Trakya Eğitim Dergisi

ISSN: 2630-6301
Yayımcı: Trakya Üniversitesi

Vol. 15, No.2, 887-918, 2025
DOI: 10.24315/tred.1509299

Araştırma Makalesi

Eğitimde Ölçmede Yapay Zekanın Entegrasyonu: Madde Tepki Kuramı Kapsamında Veri Üretiminde ChatGPT'nin Etkililiği

Hatice Gürdil[1*]
Yeşim Beril Soğuksu[2]
Salih Salihoğlu[3*]
Fatma Coşkun[4]

[1]Milli Eğitim Bakanlığı, Ankara, Türkiye gurdilhatice@gmail.com

[2]Milli Eğitim Bakanlığı, Vali Hilmi Tolun Ortaokulu, Kahramanmaraş, Türkiye berilsoguksu@gmail.com

[3]University of Miami, Endüstri ve Sistem Mühendisliği Bölümü, Florida, United States sxs4331@miami.edu

[4]Kahramanmaraş Sütçü İmam Üni., Eğitimde Ölçme ve Değerlendirme Bölümü, Kahramanmaraş, Türkiye fatmacoskuncf@gmail.com

*Sorumlu Yazar

Özet: Bu çalışmanın amacı, Madde Tepki Kuramı (MTK) kapsamında R programlama diliyle gerçekleştirilen veri üretimine yönelik algoritmaların geliştirilmesinde ChatGPT 3.5'nin etkililiğini araştırmaktır. Bu kapsamda, ChatGPT 3.5 ve araştırmacılar tarafından yazılan algoritmalarla, İki Parametreli Lojistik Modele (2PLM) göre üretilen veri setleri üzerinde geçerlik incelemeleri yapılmıştır. Yapılan incelemelerde, veri setlerinin MTK varsayımlarını ve madde parametrelerinin simülasyon koşullarını karşılama durumları dikkate alınmıştır. Sonuç olarak ChatGPT 3.5 algoritmalarının, MTK varsayımlarına uygun veri üretimi konusunda oldukça başarılı olduğu, ancak madde parametrelerinin simülasyon koşullarını karşılama hususunda, araştırmacılar tarafından geliştirilen algoritmaya kıyasla daha az etkili olduğu belirlenmiştir. Bu kapsamda ChatGPT 3.5, MTK'ya yönelik veri üretimi algoritmalarının geliştirilmesinde, araştırmacılara destek verebilecek faydalı bir araç olarak önerilmektedir.

Anahtar Kelime: ChatGPT, Madde Tepki Kuramı, Veri Üretimi, Simulasyon, R Programlama Dili

Geliş tarihi:02.07.2024
Kabul tarihi:13.08.2024
Yayım tarihi:30.04.2025

GİRİŞ

ChatGPT, Büyük Dil Modeli (Large Language Model-LLM) tabanlı bir yapay zeka sohbet robotu olup, kullanıcı girdilerine yanıt olarak metin üretmek için OpenAI'nin Üretken Ön Eğitimli Dönüştürücü 3.5 (Generative Pre-trained Transformer 3.5, GPT-3.5) dil modelini kullanmaktadır. Doğal dil işleme (Natural Language Processing-NLP) alanındaki en gelişmiş dil modellerinden biri olan GPT-3.5, kitaplar, makaleler, web sayfaları ve sosyal konuşmalar gibi çeşitli internet kaynaklarından elde edilen 175 milyardan fazla parametre ile eğitilmiştir (Uc-Cetina ve ark., 2023). Bu model, 570 GB boyutundaki bir veri seti ile eğitilerek sorulara yanıt verebilmektedir (van Dis ve ark., 2023). Daha önceki yapay zeka araçlarının (örneğin Grammarly, rTutor.ai, Research Rabbit gibi) araştırma uygulamalarını dönüştürmekte kullanılmasına rağmen (Else, 2023), ChatGPT geniş ham veri hacmi, öğrenmede kullanılan parametrelerin çokluğu, bağlama özgü mimarisi ve gelişmiş denetimli öğrenme gibi özellikleri ile önceki modellerden ayrılmaktadır (Floridi, 2023; Susnjak, 2022). Ayrıca ChatGPT, mevcut verileri analiz etme ve yeni veriler oluşturma yeteneği ile makine öğrenimi araçlarının önceki sürümlerinden farklıdır ve piyasaya sürüldükten sonraki ilk 5 gün içinde 1 milyon kullanıcıya, Ocak 2023 itibarıyla ise 100 milyondan fazla kullanıcıya ulaşarak tarihin en hızlı büyüyen tüketici uygulaması olmuştur. ChatGPT aynı zamanda kod üretimini destekleyen en popüler LLM'lerden biridir (Hu, 2023; Aljanabi ve ark., 2023; Feng ve ark., 2023; Khoury ve ark., 2023). ChatGPT 3.5 doğal dil anlama, geniş kapsamlı veri seti, çok yönlü kullanım alanları, kullanıcı dostu arayüzü, gelişmiş diyalog yetenekleri ve sürekli güncellenen yapısı sayesinde diğer yapay zeka modellerine kıyasla üstün performans sergilemektedir. Bu nedenlerle ChatGPT 3.5 bu çalışma için seçilmiş ve araştırmacılar için en uygun çözüm olarak değerlendirilmiştir.

Tüm bunlara rağmen ve hızla gelişmesine karşın ChatGPT, yanıtlarını veri izlerinden oluşturan ve mantık veya akıl yürütmeye dayanmayan bir yapıya sahiptir; olasılıksal ve rastlantısal (stokastik) bir modeldir ve eğitim setinin kalitesiyle sınırlıdır (Bender ve ark., 2021; Perrigo, 2023). Ayrıca, interneti tarama yeteneği olmadığı için, dayandığı veri ve algoritmalar hatalı veya önyargılı olduğunda, üretilen çıktılar da hatalı veya önyargılı olma eğilimindedir (Deng & Lin, 2022; McGee, 2023; OpenAI, 2023; Ortiz, 2023a; Ortiz, 2023b;

Cite as (APA 7): Gürdil H, & Soğuksu Yeşim B., Salihoğlu S., & Coşkun F. (2025).Eğitimde ölçmede yapay zekanın entegrasyonu: madde tepki kuramı kapsamında veri üretiminde ChatGPT'nin etkililiği. *Trakya Eğitim Dergisi, 15*(2),887—902. https://doi.org/10.24315/tred.1509299

Yang, 2022). Sadece 2021'e kadar olan verilerle eğitildiği için, gerçek zamanlı bilgiyi otomatik olarak entegre edememekte ve bu nedenle güncel bilgiye erişememektedir. Ek olarak, öğrencilerin ChatGPT'yi derinlemesine okuma ve konuyla ilgili eleştirel analiz yapmadan doğrudan yanıt almak için kullanması, akademik, profesyonel ve gerçek yaşam bağlamlarında gerekli olan eleştirel düşünme, problem çözme ve yaratıcılık becerilerini bastırabilmekte (O'Connor, 2023) ve akademik etik ihlallere yol açabilmektedir (Stokel-Walker, 2022). Diğer bir eleştiri, ChatGPT'ye aşırı bağımlılık riski, kötü niyetli girdilerle manipülasyon ve yaygın olarak erişilebilen platformlarda yanlış bilgi veya propaganda yayma potansiyeli gibi güvenlikle ilgili sorunları içermektedir (Deng & Lin, 2022).

Bu eleştirilere yanıt olarak, New York Şehri Eğitim Departmanı ChatGPT'nin kullanımını yasaklamış (Hirsh-Pasek & Blinkoff, 2023) ve Uluslararası Makine Öğrenimi Konferansı (2023) ise kullanımını yalnızca deneysel analizlerin bir parçası olarak sınırlandırmıştır. Tüm bu kısıtlamalara rağmen, dünyanın en hızlı büyüyen teknolojilerinden biri olan ChatGPT'nin, bilgiye erişim, öğrenme ve hatta her türlü işi yapma biçimimizi kökten değiştirme potansiyeline sahip olduğu açıkça görülmektedir ve bu durum, toplumsal bir dönüşümün eşiğinde olduğumuzu düşündürmektedir. Bu dönüşüm, giderek artan sayıda işin tehlikeye girebileceği ihtimali (Dwivedi ve ark., 2021) ve bireylerin çalışma sistemlerinde yol açacağı değişiklikler nedeniyle genellikle kullanıcı direnciyle karşılaşmaktadır (Laumer ve ark., 2016). Bu olumsuzluklar ve dirence rağmen sunduğu faydalar, kolay erişilebilirlik ve yaygın kullanımı göz önünde bulundurulduğunda, ChatGPT'yi yasaklamanın bir çözüm olmayacağı öngörülmektedir (Rosenzweig-Ziff, 2023; Springer-Nature, 2023). Teknolojinin potansiyel yıkıcı etkileri kabul edilmekle birlikte, bu yıkım aynı zamanda toplumu daha ileriye taşıyabilecek bilimsel ilerlemeler için bir fırsat olarak da görülmektedir. Bu nedenle, dijital dönüşümü yasaklarla etkisiz hale getirmeye çalışmak yerine, etkilerini faydalı yollarla kontrol etmeye ve entegre etmeye odaklanmak daha mantıklı görünmektedir. Akademinin dijital bir dönüşüm sürecine girmeye başladığı göz önünde bulundurulduğunda, ChatGPT teknolojisinin olumlu yönlerine odaklanmak ve olumsuz yönlerini olumlu sonuçlara yönlendirmek daha uygun bir yaklaşım gibi görünmektedir.

ChatGPT, geniş bir kullanım alanına yayılmış olup bilgi arama, hikâye ve rapor oluşturma, bilgisayar kodu yazma ve düzeltme, makale ve bölüm yazma ve özetleme (Baidoo-Anu & Ansah, 2023; Else, 2023; Kundalia, 2023; Rudolph ve ark., 2023; Zhai, 2022), testler yapma, veri işleme ve açıklama gibi özelliklere sahiptir (Hassani & Silva, 2023). Bu özellikler, bireylerin bilgisayarlarla olan etkileşimlerini olumlu yönde etkileyebilir (Montti, 2022), öğrencilere öğrenme süreçlerini geliştirme ve üretkenliklerini artırma fırsatları sunabilir (Dwivedi ve ark., 2021). Ancak, insanlardan farklı olarak ChatGPT, bilişsel sınırlar olmaksızın zekasını genişletebilmesine rağmen eğitim verilerinin dar alanlara veya disiplinlere sınırlı olması nedeniyle insanlar kadar uzmanlaşmış bir zekaya sahip olmayabilir. Bu bağlamda, kullanıcıların ChatGPT'yi nihai bir kaynak olarak görmemesi; aksine fikirleri tartışmak ve bakış açılarını genişletmek için bir araç olarak kullanması gerekmektedir. Bu yaklaşım, insan-makine hibrit çalışma fırsatları sunan ve insan uzmanlığının ChatGPT'yi yönlendirdiği yeni bir iş birliği türü sayesinde yapay zekadan daha iyi sonuçlar elde edilmesine yol açabilir (Mollick, 2022; van Dis ve ark., 2023). Bunun için öncelikle yapay zekanın hibrit ekiplerdeki potansiyelinin anlaşılması gereklidir. ChatGPT'nin önyargılı çıktılar üretebileceği ve verilerin gerçekliğini doğrulayamayacağı göz önünde bulundurulduğunda, temel araştırmacı görevleri olan veri analizi, yorumlama ve sonuç çıkarma gibi süreçlerin yerine geçmesine izin verilmemeli; destekleyici bir araç olarak kullanılması sağlanmalıdır (Elsevier, 2023).

ChatGPT'nin dijital izlerden bilimsel bilgi ürettiği düşünüldüğünde, özellikle karmaşık görevlerde insan müdahalesi olmadan tek başına ilerleyemeyeceği (Biswas, 2023) ve sorumluluğun tamamen kullanıcıya ait olduğu unutulmamalıdır. Bu bağlamda, kullanıcıların doğru soruları sorma gerekliliğini ve yanıtların kalitesini değerlendirme ihtiyacını anlaması, uzmanlıkları arttıkça ChatGPT'den elde ettikleri bilginin kalitesinin ve çıktıları yorumlama becerilerinin de artacağını fark etmesi gerekmektedir. Ayrıca, elde edilen katkıları eleştirel bir şekilde değerlendirmek, doğruluğunu farklı kaynaklardan araştırarak teyit etmek veya incelemeler doğrultusunda gerekli düzenlemeleri yapmak da kullanıcı sorumluluğu kapsamına girmektedir.

ChatGPT'nin kullanıcı sorumluluğu çerçevesinde kullanılmasıyla, kullanıcıların bilgiye hızlı bir şekilde erişmesi ve sıradan, tekrarlayan görevleri yerine getirmesi konusunda yardımcı olacağı düşünülmektedir. Bunun sonucunda, kullanıcılar daha üst düzey becerilere odaklanarak daha üretken ve süreç de daha verimli bir hale gelebilir. ChatGPT'nin eğitim alanındaki katkılarının da giderek artmakta olduğu ve eğitimin onlarca yıldır teknolojiye paralel olarak yeniden tasarlandığı (Baidoo-Anu & Ansah, 2023; Huang, 2019) göz önüne alındığında, bu dönüşüme hızla uyum sağlamak büyük önem taşımaktadır. Bu bağlamda, birçok alanda kullanılan ChatGPT'nin potansiyelini ve sınırlılıklarını ortaya koymak gereklidir. Bu geniş uygulama alanlarından biri yazılım ve programlama alanıdır. ChatGPT, bilgisayar programlama, programlama dilleri, algoritmalar ve veri yapıları gibi karmaşık konularda rehberlik sağlayarak kullanıcıların teknik sorunları anlamalarına ve çözmelerine yardımcı olmaktadır. Bu alanlarda geniş bir yelpazede yetenekler sunan

ChatGPT, yazılım tasarımı, oluşturma, geliştirme, test etme ve bakım süreçlerini kolaylaştırabilir; ayrıca kod yazma, tamamlama, düzeltme, tahmin etme ve hata ayıklama gibi işlevleri yerine getirebilir. Programlama zorluklarına ilişkin soruları yanıtlarken mantıksal tutarlılığı sürdürebilen ChatGPT, kod üretiminde olağanüstü özelliklere sahiptir (Chen ve ark., 2023; Dong ve ark., 2023; Liu ve ark., 2023; OpenAI, 2022; OpenAI, 2023). ChatGPT'nin bu özellikleri sayesinde, araştırmacılar kod kalitesini artırabilir, zaman ve emek tasarrufu sağlayarak bilişsel yüklerini azaltabilir ve daha yaratıcı çalışmalara odaklanarak üretkenliği artırabilir (Jaber ve ark., 2023; Biswas, 2023; Chen ve ark., 2022). Bu olumlu özellikleriyle ChatGPT, yalnızca doğal dille sınırlı değildir; C++, C#, Java, Python, R gibi ondan fazla programlama ve sorgulama dilinde iletişim kurabilmektedir (Feng ve ark., 2023). Bu yönde yapılan araştırmalar incelendiğinde, yazılımla ilgili çalışmaların (Adamson & Bägerfeldt, 2023; Biswas, 2023; Jaber ve ark., 2023), ChatGPT'nin kod üretimi ve hata ayıklama görevleri, verimliliği ve doğruluğu üzerine araştırmaların (Aljanabi ve ark., 2023; Bang ve ark., 2023; Feng ve ark., 2023; Hansson & Ellréus, 2023; Kashefi & Mukerji, 2023; Sakib ve ark., 2023; Sobania ve ark., 2023; Surameery & Shakor, 2023; Tian ve ark., 2023; White ve ark., 2023) mevcut olduğu; ancak veri üretimindeki etkinliğini inceleyen bir çalışmaya rastlanmadığı görülmektedir.

Teknolojinin hızla ilerlemesi, bilimsel çalışmaların sayısını artırmakta ve çeşitli alanlarda yeni yöntemlerin ortaya çıkmasını sağlamaktadır. Bu yeni yöntemlerin etkililiğini değerlendirmek için genellikle simülasyon çalışmaları yapılmakta ve bu çalışmalar, önceki yöntemlerle karşılaştırmaları içermektedir. Aynı zamanda, bu yöntemlerin farklı koşullar altında işlevselliğini göstermek için bazı koşulların sabit tutulup diğerlerinin değiştirildiği deneysel çalışmalara ihtiyaç duyulmaktadır. Bu deneysel çalışmaları gerçekleştirmek için öncelikle belirtilen koşullar altında veri üretilmesi gerekmektedir. Eğitim bilimlerinde veri üretimi genellikle Madde Tepki Kuramı'na (MTK) uygun olarak gerçekleştirilmektedir. MTK, bireylerin yeteneklerini ve testlerdeki performanslarını daha doğru bir şekilde değerlendirmek için kullanılan bir istatistiksel modelleme çerçevesidir. MTK'nın önemi, kişiselleştirilmiş ölçüm sunma, test maddelerinin zorluk seviyelerini belirleme, ayrıntılı madde analizleri yapma ve testlerin geçerliğini ve güvenilirliğini artırma yeteneğinde yatmaktadır. Ayrıca, MTK'nın kapsamlı, esnek, yüksek çözünürlüklü veri analizi yetenekleri ve veri üretim süreçlerinde sağladığı avantajlar nedeniyle sıkça tercih edildiği bilinmektedir. Bu nedenlerle MTK, eğitim bilimleri ve psikometrik değerlendirmelerde tercih edilen bir yöntemdir. MTK, bir bireyin bir madde üzerindeki performans olasılığına, bireyin yeteneği temelinde odaklanmaktadır. Model-veri uyumu yoluyla doğrulanabilir olan MTK modelleri, güçlü grafiksel ve matematiksel yönlere sahiptir (DeMars, 2010). Son yıllarda, popüler programlama dilleri arasında yer alan R programlama dili, MTK ile ilgili veri üretiminde kullanılmaktadır. R dilinin ChatGPT 3.5 tarafından desteklendiği bilinmektedir (Feng ve ark., 2023). Bu bağlamda, bu çalışma, R dilini kullanarak MTK tabanlı veri üretimi için kod yazma konusunda ChatGPT 3.5'in ne kadar etkili olduğunu belirlemeyi amaçlamaktadır. Bu çalışmanın, eğitimde ölçme alanına ve yapay zekânın veri üretimindeki uygulamalarına birçok önemli katkı sağlaması ve ileri bir dil modeli olan ChatGPT 3.5'in MTK için veri üretim algoritmaları geliştirme potansiyelini göstermesi beklenmektedir. Bu yenilikçi yaklaşımla, yapay zekânın eğitim ve psikolojik ölçümlerde araştırmacılara nasıl yardımcı olabileceğine ve eğitim verilerini simüle etme yeteneğine dair yeni bir bakış açısı sunması beklenmektedir. Hızla gelişen teknolojik bir ortamda, bu çalışma, ChatGPT 3.5 gibi yapay zekâ araçlarının araştırma metodolojilerine entegrasyonunun önemini vurgulamaktadır. Bu kapsamda, ChatGPT 3.5'in bu alandaki etkinliğini değerlendirmek için üretilen veri setleri, oluşturulan simülasyon tasarımında belirtilen koşulları ne ölçüde karşıladığını ve bu veri setlerinin gerçekten amaçlandığı şekilde üretilip üretilmediğini görmek için analiz edilmiştir. Ayrıca, ChatGPT 3.5 algoritmalarıyla üretilen veri setleri, araştırmacılar tarafından geliştirilen algoritmalarla üretilen veri setleriyle karşılaştırılmıştır. Bu bağlamda, bu çalışmada, ChatGPT 3.5 tarafından üretilen veri setlerinin geçerliği, tek boyutluluk, yerel bağımsızlık ve model-veri uyumu gibi temel MTK varsayımları dikkate alınarak kapsamlı bir şekilde ele alınmıştır. Bu doğrultuda, şu araştırma soruları incelenmiştir:

ChatGPT 3.5 algoritmaları kullanılarak İki Parametreli Lojistik Madde Tepki Kuramı Modeline (2PLM) göre üretilen veri setleri için;

- 2PLM ile tek boyutluluk, yerel bağımsızlık ve model-veri uyumu varsayımlarını sağlıyorlar mı?
- Simülasyon tasarımında belirtilen aralıkların dışında kalan madde parametrelerinin sayısı nedir?
- Madde parametrelerinin yanlılık ve RMSE değerleri nedir?

YÖNTEM

Araştırma Deseni

Bu çalışmada, veri üretimi için üç farklı algoritma kullanılmış ve bu algoritmalarla üretilen veri setlerinin simülasyon deseninde belirtilen koşulları ne ölçüde karşıladığı incelenmiştir. R dilinde geliştirilen algoritmalardan ikisi Ekim 2023'te ChatGPT 3.5 tarafından üretilirken, üçüncü algoritma araştırmacılar tarafından geliştirilmiştir. Bu algoritmalar kullanılarak veri setleri üretilmiş ve ardından üretilen veri setlerinin geçerlik incelemeleri gerçekleştirilmiştir. Bu yönüyle çalışma, deneysel bir niteliğe sahiptir (Morris ve ark., 2017).

Simülasyon Deseni

Bu çalışmada, önceki simülasyon çalışmalarında kullanılan (Cohen ve ark., 1996; Li ve ark., 2012) Monte Carlo simülasyonu, ChatGPT 3.5'in veri üretimindeki geçerliği üzerindeki farklı koşul değişikliklerinin etkilerini araştırmak için uygulanmıştır. Monte Carlo yaklaşımı, parametre tahminlerinin dağılımlarının elde edilmesine olanak tanır ve tek bir veri setinden mantıksız sonuçlar çıkarma olasılığını azaltır. Bu bağlamda, gerçek parametreleri yeniden örnekleme amacıyla birden fazla tekrar (replikasyon) gerçekleştirilmiştir (Harwell ve ark., 1996). Bu çalışmada, doğru ve güvenilir parametre tahminleri için önerilen tekrar sayısı (Feinberg & Rubright, 2016) dikkate alınmış ve belirtilen simülasyon koşullarına göre her bir veri seti için 100 tekrar yapılmıştır. Simülasyon tasarımında kullanılan sabit ve değişken koşullar, simülasyon çalışmalarında yaygın olarak kullanılan mantık ve literatürdeki referanslar temel alınarak belirlenmiştir. Çalışma, R programlama dilinin 4.3.2 sürümü kullanılarak yürütülmüştür. Çalışmada kullanılan tüm deneysel koşullar Tablo 1'de sunulmaktadır.

Tablo 1

Simülasyon Deseni

	Koşullar	Düzeyler	Düzey Sayısı
Sabit Koşullar	MTK Model	Tek Boyutlu 2PL	1
	Yetenek (θ)	-3≤ θ≤3	1
	Madde Ayırt Ediciliği (a)	1≤ a≤2	1
	Madde Güçlüğü (b)	-2≤ b≤2	1
Değişen Koşullar	Madde Numarası (k)	20, 40	2
	Örneklem Büyüklüğü (n)	500, 2000	2
Toplam Koşul Sayısı			2x2=4
Tekrar Sayısı			100
Toplam Veri Setleri			4x100=400

Tablo 1 incelendiğinde, çalışma kapsamında ikili (0-1) puanlanan maddeler için tek boyutlu veri üretimi amacıyla farklı koşulların dikkate alındığı görülmektedir. MTK modeli, ikili (0-1) puanlanan maddeler için madde güçlük ve ayırt edicilik parametrelerini ayrı ayrı ele alabilme yeteneği nedeniyle 2PLM'yi kullanmaktadır (DeMars, 2010). Birey yetenek parametresi genellikle ortalaması 0 ve standart sapması 1 olan normal dağılımdan çekilmektedir (Feinberg & Rubright, 2016). Bu yaklaşım, simülasyon çalışmalarında doğru parametre tahminleri elde etmek için yaygın olarak kullanılmaktadır. Madde ayırt edicilik parametreleri, literatüre göre 2PLM tahminleri için daha uygun kabul edilen, minimum 1 ve maksimum 2 arasında değişen tek biçimli (uniform) bir dağılımdan belirlenmiştir (Dekker, 2004; Hambleton ve ark., 1991). Madde güçlük parametreleri ise literatüre göre 2PLM tahminleri için uygun kabul edilen, minimum -2 ve maksimum 2 arasında değişen tek biçimli bir dağılımdan belirlenmiştir (DeMars, 2010). Monte Carlo çalışmaları, doğru madde parametre tahminleri için minimum 20 madde gerektiğini göstermektedir (De Ayala, 2013). Bu nedenle kısa testler için 20 madde, uzun testler için ise 40 madde seçilmiştir. Ayrıca, simülasyon çalışmalarında örneklem büyüklüğünün parametre tahminlerinin doğruluğunu artırdığı

bilinmektedir. Bu nedenle, küçük örneklemler için 500, büyük örneklemler için ise 2000 kişilik örneklem büyüklükleri seçilmiştir (Stone, 1992).
Birinci ve ikinci algoritmalar (A1 ve A2), veri üretimi için ChatGPT 3.5 tarafından verilen komutlar (promptlar) aracılığıyla geliştirilmiştir. ChatGPT 3.5'e yeni algoritmalar üretmesi için farklı komutlar verilmiş, ancak üretilen algoritmaların A1 ve A2'den farklı olmadığının belirlenmesi üzerine ChatGPT 3.5 ile algoritma geliştirme süreci sonlandırılmıştır. A1 ve A2 için kullanılan komutlar Ek 1 ve Ek 2'de verilmiştir. Veri üretimi için kullanılan üçüncü algoritma (A3), R dilindeki 'mirt' paketi (Chalmers, 2012) kullanılarak araştırmacılar tarafından geliştirilmiştir. Geliştirilen üç algoritmaya ait kodlar sırasıyla Ek 3, Ek 4 ve Ek 5'te yer almaktadır. Veri üretim aşamasında, geliştirilen algoritmalar R ortamında çalıştırılmış ve simülasyon desenindeki koşullar dikkate alınarak her biri için 100 tekrarlı veri setleri üretilmiştir. Sonuç olarak, 2 farklı test uzunluğu (20, 40), 2 farklı örneklem büyüklüğü (500, 2000) ve 100 tekrardan oluşan toplam 2x2x100=400 farklı veri seti üretilmiştir.

Verilerin Analizi

Araştırma kapsamında, A1, A2 ve A3 algoritmalarıyla sırasıyla üretilen veri setlerinin, tek boyutlu 2PLM'ye özgü varsayımları (tek boyutluluk, yerel bağımsızlık, model-veri uyumu) karşılaması beklenmektedir, çünkü bu veri setleri söz konusu modele uygun olarak üretilmiştir. Aksi takdirde, üretilen veriler planlanan veri üretim senaryosuna uygun olmayacak ve dolayısıyla elde edilen sonuçlar hatalı olacaktır. Ayrıca, üretilen veri setlerinde tahmin edilen parametrelerin, veri üretiminin doğruluğunu göstermek amacıyla simülasyon deseninde belirtilen koşulları karşılaması gerekmektedir. Bu nedenle, üretilen veri setlerinin MTK varsayımlarına uygunluğunu ve parametrelerin simülasyon deseninde belirtilen koşulları karşılayıp karşılamadığını kontrol etmek ilk adımdır ve bu amaç doğrultusunda aşağıdaki incelemeler gerçekleştirilmiştir.

Veri Setlerinin Tek Boyutluluk İncelemesi

Tek boyutluluk varsayımını incelemek için R programlama dilinde bulunan 'psych' (Revelle, 2020) ve 'sirt' (Robitzsch, 2019) paketleri kullanılarak Açımlayıcı Faktör Analizi (AFA) gerçekleştirilmiştir. Bu bağlamda, ikili (1-0) puanlanan veri setlerine uygun olan 'tetrachoric korelasyon matrisleri' tüm veri setleri için oluşturulmuştur. AFA varsayımları kapsamında, Kaiser-Meyer-Olkin (KMO) ve Bartlett testlerinin sonuçları incelenmiştir. Hem A3 algoritmasıyla hem de ChatGPT 3.5 algoritmaları (A1 ve A2) yardımıyla üretilen veri setlerinde, KMO test sonuçlarının her bir koşulda 0.87'den büyük olduğu ve Bartlett testlerinin anlamlı olduğu ($p<.05$) görülmüştür. Tek boyutluluk varsayımının sağlanmasının ardından, faktör sayısı Horn'un 'Paralel Analiz' yöntemi (Horn, 1965) ile belirlenmiştir. Faktör çıkarma işlemi için 'psych' paketindeki 'fa.parallel' fonksiyonu kullanılmış; faktörlere ilişkin özdeğerler ve scree plot grafikler incelenmiştir. Faktör sayısına karar verirken faktör yüklemeleri, açıklanan varyans oranları ve scree plot grafikleri dikkate alınmıştır.

Veri Setlerinin Yerel Bağımsızlık İncelemesi

MTK'nın yerel bağımsızlık varsayımı, bireylerin yetenekleri kontrol edildiğinde madde yanıtlarının birbirinden bağımsız olması gerektiğini belirtir (DeMars, 2010). Bu varsayım, Yen'in Q3 testi (Yen, 1984) kullanılarak kontrol edilmiştir. Yerel bağımsızlık varsayımını doğrulamak için, R dilindeki 'subscore' paketinden (Dai ve ark., 2022) 'Yen.Q3' fonksiyonu kullanılmıştır. Yen (1984) tarafından belirlenen, problemli maddelerin artıklar arasındaki korelasyonlarının 0.20'yi aşması durumunda yerel bağımsızlık ihlali olduğu koşulu, ihlalleri belirlemek için kriter olarak alınmıştır.

Veri Setlerinin Model-Veri Uyumu İncelemesi

Üretilen veri setlerinin 2PLM ile uyumunu kontrol etmek için M2 istatistiğinin anlamlılık düzeyleri incelenmiştir. Gözlenen ve beklenen marjinal olasılıklar arasındaki artıklar kullanılarak oluşturulan M2 istatistiği anlamlı değilse, veri setinin modelle uyumlu olduğu kabul edilmektedir (Maydeu-Olivares & Joe, 2006). M2 istatistiğinin anlamlılığını kontrol etmek için R programlama dilindeki 'mirt' paketinden (Chalmers, 2012) 'M2' fonksiyonu kullanılmıştır. Ayrıca, model-veri uyumunun değerlendirilmesinde, yuvalanmış MTK modellerinin (1PL, 2PL, 3PL) '-2 log-likelihood' değerleri ve anlamlılıkları dikkate

alınmıştır (Thissen & Steinberg, 1986). Modelleri karşılaştırmak için R programlama dilindeki 'anova' fonksiyonu kullanılmıştır.

Belirtilen Aralıklar İçinde Kalan Madde Parametrelerinin Sayısı

Üretilen veri setlerindeki madde ayırt edicilik (a) ve madde güçlük (b) parametrelerinden kaç tanesinin simülasyon deseninde belirtilen koşulları karşıladığı belirlenmiştir. Bu bağlamda, araştırmacılar, R programlama dilinde belirtilen aralıkların dışında kalan parametrelerin sayısını tespit etmek için sayaçlar oluşturmuştur. Araştırmada değişen madde sayıları ve tekrar sayıları göz önünde bulundurularak, 20 madde için 20x100=2000 ve 40 madde için 40x100=4000 durumda, belirtilen aralıkların dışında kalan madde güçlük ve madde ayırt edicilik parametrelerinin sayısı tespit edilmiştir.

Madde Parametrelerinin Yanlılık ve Hata (RMSE) Değerleri

Simülasyon deseninde belirtilen koşullar dikkate alınarak üretilen veri setlerindeki madde parametrelerinin yanlılık ve hata (RMSE) değerlerini değerlendirmek için, Denklem 1 ve 2'de yer alan formüller kullanılmıştır.

$$Yanlılık = \frac{\sum_{i=1}^{K}(\hat{X}_{i-}X_i)}{K} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{K}(\hat{X}_{i-}X_i)^2}{K}} \quad (2)$$

Denklem 1 ve 2'de, $\hat{X}_i$ madde i için parametreyi $(i = 1, 2, \ldots, K)$, X_i gerçek parameter tahminini, K ise madde sayısını ifade etmektedir. Yanlılık ve RMSE değerleri, 100 tekrarlı veri setlerindeki her bir koşul için hesaplanmış ve ardından ortalamaları alınmıştır.

BULGULAR

Bu bölümde, araştırma kapsamında elde edilen bulgular sunulmaktadır. Çalışmada, simülasyon deseninde belirtilen koşullara dayalı olarak, 2PLM için üç farklı algoritma (A1, A2 ve A3) kullanılarak veri üretimi gerçekleştirilmiştir. Üretilen veri setlerinin belirtilen simülasyon koşullarını ne ölçüde karşıladığını belirlemek amacıyla incelemeler yapılmıştır. Bu incelemelerin sonuçları Tablo 2'de gösterilmektedir.

Tablo 2

Veri Setleri Üzerine Yapılan Analizlerin Sonuçları

Koşul		Madde Sayısı	Örneklem Büyüklüğü	Faktör Sayısı	Paralel Analiz (Faktör Sayısı)	Yerel Bağımsızlık İhlali (Yen Q3)	M2 Uyum İhlali	Yuvalanmış Model Uyumu	Aralık Dışında Kalan a Parametresi	Aralık Dışında Kalan b Parametresi	Ortalama Yanlılık a	Ortalama Yanlılık b	Ortalama RMSE a	Ortalama RMSE b
1	A1	20	500	1	1 (4 x 2)	13	0 (%0)*	2PL	1549 (%77)**	0 (%0)***	-1.201	-0.093	2.108	1.145
2	A2	20	500	1	1	-	5 (%5)	2PL	241 (%12)	632 (%32)	0.012	0.206	0.21	1.936
3	A3	20	500	1	1	-	4 (%4)	2PL	222 (%11)	86 (%4)	0.018	0.04	0.189	0.137

4	A1	20	2000	1	1 (4 x 2)	2	4 (%4)	2PL	1580 (%79)	0 (%0)	-1.227	0.397	2.091	1.037
5	A2	20	2000	1	1	-	4 (%4)	2PL	78 (%3)	661 (%33)	0	-0.407	0.103	2.042
6	A3	20	2000	1	1	-	4 (%4)	2PL	84 (%4)	45 (%2)	0.007	-0.288	0.093	0.057
7	A1	40	500	1	1	9	4 (%4)	2PL	3129 (%78)	0 (%0)	-1.297	-0.013	2.142	1.215
8	A2	40	500	1	1	-	4 (%4)	2PL	495 (%12)	1280 (%32)	0.019	-0.073	0.198	2.093
9	A3	40	500	1	1	-	7 (%7)	2PL	427 (%11)	173 (%4)	0.013	0.012	0.183	0.159
10	A1	40	2000	1	1	-	6 (%6)	2PL	3181 (%79)	0 (%0)	-1.271	0.421	2.16	1.155
11	A2	40	2000	1	1	-	2 (%2)	2PL	135 (%3)	1248 (%31)	0.003	-0.186	0.095	2.133
12	A3	40	2000	1	1	-	7 (%7)	2PL	144 (%4)	79 (%2)	0.002	0.086	0.09	0.077

* Uyum ihlali yüzdesi ** Aralık dışında kalan a parametresinin yüzdesi *** Aralık dışında kalan b parametresinin yüzdesi

Tablo 2'de görüldüğü gibi, madde sayısının 20 olduğu tüm koşullarda, AFA analizinin 100 tekrarlı veri setlerinde baskın bir tek boyut bulunduğunu gösterdiği tespit edilmiştir. Paralel analizde ise yalnızca yapay zeka (YZ) algoritmasında, iki koşulda (20 madde ve 500 örneklem ile 20 madde ve 2000 örneklem) 4 veri setinde iki faktörlü bir yapı belirlenmiştir. Madde sayısının 40 olduğu tüm koşullarda hem AFA hem de Paralel Analiz sonuçları tek faktörlü bir yapıyı desteklemektedir. Sonuç olarak, küçük ihmal edilebilir istisnalar dışında, ChatGPT 3.5 algoritmalarıyla üretilen veri setlerinin beklendiği gibi tek boyutlu olduğu söylenebilir. Tablo 2 incelendiğinde, yerel bağımsızlık varsayımının ihlallerinin yalnızca ChatGPT 3.5 A1 algoritmasında bazı durumlarda görüldüğü gözlemlenmiştir. Bu algoritmada, yerel bağımsızlık ihlalleri, 20 madde ve 500 örneklem için 100 veri setinden 13'ünde, 20 madde ve 2000 örneklem için 2 veri setinde ve 40 madde ve 500 örneklem için 9 veri setinde tespit edilmiştir. Bu noktada, yerel bağımsızlık ihlali olan veri setlerinin bulunmaması dikkate alındığında, ChatGPT 3.5 A2 algoritması ve A3'ün daha başarılı olduğu söylenebilir.

Tablo 2 incelendiğinde, M2 istatistiğinin anlamlılık kontrollerine göre, model-veri uyumunu sağlamayan veri setlerinin oranının tüm algoritmalar için %0 ile %7 arasında değiştiği gözlemlenmiştir. 20 madde ve 500 örneklem büyüklüğü koşullarında, en iyi model-veri uyumu A1 ile sağlanmış ve hiçbir uyum ihlali tespit edilmemiştir (%0). Toplam madde sayısının 40 ve örneklem büyüklüğünün 2000 olduğu koşullarda, en az model-veri uyumu ihlali (%2) A2 ile gözlemlenirken, en fazla ihlal (%7) A3 ile gerçekleşmiştir. Ayrıca, yuvalanmış model uyumu dikkate alındığında, tüm koşullarda veri setlerinin 2PL modeli ile daha iyi uyum sağladığı belirlenmiştir. Genel olarak, ChatGPT 3.5 algoritmalarının (A1 ve A2), veri setlerinin 2PL modeli ile uyumu açısından, araştırmacılar tarafından geliştirilen A3 algoritması kadar başarılı olduğu söylenebilir.

Ayrıca, simülasyon deseninde belirtilen aralıklara uygun olarak parametre a için veri üretiminin, tüm koşullarda ChatGPT 3.5 A2 ve araştırmacılar tarafından geliştirilen A3 ile başarılı bir şekilde gerçekleştirildiği görülmektedir. Ancak, ChatGPT 3.5 A1'in parametre a için belirtilen aralıklar içinde veri üretme konusunda oldukça zayıf olduğu, üretilen a parametrelerinin büyük bir kısmının (%77-%79) belirtilen aralıkların dışında kaldığı gözlemlenmiştir. Örneklem büyüklüğünün 500 olduğu tüm koşullarda, A3, parametre a için belirtilen aralıkların dışında en az veri üretimi oranını gösterirken, örneklem büyüklüğünün 2000 olduğu koşullarda en düşük sapma oranı A2 ile görülmüştür. Parametre b için, belirtilen aralıklar içinde en iyi veri üretimi A1 ile gerçekleştirilmiş, bunu yakından

A3 takip etmiştir. Ancak, A2, parametre b için belirtilen aralıklar içinde veri üretimi konusunda zayıf kalmış ve b parametrelerinin bir kısmı (%31-%33) belirtilen aralıkların dışında kalmıştır. Sonuç olarak, MTK varsayımlarını karşılamada başarılı olmasına rağmen, ChatGPT 3.5 algoritmalarıyla üretilen veri setlerinde, madde parametre tahminlerinin simülasyon deseninde belirtilen koşulları karşılama konusunda A3'e kıyasla daha fazla zorluk yaşadığı söylenebilir.

Yanlılık ortalaması açısından, tüm koşullarda parametre a için en düşük yanlılık ortalamasının A1'de olduğu, A2 ve A3'ün benzer değerler ürettiği gözlemlenmiştir. Toplam madde sayısının 20 ve örneklem büyüklüğünün 500 olduğu koşulda, parametre b için en düşük yanlılık ortalaması A1'de iken, diğer tüm koşullarda bu değer A2'de görülmüştür. Örneklem büyüklüğünün 2000 olduğu koşullarda, en yüksek yanlılık ortalaması A1 ile elde edilmiştir. Örneklem büyüklüğünün 500 olduğu koşullarda, toplam madde sayısının 20 olduğu durumlarda en yüksek yanlılık ortalaması A2 ile, madde sayısının 40 olduğu durumlarda ise A3 ile elde edilmiştir. Ayrıca, tüm koşullarda en yüksek RMSE ortalama değerlerinin parametre a için A1'de, parametre b için ise A2'de olduğu belirlenmiştir. Bununla birlikte, tüm koşullarda hem parametre a hem de b için en düşük RMSE ortalama değerleri A3'te bulunmuştur. Bu nedenle, madde parametreleri açısından yanlılık ve RMSE ortalamaları bakımından en iyi sonuçların A3 ile üretilen veri setlerinde elde edildiği, ChatGPT 3.5 algoritmalarının ise daha yüksek yanlılık ve RMSE ortalamalarına yol açtığı söylenebilir.

TARTIŞMA VE SONUÇ

Bulgular değerlendirildiğinde, ChatGPT 3.5 algoritmalarının genel olarak MTK varsayımlarına uygun veri üretiminde başarılı olduğu söylenebilir. Küçük ve ihmal edilebilir birkaç istisna dışında, ChatGPT 3.5 algoritmalarıyla yerel bağımsızlık ihlali olmayan, tek boyutlu ve 2PLM ile uyumlu veri setleri üretilmiştir. Ancak, madde parametrelerinin simülasyon deseninde belirtilen aralıklara uygunluğu ile ilişkili yanlılık ve RMSE ortalamaları açısından değerlendirildiğinde, ChatGPT 3.5 algoritmalarının, araştırmacılar tarafından geliştirilen algoritmaya kıyasla daha fazla sorun yaşadığı görülmektedir.

Tablo 2'deki sonuçlara dayanarak, üç algoritma ile üretilen veri setlerinin tüm koşullarında, AFA analizinin simülasyon deseninde belirtildiği gibi tek boyutlu bir yapıyı belirlediği görülmektedir. Paralel analizde ise A1 algoritmasında toplamda sekiz veri setine denk gelen iki koşul dışında, tüm koşullarda tek boyutlu bir yapı gözlemlenmiştir. Yerel bağımsızlık kontrollerinde, A1'de üç koşul dışında tüm koşullarda varsayımın sağlandığı görülmektedir. M2 uyum ihlalleri, üç algoritmanın tümünde ihlal eden veri setlerinin varlığını ortaya koymakla birlikte, bu ihlaller nispeten azdır ve %0 ile %7 arasında değişmektedir. ChatGPT 3.5 algoritmalarındaki yerel bağımsızlık ihlalleri çeşitli nedenlerden kaynaklanabilir. Düşük kaliteli veya uzman olmayan komutlar, maddeler arasındaki bağımlılığı artırabilir. Ayrıca, ChatGPT 3.5'in kendi sınırlamaları da bu ihlallere katkıda bulunabilir. Bu tür ihlalleri azaltmak için daha gelişmiş modeller (örneğin, ChatGPT 4) kullanılabilir. Uzmanlar tarafından hazırlanmış ve yönlendirilmiş komutlar kullanılarak maddeler arasındaki bağımlılık en aza indirilebilir. Ayrıca, ChatGPT 3.5'in sınırlamaları dikkate alınarak analizler yapılmalı ve gerektiğinde insan uzmanlığından yararlanılmalıdır. Genel olarak, 20 madde ve 500 örneklem ile üretilen veri setlerinde A1 algoritması, 40 madde ve 2000 örneklem ile üretilen veri setlerinde ise A2 algoritası en iyi 2PL uyumunu sağlamıştır ve algoritmalar arasında belirgin bir uyum ihlali modeli bulunmamaktadır. Yuvalanmış modellerin 2PL ile uyumu karşılaştırıldığında, tüm algoritmalarda veri setlerinin 2PL ile iyi bir uyum gösterdiği tespit edilmiştir. Belirtilen aralıkların dışında kalan madde parametreleri a ve b'nin incelenmesinde, aralıklara en yakın sonuçların A3 ile elde edildiği görülmüştür. A1 algoritmasının istenen aralıklar içinde a parametreleri üretmede yetersiz olduğu, A2'nin ise b parametreleri için yetersiz olduğu belirlenmiştir. Bu parametrelerin yanlılık ve RMSE ortalamalarının da beklenildiği gibi bu sonucu desteklediği görülmüştür.

Sonuç olarak, ChatGPT 3.5 algoritmaları kullanılarak veri üretimi genel olarak başarılıdır; ancak simülasyon deseninin koşullarına en iyi uyum, araştırmacılar tarafından geliştirilen algoritma (A3) ile sağlanmıştır. ChatGPT 3.5 algoritmalarının desteğiyle üretilen veri setlerinin, özellikle istenen aralıklar içinde yanlılıktan arındırılmış parametreler üretme açısından, araştırmacılar tarafından ChatGPT 3.5 desteği olmaksızın üretilen veri setlerine kıyasla daha az yeterli olduğu görülmüştür. Literatürdeki bulgulara benzer şekilde, bu çalışma da ChatGPT'nin MTK varsayımlarını başarılı bir şekilde karşılayabileceğini ortaya koymakta ve yapay zekâ çıktılarının yönlendirilmesinde insan uzmanlığının vazgeçilmez rolünü vurgulamaktadır.

Literatürdeki bulgulara benzer şekilde, ChatGPT 3.5'in bilgisayar kodu üretiminde mükemmel olmadığı ve ChatGPT 3.5'in çözümlerinin insan çözümlerine kıyasla daha düşük kalitede olduğu belirlenmiştir (Adamson &

Bägerfeldt, 2023; Kundalia, 2023). Ancak ChatGPT 3.5'in birçok probleme doğru çözümler üretebildiğini, fakat bazı yönlerden yetersiz kaldığını belirten çalışmalar da mevcuttur ve bu durum bu çalışmayla da paralellik göstermektedir (Hansson & Ellréus, 2023; Sakib ve ark., 2023; Tian ve ark., 2023). Ayrıca, ChatGPT 3.5 desteğiyle üretilen veri algoritmalarından birinde (A1) tek boyutluluk ve yerel bağımsızlık varsayımlarını karşılama konusunda problemler olduğu tespit edilirken, diğerinde (A2) bu sorunların gözlemlenmediği bulunmuştur. Bu durum, iki algoritmanın farklı araştırmacılar tarafından farklı komutlar kullanılarak geliştirilmiş olmasından kaynaklanıyor olabilir. Bu çalışmanın sonucunda, literatürde belirtilen bazı çalışmalarda olduğu gibi (Surameery & Shakor, 2023), ChatGPT 3.5'in uzman seviyesine yakın veri üretebildiği ve üretilen veri setlerinin doğruluğunun uzmanlık düzeyiyle doğru orantılı olarak arttığı söylenebilir. Yapılan diğer çalışmalarda da belirtildiği gibi (Bang ve ark., 2023), ChatGPT 3.5 ile geliştirilen algoritmaların başarısı, uzmanlar tarafından verilen komutlar ve yönlendirmelerle doğrudan ilişkilidir.

Uzmanlık arttıkça ChatGPT 3.5'e girilen komutların kalitesinin de arttığı ve bu durumun çıktıyı uzman doğruluğuna daha yakın hale getirdiği ifade edilebilir. Görüldüğü üzere, endişelerin aksine ChatGPT 3.5 uzmanlık ihtiyacını ortadan kaldırmamaktadır; aksine, uzmanların işini kolaylaştıran bir araç olarak hizmet vermektedir. Ayrıca, ChatGPT'nin her versiyonunun belirli yetenek ve sınırlamalara sahip olduğu dikkate alınmalıdır. Bu çalışmada, ücretsiz ve daha fazla kişi tarafından erişilebilir olan ChatGPT 3.5'in MTK kapsamında veri üretimi için algoritma geliştirme yeterliliği incelenmiştir. Çalışma kapsamında gerçekleştirilen geçerlik kontrollerinde, ChatGPT 3.5 algoritmalarıyla üretilen veri setlerinde madde parametrelerinin simülasyon deseninde belirtilen aralıkları karşılama konusunda zayıf olduğu ve yerel bağımsızlık ihlallerinin daha sık görüldüğü gözlemlenmiştir. ChatGPT 4 veya gelecekteki versiyonlarla, bu araştırmada elde edilenlerden çok daha iyi sonuçlar elde edilebilir. Çünkü kullanılan versiyonun doğasının, araştırmada elde edilen sonuçların güvenilirliği ve doğruluğu üzerinde doğrudan bir etkisi olabileceği açıktır. Daha yeni versiyonların, daha büyük bir eğitim veri seti üzerinde eğitildikleri için daha doğru sonuçlar üretme eğiliminde olabileceği göz önünde bulundurulmalıdır.
Teknolojinin giderek ilerlediği ve önemli bir rol oynadığı bir dünyada, teknolojiyi yasaklamanın anlamlı olmadığı açıktır (King, 2023). Tüm bu sonuçlar doğrultusunda, ChatGPT 3.5'in kullanımını kısıtlamak yerine, çalışmalarda bir destek aracı olarak kullanılmasının daha uygun olduğu görülmektedir (Biswas, 2023; Deng & Lin, 2022; Dwivedi ve ark., 2023; Hansson & Ellréus, 2023; Sollie, 2009, Surameery & Shakor, 2023). Bu çalışmanın sonuçları, yapay zekanın insan yeteneklerini geliştirmek için bir destek aracı olarak kullanıldığı dengeli bir yaklaşımı ortaya koymakta ve onu insanın yerini alacak bir araç olarak değil, insanın çalışmalarını destekleyecek bir araç olarak değerlendirmektedir. Bu entegrasyonun, araştırmacıların veri üretimi ve ön analiz için yapay zekaya güvenerek daha yüksek seviyeli analitik görevlere odaklanmalarını sağlayarak daha verimli araştırma süreçlerine yol açabileceği öngörülmektedir. Bu doğrultuda, ChatGPT 3.5'in kullanımında uzmanlığın önemini dikkate alarak, araştırmacıların veri üretiminde ChatGPT 3.5'ten faydalanmaları ve veri üretiminde harcadıkları süreyi kısaltmaları tavsiye edilmektedir. Araştırmacılar, kalan zamanlarını araştırmalarının daha yaratıcı bölümleri için kullanabilirler (Hirsh-Pasek & Blinkoff, 2023). Farklı uzmanların komutlarına bağlı olarak farklı çıktılar üretilmesi durumunun, uzmanlar arası farklılıkları dikkate alan ayrı bir çalışmada incelenmesi önerilmektedir. Ayrıca, ChatGPT 3.5'in genel olarak MTK varsayımlarını karşıladığı ancak madde geri kazanımını tam olarak sağlayamadığı gerçeğini incelemek için daha ayrıntılı çalışmalar yapılması önerilmektedir. Bu bağlamda, gelecekteki çalışmalarda yalnızca ChatGPT ile sınırlı kalmayıp, ChatGPT 4, BERT ve T5 gibi daha gelişmiş modelleri de içeren farklı yapay zekâ modellerinin incelenmesi önemlidir. Bu modellerin performanslarını karşılaştırarak en uygun olanı belirlemek mümkün olacaktır. ChatGPT'nin sağladığı yanıtların kalitesini artırmak için komut tasarımını geliştirmek önemlidir; daha açık ve yönlendirici komutlar uzman rehberliğinde hazırlanabilir. Ayrıca, modeli daha büyük ve daha çeşitli veri setleriyle yeniden eğitmek, modelin bilgi tabanını genişletecek ve daha doğru ve güvenilir yanıtlar üretilmesini sağlayacaktır. Yerel bağımsızlık ihlallerini en aza indirmek için, modelin sınırlamalarını dikkate alarak özel stratejiler geliştirilebilir ve bağımlılıkların etkisi manuel kontroller yoluyla azaltılabilir.

KAYNAKÇA

Adamson, V., & Bägerfeldt, J. (2023). *Assessing the effectiveness of ChatGPT in generating Python code* [Bachelor Degree Project, University of Skövde]. Bachelor's Degree Project in Information Technology.

Aljanabi, M., Ghazi, M., Ali, A. H., Abed, S. A. & ChatGpt (2023). ChatGpt: open possibilities. *Iraqi Journal For Computer Science and Mathematics*, *4*(1), 62-64. https://doi.org/10.52866/ijcsm.2023.01.01.0018

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(*1*), 52-62.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR, abs/2302.04023, 2023.*

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March 3-10). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Canada (pp. 610–623). New York, NY: ACM. https://doi.org/10.1145/3442188.3445922

Biswas, S. (2023). Role of ChatGPT in Computer Programming.: ChatGPT in Computer Programming. *Mesopotamian Journal of Computer Science*, *2023*, 8-16. https://doi.org/10.58496/MJCSC/2023/002

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, B., Zhang, F., Nguyen, A., Zan, D., Lin, Z., Lou, J. G., & Chen, W. (2022). Code T: Code generation with generated tests. *arXiv preprint arXiv:2207.10397*.

Chen, E., Huang, R., Chen, H. S., Tseng, Y. H., & Li, L. Y. (2023). GPTutor: a ChatGPT-powered programming tool for code explanation. *arXiv preprint arXiv:2305.01863.*

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*(1), 15-26.

Dai, S., Wang, X., & Svetina, D. (2022). subscore: Computing Subscores in Classical Test Theory and Item Response Theory. R package version 3.3, <https://CRAN.R-project.org/package=subscore>.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Dekker, M. (2004). *Parameter estimation techniques*. In Baker Kim, (Eds.), *Item response theory*, New York.

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

Deng, J., & Lin, Y. (2022). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems*, *2*(2), 81-83. https://doi.org/10.54097/fcis.v2i2.4465

Dong, Y., Jiang, X., Jin, Z., & Li, G. (2023). *Self-collaboration Code Generation via ChatGPT*. *arXiv preprint arXiv:2304.07590*. https://doi.org/10.48550/arXiv.2304.07590

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, Article 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, Article 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Else, H. (2023). Abstracts written by ChatGPT fool scientists, *Nature, 613*(7944), 423-423. https://doi.org/10.1038/d41586-023-00056-7.

Elsevier, (2023). The Use of AI and AI-assisted Technologies in Scientific Writing. 〈https:// www.elsevier.com/about/policies/publishing-ethics〉 (accessed 20th Feb, 2023)

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

Feng, Y., Vanam, S., Cherukupally, M., Zheng, W., Qiu, M., & Chen, H. (2023). Investigating Code Generation Performance of Chat-GPT with Crowdsourcing Social Data. In *Proceedings of the 47th IEEE Computer Software and Applications Conference* (pp. 1-10).

Floridi, L. (2023). AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philosophy & Technology*, 36(15), 1-7. https://doi.org/10.1007/s13347-023-00621-y

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hansson, E., & Ellréus, O. (2023). *Code Correctness and Quality in the Era of AI Code Generation: Examining ChatGPT and GitHub Copilot* [Bachelor Degree Project, University of Skövde]. Digitala Vetenskapliga Arkivet.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement, 20*(2), 101-125.

Hassani, H., & Silva, E. S. (2023). The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field. *Big data and cognitive computing, 7*(*2*), 62. https://doi.org/10.3390/bdcc7020062

Hirsh-Pasek, K. & Blinkoff, E. (2023). "ChatGPT: Educational friend of foe?", Brookings Institution, https://www.brookings.edu/blog/education-plusdevelopment/2023/01/09/chatgpt-educational-friend-or-foe

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Hu, K. (2023) ChatGPT sets record for fastest-growing user base - analyst note,February 2, 〈 https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-userbase-analyst-note-2023–02-01/≥

Huang, R. (2019). *Educational technology a primer for the 21st century*. Springer Nature Singapore Pte Ltd.

Jaber, M. A., Beganovic, A., & Abd Almisreb, A. (2023). Methods and Applications of ChatGPT in Software Development: A Literature Review. *Southeast Europe Journal of Soft Computing*, *12*(1), 08-12. http://dx.doi.org/10.21533/scjournal.v12i1.251

Kashefi, A., & Mukerji, T. (2023). Chatgpt for programming numerical methods. *Apr. 2023, arXiv:2303.12093*.

Khoury, R., Avila, A. R., Brunelle, J. and Camara, B. M. (2023) "How Secure is Code Generated by ChatGPT?" *Apr. 2023, arXiv:2304.09655.*

King, M. R. (2023). Outsourcing Your Faculty Application to ChatGPT: Would this Work? Should this Work?. *Cellular and Molecular Bioengineering*, *16*(4), 423-426. https://doi.org/10.1007/s12195-023-00777-9

Kundalia, N.D. (2023). ChatGPT and the future of writing. Hindustan Times. Retrieved January 31, 2023, from 〈https://www.hindustantimes.com/books/ chatgpt-and-the-future-of-writing-101675090609362.html〉

Laumer, S., Maier, C., Eckhardt, A., & Weitzel, T. (2016). Work routines as an object of resistance during information systems implementations: Theoretical foundation and empirical evidence. *European Journal of Information Systems, 25*(4), 317–343. https://doi.org/10.1057/ejis.2016.1

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.

Liu, J., Xia, C. S., Wang, Y., & Zhang, L. (2023). Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210.* https://doi.org/10.48550/arXiv.2305.01210

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713-732. https://doi.org/10.1007/s11336-005-1295-9

McGee, R. W. (2023). What Will the United States Look Like in 2050? A ChatGPT Short Story. Working Paper, April 8, 2023. https://ssrn.com/abstract=4413442

Mollick, E. (2022). ChatGPT Is a Tipping Point for AI. Harvard Business Review. December 14.

Montti, R. (2022). What is ChatGPT and how can you use it? Search Engine Journal (Accessed from) 〈 https://www.searchenginejournal. com/what-is-chatgpt/473664/#close〉

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine, 38*(11), 2074-2102.

O'Connor, S. (2023). Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse. *Nurse Education in Practice, 66,* Article 103537. https://doi.org/10.1016/j.nepr.2022.103537

OpenAI, (2023) ChatGPT: Optimizing Language Models for Dialogue. Available at: 〈 https://openai.com/blog/chatgpt/〉

OpenAI, T. B. (2022). Chatgpt: Optimizing language models for dialogue. OpenAI. https://openai.com/blog/chatgpt/ https://openai.com/blog/chatgpt/

Ortiz, S. (2023a). ChatGPT is changing everything. But it still has its limits. ZDNet (Available online) 〈 https://www.zdnet.com/article/chatgpt-is-changing-everythin g-but-it-still-has-its-limits/〉

Ortiz, S. (2023b). What is ChatGPT and why does it matter? Here's everything you need to know. ZD Netto Innovation (Accessed from) 〈https://www.zdnet.com/article /what-is-chatgpt-and-why-does-it-matter-heres-everything-you-need-to-know/〉

Perrigo, B. (2023). OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. Time. 〈https://time.com/6247678/openai-chatgpt-kenyaworkers

Revelle, W. (2020). Psych: Procedures for psychological, psychometric, and personality research. Evanston, Illinois: Northwestern University. https://CRAN.R-project.org/package=psych.

Robitzsch, A. (2019). sirt: Supplementary item response theory models. R package version 3.1-80. https://CRAN.R-project.org/package=sirt

Rosenzweig-Ziff, D. (2023). New York City blocks use of the ChatGPT bot in its schools. The Washington Post. 〈https://www.washingtonpost.com/education/2023/01/05/n yc-schools-ban-chatgpt/

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching, 6*(*1*). https://doi.org/10.37074/jalt.2023.6.1.9

Sakib, F. A., Khan, S. H., & Karim, A. H. M. (2023). Extending the frontier of chatgpt: Code generation and debugging. *arXiv preprint arXiv:2307.08260*. https://doi.org/10.48550/arXiv.2307.08260

Sobania, D., Briesch, M., Hanna, C., & Petke, J. (2023). An analysis of the automatic bug fixing performance of chatgpt. *arXiv preprint arXiv:2301.08653*. https://doi.org/10.48550/arXiv.2301.08653

Sollie, P. (2009). On Uncertainty in Ethics and Technology. In P. Sollie, & M. Düwell (Eds.), *Evaluating New Technologies: Methodological Problems for the Ethical Assessment of Technology Developments, The International Library of Ethics, Law and Technology* (pp. 141–158). Springer.

Stokel-Walker, C. (2022). *AI bot ChatGPT writes smart essays - should professors worry?* Nature (London). https://doi.org/10.1038/d41586-022-04397-7

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16(*1), 1-16.

Surameery, N. M. S., & Shakor, M. Y. (2023). Use chat gpt to solve programming bugs. *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290*, *3*(01), 17-22. https://doi.org/10.55529/ijitc.31.17.22

Susnjak, T. (2022). ChatGPT: The End of Online Exam Integrity? *arXiv preprint arXiv: 2212.09292*. https://doi.org/10.48550/arXiv.2212.09292

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567-577.

Tian, H., Lu, W., Li, T. O., Tang, X., Cheung, S. C., Klein, J., & Bissyandé, T. F. (2023). Is ChatGPT the Ultimate Programming Assistant--How far is it?. *arXiv preprint arXiv:2304.11938*. https://doi.org/10.48550/arXiv.2304.11938

Uc-Cetina, V., Navarro-Guerrero, N., Martin-Gonzalez, A., Weber, C., & Wermter, S. (2023). Survey on reinforcement learning for language processing. *Artificial Intelligence Review, 56*(2), 1543-1575.

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. (2023). ChatGPT: five priorities for research. *Nature, 614*, 224–226. https://doi.org/10.1038/d41586- 023-00288-7

White, J., Hays, S., Fu, Q., Spencer-Smith, J., & Schmidt, D. C. (2023). Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. *arXiv preprint arXiv:2303.07839*. https://doi.org/10.48550/arXiv.2303.07839

Yang, S. (2022). The Abilities and Limitations of ChatGPT. Anaconda Perspectives. 〈https ://www.anaconda.com/blog/the-abilities-and-limitations-of-chatgpt〉

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145.

Zhai, X., (2022). ChatGPT User experience: Implications for education. (December 27, 2022). Available at SSRN: https://ssrn.com/abstract=4312418 or http://dx.doi.org/10.2139/ssrn.4312418.

EK

Ek 1. Komut 1

1- I want to generate 1-0 response data for 2 Parametric Logistic Item Response Model. My test length is 20. My sample size is 500. My ability parameters have a normal distrubition with mean 0, standard deviation 1. Item discrimination parameter will have a uniform distrubition and between 1 and 2. My difficulty parameter (b) will have a uniform distrubition and between -2 and 2.

2- My replication number is 100. I want to save 1-0 responses to an empty list. Can you give me the codes of this simulation in R language?

3- Ok, but this code does not label the columns

4- Ok. I want to estimate 2PL model parameters. First I want to estimate ability parameters via EAP method. Then I want to estimate item parameters a and b. Can you give me the codes?

5- R gives error about eap estimation code

6- Ok. One more thing. I have to save all of the simulated ability, difficulty and discrimination parameters. So the code should give me the opportunity of this.

7- No, not this. You know we simulated 1-0 response data. We generate ability, difficulty and discrimination parameters for this purpose. I need these parameters. Because after estimating model parameters I'll estimate bias and RMSE.

8- can you label the items like 'Item1', 'Item2'.

*** 23 Ekim 2023'te ChatGPT 3.5'te oluşturuldu**

Ek 2. Komut 2

1- Define the parameters that test length is 40, sample size is 2000, mean ability is 0, Sd ability is 1, minimum discrimination is 1 and maximum discrimination is 2, minimum difficulty is -2 and maximum difficulty is 2 and replications is 100 for generate 1-0 item response data for 2 Parametric Logistic Item Response Model. Generate ability parameters, discrimination parameters and item difficulty parameters. Can you give me this simulation codes in R language.

2- This code will generate ability parameters, discrimination parameters, and item difficulty parameters for each replication, calculate the probability of a correct response using the 2PL model for each item, simulate binary responses (0 or 1) based on the probabilities, assign column names to the response matrix, and store the response as a data frame in the list.

3- create a list named 'item discrimination values' to store discrimination values for each iteration in this list. And create a list named 'item difficulty values' to store discrimination values for each iteration in this list. And also create a list named ‘ability values (theta)’ to store discrimination values for each iteration in this list.

4- I want code for added the creation of a list named simulated_parameters_list to store the simulated parameters (ability, discrimination, and difficulty) for each iteration of the loop.

5- I want code which calculates the probability of a correct response using the 2PL model and Simulate binary responses (0 or 1) based on the probability

6- I want from you to give column names to the response matrix store the response matrix directly in the list item_responses for each iteration.

*** 23 Ekim 2023'te ChatGPT 3.5'te oluşturuldu**

Ek 3. Algoritma 1 (A1)

```
library(mirt)
library(eRm)

# Parameters
test_length <- 20
sample_size <- 1000
replications <- 100
mean_ability <- 0
sd_ability <- 1
min_discrimination <- 1
max_discrimination <- 2
min_difficulty <- -2
max_difficulty <- 2
response_list=list()
# Create an empty list to store the simulated parameters for each replication
simulated_parameters_list <- list()
```

```
# Loop through each replication
for (i in 1:replications) {
  # Simulate ability parameters from a normal distribution
  ability <- rnorm(sample_size, mean_ability, sd_ability)

  # Simulate discrimination parameters from a uniform distribution
  discrimination <- runif(test_length, min_discrimination, max_discrimination)

  # Simulate difficulty parameters from a uniform distribution
  difficulty <- runif(test_length, min_difficulty, max_difficulty)

  # Store the simulated parameters in the list
  simulated_parameters_list[[i]] <- list(
    ability = ability,
    discrimination = discrimination,
    difficulty = difficulty
  )
  # Simulate responses using the 2PL model
  prob <- plogis(outer(ability, -difficulty, "*") * discrimination)
  responses <- matrix(rbinom(sample_size * test_length, 1, prob), nrow = sample_size)
  # Extract the response data for the current replication
  response_df <- as.data.frame(responses)
  colnames(response_df) <- paste("Item", 1:test_length)

  # Store the simulated response data in the list
  response_list[[i]] <- response_df
}
```

Ek 4. Algoritma 2 (A2)

```
# Install and load necessary packages
install.packages("rlist")
library(rlist)
library(mirt)

set.seed(123)

# Define the parameters
test_length <- 40
sample_size <- 2000
mean_ability <- 0
sd_ability <- 1
min_discrimination <- 1
max_discrimination <- 2
min_difficulty <- -2
max_difficulty <- 2
replications <- 100

simulated_parameters_list <- list()

# Initialize empty lists to store parameters and ability values
item_discrimination_values <- list()
item_difficulty_values <- list()
theta_values <- list()
item_responses <- list()
```

```
# Generate data for each combination
for (i in 1:replications) {
  # Generate ability parameters
  ability <- rnorm(sample_size, mean = mean_ability, sd = sd_ability)

  # Generate discrimination parameters
  discrimination <- runif(test_length, min = min_discrimination, max = max_discrimination)

  # Generate item difficulty parameters
  difficulty <- runif(test_length, min = min_difficulty, max = max_difficulty)

  item_discrimination_values[[i]] <- discrimination
  item_difficulty_values[[i]] <- difficulty

  # Calculate ability values (theta) for each item
  theta_values[[i]] <- ability

  # Store the simulated parameters in the list
  simulated_parameters_list[[i]] <- list(
    ability = ability,
    discrimination = discrimination,
    difficulty = difficulty
  )

  # Initialize a matrix to store responses for this iteration
  responses <- matrix(0, nrow = sample_size, ncol = test_length)

  for (j in 1:test_length) {
    # Calculate the probability of a correct response using the 2PL model
    p_correct <- 1 / (1 + exp(-discrimination[j] * (ability - difficulty[j])))

    # Simulate binary responses (0 or 1) based on the probabilities
    responses[, j] <- rbinom(sample_size, 1, p_correct)
  }

  # Add column names to the response matrix
  colnames(responses) <- paste("Item", 1:test_length)

  # Store the response matrix for this iteration
  item_responses[[i]] <- as.data.frame(responses)
}
```

Ek 5. Algoritma 3 (A3)

```
library (mirt)
# Create empty lists to store the simulated parameters and datasets for each replication
a=list()
b=list()
ability=list()
datasets=list()
# Simulate parameters and use them to generate responses for the 2PL model
for (i in 1:100) {
    a[[i]] <-as.matrix(round(runif(20, min=1, max=2),2), ncol=1)
    b[[i]] <-as.matrix(round(runif(20, min=-2, max=2),2), ncol=1)
    ability[[i]] <-as.matrix(round(rnorm(1000, mean=0, sd=1),2), ncol=1)
    datasets[[i]] <-simdata(a=a[[i]], d=b[[i]], N=1000, itemtype='dich', Theta=ability[[i]])
    datasets[[i]] <-as.data.frame(datasets[[i]])
```

Trakya Journal of Education

ISSN : 2630-6301
Publisher : Trakya University

Vol. 15, No. 2, 887-918, 2025
DOI: 10.24315/tred.1509299

Research Article

Integration of Artificial Intelligence in Educational Measurement: Efficacy of ChatGPT in Data Generation Within The Scope of Item Response Theory

Hatice Gürdil[1*]
Yeşim Beril Soğuksu[2]
Salih Salihoğlu[3*]
Fatma Coşkun[4]

[1]Turkish Ministry of National Education, Ankara, Türkiye gurdilhatice@gmail.com

[2]Turkish Ministry of National Education, Vali Hilmi Tolun Middle School, Kahramanmaraş,Türkiye berilsoguksu@gmail.com

[3]Department of Industrial and Systems Engineering University of Miami, Florida, United States sxs4331@miami.edu

[4]Measurement and Evaluation in Education, Kahramanmaraş Sütçü İmam University, Kahramanmaraş, Türkiye fatmacoskuncf@gmail.com

[*]Corresponding Author

Received: 02.07.2024
Accepted: 13.08.2024
Available Online: 30.04.2025

Abstract: The aim of this study is to investigate the effectiveness of ChatGPT 3.5 for developing algorithms for data generation within the framework of Item Response Theory (IRT) using the R programming language. In this context, validity analyses were conducted on datasets generated according to the Two-Parameter Logistic Model (2PLM) with algorithms written by ChatGPT 3.5 and researchers. These examinations considered whether the datasets met the IRT assumptions and the simulation conditions of the item parameters. As a result, it was determined that while ChatGPT 3.5 was highly effective in generating data that met the IRT assumptions, it was less effective in meeting the simulation conditions of the item parameters compared to the algorithm developed by the researchers. In this regard, ChatGPT 3.5 is recommended as a useful tool that researchers can use in developing data generation algorithms for IRT.

Keywords: ChatGPT, Item Response Theory, Data Generation, Simulation, R Programming Language

INTRODUCTION

ChatGPT, an artificial intelligence chatbot based on the Large Language Model (LLM), uses OpenAI's Generative Pre-trained Transformer 3.5 (GPT-3.5) language model to generate text in response to user input. The GPT-3.5 model, which is among the most sophisticated language models in the field of natural language processing (NLP), has been trained with over 175 billion parameters from various internet sources such as books, articles, web pages, and social conversations (Uc-Cetina et al., 2023). It can respond to question prompts with a trained dataset that is 570 GB in size (van Dis et al., 2023). Despite the previous use of artificial intelligence tools that transformed research applications (such as Grammarly, rTutor.ai, Research Rabbit, etc.) (Else, 2023), ChatGPT stands out from earlier models with its vast volume of raw data, multitude of parameters used in learning, context-specific architecture, and advanced features like supervised learning utilized in its development (Floridi, 2023; Susnjak, 2022). Additionally, ChatGPT, different from the previous avatars of machine learning in its ability to analyze existing data and generate new data, has become the fastest-growing consumer application in history by reaching 1 million users within the first 5 days after its launch and over 100 million users as of January 2023. It is also one of the most popular LLMs supporting code generation (Hu, 2023; Aljanabi et al., 2023; Feng et al, 2023, Khoury et al., 2023). ChatGPT 3.5 demonstrates superior performance compared to other AI models due to its natural language understanding, extensive dataset, versatile usage areas, user-friendly interface, advanced dialogue capabilities, and continuously updated structure. For these reasons, ChatGPT 3.5 has been chosen for this study and is considered the most suitable solution for researchers.

For all that, despite its rapid development, ChatGPT, which constructs responses from data traces and does not rely on logic or reasoning, is probabilistic and stochastic and is limited by the quality of its training set (Bender et al., 2021; Perrigo, 2023). Moreover, since it lacks the ability to browse the internet, when the data and algorithms it depends on are flawed or biased, the outputs are likely to be flawed or biased as well (Deng & Lin, 2022, McGee, 2023, OpenAI, 2023, Ortiz, 2023a, Ortiz, 2023b, Yang, 2022). Being trained only up to 2021 data, it cannot automatically incorporate real-time information and thus cannot access current knowledge. Additionally, its use by students for direct answers without in-depth reading and critical analysis

Cite as(APA 7): Gürdil H, & Soğuksu Yeşim B., Salihoğlu S., & Coşkun F. (2025). Integration of artificial intelligence in educational measurement: efficacy of ChatGPT in data generation within the scope of item response theory. *Trakya Eğitim Dergisi, 15*(2),903-918. https://doi.org/10.24315/tred.1509299

of the subject matter can suppress critical thinking, problem-solving, and creativity skills necessary in academic, professional, and real-life contexts (O'Connor, 2023), and may lead to academic misconduct (Stokel-Walker, 2022). Another criticism concerns the risk of overreliance on ChatGPT, manipulation through malicious inputs, and the potential for spreading misinformation or propaganda on widely accessible platforms in terms of security (Deng & Lin, 2022).

In response to these criticisms, the New York City Department of Education has banned the use of ChatGPT (Hirsh-Pasek & Blinkoff, 2023), and the International Machine Learning Conference (2023) has limited its use only as part of experimental analysis. Despite all these restrictions, ChatGPT, one of the world's fastest-growing technologies, clearly has the potential to fundamentally change how we access information, learn, and even conduct all kinds of work, suggesting that we are on the cusp of a societal transformation. This shift, due to the possibility that it will endanger a growing number of jobs (Dwivedi et al., 2021) and changes it will cause in individuals' work systems, often encounters user resistance (Laumer et al., 2016).

Despite these negatives and resistance, considering its benefits, ease of accessibility, and widespread use, it is predicted that banning it will not be a solution (Rosenzweig-Ziff, 2023; Springer-Nature, 2023). While acknowledging the potential destructive effects of technology, this destruction is also seen as an opportunity for scientific advancements that can further advance society. Therefore, instead of trying to neutralize digital transformation with bans, it seems more logical to focus on controlling its effects in beneficial ways and integrating them. Given that academia has already started to undergo a digital transition, it appears more appropriate to focus on the positive aspects of ChatGPT technology and to steer its negative aspects towards positive outcomes.

ChatGPT has spread to a wide range of uses, and some of its features include searching for information on various topics, creating stories and reports, writing and correcting computer code, writing and summarizing articles and chapters (Baidoo-Anu & Ansah, 2023; Else, 2023; Kundalia, 2023; Rudolph et al., 2023; Zhai, 2022), conducting tests, processing and explaining data (Hassani & Silva, 2023). These features can positively impact individuals' interactions with computers (Montti, 2022) and provide students with opportunities to enhance their learning, and increase their productivity (Dwivedi et al., 2021). However, unlike humans, ChatGPT, while able to expand its intelligence without cognitive limits, may not possess as specialized intelligence as humans due to its training data being limited to narrow fields or disciplines. In this context, users should not rely on ChatGPT as the final source; instead, they should use it as a tool to discuss and broaden perspectives. This could lead to better outcomes from artificial intelligence through a new type of collaboration that offers human-machine hybrid work opportunities, where humans guide ChatGPT based on their expertise (Mollick, 2022; van Dis et al., 2023). For this, the potential of artificial intelligence in hybrid teams must first be understood. Considering that ChatGPT can produce biased outputs and cannot verify the reality of data, it should be used as support, ensuring that it does not replace the researcher's fundamental tasks such as data analysis, interpretation, and drawing conclusions (Elsevier, 2023).

Considering that ChatGPT generates scientific knowledge from digital traces, it should be noted that it cannot progress alone without human intervention, especially in complex tasks (Biswas, 2023), and that the responsibility lies solely with the user. In this context, users should understand the need to ask the right questions and evaluate the quality of the responses, realizing that as their expertise increases, so will the quality of the information obtained from ChatGPT and their ability to interpret its outputs. Additionally, critically viewing the contributions obtained, validating their accuracy through research from different sources, or making adjustments based on examinations also fall under user responsibility.

It is believed that using ChatGPT under user responsibility will assist users in quickly accessing information and performing mundane, repetitive tasks. Consequently, users can focus on higher-level skills, thereby becoming more productive, and the process itself can become more efficient. Considering that the contributions of ChatGPT in the field of education are also increasing, and that education has been redesigned in line with technology for decades (Baidoo-Anu & Ansah, 2023; Huang, 2019), it is crucial to adapt to this transformation promptly. In this context, it is necessary to demonstrate the potential and limitations of ChatGPT, which is used in many fields. One of these broad application areas is software and programming. ChatGPT assists users in understanding and solving technical problems by providing guidance in complex topics like computer programming, programming languages, algorithms, and data structures. Offering a wide range of capabilities in these areas, ChatGPT can facilitate the processes of designing, creating, developing, testing, and maintaining software, including writing, completing, correcting, predicting, and debugging code. ChatGPT, capable of maintaining logical consistency while answering questions related to programming challenges, has extraordinary features in code generation (Chen et al., 2023; Dong et al., 2023; Liu et al., 2023; OpenAI, 2022; OpenAI, 2023). Thanks to these features of ChatGPT, researchers can improve code quality; save time and effort, thus focusing on more creative work due to reduced cognitive

load, and enhance productivity (Jaber et al., 2023; Biswas, 2023; Chen et al., 2022). ChatGPT, with such positive features, is not limited to natural language and can communicate in more than ten programming and querying languages, including C++, C#, Java, Python, R, etc. (Feng et al., 2023). When examining research in this direction, studies related to software (Adamson & Bägerfeldt, 2023; Biswas, 2023; Jaber et al., 2023), research on ChatGPT's code generation and debugging task, efficiency, and accuracy (Aljanabi, et al., 2023; Bang et al., 2023; Feng et al., 2023; Hansson & Ellréus, 2023; Kashefi & Mukerji, 2023; Sakib et al., 2023; Sobania et al., 2023; Surameery & Shakor; 2023, Tian et al., 2023; White et al., 2023) are available; however, no study examining its effectiveness in data generation has been found.

The rapid advancement of technology is also increasing the number of scientific studies and introducing new methods in various fields. In examining the effectiveness of these new methods, simulation studies are often conducted, which frequently involve comparisons with previous methods. At the same time, experimental studies are needed to demonstrate the functionality of these methods under different conditions, where some conditions are kept constant while others are varied. To conduct these experimental studies, it is first necessary to generate data under the mentioned conditions. Data generation in educational sciences is generally carried out out in accordance with Item Response Theory (IRT). IRT is a statistical modeling framework used to evaluate individuals' abilities and their performance on tests more accurately. The importance of IRT lies in its ability to offer personalized measurement, assist in determining the difficulty levels of test items, conduct detailed item analyses, and enhance the validity and reliability of tests. Additionally, the frequent use of IRT can be attributed to its comprehensiveness, flexibility, high-resolution data analysis capabilities, and the benefits it provides in data generation processes. For these reasons, IRT is a preferred method in educational sciences and psychometric evaluations. IRT focuses on the probability of an individual's performance on an item based on their ability. IRT models, which can be falsified through model-data fit, have strong graphical and mathematical aspects (DeMars, 2010). In recent years, the R programming language, which is among the popular languages, is used in data generation related to IRT. It is known that the R language is supported by ChatGPT 3.5 (Feng et al., 2023). In this context, this study aims to determine how effective ChatGPT 3.5 is in writing code for IRT-based data generation using the R language. This work is expected to make many important contributions to the field of educational measurement and the application of artificial intelligence in data generation, and to demonstrate the potential of using ChatGPT 3.5, a state-of-the-art language model, to develop data generation algorithms for Item Response Theory (IRT). With this innovative approach, it is expected to provide a new perspective on how artificial intelligence can assist researchers in educational and psychological measurements, as well as its ability to simulate educational data. In the context of a rapidly evolving technological environment, this study underlines the importance of integrating artificial intelligence tools such as ChatGPT 3.5 into research methodologies. To evaluate ChatGPT 3.5's effectiveness in this regard, the datasets produced are analyzed to see to what extent they meet the conditions specified in the created simulation design and whether the datasets can indeed be produced as intended. Additionally, the datasets produced with the help of ChatGPT 3.5 algorithms are compared with datasets generated by algorithms developed by researchers. In this context, in this study, the validity of the datasets produced by ChatGPT 3.5 were comprehensively addressed by evaluating them according to basic IRT assumptions such as one-dimensionality, local independence and model-data fit. In this direction, the following research questions have been explored:

For datasets produced according to the Two Parameter Logistic Item Response Theory Model (2PLM) using ChatGPT 3.5 algorithms;

- Do they meet the unidimensionality, the local independence and the model-data fit assumptions with the 2PLM?
- How many item parameters fall outside the ranges specified in the simulation design?
- What are the bias and RMSE values of the item parameters?

METHOD

Research Design

In the study, three different algorithms are used for data generation, and it is investigated to what extent the datasets developed with these algorithms meet the conditions specified in the simulation pattern. Two of the algorithms developed in the R language are produced by ChatGPT 3.5 in October 2023, while the third algorithm is developed by the researchers. Using these developed algorithms, datasets are generated, followed by validity analyses of the generated datasets. In this aspect, the study is experimental in nature (Morris et al., 2019).

Simulation Pattern

In this study, the Monte Carlo simulation used in previous simulation studies (Cohen et al., 1996; Li et al., 2012) is employed to investigate the effects of various altered conditions on ChatGPT 3.5's performance in data generation validity. The Monte Carlo approach allows for the distribution of parameter estimates to be obtained, and reduces the chance of deriving unreasonable results from a single data set. In this context, multiple repetitions (replications) are conducted to resample the true parameters (Harwell et al., 1996). In this study, the recommended number of repetitions for accurate and reliable parameter estimations (Feinberg & Rubright, 2016) is considered, and 100 repetitions are conducted for each data set generated according to the specified simulation conditions. The fixed and altered conditions used in the simulation design are based on logic commonly employed in simulation studies and references in the literature. The 4.3.2 version of the R programming language is used in conducting the study. All the experimental conditions used in the study are presented in Table 1.

Table 1

Simulation Pattern

	Conditions	Levels	Number of Levels
Fixed Conditions	IRT Model	Unidimensional 2PL	1
	Ability (θ)	$-3 \leq \theta \leq 3$	1
	Item Discrimination (a)	$1 \leq a \leq 2$	1
	Item Difficulty (b)	$-2 \leq b \leq 2$	1
Altered Conditions	Item Number (k)	20, 40	2
	Sample Size (n)	500, 2000	2
Total Number of Conditions			2x2=4
Number of Repetitions			100
Total Datasets			4x100=400

When examining Table 1, it is observed that different conditions have been considered for the production of unidimensional data for items scored dichotomously (0-1) within the scope of the study. The MTK model uses 2PLM due to its ability to separately address item difficulty and discrimination parameters for binary scored items (DeMars, 2010). The individual ability parameter is typically drawn from a normal distribution with a mean of 0 and a standard deviation of 1 (Feinberg & Rubright, 2016). This approach is commonly used in simulation studies to achieve accurate parameter estimates. Item discrimination parameters are determined from a uniform distribution with a minimum of 1 and a maximum of 2, which is considered more appropriate for 2PLM estimates according to the literature (Dekker, 2004; Hambleton et al., 1991). Item difficulty parameters are also determined from a uniform distribution with a minimum of -2 and a maximum of 2, deemed suitable for 2PLM estimates as per the literature (DeMars, 2010). Monte Carlo studies indicate that a minimum of 20 items is required for accurate item parameter estimates (De Ayala, 2013). Consequently, 20 items are selected for short tests and 40 items for long tests. Additionally, sample size is

known to enhance the accuracy of parameter estimates in simulation studies. Therefore, sample sizes of 500 for small samples and 2000 for large samples are chosen (Stone, 1992).
The first and second algorithms (A1 and A2) are developed by ChatGPT 3.5 through prompts for data generation. Different prompts are used to request ChatGPT 3.5 to produce new algorithms; however, when it is determined that the generated algorithms do not differ from A1 and A2, the algorithm development process with ChatGPT 3.5 is terminated. The prompts used for A1 and A2 are given in Appendix 1 and 2. The third algorithm (A3) used for data generation is developed by researchers using the *'mirt'* package (Chalmers, 2012) in the R language. The codes for the three developed algorithms are located in Appendices 3, 4, and 5. During the data generation phase, the developed algorithms are run in the R environment, and datasets with 100 repetitions are produced considering the conditions in the simulation pattern. As a result, a total of 2x2x100=400 different datasets are produced, comprising 2 different test lengths (20, 40), 2 different sample sizes (500, 2000), and 100 repetitions.

Data Analysis

Within the scope of the research, datasets generated sequentially with A1, A2, and A3 are expected to meet the assumptions (unidimensionality, local independence, model-data fit) specific to the unidimensional 2PLM, as they are generated according to this model. Otherwise, the generated data will not conform to the intended data production scenario, and consequently, the results obtained will be erroneous. Furthermore, the parameters estimated in the generated datasets must meet the conditions specified in the simulation pattern to demonstrate the accuracy of the data generations. Therefore, the first step is to check whether the produced datasets comply with the IRT assumptions and whether the parameters meet the conditions specified in the simulation pattern, and for this purpose, the following examinations are conducted.

Unidimensionality Examination of Datasets

To examine the unidimensionality assumption, Exploratory Factor Analysis (EFA) is conducted using the *'psych'* (Revelle, 2020) and *'sirt'* (Robitzsch, 2019) packages available in the R programming language. In this context, *'tetrachoric correlation matrices'* suitable for dichotomously scored (1-0) datasets are created for all datasets. Within the scope of EFA assumptions, the results of the Kaiser-Meyer-Olkin (KMO) and Bartlett tests are examined. For both A3 and the datasets produced with the help of ChatGPT 3.5 (A1 and A2) algorithms, it is found that the KMO test results are greater than 0.87 and the Bartlett tests are significant ($p<.05$) for each condition. After meeting the assumption of unidimensionality, the number of factors is determined using Horn 's *'Parallel Analysis'* method (Horn, 1965). The *'fa.parallel'* function from the *'psych'* package is utilized for the factor extraction process; eigenvalues for the factors and scree plot graphs are examined. In deciding on the number of factors, factor loadings, explained variance ratios, and scree plots are taken into account.

Local Independence Examination of Datasets

The Item Response Theory (IRT) local independence assumption, which states that item responses should be independent of each other when individuals' abilities are controlled for (DeMars, 2010), has been checked using Yen's Q3 test (Yen, 1984). To verify the local independence assumption, the *'Yen.Q3'* function from the *'subscore'* package (Dai et al., 2022) in the R language is utilized. The condition set by (Yen, 1984) for local independence issues, where correlations between residuals for problematic items exceeding 0.20, is taken as the criterion for identifying violations.

Model-Data Fit Examination of Datasets

In checking the compatibility of the generated datasets with the 2PLM, the significance levels of the M2 statistic are examined. If the M2 statistic, constructed using the residuals between observed and expected marginal probabilities, is not significant, it is assumed that the data set is consistent with the model (Maydeu-Olivares & Joe, 2006). The *'M2'* function from the *'mirt'* package (Chalmers, 2012) in the R programming language is used to check the significance of the M2 statistic. Additionally, in assessing model-data fit, the nested IRT models' (1PL, 2PL, 3PL) *'-2 log-likelihood'* values and significances was taken into account

(Thissen & Steinberg, 1986). The *'anova'* function in the R programming language is utilized for comparing the models.

How Many Item Parameters Fall Within the Specified Ranges

It is determined how many of the item discrimination (a) and item difficulty (b) parameters in the generated datasets meet the conditions specified in the simulation pattern. In this context, the researchers create counters in the R programming language to identify the number of parameters that fall outside the specified ranges. Considering the altered item numbers and number of repetitions in the research, it is established how many item difficulty and item discrimination parameters fall outside the specified ranges, with 20 items for 20x100=2000 and 40 items for 40x100=4000.

Bias and Error (RMSE) Values of Item Parameters

In assessing the bias and error (RMSE) values of the item parameters for datasets generated considering the conditions specified in the simulation pattern, the formulas located in Equations 1 and 2 are used.

$$Bias = \frac{\sum_{i=1}^{K}\left(\hat{X}_{i-}X_i\right)}{K} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{K}\left(\hat{X}_{i-}X_i\right)^2}{K}} \quad (2)$$

In Equations 1 and 2, $\hat{X}_i$ represents the parameter for item i $(i = 1, 2, \ldots, K)$, X_i denotes the actual parameter estimate, and K indicates the number of items. Bias and RMSE values are calculated for each condition in the datasets with 100 repetitions, and then their averages are taken.

RESULTS

This section presents the findings obtained within the scope of the research. In the study, data generation for the 2PLM is carried out using three different algorithms (A1, A2, and A3) based on the conditions specified in the simulation pattern. Examinations are conducted to determine the extent to which the generated datasets meet the specified simulation conditions. The results of these examinations are shown in Table 2.

Table 2. Results of the Analyses on the Datasets

Condition		Number of Items	Sample Size	Number of Factors	Parallel Analysis (Number of Factors) Local Independence	M2 Fit Violation	Nested Model Fit	Out of Range Parameter a	Out of Range Parameter b	Average Bais a	Average Bais b	Average RMSE a	Average RMSE b
1	A1	20	500	1	1 (4 x 2) 13	0 (%0)*	2PL	1549 (%77)**	0 (%0)***	-1.201	-0.093	2.108	1.145
2	A2	20	500	1	1 -	5 (%5)	2PL	241 (%12)	632 (%32)	0.012	0.206	0.21	1.936
3	A3	20	500	1	1 -	4	2PL	222	86		0.04	0.189	0.137

							(%4)		(%11)	(%4)	0.018			
4	A1	20	2000	1	1 (4 x 2)	2	4 (%4)	2PL	1580 (%79)	0 (%0)	-1.227	0.397	2.091	1.037
5	A2	20	2000	1	1	-	4 (%4)	2PL	78 (%3)	661 (%33)	0	-0.407	0.103	2.042
6	A3	20	2000	1	1	-	4 (%4)	2PL	84 (%4)	45 (%2)	0.007	-0.288	0.093	0.057
7	A1	40	500	1	1	9	4 (%4)	2PL	3129 (%78)	0 (%0)	-1.297	-0.013	2.142	1.215
8	A2	40	500	1	1	-	4 (%4)	2PL	495 (%12)	1280 (%32)	0.019	-0.073	0.198	2.093
9	A3	40	500	1	1	-	7 (%7)	2PL	427 (%11)	173 (%4)	0.013	0.012	0.183	0.159
10	A1	40	2000	1	1	-	6 (%6)	2PL	3181 (%79)	0 (%0)	-1.271	0.421	2.16	1.155
11	A2	40	2000	1	1	-	2 (%2)	2PL	135 (%3)	1248 (%31)	0.003	-0.186	0.095	2.133
12	A3	40	2000	1	1	-	7 (%7)	2PL	144 (%4)	79 (%2)	0.002	0.086	0.09	0.077

*Percentage of fit violation ** Percentage of a parameter outside the range *** Percentage of a parameter outside the range

As can be seen in Table 2, it is found that in all conditions where the number of items is 20, the EFA analysis indicates the presence of a dominant single dimension in the datasets with 100 repetitions. In the parallel analysis, only in A1, two conditions (20 items with 500 samples, 20 items with 2000 samples) reveal a two-factor structure for 4 datasets. In all conditions where the number of items is 40, both EFA and Parallel Analysis results support a single-factor structure. Consequently, it can be said that the datasets generated with ChatGPT 3.5 algorithms are unidimensional as expected, with a few negligible exceptions.

Upon examining Table 2, it is observed that violations of the local independence assumption are seen only in some cases with the ChatGPT 3.5 A1 algorithm. In this algorithm, violations of local independence in item pairs are identified in 13 out of 100 datasets for 20 items with 500 samples, in 2 datasets for 20 items with 2000 samples, and in 9 datasets for 40 items with 500 samples. At this point, considering the absence of datasets with local independence violations, it can be said that the ChatGPT 3.5 A2 algorithm and A3 are more successful.

It is observed that, according to the significance checks of the M2 statistic, the proportion of datasets that do not meet the model-data fit varies between 0% and 7% for all algorithms upon examining Table 2. In conditions with 20 items and a sample size of 500, the best model-data fit is achieved with A1, with no fit violation observed (0%). When the total number of items is 40 and the sample size is 2000, it is determined that the least model-data fit violation (2%) is with A2, while the most violations (7%) occur with A3. Additionally, considering the nested model fit, it is determined that in all conditions, the datasets are better fitted with the 2PL model. Generally, it can be said that ChatGPT 3.5 algorithms (A1 and A2) are at least as successful as the researcher-developed A3 in terms of the compatibility of the datasets with the 2PL model.

Furthermore, it is evident that the data generation for the parameter a, fitting the ranges specified in the simulation pattern, is successfully conducted with ChatGPT 3.5 A2 and the researcher-developed A3 across all conditions. ChatGPT 3.5 A1, however, appears quite weak in generating data within the specified ranges for the parameter a, with a majority of the produced a parameters (77%-79%) falling outside the determined ranges. In all conditions with a sample size of 500, A3 shows the least percentage of data generation outside the specified ranges for the parameter a, while in conditions with a sample size of 2000, A2 has the least percentage of deviation. For the parameter b, the best data generation within the specified ranges is achieved with A1, followed closely by A3. A2, however, is weak in generating data within the specified ranges for parameter b, with a portion of the b parameters (31%-33%) falling outside the specified ranges. In conclusion, although successful in meeting the IRT assumptions, it can be said that the item parameter estimations in the datasets generated with ChatGPT 3.5 algorithms have more difficulties in meeting the conditions specified in the simulation pattern compared to A3.

As for bias average, it is observed that for the parameter a across all conditions, the lowest bias average is found in A1, while A2 and A3 produce similar values. In the condition with a total item count of 20 and a sample size of 500, the lowest bias average for parameter b is in A1, while in all other conditions, it is in A2. In conditions with a sample size of 2000, the highest bias average is obtained with A1. For conditions with a sample size of 500, the

highest bias average for a total item count of 20 is with A2, and for 40, it is with A3. Additionally, the highest RMSE average values across all conditions are determined to be for parameter a in A1 and for parameter b in A2. Furthermore, across all conditions, the lowest RMSE average values for both parameters a and b are found in A3. Therefore, it can be said that the best results in terms of bias and RMSE averages for item parameters are obtained in datasets produced with A3, while ChatGPT 3.5 algorithms lead to higher bias and RMSE averages.

DISCUSSION AND CONCLUSION

When the findings are evaluated, it can be stated that ChatGPT 3.5 algorithms are generally successful in generating data that conforms to IRT assumptions. Excluding a few negligible exceptions, unidimensional datasets with no local independence violations and compatible with the 2PLM are generated using ChatGPT 3.5 algorithms. However, when evaluating in terms of the compliance of item parameters with the ranges specified in the simulation pattern and the associated bias and RMSE averages, it appears that ChatGPT 3.5 algorithms have more issues compared to the algorithm developed by researchers.

Based on the results in Table 2, it is found that in all conditions, EFA analysis of the datasets generated with the three algorithms determines a unidimensional structure, as specified in the simulation pattern. In the parallel analysis, except for two conditions in A1, which account for a total of eight datasets, a unidimensional structure is observed in all conditions. The local independence checks show that the assumption is met in all conditions except for three in A1. M2 fit violations in all three algorithms reveal the presence of violating datasets, but these violations are relatively few, varying between 0% and 7%. Local independence violations in ChatGPT 3.5 algorithms can stem from various reasons. Suboptimal prompt design may increase dependency between items. The limitations of ChatGPT 3.5 itself may also contribute to these violations. To reduce such violations, more advanced models (e.g., ChatGPT 4) can be employed. Using prompts prepared and guided by experts can help minimize dependency between items. Additionally, analyses should be conducted considering the limitations of ChatGPT 3.5, and human expertise should be sought when necessary. Generally, the datasets that best fit the 2PL are those produced with A1 for 20 items with 500 samples, and those produced with A2 for 40 items with 2000 samples, with no clear pattern of fit violations among the algorithms. When nested models are compared for fit with 2PL, datasets in all algorithms are found to align well with 2PL. In the examination of item parameters a and b falling outside the specified ranges, it is seen that the closest results to the specified ranges are obtained with A3. The A1 algorithm is found to be inadequate in producing a parameters within the desired ranges, and A2 is inadequate for b parameters.

The bias and RMSE averages for these parameters are found to support this result as expected.

In conclusion, data generation using ChatGPT 3.5 algorithms is generally successful, but the best compliance with the conditions of the simulation pattern is achieved with the algorithm developed by researchers (A3). Datasets generated with the support of ChatGPT 3.5 algorithms are found to be less adequate than those generated by researchers without ChatGPT 3.5 support, especially in terms of generating unbiased parameters facilitate the expert's work within the desired ranges. Similar to literature findings, the findings reveal that ChatGPT can successfully comply with IRT assumptions and underline the indispensable role of human expertise in driving AI outcomes.

Similar to literature findings, ChatGPT 3.5 was found to be not perfect at generating computer code and ChatGPT 3.5's solutions were found to be of lower quality than human solutions (Adamson & Bägerfeldt, 2023; Kundalia, 2023). However, there are also studies stating that ChatGPT 3.5 can produce correct solutions to many problems but falls short in some aspects, similar to this study (Hansson & Ellréus, 2023; Sakib et al., 2023; Tian et al., 2023). Additionally, it is found that there are problems in meeting unidimensionality and local independence assumptions in one of the data algorithms produced with the support of ChatGPT 3.5 (A1), while these problems are not observed in the other (A2). This may be attributed to the fact that the two algorithms are developed by different researchers using different prompts. As a result of this study, it can be said that ChatGPT 3.5 can produce data close to an expert level, as stated in some studies in the literature (Surameery & Shakor, 2023) and the accuracy of the generated data set will increase in line with the expertise. As mentioned in the studies (Bang et al., 2023), the success of algorithms developed with ChatGPT 3.5 is influenced by the prompts and guidance provided by experts. It can be stated that as expertise increases, the quality of prompts entered into ChatGPT 3.5 also increases, thus bringing the output closer to expert accuracy. As seen, contrary to fears, ChatGPT 3.5 does not eliminate the need for expertise; rather, it serves as a tool to facilitate the expert's work. In addition, it should be taken into consideration that ChatGPT has certain capabilities and limitations in each version. In this study, the adequacy of ChatGPT 3.5, which is free of charge and accessible to more people, in developing algorithms for data generation within the scope of IRT was examined. In the validity checks carried out within the scope of the study, it was observed that in the datasets produced with ChatGPT 3.5 algorithms, the item parameters were weak

in meeting the ranges specified in the simulation pattern and local independence violations were more common. ChatGPT 4 version or in the future versions, much better results can be obtained than those obtained in the research. Because it is possible that the nature of the version used may have a direct impact on the reliability and accuracy of the results obtained in the research. It should be taken into account that more recent versions may tend to produce more accurate results as they are trained on a larger training data set.

In a world where technology is increasingly advancing and playing a major role, banning technology is clearly not meaningful (King, 2023). In line with all these results, it seems more appropriate to use ChatGPT 3.5 as a support tool in research, as suggested in the studies (Biswas, 2023; Deng & Lin, 2022, Dwivedi et al., 2023; Hansson & Ellréus, 2023; Sollie, 2009, Surameery & Shakor, 2023), rather than restricting its use. The results of this study reveal a balanced approach in which artificial intelligence is used as a supporting tool to enhance human capabilities rather than replace them. It is envisioned that this integration could lead to more efficient research processes, allowing researchers to focus on higher-level analytical tasks while relying on AI for data generation and preliminary analysis. In this direction, considering the importance of expertise in the use of ChatGPT 3.5, it is advisable for researchers to benefit from ChatGPT 3.5 in data generation and shorten the time they spend on data generation. Researchers can use the remaining time for more creative parts of their research, as suggested in the studies (Hirsh-Pasek & Blinkoff, 2023). The situation of different outputs being produced depending on the commands of different experts can be examined in a different study that takes into account inter-expert variations. Additionally, it is recommended to conduct further in-depth studies to examine the fact that ChatGPT 3.5 generally meets IRT assumptions but does not fully achieve item recovery. In this context, it is important to explore different AI models for future studies; not limited to ChatGPT alone, but also including advanced models such as ChatGPT 4, BERT, and T5. By comparing the performances of these models, the most suitable one can be identified. Improving prompt design is crucial to enhance the quality of responses provided by ChatGPT; clearer and more directive prompts can be prepared under expert guidance. Additionally, re-training the model using larger and more diverse datasets can expand the model's knowledge base, resulting in more accurate and reliable responses. To minimize local independence violations, special strategies can be developed considering the model's limitations, and the impact of dependencies can be reduced through manual checks.

REFERENCES

Adamson, V., & Bägerfeldt, J. (2023). *Assessing the effectiveness of ChatGPT in generating Python code* [Bachelor Degree Project, University of Skövde]. Bachelor's Degree Project in Information Technology.

Aljanabi, M., Ghazi, M., Ali, A. H., Abed, S. A. & ChatGpt (2023). ChatGpt: open possibilities. *Iraqi Journal For Computer Science and Mathematics*, *4*(1), 62-64. https://doi.org/10.52866/ijcsm.2023.01.01.0018

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*(*1*), 52-62.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR, abs/2302.04023, 2023.*

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March 3-10). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Canada (pp. 610–623). New York, NY: ACM. https://doi.org/10.1145/3442188.3445922

Biswas, S. (2023). Role of ChatGPT in Computer Programming.: ChatGPT in Computer Programming. *Mesopotamian Journal of Computer Science*, *2023*, 8-16. https://doi.org/10.58496/MJCSC/2023/002

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, B., Zhang, F., Nguyen, A., Zan, D., Lin, Z., Lou, J. G., & Chen, W. (2022). Code T: Code generation with generated tests. *arXiv preprint arXiv:2207.10397*.

Chen, E., Huang, R., Chen, H. S., Tseng, Y. H., & Li, L. Y. (2023). GPTutor: a ChatGPT-powered programming tool for code explanation. *arXiv preprint arXiv:2305.01863.*

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*(1), 15-26.

Dai, S., Wang, X., & Svetina, D. (2022). subscore: Computing Subscores in Classical Test Theory and Item Response Theory. R package version 3.3, <https://CRAN.R-project.org/package=subscore>.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Dekker, M. (2004). *Parameter estimation techniques*. In Baker Kim, (Eds.), *Item response theory*, New York.

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

Deng, J., & Lin, Y. (2022). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems*, *2*(2), 81-83. https://doi.org/10.54097/fcis.v2i2.4465

Dong, Y., Jiang, X., Jin, Z., & Li, G. (2023). *Self-collaboration Code Generation via ChatGPT. arXiv preprint arXiv:2304.07590.* https://doi.org/10.48550/arXiv.2304.07590

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, Article 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, Article 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Else, H. (2023). Abstracts written by ChatGPT fool scientists, *Nature, 613*(7944), 423-423. https://doi.org/10.1038/d41586-023-00056-7.

Elsevier, (2023). The Use of AI and AI-assisted Technologies in Scientific Writing. 〈https://www.elsevier.com/about/policies/publishing-ethics〉 (accessed 20th Feb, 2023)

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

Feng, Y., Vanam, S., Cherukupally, M., Zheng, W., Qiu, M., & Chen, H. (2023). Investigating Code Generation Performance of Chat-GPT with Crowdsourcing Social Data. In *Proceedings of the 47th IEEE Computer Software and Applications Conference* (pp. 1-10).

Floridi, L. (2023). AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philosophy & Technology*, 36(15), 1-7. https://doi.org/10.1007/s13347-023-00621-y

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hansson, E., & Ellréus, O. (2023). *Code Correctness and Quality in the Era of AI Code Generation: Examining ChatGPT and GitHub Copilot* [Bachelor Degree Project, University of Skövde]. Digitala Vetenskapliga Arkivet.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement, 20*(2), 101-125.

Hassani, H., & Silva, E. S. (2023). The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field. *Big data and cognitive computing, 7*(*2*), 62. https://doi.org/10.3390/bdcc7020062

Hirsh-Pasek, K. & Blinkoff, E. (2023). "ChatGPT: Educational friend of foe?", Brookings Institution, https://www.brookings.edu/blog/education-plusdevelopment/2023/01/09/chatgpt-educational-friend-or-foe

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Hu, K. (2023) ChatGPT sets record for fastest-growing user base - analyst note,February 2, 〈 https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-userbase-analyst-note-2023–02-01/>

Huang, R. (2019). *Educational technology a primer for the 21st century*. Springer Nature Singapore Pte Ltd.

Jaber, M. A., Beganovic, A., & Abd Almisreb, A. (2023). Methods and Applications of ChatGPT in Software Development: A Literature Review. *Southeast Europe Journal of Soft Computing, 12*(1), 08-12. http://dx.doi.org/10.21533/scjournal.v12i1.251

Kashefi, A., & Mukerji, T. (2023). Chatgpt for programming numerical methods. *Apr. 2023, arXiv:2303.12093*.

Khoury, R., Avila, A. R., Brunelle, J. and Camara, B. M. (2023) "How Secure is Code Generated by ChatGPT?" *Apr. 2023, arXiv:2304.09655.*

King, M. R. (2023). Outsourcing Your Faculty Application to ChatGPT: Would this Work? Should this Work?. *Cellular and Molecular Bioengineering*, *16*(4), 423-426. https://doi.org/10.1007/s12195-023-00777-9

Kundalia, N.D. (2023). ChatGPT and the future of writing. Hindustan Times. Retrieved January 31, 2023, from 〈https://www.hindustantimes.com/books/ chatgpt-and-the-future-of-writing-101675090609362.html〉

Laumer, S., Maier, C., Eckhardt, A., & Weitzel, T. (2016). Work routines as an object of resistance during information systems implementations: Theoretical foundation and empirical evidence. *European Journal of Information Systems, 25*(4), 317–343. https://doi.org/10.1057/ejis.2016.1

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.

Liu, J., Xia, C. S., Wang, Y., & Zhang, L. (2023). Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210*. https://doi.org/10.48550/arXiv.2305.01210

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713-732. https://doi.org/10.1007/s11336-005-1295-9

McGee, R. W. (2023). What Will the United States Look Like in 2050? A ChatGPT Short Story. Working Paper, April 8, 2023. https://ssrn.com/abstract=4413442

Mollick, E. (2022). ChatGPT Is a Tipping Point for AI. Harvard Business Review. December 14.

Montti, R. (2022). What is ChatGPT and how can you use it? Search Engine Journal (Accessed from) 〈 https://www.searchenginejournal. com/what-is-chatgpt/473664/#close〉

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, *38*(11), 2074-2102.

O'Connor, S. (2023). Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse. *Nurse Education in Practice, 66*, Article 103537. https://doi.org/10.1016/j.nepr.2022.103537

OpenAI, (2023) ChatGPT: Optimizing Language Models for Dialogue. Available at: 〈 https://openai.com/blog/chatgpt/〉

OpenAI, T. B. (2022). Chatgpt: Optimizing language models for dialogue. OpenAI. https://openai.com/blog/chatgpt/ https://openai.com/blog/chatgpt/

Ortiz, S. (2023a). ChatGPT is changing everything. But it still has its limits. ZDNet (Available online) 〈 https://www.zdnet.com/article/chatgpt-is-changing-everythin g-but-it-still-has-its-limits/〉

Ortiz, S. (2023b). What is ChatGPT and why does it matter? Here's everything you need to know. ZD Netto Innovation (Accessed from) 〈https://www.zdnet.com/article /what-is-chatgpt-and-why-does-it-matter-heres-everything-you-need-to-know/〉

Perrigo, B. (2023). OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. Time. 〈https://time.com/6247678/openai-chatgpt-kenyaworkers

Revelle, W. (2020). Psych: Procedures for psychological, psychometric, and personality research. Evanston, Illinois: Northwestern University. https://CRAN.R-project.org/package=psych.

Robitzsch, A. (2019). sirt: Supplementary item response theory models. R package version 3.1-80. https://CRAN.R-project.org/package=sirt

Rosenzweig-Ziff, D. (2023). New York City blocks use of the ChatGPT bot in its schools. The Washington Post. 〈https://www.washingtonpost.com/education/2023/01/05/n yc-schools-ban-chatgpt/

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching, 6*(*1*). https://doi.org/10.37074/jalt.2023.6.1.9

Sakib, F. A., Khan, S. H., & Karim, A. H. M. (2023). Extending the frontier of chatgpt: Code generation and debugging. *arXiv preprint arXiv:2307.08260*. https://doi.org/10.48550/arXiv.2307.08260

Sobania, D., Briesch, M., Hanna, C., & Petke, J. (2023). An analysis of the automatic bug fixing performance of chatgpt. *arXiv preprint arXiv:2301.08653*. https://doi.org/10.48550/arXiv.2301.08653

Sollie, P. (2009). On Uncertainty in Ethics and Technology. In P. Sollie, & M. Düwell (Eds.), *Evaluating New Technologies: Methodological Problems for the Ethical Assessment of Technology Developments, The International Library of Ethics, Law and Technology* (pp. 141–158). Springer.

Stokel-Walker, C. (2022). *AI bot ChatGPT writes smart essays - should professors worry?* Nature (London). https://doi.org/10.1038/d41586-022-04397-7

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16(*1), 1-16.

Surameery, N. M. S., & Shakor, M. Y. (2023). Use chat gpt to solve programming bugs. *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290*, *3*(01), 17-22. https://doi.org/10.55529/ijitc.31.17.22

Susnjak, T. (2022). ChatGPT: The End of Online Exam Integrity? *arXiv preprint arXiv: 2212.09292*. https://doi.org/10.48550/arXiv.2212.09292

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567-577.

Tian, H., Lu, W., Li, T. O., Tang, X., Cheung, S. C., Klein, J., & Bissyandé, T. F. (2023). Is ChatGPT the Ultimate Programming Assistant--How far is it?. *arXiv preprint arXiv:2304.11938*. https://doi.org/10.48550/arXiv.2304.11938

Uc-Cetina, V., Navarro-Guerrero, N., Martin-Gonzalez, A., Weber, C., & Wermter, S. (2023). Survey on reinforcement learning for language processing. *Artificial Intelligence Review*, *56*(2), 1543-1575.

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. (2023). ChatGPT: five priorities for research. *Nature, 614*, 224–226. https://doi.org/10.1038/d41586- 023-00288-7

White, J., Hays, S., Fu, Q., Spencer-Smith, J., & Schmidt, D. C. (2023). Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. *arXiv preprint arXiv:2303.07839*. https://doi.org/10.48550/arXiv.2303.07839

Yang, S. (2022). The Abilities and Limitations of ChatGPT. Anaconda Perspectives. 〈https ://www.anaconda.com/blog/the-abilities-and-limitations-of-chatgpt〉

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145.

Zhai, X., (2022). ChatGPT User experience: Implications for education. (December 27, 2022). Available at SSRN: https://ssrn.com/abstract=4312418 or http://dx.doi.org/10.2139/ssrn.4312418.

APPENDIX

Appendix 1. Prompt 1

1- I want to generate 1-0 response data for 2 Parametric Logistic Item Response Model. My test length is 20. My sample size is 500. My ability parameters have a normal distrubition with mean 0, standard deviation 1. Item discrimination parameter will have a uniform distrubition and between 1 and 2. My difficulty parameter (b) will have a uniform distrubition and between -2 and 2.

2- My replication number is 100. I want to save 1-0 responses to an empty list. Can you give me the codes of this simulation in R language?

3- Ok, but this code does not label the columns

4- Ok. I want to estimate 2PL model parameters. First I want to estimate ability parameters via EAP method. Then I want to estimate item parameters a and b. Can you give me the codes?

5- R gives error about eap estimation code

6- Ok. One more thing. I have to save all of the simulated ability, difficulty and discrimination parameters. So the code should give me the opportunity of this.

7- No, not this. You know we simulated 1-0 response data. We generate ability, difficulty and discrimination parameters for this purpose. I need these parameters. Because after estimating model parameters I'll estimate bias and RMSE.

8- can you label the items like 'Item1', 'Item2'.

*** Created on October 23, 2023 in ChatGPT 3.5**

Appendix 2. Prompt 2

1- Define the parameters that test length is 40, sample size is 2000, mean ability is 0, Sd ability is 1, minimum discrimination is 1 and maximum discrimination is 2, minimum difficulty is -2 and maximum difficulty is 2 and replications is 100 for generate 1-0 item response data for 2 Parametric Logistic Item Response Model. Generate ability parameters, discrimination parameters and item difficulty parameters. Can you give me this simulation codes in R language.

2- This code will generate ability parameters, discrimination parameters, and item difficulty parameters for each replication, calculate the probability of a correct response using the 2PL model for each item, simulate binary responses (0 or 1) based on the probabilities, assign column names to the response matrix, and store the response as a data frame in the list.

3- create a list named 'item discrimination values' to store discrimination values for each iteration in this list. And create a list named 'item difficulty values' to store discrimination values for each iteration in this list. And also create a list named ‘ability values (theta)’ to store discrimination values for each iteration in this list

4- I want code for added the creation of a list named simulated_parameters_list to store the simulated parameters (ability, discrimination, and difficulty) for each iteration of the loop.

5- I want code which calculates the probability of a correct response using the 2PL model and Simulate binary responses (0 or 1) based on the probability

6- I want from you to give column names to the response matrix store the response matrix directly in the list item_responses for each iteration.

*** Created on October 23, 2023 in ChatGPT 3.5**

Appendix 3. Algorithm 1 (A1)

```
library(mirt)
library(eRm)

# Parameters
test_length <- 20
sample_size <- 1000
replications <- 100
mean_ability <- 0
sd_ability <- 1
min_discrimination <- 1
max_discrimination <- 2
min_difficulty <- -2
max_difficulty <- 2
response_list=list()
# Create an empty list to store the simulated parameters for each replication
simulated_parameters_list <- list()

# Loop through each replication
for (i in 1:replications) {
  # Simulate ability parameters from a normal distribution
  ability <- rnorm(sample_size, mean_ability, sd_ability)
```

```
  # Simulate discrimination parameters from a uniform distribution
  discrimination <- runif(test_length, min_discrimination, max_discrimination)

  # Simulate difficulty parameters from a uniform distribution
  difficulty <- runif(test_length, min_difficulty, max_difficulty)

  # Store the simulated parameters in the list
  simulated_parameters_list[[i]] <- list(
    ability = ability,
    discrimination = discrimination,
    difficulty = difficulty
  )
  # Simulate responses using the 2PL model
  prob <- plogis(outer(ability, -difficulty, "*") * discrimination)
  responses <- matrix(rbinom(sample_size * test_length, 1, prob), nrow = sample_size)
  # Extract the response data for the current replication
  response_df <- as.data.frame(responses)
  colnames(response_df) <- paste("Item", 1:test_length)

  # Store the simulated response data in the list
  response_list[[i]] <- response_df
}
```

Appendix 4. Algorithm 2 (A2)

```
# Install and load necessary packages
install.packages("rlist")
library(rlist)
library(mirt)

set.seed(123)

# Define the parameters
test_length <- 40
sample_size <- 2000
mean_ability <- 0
sd_ability <- 1
min_discrimination <- 1
max_discrimination <- 2
min_difficulty <- -2
max_difficulty <- 2
replications <- 100

simulated_parameters_list <- list()

# Initialize empty lists to store parameters and ability values
item_discrimination_values <- list()
item_difficulty_values <- list()
theta_values <- list()
item_responses <- list()

# Generate data for each combination
for (i in 1:replications) {
  # Generate ability parameters
  ability <- rnorm(sample_size, mean = mean_ability, sd = sd_ability)
```

```
  # Generate discrimination parameters
  discrimination <- runif(test_length, min = min_discrimination, max = max_discrimination)

  # Generate item difficulty parameters
  difficulty <- runif(test_length, min = min_difficulty, max = max_difficulty)

  item_discrimination_values[[i]] <- discrimination
  item_difficulty_values[[i]] <- difficulty

  # Calculate ability values (theta) for each item
  theta_values[[i]] <- ability

  # Store the simulated parameters in the list
  simulated_parameters_list[[i]] <- list(
    ability = ability,
    discrimination = discrimination,
    difficulty = difficulty
  )

  # Initialize a matrix to store responses for this iteration
  responses <- matrix(0, nrow = sample_size, ncol = test_length)

  for (j in 1:test_length) {
    # Calculate the probability of a correct response using the 2PL model
    p_correct <- 1 / (1 + exp(-discrimination[j] * (ability - difficulty[j])))

    # Simulate binary responses (0 or 1) based on the probabilities
    responses[, j] <- rbinom(sample_size, 1, p_correct)
  }

  # Add column names to the response matrix
  colnames(responses) <- paste("Item", 1:test_length)

  # Store the response matrix for this iteration
  item_responses[[i]] <- as.data.frame(responses)
}
```

Appendix 5. Algorithm 3 (A3)

```
library (mirt)
# Create empty lists to store the simulated parameters and datasets for each replication
a=list()
b=list()
ability=list()
datasets=list()
# Simulate parameters and use them to generate responses for the 2PL model
for (i in 1:100) {
    a[[i]] <-as.matrix(round(runif(20, min=1, max=2),2), ncol=1)
    b[[i]] <-as.matrix(round(runif(20, min=-2, max=2),2), ncol=1)
    ability[[i]] <-as.matrix(round(rnorm(1000, mean=0, sd=1),2), ncol=1)
    datasets[[i]] <-simdata(a=a[[i]], d=b[[i]], N=1000, itemtype='dich', Theta=ability[[i]])
    datasets[[i]] <-as.data.frame(datasets[[i]])
```