DİJİTED
DİJİTAL TEKNOLOJİLER VE EĞİTİM DERGİSİ

Dijital Teknolojiler ve Eğitim Dergisi, Yıl: 2024, Cilt: 1, Sayı: 1, Sayfa : 1-12
Journal of Digital Technologies and Education, Year: 2024, Volume: 3, Issue: 1, Pages : 1-12
Araştırma Makalesi

Gönderim Tarihi : 01.06.2024
Kabul Tarihi : 29.06.2024

DOI: 10.5281/zenodo.12637347

Gülgün Bulut[1]
Murat Akyıldız[1]

Yapay Zekâ ile Üretilen Soruların ve Madde Parametrelerinin MST Test Koşullarında Karşılaştırılması

Özet

Akıllı teknolojilerin hızla yaygınlaşmasının eğitim dünyasındaki önemli yansımalarından biri de ölçme ve değerlendirme alanında gerçekleşmektedir. Doğası gereği ölçmenin gerçeğe yakın doğrulukta sonuçlar verebilmesi için öncelikle güvenli bir ortamda gerçekleştirilmesi temel koşul olarak belirtilmektedir. Günümüzde yeni teknolojilerin gerek hayatın bir parçası haline gelmesi gerekse eğitim sisteminin birçok aşamasına entegre edilmiş olması sebebiyle ölçme ve değerlendirme işlemlerinin de aynı gelişmişlik düzeyinde gerçekleştirilmesi gerekmektedir. Akıllı teknolojilerin paradigma değişimi olarak nitelendirilecek düzeyde eğitim öğretim süreçlerine dahil edildiği bir sistemde ölçme işleminin geleneksel yöntemlerle yapılması sağlıklı sonuçlar elde edilmesinin önünde engel durumundadır. Mevcut sistemde yaygın biçimde uygulanan ölçme ve değerlendirme yönteminin geleneksel yöntem olması alanda bu yönde yapılacak çalışmalara duyulan ihtiyacın göstergesidir. Bu kapsamda ölçme ve değerlendirme işleminin hassas ölçümlere olanak tanıyan yöntem ve teknolojilerle yapılması önemli bir adımdır. Araştırmanın amacı ölçme ve değerlendirme süreçlerinde akıllı teknolojilerin kullanımını somut bir biçimde ortaya koymaya yönelik olarak tasarlanmıştır. Bu sebeple araştırma yeni akıllı teknolojilerle desteklenen ve hassas ölçümlere olanak tanıyan modern test sunum yöntemlerinin bir araya getirilme sürecini kapsamaktadır. Bu süreçte ilk olarak çok aşamalı testler (Multistage Testing) ve ChatGPT teorik olarak ele alınmıştır. Bir sonraki aşamada araştırma sınırlılıklarında ChatGPT ile soru üretimi yapılarak üretilen soruların b parametreleri ChatGPT'ye tahmin ettirilmiştir. Araştırmanın bir diğer aşamasında ise aynı soruların multistage yöntem ile test montajı sağlanarak b parametreleri hesaplanmıştır. ChatGPT ve MST ile elde edilen b parametreleri sonuçları karşılaştırılmıştır. Elde edilen bulgulara göre yanılma payının çok yüksek olmadığı ChatGPT'nin MST yönteminde gözetimli olarak kullanılmasının uygun olduğu tespit edilmiştir.

Anahtar Kelimeler: Ölçme, Yapay Zekâ, Çok Aşmalı Testler, ChatGPT

Comparison of Artificial Intelligence-Generated Questions and Item Parameters Under MST Test Conditions

Abstract

One of the major implications of the rapid proliferation of smart technologies in the world of education is in the area of measurement and evaluation. By its very nature, it is a basic requirement that measurement should be carried out in a safe environment in order to provide results that are close to the truth. Today, as new technologies have become part of life and have been integrated into many stages of the education system, measurement and evaluation processes should be carried out at the same level of development. In a system where smart technologies are integrated into education and training processes at a level that can be described as a paradigm shift, the traditional methods of measurement are an obstacle to achieving sound results. The fact that the measurement and evaluation method widely used in the current system is the traditional method is an indication of the need for studies in this area. In this context, it is an important step to carry out the measurement and evaluation process with methods and technologies that allow precise measurements. The aim of the research is to demonstrate the use of smart technologies in measurement and evaluation processes in a concrete way. For this reason, the research covers the process of bringing together modern test presentation methods supported by new smart technologies and enabling precise measurements. In this process, Multistage Testing and ChatGPT were first discussed theoretically. In the next stage, questions were created with ChatGPT within the constraints of the research and the b-parameters of the created questions were estimated by ChatGPT. In another stage of the research, the b-parameters of the same questions were calculated by test assembly using the multistage method. The results of the b-parameters obtained by ChatGPT and MST were compared. According to the results, it was determined that the margin of error is not very high and it is appropriate to use ChatGPT supervised in the MST method.

Keywords: Assessment, Artificial Intelligence, Multistage Testing, ChatGPT

[1]Anadolu Üniversitesi

Giriş

Ölçmenin doğası gereği gerçeğe yakın doğrulukta sonuçlar verebilmesi için öncelikli koşul güvenli bir ortamda gerçekleştirilmesi gerekir. Gelişen teknoloji yaşamın pek çok alanında olduğu gibi ölçme ve değerlendirmede güvenliğin sağlanması konusunda da değişim ve dönüşümün gerekliliğini beraberinde getirmektedir. Özellikle son dönemlerde yaşanan pandemi, deprem gibi toplumun bütünü etkileyen olaylar sonucunda hızla uygulanmaya başlanan çevrimiçi sınavlar bu durumu net bir biçimde özetlemektedir. Çünkü bilinen geleneksel sınavlarda uygulanan güvenlik önlemleri çevrim içi sınavlarda anlamını yitirmiş ve yetersiz kalmıştır. Diğer bir ifadeyle mevcut yöntem yeni teknolojinin uygulandığı bir süreçle aynı gelişmişlik hızında olmadığı için çağa ayak uyduramamıştır. Bu noktada bir diğer önemli konu ise akıllı teknolojilerin hali hazırda bir süredir eğitim süreçlerinde uygulanıyor olması ve ölçme süreçlerinin de aynı gelişmişlik hızında uygulanması gereğidir. Ölçme süreçlerinin eğitim öğretim süreçleri ile aynı gelişmişlik hızını yakalayabilmesi için akıllı teknolojilerin işe koşulduğu bir değerlendirme sistemine duyulan ihtiyaç kaçınılmazdır. Günümüzde yaygın biçimde uygulanan ölçme ve değerlendirme yöntemi ise geleneksel yöntemdir. Akıllı teknolojilerin paradigma değişimi olarak nitelendirilecek düzeyde eğitim öğretim süreçlerine dahil edildiği bir sistemde ölçme işleminin geleneksel yöntemlerle yapılması sağlıklı sonuçlar elde edilmesinin önünde engel durumundadır. Bu sebeple ölçme ve değerlendirme süreçlerinin çağın gerekleri doğrultusunda hassas ölçümlere olanak tanıyan yöntemler işe koşularak gerçekleştirilmesinin sağlanması gerekmektedir. Çalışmanın çıkış noktasını tam da bu durum belirlemiştir ve araştırma tasarımı bu yönde yapay zekâ ile hassas ölçümlere olanak tanıyan test sunum yöntemi MST (Multistage Testing- Çok Aşamalı Test Etme)'nin işe koşulduğu bir çalışma olarak şekillendirilmiştir.

Yapay zekâ (YZ) son yıllarda hızla gelişerek eğitimin de aralarında olduğu (Naidu & Sevnarayan, 2023; Zawacki-Richter, Marín, Bond & Gouverneur, 2019) farklı disiplinlerde çeşitli uygulamalara yol açmıştır. YZ sistemleri, insan beynini simüle etmek ve büyük miktarda veri kullanarak rutin işleri yürütmek için eğitilebilen araçlardır (Bengio, Lecun & Hinton, 2021). Öte yandan derin makine öğrenimi, platformların anında insan benzeri metinler üretebileceği bir gelişmişlik düzeyindedir. Bu platformlar arasında ChatGPT yapay zekâ uygulamasının kullanıcı dostu arayüzü kısa sürede çok sayıda kullanıcıya ulaşmıştır (Gleason, 2022). Değerlendirmenin kaliteli eğitimin bel kemiği olduğu göz önüne alındığında (Sok & Heng, 2023) bu yapay zekâ aracının (ChatGPT) MST yöntemle birlikte ölçme ve değerlendirme süreçlerinde kullanım durumunu incelemenin/analiz etmenin alana önemli katkılar getireceği düşünülmektedir. Bu sebeple yapay zekâ ile üretilen maddelerin ve madde parametrelerinin MST ile test sunumuna uygunluğunun ne kadar olduğunun belirlemesi bu çalışmanın problemini oluşturmaktadır. Araştırma kapsamı ChatGPT ve MST yöntemlerinin koşulları ile sınırlıdır. Modern test sunum yöntemleri ve bu yöntemlere yapay zekâ araçları dahil edildiğindeki farkı gözlemlemek bu çalışmanın amacını oluşturmaktadır.

ChatGPT

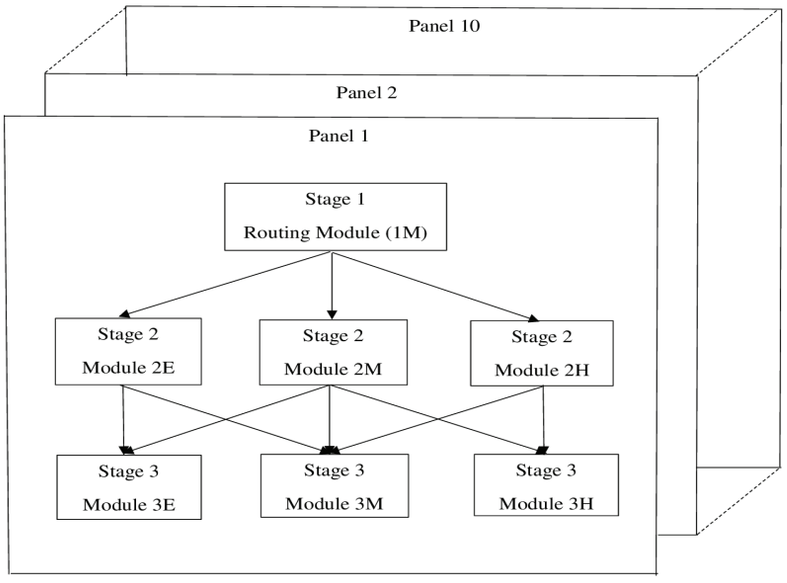
ChatGPT, Kasım 2022'de OpenAI adlı bir Amerikan yapay zekâ araştırma laboratuvarı tarafından büyük dil modelleri (Kohnke vd., 2023) kullanılarak piyasaya sürülen bir yapay zekâ aracıdır (Sok & Heng, 2023). Generative Pre-trained Transformer (GPT) teknolojisi ve yetenekleri, geleneksel öğrenme ve yazma yöntemlerinde devrim yaratan ChatGPT teknolojisinin temelini oluşturmaktadır. Açık erişimli ve kamuya açık bir araç olan ChatGPT, GPT dil modeli teknolojisine göre çalışan oldukça sofistike bir sohbet robotudur (Imran & Almusharraf, 2023, s. 2; Kirmani, 2022, s. 574-576; Newton & Xiromeriti, 2024). Bu model istemlere ve takip sorularına hızlı bir şekilde ayrıntılı yanıtlar oluşturabilmektedir. Verilen bir ipucuna veya sohbete dayalı olarak insan benzeri metin üretmek için tasarlanmıştır ve çok çeşitli konularda doğal, açık uçlu konuşmalar yapma yeteneğine sahiptir (OpenAI, 2023). Model, internetten alınan çeşitli metin verileri üzerinde eğitilerek, verilen bilgilere dayalı olarak tutarlı ve anlaşılır metinler üretmesine olanak sağlamakta (Jukiewicz, 2024), soruları yanıtlayabilmekte, uzun yazılar oluşturabilmekte ve bir insana benzer şekilde yaratıcı bir şekilde yazabilmektedir (Lund ve Wang, 2023).

ChatGPT, doğal dilde konuşmaları mümkün kılmak için yapay zekâyı (AI) kullanan yeni bir teknolojidir. İnsan konuşmalarının geniş veri kümeleri üzerinde eğitilmiş derin bir öğrenme modeli kullanılarak çalışır. Bu model daha sonra kullanıcılar tarafından sorulan sorulara doğal görünen yanıtlar üretmek için kullanılır (Gleason, 2022). Büyük dil modelleri, son yıllarda doğal dil işleme (NLP) alanında önemli ilerlemeler kaydetmiştir. Bu modeller büyük miktarda metin verisi üzerinde eğitilmekte ve insan benzeri metinler oluşturabilmekte, soruları yanıtlayabilmekte ve dille ilgili diğer görevleri yüksek doğrulukla tamamlayabilmektedir (Kanesci vd., 2023; Rasul vd., 2023).

Kasım 2022'de piyasaya sürülmesinden bu yana, akademisyenler arasında en çok tartışılan araç haline gelen ChatGPT farklı alanlardan birçok kullanıcı tarafından da kullanılmaktadır (Imran & Almusharraf, 2023). ChatGPT Ocak 2023'te büyük bir başarı elde etti ve sadece iki ay içinde 100 milyondan fazla aktif kullanıcıyla tarihin en hızlı büyüyen ve en çok memnuniyetle karşılanan yapay zekâ (AI) teknolojik araçlarından biri haline geldi (Hu, 2023). ChatGPT yapay zekâ aracı başta eğitim sektörü olmak üzere birçok sektörde alarm zillerinin çalmasına neden oldu. Araştırmacıların, soruları birkaç dakikadan kısa bir sürede çözen ChatGPT'nin piyasaya sürülmesinin ardından eğitim hizmeti kapsamında yer alan paydaşları yeni değerlendirme biçimleri geliştirmeye davet etmesi bu durumun önemli göstergelerindendir (Adeshola & Adepoju, 2023). Eğitimde büyük dil modellerinin kullanımı, sundukları çok çeşitli uygulamalar nedeniyle potansiyel bir ilgi alanı olarak belirlenmiştir (Kanesci vd., 2023, s. 1-3). Gelişmekte olan bu yapay zekâ aracının öğrenme hedefleri, öğrenme etkinlikleri ve ölçme değerlendirme uygulamaları üzerindeki etkisinin bu alanlarda değişikliklere yol açabileceği için çok önemli olacağı düşünülmektedir (Zhai, 2022).

Çok Aşamalı Testler (MST: Multistage Testing)

MST, CAT (Computerized Adaptive Testing) test sunum yöntemine alternatif olarak geliştirilmiş çok aşamalı bir test sunum yöntemidir (Han, 2013). Çok aşamalı testler, alan yazında en genel biçimde "çeşitli potansiyel amaçlar için bağımsız, otomatik uyarlanabilir birimleri tasarlamanın, birleştirmenin ve yönetmenin modern bir yolu" olarak ifade edilmektedir (Han & Guo, 2013; Yan, Davier & Lewis, 2014, s. 21-27). MST test üretme ve uygulanma tekniği bakımından CAT yönteminde olduğu gibi madde bazlı değil modül adı verilen madde kümeleri aracılığıyla sunulan bir uyarlamalı yapıyı içermektedir (Ariel, Veldkamp & Breithaupt, 2006; Berger, Verschoor, Eggen, & Moser, 2019; Gierl, Lai, & Li, 2011; Han, 2013). Aşağıda yer alan Şekil 1'de "MST Tasarımı Örneği" ana hatları ile sunulmuştur:



Kaynak: Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing* (Order No. 10273809). Available from ProQuest Dissertations & Theses Global. (1901897901). Retrieved from https://www.proquest.com/dissertations-theses/fair-comparison-performance-computerized-adaptive/docview/1901897901/se-2

Şekil 1. Çok Aşamalı Test (MST: Multistage Testing) Tasarımı Örneği

Şekil 1'de sunulan MST panel tasarımı üç aşamadan ve yedi modülden oluşmakta ve her bir modülün zorluk (madde güçlüğü) seviyeleri kolay, orta ve zor olarak sınıflandırılmıştır. Şekil 1'deki modülleri birleştiren çizgiler ise bir sınav katılımcısının izleyebileceği olası yolları temsil etmektedir (Ariel, Veldkamp & Breithaupt, 2006). Adayların yetenek düzeyi ile maddelerin veya madde takımlarının (modül) zorluk profili arasındaki ilişki, madde tepki kuramı (IRT) modelleri tarafından tanımlanmaktadır (Hambleton, Swaminathan & Jones, 1991).

MST test sunumu "Routing" ve "Shaping" olmak üzere iki farklı yapıda gerçekleştirilmektedir. "Routing" yönteminde her aşamanın modülü önceden monte edilerek katılımcının performansına göre hazır yapı içerisinden sunulmaktadır (Han & Guo, 2013; Luecht & Nungester, 1998; Yan, Davier & Lewis, 2014). "Shaping" yönteminde ise test modüllerinin montajı anlık olarak katılımcının performansına göre şekillendirilmektedir (Han, 2013; Yan, Davier & Lewis, 2014, s. 411-420).

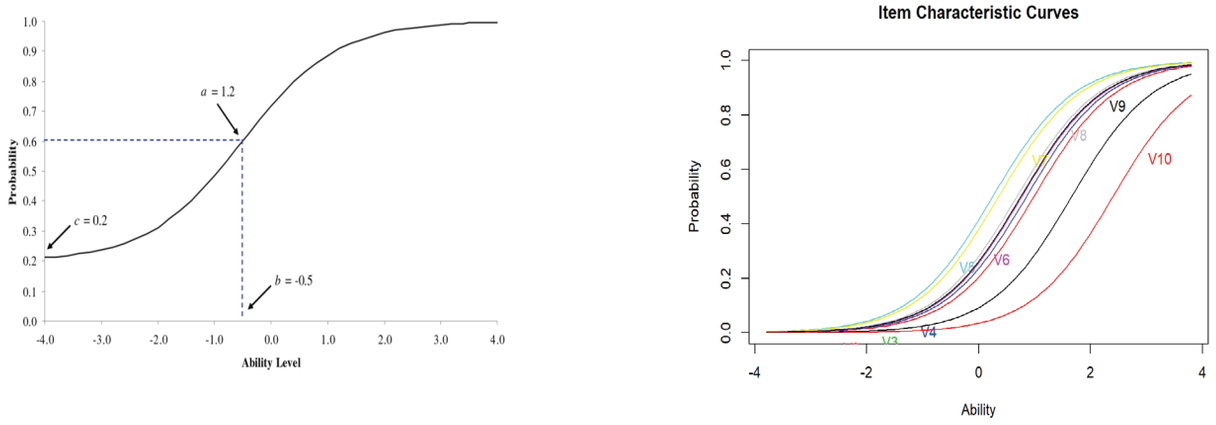
Madde Parametreleri

Kökeni 1920'li yıllara kadar uzanan Madde Tepki Kuramı (MTK)/Item Response Theory (IRT) alan yazında "Örtük Özellikler Kuramı" olarak da bilinmektedir (Bock, 1997; Crocker & Algina, 1986). Kuramla ilgili asıl çalışmalar 1950'li yıllarda Lord tarafından uygulanmaya başlanmış ve özellikle 1970'lerden sonra yapılan test geliştirme çalışmalarında KTK (Klasik Test Teorisi)'nın puanlama yöntemlerine alternatif olarak uygulama alanı bulmuştur (Hambleton& Linden, 1997). Madde Tepki Kuramında puanlama yapılırken madde parametrelerinden yararlanılmaktadır. Bu parametreler,

- a parametresi: madde ayırt ediciliği (maddenin ölçülmeye çalışılan beceriye sahip olanlarla olmayanları ayırt etme gücü)
- b parametresi: madde güçlüğü (bir maddenin uygulandığı grupta doğru cevaplanma yüzdesi)
- c parametresi: şans başarısı (sorunun şansla doğru cevaplanma yüzdesi)
- u parametresi: yüksek bilgiye rağmen soruyu doğru yanıtlayamama ihtimali

olmak üzere yaygın olarak 4 tanedir (Adedoyin & Mokobi, 2013; Hambleton, & Linden, 1997; Hambleton, Swaminathan & Rogers, 1991). Her bir maddede puanlama yapılırken bu parametrelerden en az bir tanesi devreye girmektedir. Puanlama sırasında kullanılan parametre sayısına bağlı olarak başarı testlerinde 1, 2, 3 ve 4 parametreli olmak üzere 4 farklı model bulunmaktadır.

Madde Tepki Kuramında yer alan madde parametreleri Şekil 2' de görsel olarak sunulmuştur.



Şekil 2. MTK'da Madde Parametreleri

Bu çalışmada sadece b parametresi kullanılarak kestirimler yapılarak MST koşulları üretilmiştir.

Yöntem

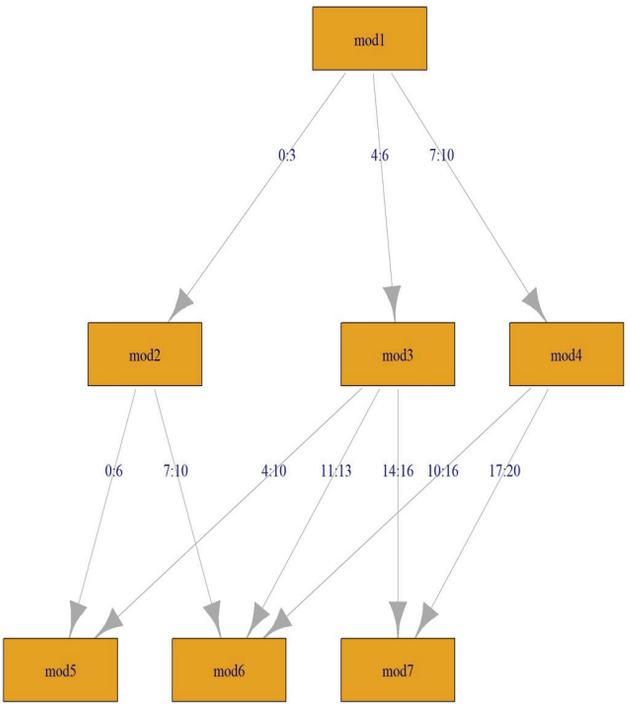
Çalışmada sayısal veriler kullanılarak istatistiksel analizler aracılığı ile kestirimler yapılmıştır. Bu haliyle çalışma nicel bir araştırmadır. Simülatif veriler kullanılarak kontrollü durumlar oluşturulduğu için simülasyona dayalı deneysel araştırma desenindedir.

Çalışmada üretilen sorulara, parametrelerine ve analizlere ilişkin bilgiler aşağıda sunulmuştur.

Verilerin Üretilmesi

Araştırmada kullanılmak amacıyla ChatGPT 4o modeli kullanılarak 70 soru üretilmiştir. Sorular ürettirilirken ChatGPT 4o modeline çalışmanın amacı anlatılmış, bu amaç doğrultusunda 7. sınıf düzeyinde Milli Eğitim Bakanlığı'nın araştırmanın yapıldığı tarihte yürürlükte olan müfredatının tüm konularını kapsayan 70 matematik sorusu üretmesi istenmiştir. Ardından ChatGPT 4o modelinden, bu sorular 7. sınıf öğrencilerine uygulanacak olsa b parametrelerinin ne kadar olabileceğini tahmin etmesi istenmiştir. Elde edilen veriler tablolaştırılmıştır.

Çalışmanın ikinci aşamasında ChatGPT'den elde edilen madde parametreleri R ortamında (R Core Team, 2024) MST test simülasyon ortamları üretmek için kullanılabilen DexterMST (Bechger vd., 2023) paketine yüklenmiştir. DexterMST paketi kendisine verilen sayıda soru ve bu sorulara ait parametreler için istenilen rotalama haritasına uygun simülatif data üretmektedir. Bu datayı üretirken kendisine verilen madde parametrelerine mümkün olan üst düzeyde uyum sağlamaya çalışmaktadır. DexterMST paketi ile 3000 kişi için veri üretilmiştir. Bu veriler üretilirken ChatGPT'nin bildirdiği b parametrelerinin en düşük ve en yüksek değerleri DexterMST'nin ihtiyaç duyduğu theta dağılımının uç sınırları olacak şekilde kullanılmıştır. Simülatif verilerin ürettirilmesi aşamasında kullanılan MST rotaları aşağıda Şekil 3'te verilmiştir.



Şekil 3. MST (Multistage Testing) Rotalama Haritası

Şekil 3'te görüldüğü üzere MST 1-3-3 desen rotalama kullanılmıştır. Rotalamaya birkaç örnekle anlatılacak olursa birinci modülde 0-3 tane arasında doğru cevaplama durumunda yönlendirme modül 2'ye yapılmaktadır. 4-6 tane arasında ise modül 3' yapılmaktadır. Görüleceği gibi tüm rotalama toplamda 70 madde ile gerçekleştirilmiştir.

Verilerin Analizi

Verilerin analizi aşamasında ChatGPT 4o'dan elde edilen b parametreleri ile DexterMST'ten elde edilen b parametreleri karşılaştırılmıştır. Bu amaçla her bir sorunun ChatGPT tarafından tahmin edilen b parametresi ile DexterMST'ten elde edilen b parametresi arasındaki fark alınmıştır. Böylece tüm sorular için b parametrelerinin tahmin edilen ve elde edilen değerlerinin farkı dağılımı elde edilmiştir.

Bulgular

ChatGPT yapay zekâ aracı ile üretilen sorular ve b parametrelerinin MST test koşullarında karşılaştırılmasına yönelik gerçekleştirilen analizler sonucunda elde edilen bulgular aşağıda sunulmuştur.

Tahmin edilen b değerlerinin ve MST sonrası elde edilen b değerlerinin ortalamaları, minimum ve maksimum değerleri aşağıda Tablo 1'de verilmiştir.

Tablo 1. ChatGPT Tarafından Tahmin Edilen ve DexterMST ile Elde Edilen b Değerlerine İlişkin Bilgiler

Elde etme yöntemi	Ortalama	Minimum	Maksimum
ChatGPT 4o (Tahmin)	1.64	1	2.3
DexterMST (Elde edilen)	1.33	0.44	2.33

Görüleceği gibi elde edilen b değerlerinin ortalaması ChatGPT için 1.64, deneysel olarak elde edilen DexterMST için 1.33'tür. b değerlerinin en küçük ve en yüksek değerleri bakımından en yüksek değerler açısından benzerlik yakınken en düşük değer açısından farklılık gözlenmektedir. Aşağıda Tablo 2'de soruların ChatGPT tarafından tahmin edilen b değerleri ile DexterMST'den deneysel olarak elde edilen b değerleri ve bunlar arasındaki farklar gösterilmiştir.

Tablo 2. Deneysel Olarak Elde Edilen b Değerleri ile ChatGPT'den Elde Edilen b Değerleri ve Farkları

Soru No	DexterMST b değeri	ChatGPT b değeri	Fark	Soru No	DexterMST b değeri	ChatGPT b değeri	Fark
item01	1.00	1.00	0.00	item36	1.18	1.60	-0.42
item02	1.42	1.10	0.32	item37	2.09	1.60	0.49
item03	2.25	1.20	1.05	item38	0.58	1.70	-1.12
item04	0.93	1.20	-0.27	item39	1.25	1.70	-0.45
item05	2.31	1.20	1.11	item40	1.87	1.70	0.17
item06	0.58	1.20	-0.62	item41	1.10	1.70	-0.60
item07	1.25	1.20	0.05	item42	1.60	1.70	-0.10
item08	1.69	1.20	0.49	item43	0.99	1.70	-0.71
item09	1.13	1.20	-0.07	item44	1.38	1.80	-0.42

item10	0.44	1.30	-0.86	item45	0.89	1.80	-0.91
item11	1.45	1.30	0.15	item46	1.79	1.80	-0.01
item12	0.97	1.30	-0.33	item47	1.26	1.80	-0.54
item13	2.33	1.30	1.03	item48	1.78	1.80	-0.02
item14	0.91	1.30	-0.39	item49	0.75	1.80	-1.05
item15	1.68	1.30	0.38	item50	1.29	1.90	-0.61
item16	1.16	1.30	-0.14	item51	1.82	1.90	-0.08
item17	2.07	1.30	0.77	item52	1.12	1.90	-0.78
item18	0.73	1.40	-0.67	item53	2.16	1.90	0.26
item19	1.25	1.40	-0.15	item54	0.99	1.90	-0.91
item20	2.05	1.40	0.65	item55	1.47	1.90	-0.43
item21	1.26	1.40	-0.14	item56	0.90	2.00	-1.10
item22	1.50	1.40	0.10	item57	1.54	2.00	-0.46
item23	0.82	1.40	-0.58	item58	0.98	2.00	-1.02
item24	1.35	1.50	-0.15	item59	1.31	2.00	-0.69
item25	1.68	1.50	0.18	item60	0.69	2.00	-1.31
item26	0.99	1.50	-0.51	item61	1.35	2.00	-0.65
item27	1.49	1.50	-0.01	item62	1.88	2.00	-0.12
item28	0.80	1.50	-0.70	item63	1.21	2.10	-0.89
item29	1.91	1.50	0.41	item64	2.27	2.20	0.07
item30	1.00	1.50	-0.50	item65	1.08	2.20	-1.12
item31	1.79	1.60	0.19	item66	1.35	2.20	-0.85
item32	1.04	1.60	-0.56	item67	0.88	2.20	-1.32
item33	1.50	1.60	-0.10	item68	1.34	2.20	-0.86
item34	0.84	1.60	-0.76	item69	1.06	2.20	-1.14
item35	1.78	1.60	0.18	item70	0.73	2.30	-1.57

Tablo 2'den görüleceği gibi ChatGPT'nin b değeri tahminlerinin 1 ve daha fazlası olacak şekilde yanıldığı soru sayısı 11 tanedir. ChatGPT kendi yazdığı soruların gerçekte olduğundan daha zor olduğunu düşünmektedir. 70 sorunun 50 tanesini olduğundan daha zor olarak tahmin etmiştir. 20 soruyu ya olması gerektiği zorlukta veya olduğundan daha kolay tahmin etmiştir. ChatGPT'nin b değeri açısından yanılma ortalaması mutlak değer cinsinden 0.54'tür. Buna göre ChatGPT kendi sorularını ortalamada 0.54 b değeri kadar daha zor algılamaktadır. Bir başka ifadeyle ChatGPT, kendi yazdığı maddelerden alınacak olan puanları 0.54 standart sapma kadar daha az puanlanacak şekilde değerlendirmektedir. Farkların standart sapması 0.39'dur. Buna göre her bir maddede b değerlerinin tahmin edilen değerleri ile DexterMST'ten elde edilen deneysel değerleri arasındaki farklar birbirinden önemli ölçüde farklılaşabilmektedir. Bir başka ifadeyle her maddede ChatGPT değişen oranlarda yanılmaktadır. DexterMST'ten elde edilen deneysel b değerleri ile ChatGPT'nin tahmin ettiği b değerleri arasındaki farklara ilişkin dağılım aşağıda Şekil 4'te verilmiştir.

Şekil 4. DexterMST ve ChatGPT'den Elde Edilen b Parametrelerinin Farklarının Dağılımı

Şekil 4'te verilen fark dağılımı incelendiğinde ChatGPT'nin tahminlerinin yanılma payının en çok 0.677 civarında yoğunlaştığı görülmektedir. 16 soruda 0.677 ve yakında derecelerde yanılmıştır. 1 soruda 1.11 ve üzerinde yanılma payına sahiptir. 14 soruda 0.08 ve civarında yanılma payına sahiptir.

Genel olarak değerlendirildiğinde ChatGPT

Kendi yazdığı soruları olduklarından daha zor tahmin etme eğilimindedir.

- 70 maddenin 50 tanesini olduğundan daha zor tahmin etmektedir.
- 1 maddeyi tam doğru tahmin etmektedir.
- 19 maddeyi olduğundan daha kolay tahmin etmektedir.

ChatGPT tahminlerinin yanılma miktarının büyük olduğu durumların sayısı görece azdır.

- "b parametresi" bakımından 1 ve daha fazla yanıldığı maddelerin sayısı 12 tanedir.
- 58 maddede "b parametresi" bakımından yanılma payı 1'in altındadır.
- Yanılma değerlerinin ortalaması 0.54
- Yanılma değerlerinin standart sapması 0.39'dur.

Sonuç ve Tartışma

ChatGPT, geleneksel yaklaşımları bozma ve öğretme ve değerlendirme şeklimizi dönüştürme potansiyeli ile çevrimiçi değerlendirme alanında önemli bir gelişmeyi temsil etmektedir. Yıkıcı yenilik teorisi, akademik bütünlüğü korurken değişen teknolojilere uyum sağlamanın ve bunları değerlendirme uygulamalarına dahil etmenin önemini vurgulamaktadır. Bu anlamda ChatGPT'nin geliştirilmesi ve değerlendirmeye entegre edilmesi hem öğretenler hem de öğrenenler için heyecan verici fırsatlar ve zorluklar sunmaktadır (Naidu & Sevnarayan, 2023).

Alan yazında yer alan çalışmalar ChatGPT gibi yapay zekâ araçlarının eğitim süreçlerinin farklı aşamalarında tüm paydaşlar (öğrenen, öğreten, karar alıcılar vb.) açısından otomatikleştirme sağladığını kabul ederken (Kasneci ve diğerleri, 2023) uzmanlığının yerini alamayacağını vurgulamaktadır (Fabiyi, 2024, s. 9-12). Bang ve arkadaşları (2023) ChatGPT'nin bazen doğru olmayan veya anlamsız bilgiler üretebildiği ve bunun nedenlerinin model sınırlamaları, eğitim verilerinin kalitesi ve belirli kullanıcı sorguları ile ilgili olabileceğini belirtmektedirler. Qadir (2023) ise soruların bağlamı ve sorulma şeklinin yanıtları etkilediğini bu sebeple ChatGPT kullanılırken dikkatli olunması gerektiğini vurgulamaktadır. Lo (2023)' benzer şekilde yaptığı literatür taramasında ChatGPT'nin öğretmen ve öğrenciler için süreçleri kolaylaştırıcı değerli bir araç olduğu ancak dikkatli kullanılması gerektiğine dikkat çekmiştir. Bu çalışmaların araştırma bulguları ile örtüşen yönleri ChatGPT gibi yapay zekâ araçlarının gözetimli olarak dikkatli bir biçimde kullanılmasının faydalı olacağıdır.

Tıp eğitimi alanında Kung ve arkadaşları (2022) tarafından yapılan çalışmada, ChatGPT'nin Amerika Birleşik Devletleri Tıp Lisanslama Sınavı'ndaki performansını araştırmıştır. Değerlendirme sonuçlarına göre, ChatGPT'nin bu testteki performansı, herhangi bir alan ince ayarı yapılmaksızın geçme eşiğinde veya yakınında olmuştur. Bu sonuçlara dayanarak yazarlar, büyük dil modellerinin tıp eğitimine ve nihayetinde klinik karar verme süreçlerine yardımcı olmak için güçlü bir araç olabileceğini savunmaktadır. Jukiewicz (2024) çalışmasında ChatGPT tarafından verilen notlar, öğreticiler tarafından verilenlerle karşılaştırılmıştır. Sonuçlar, ChatGPT ile öğreticiler arasında güçlü bir pozitif korelasyon (uyum) olduğunu göstermektedir. Latif ve Zhai (2023) ise araştırmalarında ChatGPT (GPT-3,5)'nin tüm etiketlerde BERT (Google benzer yapıdaki uygulamasıdır.)'den önemli ölçüde daha yüksek puanlama doğruluğu elde ettiği sonucuna ulaşmışlardır (Latif & Zhai, 2023). Araştırma sonuçları açısından değerlendirildiğinde de ChatGPT gerçek değerlere önemli ölçüde uyum göstermiştir.

ChatGPT'nin değerlendirme aşamasına dahil edilmesi konusu alan yazında çoğu araştırmacı tarafından sıklıkla gündeme getirilmektedir. Bu bulgular ise araştırma bulguları ile örtüşmektedir. Sharples (2022) uzaktan eğitim üniversitelerinin öğrenmeyi teşvik etmek ve öğretim elemanlarının teknoloji konusunda güncel kalmasına yardımcı olmak için ChatGPT'yi değerlendirme uygulamalarına dahil etmelerinin faydalı olacağına dikkat çekmektedir. Zhai (2023), öğretmenlerin ChatGPT'yi kullanarak öğrenme değerlendirme öğeleri oluşturabileceğini, zamandan ve emekten tasarruf edebileceğini ve standart bir çerçeveye bağlı kalarak soruların kalitesini potansiyel olarak artırabileceğini iddia etmiştir. ChatGPT'nin sunduğu yetenekler sayesinde öğretmenler, öğretim derslerinin öğrenme hedefleri ve başarı kriterleriyle uyumlu açık uçlu soru istemleri geliştirebilmektedir (Baidoo-Anu & Ansah, 2023). Öğretmenlerin çoğunun kısa sınavlar, aylık testler ve sınavlar oluşturmak için çok fazla zaman harcadığı göz önüne alındığında, eğitimcilerin ChatGPT'den yardım alarak değerlendirme baskısını azaltma fırsatı olduğu görülmektedir.

Bulgular, eğitimde yapay zekânın (YZ) kullanımının devam eden gelişim sürecinin bir parçası olduğunu ve ChatGPT'nin de bunun bir uzantısı olduğunu göstermektedir (Imran & Almusarraf, 2023). Bu nedenle, eğitim süreçlerinde ChatGPT'yi bir yazma asistanı olarak benimseme konusu hem fırsatlara hem de zorluklara sahiptir. Burada ihtiyaç, ChatGPT'nin hem öğrenciler hem de eğitmenler için bir yardımcı ve kolaylaştırıcı olarak rolünü anlamaktır, çünkü "chatbotlar" akademik süreci kolaylaştırmak ve desteklemek için nispeten faydalı araçlardır.

Ölçme ve değerlendirmede akıllı teknolojilerin ve modern test sunum yöntemlerinin bir arada işe koşulduğu bir sürecin nasıl işlediğine dair yürütülen bu araştırmada önemli sonuçlara ulaşılmıştır. ChatGPT ve MST yöntemin simülatif olarak analiz edilmesi ve ulaşılan sonuçlar çalışmanın özgün yönünü ortaya koymaktadır. ChatGPT ve MST'nin birlikte ele alındığı bu araştırmada ulaşılan sonuçlar toplu olarak değerlendirildiğinde;

- Yanılma payının çok yüksek olmadığı,
- Görece az sayıda maddede hatalı tahminde bulunduğu,
- Bu hatanın soruların "b parametrelerini (zorluklarını)" olduklarından daha yüksek tahmin etme yönünde olduğu gözlenmiştir.
- ChatGPT'nin bu haliyle MST yönteminde gözetimli olarak kullanılmasının uygun olduğu tespit edilmiştir.

İlerleyen çalışmalarda;

- daha yüksek sayılarda madde içeren büyük soru havuzlarında çalışılması
- MST data yapısına uygun analizleri gerçekleştirebilecek yazılım programlarının geliştirilmesi önerilebilir.

Kaynakça

Adedoyin, O. ve Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. International Journal of Asian Social Science, 3(4), 992–1011. Erişim adresi: https://archive.aessweb.com/index.php/5007/article/view/2471

Adeshola, I. & Adepoju, A. P. (2023). The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments, 0*(0), 1–14. https://doi.org/10.1080/10494820.2023.2253858

Ariel, A., Veldkamp, B. P. & Breithaupt, K. (2006). Optimal testlet pool assembly for multistage testing designs. Applied Psychological Measurement, 30(3), 204–215. doi:10.1177/0146621605284350

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. http://dx.doi.org/10.2139/ssrn.4337484

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., ... & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. Erişim adresi: http://arxiv.

org/pdf/2302.04023.

Bechger T, Koops J, Partchev I, Maris G (2023). dexterMST: CML and Bayesian Calibration of Multistage Tests. R package version 0.9.6, URL: https://CRAN.R-project.org/package=dexterMST

Bengio, Y., Lecun, Y. & Hinton, G. (2021). Deep learning for AI. *Communications of the ACM, 64*(7), 58-65. doi:10.1145/3448250

Berger, S., Verschoor, A. J., Eggen, T. J. H. M. ve Moser, U. (2019). Improvement of measurement efficiency in multistage tests by targeted assignment. Frontiers in Education, 4(January). https://doi.org/10.3389/feduc.2019.00001

Bock, R. D. (1997). A brief history of item response theory. Educational Measurement: Issues and practice, 16, 21–23. doi:10.1111/j.1745-3992.1997.tb00605.x.

Cotton, D. R., Cotton, P. A. & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228-239. https://doi.org/10.1080/14703297.2023.2190148

Crocker, L. ve Algina, J. (2008). Introduction to classical and modern test theory. In M. Baird, M., Staudt, M. & Strans (Ed.), Cengage Learning. USA: Cengage Learning.

Elkins, K. & Chun, J. (2020). Can GPT-3 pass a writer's turing test?. *Journal of Cultural Analytics, 5*(2), 1-16. doi: 10.22148/001c.17212

Fabiyi, S. D. (2024). What can ChatGPT not do in education? Evaluating its effectiveness in assessing educational learning outcomes. Innovations in Education and Teaching International, 0(0), 1–15. https://doi.org/10.1080/14703297.2024.2333395

Gierl, M., Lai, H. & Li, J. (2011). Evaluating the performance of CATSIB in a multistage adaptive testing Environment. Erişim adresi: https://mcc.ca/wpcontent/uploads/Technical-Reports-Gierl-Lai-Li-2011.pdf

Gleason, N. (2022). ChatGPT and the rise of AI writers: How should higher education respond? THE Campus Learn, Share, Connect. Erişim adresi https://www.timeshighereducation.com/campus/chatgpt-and-rise-aiwriters-how-should-higher-education-respond

Hambleton, R. K. ve Linden, W. J. (1997). Handbook of modern item response theory (1st ed.). USA: Springer. https://doi.org/10.1007/978-1-4757-2691-6

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). Fundamentals of item response theory library (1st ed.; D. Foster, ed.). London: SAGE.

Han, K. T. (2013). MSTGen: Simulated data generator for multistage testing. Applied Psychological Measurement, 37, 666–668. doi: 10.1177/0146621613499639

Han, K. T. ve Guo, F. (2013). An approach to assembling optimal multistage testing modules on the fly. GMAC Research Reports (Report No: RR-13-01). Erişim adresi: https://www.gmac.com/-/media/files/gmac/research/research-report-series/rr-13-01-moduleassemblyonthefly.pdf

Hu, K. (2023, Şubat). ChatGPT sets record for fastest-growing user base - analyst note. Erişim adresi: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Imran, M. & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology, 15*(4), ep464. https://doi.org/10.30935/cedtech/13605

Jukiewicz, M. (2024). The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Thinking Skills and Creativity, 52*(101522), 1-9. https://

doi.org/10.1016/j.tsc.2024.101522

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kirmani, A. R. (2022). Artificial intelligence-enabled science poetry. *ACS Energy Letters, 8*(1), 574-576. https://doi.org/10.1021/acsenergylett.2c02758

Kohnke, L., Moorhouse, B. L. & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 1–14. https://doi.org/10.1177/00336882231162868

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digital Health, 2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Latif, E. & Zhai, X. (2024). Fine-tuning chatgpt for automatic scoring. Computers and Education: *Artificial Intelligence, 6*(100210), 1-10. https://doi.org/10.1016/j.caeai.2024.100210

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 1-15. https://doi.org/10.3390/educsci13040410

Luecht, R. M. & Nungester, R. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*(3), 239–249. https://doi.org/10.1111/j.1745-3984.1998.tb00537.x

Naidu, K. & Sevnarayan, K. (2023). ChatGPT: An ever-increasing encroachment of artificial intelligence in online assessment in distance education. *Online Journal of Communication and Media Technologies, 13*(3), e202336. https://doi.org/10.30935/ojcmt/13291

Newton, P. & Xiromeriti, M. (2024). ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. *Assessment & Evaluation in Higher Education*, *0*(0), 1–18. https://doi.org/10.1080/02602938.2023.2299059

OpenAI. (2023). ChatGPT: Optimizing language models for dialogue. Erişim Adresi: https://openai.com/research

Qadir, J. (2023). Engineering education in the era of chatGPT: Promise and pitfalls of generative AI for education. 2023 IEEE Global engineering education conference (EDUCON) (pp. 1–9). IEEE. Erişim adresi: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10125121

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W. J., Sun, M., Day, I., Rather, R. A., & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching, 6*(1). https://doi.org/10.37074/jalt.2023.6.1.29

Sharples, M. (2022). Automated essay writing: An AIED opinion. International Journal of Artificial Intelligence in Education, 32(4), 1119-1126. https://doi.org/10.1007/s40593-022-00300-7

Sok, S. & Heng, K. (March 6, 2023). ChatGPT for education and research: A review of benefits and risks. http://dx.doi.org/10.2139/ssrn.4378735

Wang, K. (2017). A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing (Doktora Tezi). ProQuest Dissertations & Theses Global veri tabanında erişildi (Order No. 10273809). Erişim adresi: https://www.proquest.com/dissertations-theses/fair-comparison-performancecomputerized-adaptive/docview/1901897901/se-2

Yan, D., Davier, A. A. & Lewis, C. (2014). Computerized multistage testing: Theory and application (1st ed.). USA: CRC Press. doi: 10.1201/b16858

Zhai, X. (December 27, 2022). ChatGPT user experience: Implications for education. http://dx.doi.org/10.2139/ssrn.4312418

Zhai, X. (2023). ChatGPT for next generation science learning. *XRDS: Crossroads, The ACM Magazine for Students, 29*(3), 42-46. doi: 10.1145/358964

Zhang, X., Li, D., Wang, C., Jiang, Z., Ngao, A. I., Liu, D., Peters, M. A., & Tian, H. (2023). From ChatGPT to China' Sci-Tech: Implications for Chinese Higher Education. *Beijing International Review of Education*, *5*(3), 296-314. https://doi.org/10.1163/25902539-05030007

Zawacki-Richter, O., Marín, V. I., Bond, M. & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators?. International *Journal of Educational Technology in Higher Education, 16*(1), 1-27. https://doi.org/10.1186/s41239-019-0171-0