF inceleme çalışmalarının yanı sıra psikometrinin farklı konu alanlarında yapılacak olan simülasyon çalışmalarında da simülasyonun geçerlik kanıtlarının sunulması gerektiği önerilmiştir.

Anahtar Kelimeler: Simülasyon, simülasyonun geçerliği, değişen madde fonksiyonu, madde güçlük parametresi, madde ayırt edicilik parametresi

Giriş

Bilgisayar teknolojisindeki gelişmeler sayesinde, simülasyon çalışmaları sağlık bilimleri, davranış bilimleri, eğitim bilimleri ve psikometri gibi alanlardaki araştırmalarda giderek daha büyük bir öneme sahip olmaktadır (Berends ve Romme, 1999; Feinberg ve Rubright, 2016; Harwell vd., 2017; Morris vd., 2019; Tureson ve Odland, 2018). Simülasyon çalışmalarında, bilgisayar yazılımları aracılığıyla oluşturulan kodlar kullanılarak veri üretilmesi, bu çalışmaların gerçek durumları taklit edebilen deneysel araştırmalar olarak da dikkate alınmasını sağlamaktadır (Harwell vd., 1996; Morris vd., 2019). Özellikle yeni geliştirilen tekniklerin, modellerin veya yaklaşımların performanslarını karşılaştırmak (Hallgren, 2013; Harwell vd., 1992; Harwell vd., 1996) ya da parametre kestirimlerinin doğruluğunu araştırmak (Harwell vd., 1996) amacıyla yapılan simülasyonlarda deneysel kontrolün sağlanabilmesi önemli bir üstünlük olarak kabul edilmektedir. Simülasyon çalışmalarının bu üstünlüğü öncelikle dikkate alınan faktörler ve koşulların gerçek uygulamalardaki durumları yansıtmasına bağlıdır (Harwell vd., 1996; Scott, 2014). Berends ve Romme (1999) ile Sigal ve Chalmers (2016) tarafından açıklandığı üzere simülasyon çalışmalarında gerçek uygulamalarda karşılaşılan durumları yansıtan faktörlere ve koşullara odaklanılması, çalışılan konu alanı üzerinde derin bir bakış açısı oluşmasını sağlamaktadır. Ancak simülasyon sonuçlarının dikkate alınabilmesi, yapılan hesaplamaların doğru, güvenilir ve geçerli olduğunun kanıtlanmasına da bağlıdır (Law, 2003; Wood, 2011). Diğer bir deyişle, deneysel işlemlerin öncesinde yazılan kodların doğruluğunun ve gerçekleştirilen simülasyonun geçerliğinin (Feinberg ve Rubright, 2016) kanıtlanması büyük önem taşımaktadır. Simülasyon çalışmalarında sürecin titizlikle takip edilmesi halinde elde edilen sonuçların yorumlanabileceği, aksi takdirde bu sonuçlara dayanarak yapılan çıkarımların yanıltıcı olabileceği belirtilmektedir (Davis vd., 2007; Law, 2003; Spence, 1983).

Simülasyon çalışmalarında bilimsel olarak gerçek sonuçlara ulaşmak için simülasyon sonuçlarının geçerliğinin farklı yöntemlerle incelenmesi gerekmekte (Berends ve Romme, 1999) ve alanyazında genellikle görünüş geçerliği ile istatistiksel geçerlik olmak üzere iki temel geçerlik türü üzerinde durulmaktadır (Scott, 2014). Görünüş geçerliğinin önemli ancak öznel bir nitelik taşıması nedeniyle, simülasyonun geçerliğini kanıtlamak için tek başına yeterli olmadığı, istatistiksel geçerliğin sağlanmasının da son derece önemli olduğu belirtilmektedir (Chung, 2004). Bu

bağlamda araştırmalarda simülasyonun geçerliğinden emin olmak ve elde edilen sonuçlara güvenilip güvenilmeyeceğine karar verebilmek amacıyla her iki geçerlik türüne ilişkin kanıtların sunulmasının ve tüm bu kanıtların birlikte değerlendirilmesinin büyük önem taşıdığı düşünülmektedir.

İlgili alanyazında, simülasyonun görünüş geçerliği incelenirken genellikle iki aşamalı bir sürecin izlendiği gözlemlenmiştir. İlk aşamada, bir tekrar ile oluşturulan örnek bir veri kullanılarak deneme çalışması yapılması önerilmiştir. Bu deneme sürecinde, sırasıyla yazılan kodların gözden geçirilmesi (Davis vd., 2007; Li vd., 2012), simülasyon sırasında herhangi bir hata uyarısı alınıp alınmadığının kontrol edilmesi (Sandilands, 2014), simülasyonun bitiş süresinin tespit edilmesi (Law, 2003), üretilen veri setindeki olası eksik veya hatalı değerlerin belirlenmesi (Lopez Rivas, 2012), hem referans hem de odak gruplarda hesaplanan madde parametrelerinin farklı DMF miktarları için incelenmesi (Sandilands, 2014; Scott, 2014) ve hem DMF'li hem de DMF'li olmayan maddelerin güçlük ve ayırt edicilik parametreleri için Madde Karakteristik Eğrilerinin (MKE) çizdirilip kontrol edilmesi (Lopez Rivas, 2012) gibi çok sayıda işlemin gerçekleştirildiği görülmüştür. Görünüş geçerliği çalışmalarının ikinci aşamasında ise, dikkate alınan her bir koşul için gerçekleştirilen en az 100 tekrarlı veri setindeki satır ve sütun sayılarının kontrol edilmesi (Sandilands, 2014) ve ulaşılan sonuçların alanyazındaki ilgili araştırmalarla tutarlı ve makul olup olmadığının değerlendirilmesi (Davis vd., 2007; Feinberg ve Rubright, 2016; Li ve Zumbo, 2009; Paxton vd., 2001) gibi süreçlerin takip edildiği belirlenmiştir.

Simülasyon çalışmalarında gerçekleştirilen hesaplamaların doğruluğunu değerlendirmek için ikinci olarak istatistiksel geçerlik incelenmektedir (Sandilands, 2014; Scott, 2014). Madde Tepki Kuramı'na (MTK) dayalı verilerin üretildiği simülasyon çalışmalarında, madde parametrelerine ilişkin kestirimlerin doğruluğunun büyük önem taşıması nedeniyle (Bolt vd., 2002; Wang ve Chen, 2005; Wen, 2014), araştırmada hem güçlük hem de ayırt edicilik parametreleri için gerçekleştirilen kestirimler ile gerçek parametreler arasındaki farklılaşmanın hesaplanması gerektiği açıklanmıştır (DeMars ve Lau, 2011; Sandilands, 2014). Bu hesaplamalar sayesinde, parametre kestirimlerinin ne kadar doğru olduğunun anlaşılabileceği ifade edilmiştir (Finch, 2016).

Alanyazındaki simülasyon çalışmalarında, yapılan hesaplamaların simülasyonun temelindeki kuramsal yapıyı doğru bir şekilde yansıtması gerektiği vurgulanmaktadır (Davis vd., 2007). Ancak bu açıklamaların aksine Harwell ve diğerleri (2018), yaptıkları araştırmada 1985-2012 yılları arasında 6 dergide simülasyon çalışmalarıyla ilgili yayımlanan 677 makaleyi incelemiş ve üretilen veri setlerinin istenilen özellikleri sağladığını ispatlayan çalışma sayısının sadece 15 (%2.2) olduğunu tespit etmiştir. Bu bulgu, simülasyon çalışmalarında genellikle yazılan bilgisayar kodlarının ve yapılan hesaplamaların doğru kabul edilerek herhangi bir kontrolden geçirilmediğini ortaya koymuştur (Wood, 2011).

Yapılan bu çalışma kapsamında da verilerin doğru bir şekilde üretilerek, simülasyon sürecinin beklenen şekilde gerçeğe yakın sonuçlar ortaya koyduğundan emin olmak amacıyla DMF'li maddeler içeren veri üretilmiştir. Kim'in (2010) çalışmasında açıklandığı üzere DMF çalışmalarında örneklemin büyüklüğü, testin uzunluğu, grupların yetenek düzeyi, MTK modeli veya DMF türü gibi faktörler sıklıkla dikkate alınmıştır. Bu araştırma kapsamında da referans grubun örneklem büyüklüğü, odak grupların örneklem büyüklüklerinin referans gruba oranı, gruplara eklenen DMF miktarı ve DMF tekniği faktörleri ele alınarak simülasyon sürecinin beklenen şekilde gerçeğe yakın sonuçlar ortaya çıkarıp çıkarmadığının belirlenmesi amacıyla görünüş ve istatistiksel geçerlik kanıtlarının sunulması amaçlanmıştır.

Yöntem

Araştırmanın Modeli

Simülasyon çalışmaları, rastgele veri setleri üretmek için farklı parametrelerin dikkate alındığı bilgisayar destekli araştırmalardır (Mooney, 1997; Morris vd., 2019; Rubinstein ve Kroese, 2017). Kuramsal sorunların çözümünde simülasyon çalışmalarının etkili bir yöntem olduğu ve bu tür çalışmalar ile kuramların ilerlemesine önemli katkılar sunulabileceği ifade edilmiştir (Harwell vd., 2017; Morris vd., 2019). Bu nitelikleri sebebiyle, simülasyon çalışmaları genellikle metodolojik bir yaklaşım olarak değerlendirilmektedir (Davis vd., 2007; Happach ve Tilebein, 2015). Bu bağlamda bu araştırma da DMF'li madde içerecek şekilde veri üretilen bir simülasyon çalışması olarak planlanmıştır.

Simülasyonun Deseni

Değişimlenen Koşullar

Bu simülasyon çalışmasının görünüş geçerliği kapsamında grup sayısı üç olarak belirlenmiş ve referans grubun örneklem büyüklüğü (2 [600, 1800]), odak grupların örneklem büyüklüklerinin referans gruba oranı (9 [1:1/1,

1:1/2, 1:2/3]), gruplara eklenen DMF miktarı (25 [a(0.0, 0.4, 0.7), b(0.0, 0.4, 0.7)]) ve DMF tekniği (5 [Genelleştirilmiş Lord Ki-kare (GLK), Genelleştirilmiş Lojistik Regresyon-Olabilirlik Oran Testi (GLR-LRT), Genelleştirilmiş Lojistik Regresyon-Wald Testi (GLR-Wald), Genelleştirilmiş Mantel-Haenszel (GMH), Logistic Ordinal Regression Differential Item Functioning Using IRT (Lordif)]) olmak üzere toplamda 2250 farklı koşul ele alınmıştır. İstatistiksel geçerlik için referans grubun örneklem büyüklüğü, odak grupların örneklem büyüklüklerinin referans gruba oranı ve gruplara eklenen DMF miktarı faktörleri ile analizler gerçekleştirilmiştir.

Sabit Tutulan Koşullar

İkili puanlanan maddeler üzerinde DMF belirlemek amacıyla gerçekleştirilen simülasyon çalışmalarında, hem Tek Biçimli (TB) hem de Tek Biçimli Olmayan (TBO) DMF için genellikle 2PLM kullanılarak veri üretilmiştir (Finch, 2016; Rollins III, 2018; Svetina ve Rutkowski, 2014). Bu çalışmada da Finch'in (2016) araştırmasında gerçek test verilerinden elde edilen madde parametrelerinin dağılımları temel alınarak güçlük parametresi için b~U(-1.06, 1.39), ayırt edicilik parametresi için ise a~U(0.84, 1.53) şeklinde tekdüze (uniform) dağılım ile 20 madde için 2PLM'ye dayalı veri üretilmiştir. Çalışma kapsamında, DMF'li madde oranı sabit bir koşul olarak belirlenmiş ve testteki maddelerin %20'sinin DMF içermesi sağlanmıştır.

Verilerin Üretilmesi

Simülasyonun geçerliğini kanıtlamak amacıyla gerçekleştirilen bu çalışmada, elde edilen sonuçların alanyazındaki benzer araştırmalarla uyumlu ve makul olup olmadığını değerlendirmek için, simülasyon deseninde belirtilen koşullar ele alınarak grup sayısının üç olduğu senaryo için veri seti oluşturulmuştur. Veri üretimi sırasında; dört maddenin güçlük ve ayırt edicilik parametrelerine DMF miktarları (0.0, 0.40, 0.70) eklenerek odak gruplar için madde parametreleri oluşturulmuştur. Tam çapraz faktöriyel desen temel alınarak iki odak grubun güçlük ve ayırt edicilik parametrelerine belirtilen DMF miktarlarının kombinasyonu eklendikten sonra yürütülen veri üretim sürecinde 100 tekrar yapılmıştır.

Verilerin Analizi

Veri üretimi aşamasında çeşitli yazılımlar kullanılabilse de (Morris vd., 2019), psikometri alanındaki araştırmalarda R programlama dilinin sıkça tercih edilmesi nedeniyle, bu simülasyon çalışmasında da tüm kodlar R'ın 3.6.1 sürümünde (R Core Team, 2019) geliştirilmiş ve veri üretimi için farklı paketlerdeki fonksiyonlardan (Choi ve Asilkalkan, 2019; Li vd., 2012; Morris vd., 2019; Rusch vd., 2013) yararlanılmıştır. Simülasyonun işlem sürecini hızlandırmak amacıyla, doParallel 1.0.15 paketindeki fonksiyonlar kullanılarak paralel hesaplamalar yapılmıştır (Alfons vd., 2010). Veri üretimi sürecinde TÜBİTAK ULAKBİM Yüksek Başarımlı ve Grid Hesaplama Merkezi tarafından desteklenen Türk Ulusal Bilim e-Altyapısı'ndan (TRUBA) yararlanılmıştır.

Değerlendirme Ölçütleri

Bu çalışmada, parametre kestirimlerinin doğruluğunu analiz etmek için mutlak yanlılık ve RMSE (Root Mean Square Error) birlikte ele alınmıştır (DeMars, 2003; Sigal ve Chalmers, 2016). Veri üretimi tamamlandıktan sonra, güçlük ve ayırt edicilik parametrelerine ilişkin hem madde düzeyinde hem de ilgili faktörler dikkate alınarak ayrı ayrı mutlak yanlılık ve RMSE hesaplanmıştır. Alanyazında yer alan pek çok araştırmada (Feinberg ve Rubright, 2016; Harwell vd., 1996; Harwell vd., 2018; Yuan vd., 2015), madde parametre kestirimlerinin doğruluğunu etkileyen faktörleri tespit etmek amacıyla Faktöriyel ANOVA kullanılması nedeniyle (Rockoff, 2018; Seybert ve Stark, 2012; Wood, 2011), bu çalışmada da hangi faktörlerin ortalama mutlak yanlılık ve ortalama RMSE üzerinde etkili olduğunu belirlemek için Faktöriyel ANOVA yapılmıştır. Faktöriyel ANOVA'da ana etkiler için eta kare, etkileşim etkileri için ise kısmi eta kare katsayıları yorumlanmıştır (Keppel ve Wickens, 2004). Bu kapsamda hesaplanan katsayılar, Gray ve Kinnear (2012) tarafından belirtilen aralıklara [$.01 \leq \eta_p^2 < .06$ (düşük), $.06 \leq \eta_p^2 < .14$ (orta) ve $\eta_p^2 \geq .14$ (yüksek)] göre değerlendirilmiştir.

Bulgular

Görünüş Geçerliği İçin Elde Edilen Bulgular

Simülasyonların görünüş geçerliği kapsamında simülasyon sürecinin beklenildiği gibi gerçekleştiğinden emin olmak için Feinberg ve Rubright (2016) ile Sandilands (2014) tarafından önerilen adımlar dikkate alınmıştır. Bu doğrultuda, toplam grup sayısının üç olduğu durum için tek bir örnek veri seti kullanılmış ve araştırma kapsamında gerçekleştirilen tüm işlemler adım adım sunulmuştur.

İlk olarak grup sayısının üç olduğu durumda yazılan tüm kodlar gözden geçirilmiş ve bir tekrar içeren örnek bir veri üretilmiştir. Bu deneme süreci esnasında, veri setinin üretilme süresinin makul bir zaman diliminde olduğu ve simülasyon sürecinde herhangi bir hata uyarısı alınmadığı tespit edilmiştir. Bunun yanı sıra, gerçekleştirilen

incelemeler sonucunda veri setinde eksik veya normal olmayan bir değere rastlanmamıştır. İkinci aşamada odak grupların madde parametrelerine üç farklı (0, 0.40, 0.70) DMF miktarı eklenmiş ve referans ile odak grupların örneklem büyüklüklerinin eşit (N_R = 600, N_{O1} = 600, N_{O2} = 600) olduğu koşullar için ele alınan grupların güçlük ve ayırt edicilik parametrelerinin incelenmesi hedeflenmiştir. Bu amaçla, öncelikle yalnızca bir veri seti üretilerek 20 maddenin referans gruptaki madde parametreleri incelenmiştir. Ardından DMF'li olan dört madde (M5, M6, M9, M14) için referans ve odak gruplardaki madde parametreleri kontrol edilmiş ve bu parametrelerin doğruluğunu incelemek için MKE'ler çizdirilmiştir (EK 1). Bu sayede, yalnızca referans grubun MKE'leri ile DMF'li olan maddelerin referans ve odak gruplardaki MKE'leri karşılaştırılarak, madde parametrelerinin doğru bir şekilde hesaplandığı belirlenmiştir. Bununla birlikte, DMF'li olarak belirlenen dört maddenin referans gruptaki MKE'leri ile bu maddelere 0.70 DMF miktarı eklendikten sonraki süreçte odak gruplardaki MKE'leri çizdirilerek görsel kontroller gerçekleştirilmiştir. Yapılan incelemeler sonucunda, DMF'li olarak belirlenen dört maddenin odak gruplar için daha zor ve daha ayırt edici olduğu tespit edilmiş ve beklenen yönde olan bu sonucun görünüş geçerliğine kanıt olarak dikkate alınabileceği ortaya konulmuştur.

Üçüncü aşamada, 100 tekrar sayısına dayalı olarak üretilen verilerin beklenen satır ve sütun sayısına sahip olup olmadığı kontrol edilmiştir. Grup sayısının üç olduğu durumda, 2 (referans grubun örneklem büyüklüğü) × 9 (odak grupların örneklem büyüklükleri oranı) × 25 (DMF miktarı) × 5 (DMF tekniği) × 100 (tekrar sayısı) = 225.000 büyüklüğünde veri setinin oluşturulduğu tespit edilmiştir. Yapılan kontroller sonucunda, nihai veri setlerinde hatalı veya eksik verilerin bulunmadığı ve bu nedenle veri setlerinin analiz için uygun olduğu ortaya konulmuştur. Dördüncü ve son aşamada ise araştırmadan elde edilen bulgular ile alanyazındaki benzer çalışmaların sonuçları arasındaki tutarlılık değerlendirilmiştir. Bu süreçte öncelikle araştırma kapsamında ele alınan GLR-LRT ve GLR-Wald tekniklerine ilişkin sonuçların, tüm simülasyon koşullarında ilgili araştırmalarda vurgulandığı üzere (Gao, 2019; Magis vd., 2011) birbirine paralel olduğu tespit edilmiştir.

İstatistiksel Geçerlik İçin Elde Edilen Bulgular

İstatistiksel geçerliğe kanıt sunmak amacıyla üretilen 20 madde için madde düzeyinde hesaplanan mutlak yanlılık ve RMSE değerleri Tablo 1'de verilmiştir.

Tablo 1

Hesaplanan Ortalama Mutlak Yanlılık ve Ortalama RMSE Değerleri

Güçlük Parametresi			Ayırt Edicilik Parametresi		
Madde	Ortalama Mutlak Yanlılık	Ortalama RMSE	Madde	Ortalama Mutlak Yanlılık	Ortalama RMSE
b_1	0.0870	0.1112	a_1	0.1212	0.1549
b_2	0.1092	0.1412	a_2	0.0965	0.1224
b_3	0.1185	0.1537	a_3	0.0935	0.1183
b_4	0.0894	0.1142	a_4	0.1109	0.1411
b_5	0.0855	0.1085	a_5	0.1118	0.1449
b_6	0.1495	0.2007	a_6	0.1148	0.1497
b_7	0.0903	0.1156	a_7	0.0864	0.1099
b_8	0.1114	0.1429	a_8	0.1012	0.1289
b_9	0.1188	0.1571	a_9	0.1430	0.1873
b_{10}	0.1326	0.1733	a_{10}	0.0981	0.1245
b_{11}	0.0754	0.0957	a_{11}	0.0951	0.1203
b_{12}	0.0832	0.1064	a_{12}	0.0918	0.1159
b_{13}	0.0727	0.0922	a_{13}	0.1005	0.1277
b_{14}	0.0661	0.0832	a_{14}	0.1478	0.1929
b_{15}	0.0762	0.0972	a_{15}	0.1162	0.1474
b_{16}	0.0555	0.0701	a_{16}	0.1201	0.1531
b_{17}	0.0589	0.0741	a_{17}	0.1112	0.1416
b_{18}	0.1030	0.1332	a_{18}	0.1029	0.1307
b_{19}	0.0837	0.1070	a_{19}	0.1140	0.1452
b_{20}	0.0887	0.1127	a_{20}	0.0974	0.1238

Tablo 1'de 20 maddenin güçlük ve ayırt edicilik parametrelerine ilişkin 2PLM temelinde yapılan kestirimler

için hesaplanmış olan ortalama mutlak yanlılık ve ortalama RMSE değerleri sunulmuştur. Ortalama mutlak yanlılık değerleri; güçlük parametresi için 0.0555 ile 0.1495 arasında, ayırt edicilik parametresi için ise 0.0864 ile 0.1478 arasında değişim göstermiştir. Ortalama RMSE değerleri ise; güçlük parametresi için 0.0701 ile 0.2007 arasında, ayırt edicilik parametresi için 0.1099 ile 0.1929 arasında farklılaşmıştır. Elde edilen bulgular, güçlük ve ayırt edicilik parametreleri için madde düzeyinde hesaplanan ortalama mutlak yanlılık ile ortalama RMSE değerlerinin birbirine paralel olduğunu ve RMSE değerlerinin biraz daha yüksek çıktığını ortaya koymuştur. Alanyazındaki bir araştırmada da benzer bir bulgu elde edilmiş, 20 madde için hesaplanan ortalama RMSE değerlerinin; güçlük parametresinde 0.1422 ile 0.4211 arasında, ayırt edicilik parametresinde ise 0.1126 ile 0.2770 arasında olduğu ortaya konulmuştur (Atar, 2007). Benzer şekilde, Sandilands (2014) tarafından yapılan çalışmada, 90 madde için güçlük parametresine ilişkin ortalama RMSE değerlerinin 0.064 ile 0.290 aralığında olduğu belirlenmiştir. Başka bir çalışmada ise güçlük parametresi için hesaplanan ortalama RMSE değerlerinin 0.120 ile 0.210 arasında değiştiği ifade edilmiştir (Lu ve Jiao, 2009, akt., DeMars ve Lau, 2011). Bu bilgiler doğrultusunda, yapılan bu simülasyon çalışmasından elde edilen sonuçların, alanyazındaki diğer araştırmalarla önemli ölçüde tutarlı olduğu görülmüştür. Bunun yanı sıra, her iki madde parametresi için madde düzeyinde hesaplanan ortalama RMSE ve ortalama mutlak yanlılık değerlerinin sıfıra yakın çıkması, kestirim hatalarının oldukça küçük olduğunu ortaya koymuştur. Bu durum, madde düzeyinde kestirilen parametreler ile gerçek parametre değerleri arasında önemli bir farklılık bulunmadığını göstererek, parametre kestirimlerinin güvenirliğini ve etkinliğini desteklemiştir.

Bu simülasyon araştırmasında, üç grup (referans-odak1-odak2) söz konusu olduğunda, referans grubun örneklem büyüklüğü, odak grupların örneklem büyüklüğü oranları ve odak gruplara eklenen DMF miktarları bağımsız değişkenler olarak belirlenirken, ortalama mutlak yanlılık ve ortalama RMSE değerleri bağımlı değişkenler olarak kabul edilmiş ve bu koşullar altında ilgili faktörlerin ele alındığı Faktöriyel ANOVA uygulanmıştır. Analizlerde ana etkilerin yanı sıra bağımsız değişkenler arasındaki ikili etkileşimler de analiz edilmiş, ancak sonuçlarda istatistiksel açıdan anlamlı bir bulguya rastlanmadığı için tablolarda sadece ana etkiler raporlanmıştır. Güçlük parametresine ilişkin ortalama mutlak yanlılık değerleri temel alınarak gerçekleştirilen Faktöriyel ANOVA bulguları Tablo 2'de verilmiştir.

Tablo 2

Güçlük Parametresi İçin Faktöriyel ANOVA Sonuçları (Ortalama Mutlak Yanlılık)

	Kareler Toplamı	sd	*F*	*p*	η^2	η_p^2
Referans örneklem büyüklüğü (R)	45.499	1	237.298	.000	.643	.841
Odak1 örneklem oranı (O_1)	4.603	2	12.003	.000	.065	.348
Odak2 örneklem oranı (O_2)	9.840	2	25.661	.000	.139	.533
Odak1 DMF miktarı (b_1)	0.086	1	0.451	.505	.001	.010
Odak2 DMF miktarı (b_2)	0.522	1	2.725	.106	.007	.057
Hata	8.628	45			.122	
Toplam	70.737					

Tablo 2'de verilen güçlük parametresine ilişkin ortalama mutlak yanlılık sonuçları incelendiğinde, referans grubun örneklem büyüklüğünün ana etkisinin ($F_{(1,45)}$ = 237.298, p < .001, η^2 = .643) istatistiksel açıdan anlamlı ve yüksek bir etki büyüklüğüne sahip olduğu görülmektedir. Aynı şekilde, birinci odak grubun örneklem büyüklüğü oranı ($F_{(2,45)}$ = 12.003, p < .001, η^2 = .065) ve ikinci odak grubun örneklem büyüklüğü oranı ($F_{(2,45)}$ = 25.661, p < .001, η^2 = .139) için ana etkilerin istatistiksel olarak anlamlı olduğu, ancak etki büyüklüklerinin orta düzeyde kaldığı belirlenmiştir. Çalışmada, odak gruplara eklenen DMF miktarlarının ana etkisi ve ikili etkileşimler için elde edilen sonuçlar ise istatistiksel olarak anlamlı bulunmamıştır. Bulgulara göre, madde güçlük parametresi için ortalama mutlak yanlılık değerleri açısından, referans grubun örneklem büyüklüğü ile odak grupların örneklem büyüklüğü oranlarının farklı düzeyleri arasında sırasıyla yüksek ve orta düzeyde farklılıklar olduğu sonucuna ulaşılmıştır. Faktörlerin düzeyleri arasındaki farklılıkları belirlemek amacıyla madde güçlük parametresi için hesaplanan ortalama mutlak yanlılık değerleri EK 2'de paylaşılmıştır.

Parametre kestirimlerinin doğruluğunu değerlendirmek amacıyla, RMSE de hesaplanmış ve güçlük parametresi için ortalama RMSE'yi etkileyen faktörleri belirlemek üzere gerçekleştirilen Faktöriyel ANOVA sonuçları Tablo 3'te sunulmuştur.

Tablo 3

Güçlük Parametresi İçin Faktöriyel ANOVA Sonuçları (Ortalama RMSE)

	Kareler Toplamı	sd	*F*	*p*	η^2	η_p^2
Referans örneklem büyüklüğü (R)	63.602	1	1255.154	.000	.896	.965
Odak1 örneklem oranı (O_1)	1.456	2	14.366	.000	.021	.390
Odak2 örneklem oranı (O_2)	3.015	2	29.748	.000	.042	.569
Odak1 DMF miktarı (b_1)	0.029	1	0.565	.456	.000	.012
Odak2 DMF miktarı (b_2)	0.112	1	2.212	.144	.002	.047
Hata	2.280	45			.032	
Toplam	70.999					

Tablo 3'te yer alan güçlük parametresi için ortalama RMSE sonuçları incelendiğinde, referans grubun örneklem büyüklüğünün ana etkisinin ($F_{(1,45)}$ = 1255.154, p < .001, η^2 = .896) istatistiksel açıdan anlamlı ve yüksek bir etki büyüklüğüne sahip olduğu görülmektedir. Aynı şekilde, birinci odak grubun örneklem büyüklüğü oranı ($F_{(2,45)}$ = 14.366, p < .001, η^2 = .021) ve ikinci odak grubun örneklem büyüklüğü oranı ($F_{(2,45)}$ = 29.748, p < .001, η^2 = .042) için ana etkilerin istatistiksel olarak anlamlı olduğu, ancak etki büyüklüklerinin düşük düzeyde kaldığı belirlenmiştir. Çalışmada, odak gruplara eklenen DMF miktarlarının ana etkisi ve ikili etkileşimler için elde edilen sonuçlar ise istatistiksel olarak anlamlı bulunmamıştır. Bulgulara göre, madde güçlük parametresi için ortalama RMSE değerleri açısından, referans grubun örneklem büyüklüğü ile odak grupların örneklem büyüklüğü oranlarının farklı düzeyleri arasında sırasıyla yüksek ve düşük düzeyde farklılıklar olduğu sonucuna ulaşılmıştır. Faktörlerin hangi düzeyleri arasında farklılıkların bulunduğunu belirlemek amacıyla madde güçlük parametresi için hesaplanan ortalama RMSE değerleri EK 2'de paylaşılmıştır.

Çalışma kapsamında, güçlük parametresinin yanı sıra ayırt edicilik parametresi için de ortalama mutlak yanlılık ve ortalama RMSE değerleri hesaplanarak analiz edilmiştir. Ayırt edicilik parametresine yönelik ortalama mutlak yanlılık değerleri dikkate alınarak yürütülen Faktöriyel ANOVA sonuçları Tablo 4'te verilmiştir.

Tablo 4

Ayırt Edicilik Parametresi İçin Faktöriyel ANOVA Sonuçları (Ortalama Mutlak Yanlılık)

	Kareler Toplamı	sd	*F*	*p*	η^2	η_p^2
Referans örneklem büyüklüğü (R)	45.499	1	658.464	.000	.643	.936
Odak1 örneklem oranı (O_1)	10.547	2	76.315	.000	.149	.772
Odak2 örneklem oranı (O_2)	10.536	2	76.236	.000	.149	.772
Odak1 DMF miktarı (a_1)	0.133	1	1.926	.172	.002	.041
Odak2 DMF miktarı (a_2)	0.067	1	0.966	.331	.001	.021
Hata	3.109	45			.044	
Toplam	70.738					

Tablo 4'de yer alan ayırt edicilik parametresine ilişkin ortalama mutlak yanlılık sonuçları incelendiğinde, referans grubun örneklem büyüklüğünün ana etkisinin ($F_{(1,45)}$ = 658.464, p < .001, η^2 = .643) istatistiksel açıdan anlamlı ve yüksek bir etki büyüklüğüne sahip olduğu görülmektedir. Aynı şekilde, birinci odak grubun örneklem büyüklüğü oranı ($F_{(2,45)}$ = 76.315, p < .001, η^2 = .149) ve ikinci odak grubun örneklem büyüklüğü oranı ($F_{(2,45)}$ = 76.236, p < .001, η^2 = .149) için ana etkilerin istatistiksel olarak anlamlı olduğu ve yüksek düzeyde bir etki gösterdiği belirlenmiştir. Çalışmada, odak gruplara eklenen DMF miktarlarının ana etkisi ve ikili etkileşimler için elde edilen sonuçlar ise istatistiksel olarak anlamlı bulunmamıştır. Bulgulara göre, madde ayırt edicilik parametresi için ortalama mutlak yanlılık değerleri açısından, referans grubun örneklem büyüklüğü ile odak grupların örneklem büyüklüğü oranlarının farklı düzeyleri arasında yüksek düzeyde farklılıklar olduğu sonucuna ulaşılmıştır. Faktörlerin hangi düzeyleri arasında farklılıkların bulunduğunu belirlemek amacıyla madde ayırt edicilik parametresi için hesaplanan ortalama mutlak yanlılık değerleri ise EK 3'te paylaşılmıştır.

Araştırmada, parametre kestirimlerinin doğruluğunu değerlendirebilmek için RMSE de hesaplanmış ve ayırt edicilik parametresinin ortalama RMSE değerlerini hangi faktörlerin etkilediğini belirlemek için yapılan Faktöriyel ANOVA bulguları, Tablo 5'te yer almıştır.

Tablo 5

Ayırt Edicilik Parametresi İçin Faktöriyel ANOVA Sonuçları (Ortalama RMSE)

	Kareler Toplamı	sd	*F*	*p*	η^2	η_p^2
Referans örneklem büyüklüğü (R)	45.499	1	637.780	.000	.643	.934
Odak1 örneklem oranı (O_1)	10.835	2	75.943	.000	.153	.771
Odak2 örneklem oranı (O_2)	10.434	2	73.126	.000	.147	.765
Odak1 DMF miktarı (a_1)	0.100	1	1.397	.243	.001	.030
Odak2 DMF miktarı (a_2)	0.056	1	0.779	.382	.001	.017
Hata	3.210	45			.045	
Toplam	70.737					

Tablo 5'te yer alan ayırt edicilik parametresine ilişkin ortalama RMSE sonuçları incelendiğinde, referans grubun örneklem büyüklüğünün ana etkisinin ($F_{(1,45)}$ = 637.780, p < .001, η^2 = .643) istatistiksel açıdan anlamlı ve oldukça güçlü bir etkiye sahip olduğu görülmektedir. Benzer şekilde, birinci odak grubun ($F_{(2,45)}$ = 75.943, p < .001, η^2 = .153) ve ikinci odak grubun ($F_{(2,45)}$ = 73.126, p < .001, η^2 = .147) örneklem büyüklüğü oranlarının ana etkileri de istatistiksel olarak anlamlı ve yüksek düzeyde etkili bulunmuştur. Öte yandan, odak gruplara eklenen DMF miktarlarının ana etkisi ve ikili etkileşimlerin sonuçları istatistiksel açıdan anlamlı bulunmamıştır. Bulgular, madde ayırt edicilik parametresinin ortalama RMSE değerleri üzerinde referans grubun örneklem büyüklüğü ile odak grupların örneklem büyüklüğü oranlarının düzeyleri arasında belirgin ve önemli farklılıklar olduğunu ortaya koymaktadır. Madde ayırt edicilik parametresine ilişkin tüm simülasyon koşulları için hesaplanan ortalama RMSE değerleri EK 3'te paylaşılmıştır.

Tartışma, Sonuç ve Öneriler

Bu simülasyon çalışmasında, DMF'li veri setinde simülasyonun geçerliğine kanıt sunmak amaçlanmış ve grup sayısının üç olduğu durumda, 2 (referans grubun örneklem büyüklüğü) × 9 (odak grupların örneklem büyüklükleri oranı) × 25 (DMF miktarı) × 5 (DMF tekniği) × 100 (tekrar sayısı) ile sınırlı tutulan veri simüle edilmiştir. Araştırmada simülasyonun geçerliğini test etmek amacıyla, öncelikle görünüş geçerliği çalışması yürütülmüş ve yapılan incelemeler sonucunda; veri üretilme sürecinin doğru bir şekilde gerçekleştirildiği, veri setinde hata olmadığı, sonuçların alanyazındaki çalışmalarla tutarlı ve beklenildiği gibi olduğu sonucuna ulaşılmıştır. Ayrıca yapılan bu çalışmada parametre kestirimlerinin gerçek değerleri ne ölçüde doğru temsil ettiği incelenmiş ve güçlük ile ayırt edicilik parametreleri için ortalama mutlak yanlılık ve ortalama RMSE değerlerini etkileyen faktörler tespit edilmiştir. Madde parametrelerinin 2PLM kullanılarak kestirilmesi sürecinde hata oranının minimize edilmesi kritik bir öneme sahip olduğundan (Tay vd., 2015), çalışmada güçlük ve ayırt edicilik parametrelerine dair ortalama mutlak yanlılık ve ortalama RMSE değerleri birlikte değerlendirilmiştir. Yapılan analizler neticesinde, elde edilen sonuçların hem mutlak yanlılık ve RMSE hem de madde parametreleri açısından birbiriyle uyumlu ve tutarlı olduğu ortaya konulmuştur.

Güçlük ve ayırt edicilik parametrelerine ilişkin ortalama mutlak yanlılık ve ortalama RMSE değerlerinin düşük ve sıfıra yakın çıkması, kestirilen parametreler ile gerçek değerler arasındaki farkın az olduğunu, dolayısıyla kestirim hatalarının düşük ve sonuçların daha isabetli olduğunu göstermiştir. Bu sonuç Liu ve diğerleri (2015) ile Yuan ve diğerleri (2015) tarafından yapılan araştırmalarla desteklenmiştir. Çalışma kapsamında, tüm koşullar altında hesaplanan RMSE değerlerinin mutlak yanlılıktan genellikle daha yüksek olduğu tespit edilmiştir. Bunun yanı sıra, ayırt edicilik parametresi için hesaplanan mutlak yanlılık ve RMSE değerlerinin, güçlük parametresine kıyasla daha yüksek olduğu sonucuna varılmıştır. Hem güçlük hem de ayırt edicilik parametreleri açısından, toplam örneklem büyüklüğü arttıkça ortalama mutlak yanlılık ve ortalama RMSE değerlerinin sıfıra yaklaşma eğilimi gösterdiği gözlemlenmiştir.

Bu çalışmayla uyumlu olarak, alanyazında yer alan birçok araştırma hem güçlük hem de ayırt edicilik parametreleri için mutlak yanlılık ve RMSE değerlerinin örneklem büyüklüğünden etkilendiğini ortaya koymuştur (DeMars, 2003; Sandilands, 2014; Sigal ve Chalmers, 2016; Wang ve Chen, 2005; Woods, 2009). Özellikle MTK kullanılarak yapılan çalışmalarda, madde parametrelerinin güvenilir bir şekilde kestirilebilmesi için genellikle 500'ün üzerinde bir örneklem büyüklüğü önerilmektedir (Hulin vd., 1982). Hulin ve diğerleri (1982), 2PLM kullanılarak yapılan kestirimlerde, örneklem büyüklüğü ve testteki madde sayısı arttıkça RMSE değerlerinin sıfıra daha fazla yaklaştığını göstermiştir. Wang ve Chen (2005), testteki madde sayısının 20 veya daha fazla olması durumunda, örneklem büyüklüğü arttıkça yanlılığın

azaldığını belirtmiştir. Benzer şekilde, Sandilands (2014) ile Bulut ve Sünbül (2017) araştırmalarında test uzunluğu ve örneklem büyüklüğü arttıkça yanlılık ve RMSE değerlerinin düştüğünü ifade etmişlerdir. Socha ve diğerleri (2015) tarafından yapılan bir çalışmada, güçlük parametresi için grupların örneklem büyüklüğü arttıkça yanlılık ve RMSE değerlerinin azaldığı bildirilmiştir. Atar (2007) ise güçlük ve ayırt edicilik parametreleri için örneklem büyüklüğü arttıkça RMSE ($p < .001$) ve yanlılığın karesi ($p < .001$) değerlerinin azaldığını tespit etmiştir. Ancak, DMF miktarı ve toplam örneklem büyüklüğü sabit tutulduğunda, grupların örneklem büyüklüğü dengesiz olduğunda, ayırt edicilik parametresi için RMSE ($p < .001$) ve yanlılığın karesi ($p < .01$) değerlerinin arttığı, güçlük parametresi için ise RMSE ($p = .015$) değerlerinin artmasına rağmen yanlılığın karesi ($p = .042$) değerlerinde belirgin bir değişim olmadığı gözlemlenmiştir. Ayrıca, DMF miktarındaki farklılaşmanın sonuçlar üzerinde istatistiksel olarak anlamlı bir etki yaratmadığı sonucuna varılmıştır (Atar, 2007).

Alanyazındaki benzer çalışmalara paralel olarak, bu araştırmada da grupların örneklem büyüklüklerinin artması, RMSE ve yanlılık değerlerinin sıfıra yaklaşmasına yol açarak kestirimlerin daha isabetli hale geldiğini ortaya koymuştur. Bunun yanı sıra hem güçlük hem de ayırt edicilik parametreleri için önemli faktörlere bağlı olarak hesaplanan mutlak yanlılık ve RMSE değerlerinin, alanyazındaki diğer çalışmalarla (Atar, 2007; Hulin vd., 1982; Sandilands, 2014; Socha vd., 2015) büyük ölçüde tutarlı olduğu görülmüştür. Simülasyonun geçerliğini test etmek amacıyla yapılan tüm incelemeler sonucunda, bu araştırmada üretilen verilerin yüksek düzeyde güvenilir ve araştırmanın amaçlarına uygun olduğu sonucuna varılmıştır. Yapılan bu araştırmadan hareketle DMF inceleme çalışmalarının yanı sıra psikometrinin farklı konu alanlarındaki simülasyon çalışmalarında da simülasyonun geçerlik kanıtlarının sunulmasının önemli olduğu ortaya konulmuştur.

Etik Komite Onayı: Bu araştırma bir simülasyon çalışması olduğu için etik kurul onayı alınmamıştır.

Hakem Değerlendirmesi: Dış bağımsız.

Çıkar Çatışması: Yazar, çıkar çatışması olmadığını beyan etmiştir.

Finansal Destek: Yazar, bu çalışma için finansal destek almadığını beyan etmiştir.

Ethics Committee Approval: Because this study was a simulation study, ethics committee approval was not obtained.

Peer-review: Externally peer-reviewed.

Conflict of Interest: The author has no conflicts of interest to declare.

Financial Disclosure: The author declared that this study has received no financial support.

References

Alfons, A., Templ, M., & Filzmoser, P. (2010). An object-oriented framework for statistical simulation: The r package simFrame. *Journal of Statistical Software, 37*(3), 1-35. https://doi.org/10.18637/jss.v037.i03

Atar, B. (2007). *Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures* [Doctoral dissertation]. Florida State University, Florida.

Berends, P., & Romme, G. (1999). Simulation as a research tool in management studies. *European Management Journal, 17*(6), 576-583. https://doi.org/10.1016/S0263-2373(99)00048-1

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*(4), 331-348. https://doi.org/10.1111/j.1745-3984.2002.tb01146.x

Bulut, O., & Sünbül, Ö. (2017). Monte carlo simulation studies in item response theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology, 8*(3), 266-287. https://doi.org/10.21031/epod.305821

Choi, Y. J., & Asilkalkan, A. (2019) R packages for item response theory analysis: Descriptions and features. *Measurement: Interdisciplinary Research and Perspectives, 17*(3), 168-175, https://doi.org/10.1080/15366367.2019.1586404

Chung, C. A. (2004). *Simulation modeling handbook: A practical approach*. CRC Press.

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007). Developing theory through simulation methods. *Academy of Management Review, 32*(2), 480-499. https://doi.org/10.5465/amr.2007.24351453

DeMars, C. E. (2003). Sample size and recovery of nominal response model item parameters. *Applied Psychological Measurement, 27*(4), 275-288. https://doi.org/10.1177/0146621603027004003

DeMars, C. E., & Lau, A. (2011). Differential item functioning detection with latent classes: How accurately can we detect who is responding differentially? *Educational and Psychological Measurement, 71*(4), 597-616. https://doi.org/10.1177/0013164411404221

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36-49. https://doi.org/10.1111/emip.12111

Finch, W. H. (2016). Detection of item functioning for more than two groups: A monte carlo comparison of methods. *Applied Measurement in Education, 29*(1), 30-45. https://doi.org/10.1080/08957347.2015.1102916

Gao, X. (2019). *A comparison of six DIF detection methods* [Master thesis]. University of Connecticut Graduate School.

Gray, C. D., & Kinnear, P. R. (2012). *IBM SPSS statistics 19 made simple*. Psychology Press, Taylor & Francis Group.

Hallgren, K. A. (2013). Conducting simulation studies in the R programming environment. *Tutorials in Quantitative Methods for Psychology, 9*(2), 43-60. https://doi.org/10.20982/tqmp.09.2.p043

Happach, R. M., & Tilebein, M. (2015). Simulation as research method: Modeling social interactions in management science. In C. Misselhorn (Ed.), *Collective agency and cooperation in natural and articial systems* (pp. 239-259). Springer

Harwell, M. R., Kohli, N., & Peralta, Y. (2017). Experimental design and data analysis in computer simulation studies in the behavioral sciences. *Journal of Modern Applied Statistical Methods, 16*(2), 3-28. https://doi.org/10.22237/jmasm/1509494520

Harwell, M. R., Kohli, N., & Peralta-Torres, Y. (2018). A survey of reporting practices of computer simulation studies in statistical research. *The American Statistician, 72*(4), 321-327. https://doi.org/10.1080/00031305.2017.1342692

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing monte carlo results in methodological research: The one-and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics, 17*(4), 315-339. https://doi.org/10.2307/1165127

Harwell, M. R., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125. https://doi.org/10.1177/014662169602000201

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement, 6*(3), 249-260. https://doi.org/10.1177/014662168200600301

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Pearson.

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

Law, A. M. (2003). *How to conduct a successful simulation study*. Proceedings of the 2003 Winter Simulation Conference, New Orleans, LA. U.S.A.

Li, Y., Brooks, G. P., & Johanson, G. A. (2012). Item discrimination and Type I error in the detection of differential item functioning. *Educational and Psychological Measurement, 72*(5), 847-861. https://doi.org/10.1177/0013164411432333

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicológica, 30*(2), 343-370. https://psycnet.apa.org/record/2009-18227-011

Liu, H., Zhang, Y., & Luo, F. (2015). Mediation analysis for ordinal outcome variables. In Millsap, Bolt, Ark & Wang, (Eds.), *Quantitative psychology research* (pp. 429-450). Springer International Publishing.

Lopez Rivas, G. E. (2012). *Detection and classification of DIF types using parametric and nonparametric methods: A comparison of the IRT-likelihood ratio test, crossing-SIBTEST, and logistic regression procedures* [Doctoral

dissertation]. University of South Florida, Florida.

Magis, D., Raîche, G., Béland, S., & Gérard, P. (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing, 11*(4), 365-386. https://doi.org/10.1080/15305058.2011.602810

Mooney, C. Z. (1997). *Monte carlo simulation*. Sage.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Tutorial in Biostatistics, 38*(11), 2074-2102. https://doi.org/10.1002/sim.8086

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte carlo experiments: Design and implementation. *Structural Equation Modeling, 8*(2), 287-312. https://doi.org/10.1207/S15328007SEM0802_7

R Core Team (2019). R*: A language and environment for statistical computing*. Vienna, Austria. URL http://www.R-project.org/

Rockoff, D. (2018). *A randomization test for the detection of differential item functioning* [Doctoral dissertation]. The University of Arizona, Arizona.

Rollins III, J. D. (2018). *A comparison of observed score approaches to detecting differential item functioning among multiple groups* [Doctoral dissertation]. The University of North Carolina at Greensboro, Greensboro.

Rubinstein, R. Y., & Kroese, D. P. (2017). *Simulation and the monte carlo method*. John Wiley & Sons.

Rusch, T., Mair, P., & Hatzinger, R. (2013). *Psychometrics with R: A review of CRAN packages for item response theory* (Discussion Paper). Center for Empirical Research Methods.

Sandilands, D. A. (2014). *Accuracy of differential item functioning detection methods in structurally missing data due to booklet design* [Doctoral dissertation]. The University of British Columbia, Vancouver.

Scott, L. (2014). *Controlling Analytic selection of a valid subtest for DIF analysis when DIF has multiple potential causes among multiple groups* [Doctoral dissertation]. Arizona State University, Arizona.

Seybert, J., & Stark, S. (2012). Iterative linking with the differential functioning of items and tests (DFIT) method: Comparison of testwide and item parameter replication (IPR) critical values. *Applied Psychological Measurement, 36*(6), 494-515. https://doi.org/10.1177/0146621612445182

Sigal, M. J., & Chalmers, R. P. (2016). Play it again: Teaching statistics with monte carlo simulation. *Journal of Statistics Education, 24*(3), 136-156. https://doi.org/10.1080/10691898.2016.1246953

Socha, A., DeMars, C. E., Zilberberg, A., & Phan, H. (2015). Differential item functioning detection with the Mantel-Haenszel procedure: The effects of matching types and other factors. *International Journal of Testing, 15*(3), 193-215. https://doi.org/10.1080/15305058.2014.984066

Spence, I. (1983). Monte carlo simulation studies. *Applied Psychological Measurement, 7*(4), 405-425. https://doi.org/10.1177/014662168300700403

Svetina, D., & Rutkowski, L. (2014). Detecting differential item functioning using generelized logistic regression in the context of large-scale assessments. *Large Scale Assessments in Education, 2*(4), 1-17. https://doi.org/10.1186/s40536-014-0004-5

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18*(1), 3-46. https://doi.org/10.1177/1094428114553062

Tureson, K., & Odland, A. (2018). Monte Carlo simulation studies. In Bruce B. Frey (Ed.). *The SAGE encylopedia of educational research, measurement, and evaluation* (pp. 1085-1089). SAGE Publications, Inc.

Wang, W. C., & Chen, C. T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the winsteps program for the family of Rasch models. *Educational and Psychological Measurement, 65*(3), 376-404. https://doi.org/10.1177/0013164404268673

Wen, Y. (2014). *DIF analyses in multilevel data: Identification and effects on ability estimates* [Doctoral dissertation]. The University of Wisconsin, Milwaukee.

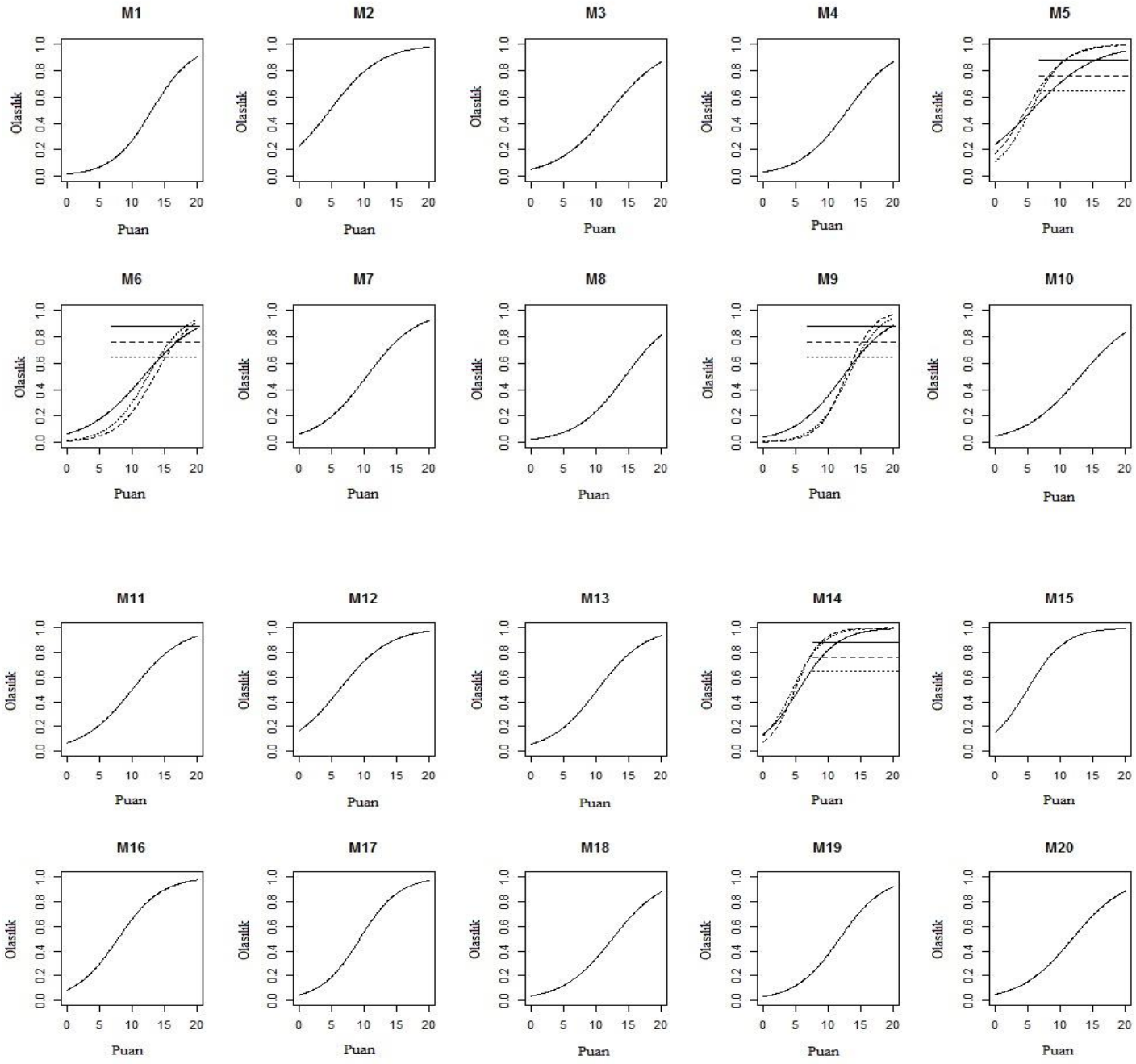
Wood, W. S. (2011). *Differential item functioning procedures for polytomous items when examinee sample sizes are small* [Doctoral dissertation]. Graduate College of The University of Iowa, Iowa.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*(1), 1-27. https://doi.org/10.1080/00273170802620121

Yuan, K.-H., Tong, X., & Zhang, Z. (2015). Bias and efficiency for SEM with missing data and auxiliary variables: Two-stage robust method versus two-stage ML. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(2), 178-192. https://doi.org/10.1080/10705511.2014.935750

Ekler

Ek 1. Testteki 20 Maddenin Referans ve Odak Gruplar İçin Madde Karakteristik Eğrileri



Ek 2. Madde Güçlük Parametresine İlişkin Ortalama Mutlak Yanlılık ve Ortalama RMSE Sonuçları

Madde Güçlük Parametresi İçin Mutlak Yanlılık ve RMSE Sonuçları

Referans Örneklem Büyüklüğü	Odak1 Örneklem Büyüklüğü	Odak2 Örneklem Büyüklüğü	Odak1 DMF Miktarı	Odak2 DMF Miktarı	Ortalama Mutlak Yanlılık	Ortalama RMSE
600	300	300	0.40	0.40	0.1284	0.1708
600	300	300	0.70	0.40	0.1422	0.1905
600	300	300	0.40	0.70	0.1398	0.1787
600	300	300	0.70	0.70	0.1288	0.1725
600	300	400	0.40	0.40	0.1316	0.1738
600	300	400	0.70	0.40	0.1255	0.1671
600	300	400	0.40	0.70	0.1172	0.1540
600	300	400	0.70	0.70	0.1195	0.1544
600	300	600	0.40	0.40	0.1100	0.1446
600	300	600	0.70	0.40	0.1151	0.1475
600	300	600	0.40	0.70	0.1135	0.1489
600	300	600	0.70	0.70	0.1121	0.1441
600	400	300	0.40	0.40	0.1227	0.1565
600	400	300	0.70	0.40	0.1188	0.1504
600	400	300	0.40	0.70	0.1251	0.1657
600	400	300	0.70	0.70	0.1295	0.1777
600	400	400	0.40	0.40	0.1156	0.1479
600	400	400	0.70	0.40	0.1088	0.1372
600	400	400	0.40	0.70	0.1191	0.1525
600	400	400	0.70	0.70	0.1274	0.1671
600	400	600	0.40	0.40	0.1370	0.1748
600	400	600	0.70	0.40	0.1031	0.1294
600	400	600	0.40	0.70	0.1077	0.1351
600	400	600	0.70	0.70	0.1065	0.1389
600	600	300	0.40	0.40	0.1218	0.1533
600	600	300	0.70	0.40	0.1062	0.1374
600	600	300	0.40	0.70	0.1139	0.1487
600	600	300	0.70	0.70	0.1213	0.1596
600	600	400	0.40	0.40	0.1051	0.1359
600	600	400	0.70	0.40	0.1102	0.1427
600	600	400	0.40	0.70	0.1160	0.1466
600	600	400	0.70	0.70	0.1102	0.1406
600	600	600	0.40	0.40	0.1054	0.1341
600	600	600	0.70	0.40	0.1003	0.1294
600	600	600	0.40	0.70	0.1063	0.1363
600	600	600	0.70	0.70	0.1064	0.1367
1800	900	900	0.40	0.40	0.0747	0.0952
1800	900	900	0.70	0.40	0.0699	0.0905
1800	900	900	0.40	0.70	0.0736	0.0935
1800	900	900	0.70	0.70	0.0762	0.0965
1800	900	1200	0.40	0.40	0.0683	0.0866
1800	900	1200	0.70	0.40	0.0669	0.0850
1800	900	1200	0.40	0.70	0.0701	0.0910
1800	900	1200	0.70	0.70	0.0703	0.0908
1800	900	1800	0.40	0.40	0.0665	0.0859
1800	900	1800	0.70	0.40	0.0666	0.0867
1800	900	1800	0.40	0.70	0.0700	0.0883

(Devam ediyor)

Madde Güçlük Parametresi İçin Mutlak Yanlılık ve RMSE Sonuçları (Devam)

Referans Örneklem Büyüklüğü	Odak1 Örneklem Büyüklüğü	Odak2 Örneklem Büyüklüğü	Odak1 DMF Miktarı	Odak2 DMF Miktarı	Ortalama Mutlak Yanlılık	Ortalama RMSE
1800	900	1800	0.70	0.70	0.0655	0.0835
1800	1200	900	0.40	0.40	0.0793	0.1000
1800	1200	900	0.70	0.40	0.0706	0.0926
1800	1200	900	0.40	0.70	0.0757	0.0950
1800	1200	900	0.70	0.70	0.0800	0.0991
1800	1200	1200	0.40	0.40	0.0651	0.0823
1800	1200	1200	0.70	0.40	0.0658	0.0848
1800	1200	1200	0.40	0.70	0.0746	0.0954
1800	1200	1200	0.70	0.70	0.0659	0.0827
1800	1200	1800	0.40	0.40	0.0589	0.0746
1800	1200	1800	0.70	0.40	0.0626	0.0793
1800	1200	1800	0.40	0.70	0.0611	0.0787
1800	1200	1800	0.70	0.70	0.0648	0.0829
1800	1800	900	0.40	0.40	0.0766	0.1028
1800	1800	900	0.70	0.40	0.0717	0.0896
1800	1800	900	0.40	0.70	0.0671	0.0868
1800	1800	900	0.70	0.70	0.0729	0.0922
1800	1800	1200	0.40	0.40	0.0604	0.0762
1800	1800	1200	0.70	0.40	0.0639	0.0809
1800	1800	1200	0.40	0.70	0.0674	0.0856
1800	1800	1200	0.70	0.70	0.0631	0.0806
1800	1800	1800	0.40	0.40	0.0545	0.0686
1800	1800	1800	0.70	0.40	0.0586	0.0735
1800	1800	1800	0.40	0.70	0.0670	0.0831
1800	1800	1800	0.70	0.70	0.0658	0.0825

Ek 3. Madde Ayırt Edicilik Parametresine İlişkin Ortalama Mutlak Yanlılık ve Ortalama RMSE Sonuçları

Madde Ayırt Edicilik Parametresi İçin Mutlak Yanlılık ve RMSE Sonuçları

Referans Örneklem Büyüklüğü	Odak1 Örneklem Büyüklüğü	Odak2 Örneklem Büyüklüğü	Odak1 DMF Miktarı	Odak2 DMF Miktarı	Ortalama Mutlak yanlılık	Ortalama RMSE
600	300	300	0.40	0.40	0.1618	0.2100
600	300	300	0.70	0.40	0.1502	0.1922
600	300	300	0.40	0.70	0.1584	0.2037
600	300	300	0.70	0.70	0.1530	0.1984
600	300	400	0.40	0.40	0.1461	0.1899
600	300	400	0.70	0.40	0.1470	0.1868
600	300	400	0.40	0.70	0.1486	0.1885
600	300	400	0.70	0.70	0.1508	0.1961
600	300	600	0.40	0.40	0.1334	0.1734
600	300	600	0.70	0.40	0.1425	0.1822
600	300	600	0.40	0.70	0.1350	0.1721
600	300	600	0.70	0.70	0.1379	0.1779
600	400	300	0.40	0.40	0.1423	0.1828
600	400	300	0.70	0.40	0.1439	0.1841
600	400	300	0.40	0.70	0.1427	0.1840
600	400	300	0.70	0.70	0.1455	0.1893
600	400	400	0.40	0.40	0.1363	0.1724
600	400	400	0.70	0.40	0.1446	0.1826
600	400	400	0.40	0.70	0.1416	0.1793
600	400	400	0.70	0.70	0.1411	0.1821
600	400	600	0.40	0.40	0.1282	0.1636
600	400	600	0.70	0.40	0.1379	0.1754
600	400	600	0.40	0.70	0.1280	0.1629
600	400	600	0.70	0.70	0.1324	0.1724
600	600	300	0.40	0.40	0.1334	0.1711
600	600	300	0.70	0.40	0.1368	0.1764
600	600	300	0.40	0.70	0.1378	0.1778
600	600	300	0.70	0.70	0.1354	0.1749
600	600	400	0.40	0.40	0.1279	0.1644
600	600	400	0.70	0.40	0.1287	0.1652
600	600	400	0.40	0.70	0.1357	0.1741
600	600	400	0.70	0.70	0.1342	0.1691
600	600	600	0.40	0.40	0.1203	0.1528
600	600	600	0.70	0.40	0.1255	0.1590
600	600	600	0.40	0.70	0.1235	0.1552
600	600	600	0.70	0.70	0.1181	0.1501
1800	900	900	0.40	0.40	0.0862	0.1091
1800	900	900	0.70	0.40	0.0864	0.1111
1800	900	900	0.40	0.70	0.0947	0.1230
1800	900	900	0.70	0.70	0.0927	0.1182
1800	900	1200	0.40	0.40	0.0814	0.1038
1800	900	1200	0.70	0.40	0.0826	0.1056
1800	900	1200	0.40	0.70	0.0819	0.1042
1800	900	1200	0.70	0.70	0.0903	0.1150
1800	900	1800	0.40	0.40	0.0760	0.0984
1800	900	1800	0.70	0.40	0.0782	0.1007
1800	900	1800	0.40	0.70	0.0784	0.1013

(Devam ediyor)

Madde Ayırt Edicilik Parametresi İçin Mutlak Yanlılık ve RMSE Sonuçları (Devam)

Referans Örneklem Büyüklüğü	Odak1 Örneklem Büyüklüğü	Odak2 Örneklem Büyüklüğü	Odak1 DMF Miktarı	Odak2 DMF Miktarı	Ortalama Mutlak yanlılık	Ortalama RMSE
1800	900	1800	0.70	0.70	0.0779	0.0993
1800	1200	900	0.40	0.40	0.0828	0.1059
1800	1200	900	0.70	0.40	0.0828	0.1050
1800	1200	900	0.40	0.70	0.0866	0.1091
1800	1200	900	0.70	0.70	0.0853	0.1091
1800	1200	1200	0.40	0.40	0.0789	0.1008
1800	1200	1200	0.70	0.40	0.0804	0.1029
1800	1200	1200	0.40	0.70	0.0791	0.0994
1800	1200	1200	0.70	0.70	0.0784	0.0998
1800	1200	1800	0.40	0.40	0.0713	0.0905
1800	1200	1800	0.70	0.40	0.0719	0.0908
1800	1200	1800	0.40	0.70	0.0746	0.0944
1800	1200	1800	0.70	0.70	0.0756	0.0954
1800	1800	900	0.40	0.40	0.0793	0.1008
1800	1800	900	0.70	0.40	0.0770	0.0981
1800	1800	900	0.40	0.70	0.0773	0.1003
1800	1800	900	0.70	0.70	0.0786	0.1005
1800	1800	1200	0.40	0.40	0.0765	0.0976
1800	1800	1200	0.70	0.40	0.0721	0.0914
1800	1800	1200	0.40	0.70	0.0720	0.0910
1800	1800	1200	0.70	0.70	0.0764	0.0973
1800	1800	1800	0.40	0.40	0.0692	0.0865
1800	1800	1800	0.70	0.40	0.0680	0.0852
1800	1800	1800	0.40	0.70	0.0678	0.0850
1800	1800	1800	0.70	0.70	0.0721	0.0913