Disiplinlerarası Eğitim Araştırmaları Dergisi

ISSN: 2602-2516
Disiplinlerarası Eğitim Araştırmaları Dergisi
Journal of Interdisciplinary Educational Research
2023; 7(15);202-210

Madde Güçlüklerinin Tahmin Edilmesinde Uzman Görüşleri ve ChatGPT Performansının Karşılaştırılması

Comparison of Expert Opinions and ChatGPT Performance in Predicting Item Difficulties

Erdem BODUROĞLU[1] Oğuz KOÇ[2] Mahmut Sami YİĞİTER[3]

Atıf:
Boduroğlu, E., Koç, O., Yiğiter, M. S.,(2023). Madde Güçlüklerinin Tahmin Edilmesinde Uzman Görüşleri ve ChatGPT Performansının Karşılaştırılması. *Disiplinlerarası Eğitim Araştırmaları Dergisi*. 7(15), 202-210, DOI: 10.57135/jier. 1296255

Öz
Bu çalışmada ChatGPT'nin çoktan seçmeli test maddelerini yanıtlama ve bu maddelerin madde güçlük düzeylerini sınıflama performansı incelenmiştir. 20 maddeden oluşan beş seçenekli çoktan seçmeli test maddesine 4930 öğrencinin verdiği yanıtlara göre madde güçlük düzeyleri belirlenmiştir. Bu güçlük düzeyleri ile ChatGPT'nin ve uzmanların yaptığı sınıflandırmalar arasındaki ilişkiler incelenmiştir. Elde edilen bulgulara göre ChatGPT'nin çoktan seçmeli maddelere doğru yanıt verme performansının orta düzeyde olduğu (%55) görülmüştür. Ancak madde güçlük düzeylerini sınıflandırma konusunda ChatGPT; gerçek madde güçlük düzeyleri ile 0.748, uzman görüşleri ile 0.870 korelasyon göstermiştir. Bu sonuçlara göre deneme uygulamasının yapılamadığı veya uzman görüşlerine başvurulamadığı durumlarda ChatGPT'den test geliştirme aşamalarında destek alınabileceği düşünülmektedir. Geniş ölçekli sınavlarda da uzman gözetiminde ChatGPT benzeri yapay zekâ teknolojilerinden faydalanılabilir.

Anahtar Kelimeler: ChatGPT, yapay zekâ, madde güçlüğü, uzman görüşü

Abstract
In this study, ChatGPT's performance in answering multiple-choice test items and classifying the item difficulty levels of these items was examined. Item's actual difficulty levels were determined according to the responses of 4930 students to the five-choice multiple-choice test items consisting of 20 items. The relationships between these difficulty levels and the classifications made by ChatGPT and experts were tested. The findings demonsrated that ChatGPT's performance in giving correct answers to multiple-choice items was at moderate level (55%). However, in terms of classifying item difficulty levels, ChatGPT showed a correlation of 0.748 with actual item difficulty levels and 0.870 with expert opinions. According to these results, it is thought that ChatGPT can be used to support test development in cases where trial application cannot be conducted or expert opinions cannot be consulted. In large-scale exams, ChatGPT-like artificial intelligence technologies can be utilized under expert supervision.

Keywords: ChatGPT, artificial intelligence, item difficulties, expert opinion

[1] Müdür Yrd., Niğde Ölçme Değerlendirme Merkezi, Milli Eğitim Bakanlığı, Niğde-Türkiye,erdemboduroglu@gmail.com, orcid.org/0000-0001-8318-4914

[2]Öğretmen, Niğde Ölçme Değerlendirme Merkezi, Milli Eğitim Bakanlığı, Niğde-Türkiye, oguzkoc20@hotmail.com, orcid.org/0000-0002-8656-6069

[3] Öğr. Gör., Ankara Sosyal Bilimler Üniversitesi, Uzaktan Eğitim Uygulama ve Araştırma Merkezi, Ankara-Türkiye, mahmutsami.yigiter@asbu.edu.tr, orcid.org/ 0000-0002-2896-0201

Geliş Tarihi : 12/05/2023
Kabul Tarihi: 25/07/2023
Araştırma Makalesi

GİRİŞ

OpenAI (2023) kar amacı gütmeyen bir yapay zekâ araştırma şirketidir ve amacı, insanlık için en yararlı olacak şekilde dijital zekâyı geliştirip finansal getiri kaygısı olmadan çalışmaktır. Bu kapsamda şirketin en ünlü ürünü ChatGPT 30 Kasım 2022'de piyasaya sürülmüştür. ChatGPT, yapay zekâ konusunda eğitilmiş bir prototip sohbet robotudur ve özellikle diyalog konusunda uzmanlaşmıştır. ChatGPT, bağlamsal olarak uygun cevaplar üretme ve doğal bir şekilde konuşma yapabilme yeteneği olan büyük bir dil modelidir (Deng & Lin, 2022). Büyük bir başarı yakalayan dil modeli; sağlık, eğitim ve çeşitli disiplinlerde farklı uygulamalara konu olmuştur. Örneğin, Minnesota Üniversitesi Hukuk Fakültesi'nde dört ayrı sınavı büyük bir başarıyla tamamlamıştır (CNN, 2023). Puanları şimdilik yüksek olmasa da, sonuçlar bu yapay zekâ uygulamasının bir üniversite diploması alabilecek kapasitede olduğunu ortaya koymaktadır (Choi, Hickman, Monahan ve Schwarcz, 2023).

Lo (2023) ChatGPT ile gerçekleştirilen 50 farklı çalışmayı incelemiş ve eğitmenler ile ilgili olarak beş ana işlevi iki temada ortaya koymuştur. Buna göre öğretim hazırlığı (materyal geliştirme, öneriler, çeviri vb.) ve değerlendirme (görevler oluşturma ve performans değerlendirme) temalarını belirtmişlerdir. İlgili çalışmalarda temalar oluşturulmuş olsa da hem öğretim hem de değerlendirme sürecine yardımcı olma konusunda hala kısıtlı sayıda çalışma olduğu görülmektedir. Zhai (2023), öğretmenlerin ChatGPT kullanarak değerlendirme çalışmalarını zamandan ve emekten tasarruf ederek gerçekleştirebileceklerini belirtmektedir. Öğretmenlerin birçoğu kısa sınavlar, aylık testler ve sınavlar hazırlamak için çok zaman harcadığından eğitimcilerin ChatGPT'den yardım alarak değerlendirme baskısını hafifletme fırsatı bulabileceği ifade edilmektedir (Sok ve Heng, 2023). ChatGPT, öğrencilerin öğrenme çıktılarını artırmak için gerekli geri bildirimleri içeren bir otomatik notlandırma sistemini de sunabilir. ChatGPT öğrencilerin hem zayıf hem de güçlü yönlerini belirleyerek notlandırmayı yarı otomatik hale getirebilir. Bu şekilde öğrencilerin çalışmalarına daha iyi bir not verilmesine yardımcı olabilir (Kasneci vd., 2023). Ayrıca öğretmenlerin yüklerini azaltarak zamandan tasarruf etmelerini sağlayabilir (Bozkurt vd., 2023).

Günümüzde eğitim, insan davranışlarını geliştiren bir sistem olarak görülmektedir. Bu sistemin bileşenlerinin işleyip işlemediği, eksik ve güçlü yanları ölçme ve değerlendirme öğesi vasıtasıyla ortaya konulur (Baykul, 2015). Ölçme ve değerlendirme çalışmalarında ise testler önemli bir yer tutar. Geliştirilen bir testin niteliği hazırlanan maddelerin niteliği ile orantılıdır. Maddelerin niteliğini ortaya koymak için ise çeşitli madde göstergeleri hesaplanır. Madde güçlük indeksi, madde ayırt edicilik indeksi, çeldirici analizi gibi istatistikler ile maddelerin niteliği ortaya koyulur. Klasik test kuramına göre bu göstergelerden birisi olan madde güçlük indeksi; bir maddeyi doğru yanıtlayan birey sayısının gruptaki tüm birey sayısına bölünmesi ile elde edilir. Kısaca bu indeks katılımcılar tarafından bir maddenin doğru cevaplanma yüzdesini ifade eder. Bilen ile bilmeyeni en iyi ayırt edecek testlerin madde güçlük indeksi ortalamasının 0.50 civarında olması beklenir. Bir maddenin güçlük indeksi 0.00-0.20 arası ise bu madde "çok zor", 0.21-0.40 arası "zor",0.41-0.60 arası "orta, 0.61-0.80 arası "kolay" ve 0.81-1.00 arası ise "çok kolay" olarak yorumlanır (Baykul, 2015; Crocker & Algina, 1986; Urbina, 2014). Alanyazın incelendiğinde, uzman görüşlerine dayalı olarak belirlenen madde güçlükleri ile gerçek madde güçlük indeksleri arasındaki uyumun incelenmesine yönelik çok sayıda araştırma olduğu ifade edilebilir (Anıl, 2002; Baykul & Sezer, 1993; Güler, İlhan & Teker, 2021; Impara & Plake, 1998; Lorge & Diamon, 1954; Quereshi & Fisher, 1977; Tinkelman, 1947). Ancak madde güçlük indekslerinin belirlenmesinde yapay zekânın kullanıldığı çalışmaların sınırlı sayıda olduğu görülmektedir. Uzmanlara ulaşmada zorlanıldığı veya deneme uygulamasının yapılamadığı durumlarda nitelikli ölçme değerlendirme araçları geliştirebilmek için ChatGPT'nin nasıl bir performans sergileyeceğine yönelik çalışmaların literatüre katkı sağlayacağı söylenebilir.

Araştırmanın Amacı

Test geliştirme sürecinde öncelikle testin amacı ve kapsamı belirlenir ve sonrasında da test maddelerinin hazırlanması aşamasına geçilir. Test maddelerinin testin genel amacına uygun olarak hazırlanması, gözlenen davranışları ortaya çıkaracak nitelikte olması, bilimsel açıdan ve dil yönünden hatalardan arınık olması gereklidir. Maddelerin bu özelliklere sahip olup olmadığını belirlemek için uzman görüşlerine başvurulur ve deneme uygulamaları gerçekleştirilir (Baykul, 2015). Test geliştirme süreçlerinde deneme uygulamasının yapılamadığı veya sınav güvenliği sebebiyle uzman görüşlerine başvurulamadığı durumlarda birtakım sorunlarla karşılaşılabilir. Böyle durumlarda madde ve test istatistikleri test geliştiricilerin veya madde yazarlarının sınırlı tahminleri ile kestirilmeye çalışılır ve hatalı sonuçlar elde edilebilir (Sezer, 1992). Bu sorunları en aza indirgemek için ChatGPT teknolojisinin eğitim alanında destekleyici bir rolde kullanılıp kullanılamayacağına yönelik bir araştırmanın yürütülmesi hedeflenmiştir. Bu araştırmada ChatGPT'nin matematik alanında çoktan seçmeli test maddelerine verdiği yanıtların incelenmesi ve maddelerin ChatGPT tarafından madde güçlük düzeylerine göre sınıflandırılma performansının uzman görüşleri ve gerçek madde güçlük düzeyleri ile karşılaştırılması amaçlanmıştır. Bu amaçlarla araştırmada aşağıdaki iki alt probleme cevap aranmıştır:

1- ChatGPT'nin çoktan seçmeli test maddelerini yanıtlamadaki performansı nasıldır?

2- Madde güçlük düzeylerinin belirlenmesinde; uzman görüşleri, ChatGPT sınıflaması ve gerçek madde güçlük düzeyleri arasındaki ilişki nasıldır?

YÖNTEM

Araştırma Modeli

Bu araştırma nicel araştırma yöntemlerinden olan ilişkisel araştırma desenindedir. İlişkisel araştırmalar iki ya da daha fazla değişken arasındaki ilişkileri olduğu gibi ortaya koymayı amaçlar. İlişkisel araştırmalar değişkenler arasındaki var olan bir ilişkiyi tanımladığı için bazen betimleyici araştırmaların bir türü olarak da adlandırılır (Fraenkel, Wallen ve Hyun, 2012). Bu araştırmada ChatGPT ve uzman görüşlerine göre elde edilen madde güçlük düzeyleri sınıflandırması ile gerçek madde güçlük indeksleri arasındaki ilişkiler incelenmiştir. Bu sebeple araştırmanın ilişkisel araştırma deseninde olduğu ifade edilebilir.

Araştırma Verileri

Araştırmada kullanılan madde istatistikleri Niğde ili genelinde 9.sınıf düzeyinde öğrenim gören 4930 öğrenciye uygulanan beş seçenekli ve çoktan seçmeli 20 maddelik Matematik testinden elde edilmiştir. Uygulanan bu test; beş alan uzmanı, bir dil uzmanı ve bir ölçme değerlendirme uzmanı tarafından geliştirilmiştir. Geliştirilen testin pilot uygulamaları yapılmış ve elde edilen test ve madde istatistiklerine göre nihai maddeler oluşturulmuştur. Test geliştirme sürecinin tüm prosedürlerine uygun olarak geliştirilen bu test maddelerine ChatGPT'nin verdiği yanıtlar ve madde güçlük düzeyi sınıflandırmaları ile üç alan uzmanı tarafından yapılan madde güçlük sınıflamaları araştırmanın verilerini oluşturmaktadır.

Veri Toplama Yöntemleri

Araştırmada incelenen test maddelerinin ve bu maddelere dair madde güçlük indekslerinin kullanılabilmesi için Niğde İl Milli Eğitim Müdürlüğü'nden resmi izin alınmış ve verilere erişim sağlanmıştır. Bu veriler Niğde ili genelinde öğrenim gören 9.sınıf öğrencilerinin tamamına uygulanan geniş ölçekli ortak sınav uygulamasından elde edilmiş olup araştırmada kullanılacak diğer veriler için bir ölçüt niteliği taşımaktadır. Bu çoktan seçmeli maddeler seçenekleri ile birlikte ChatGPT 3.5 uygulamasına verilmiş ve bu maddelerin yanıtlanması istenilmiştir. Ayrıca ChatGPT'den yanıtlanan her bir maddeyi madde güçlük düzeylerine göre Çok Kolay-Kolay-Orta-Zor-Çok Zor olarak sınıflandırması ve gerekçe sunması istenmiştir. Sonrasında madde

güçlük düzeyi sınıflandırması için uzman görüşlerine başvurulmuştur. Devlet okullarında görev yapmakta olan ve test geliştirme konusunda yetkin üç matematik öğretmeni uzman ekip olarak belirlenmiştir. Bu uzmanlardan testteki 20 maddeyi inceleyerek maddeleri güçlük düzeylerine göre gerekçesi ile birlikte sınıflandırması istenmiştir. Bu sınıflandırmadan sonra veri toplama işlemi tamamlanmıştır.

Verilerin Analizi

Araştırmanın birinci alt probleminde 20 çoktan seçmeli test maddesine ChatGPT 3.5 versiyonu tarafından 01.03.2023 tarihinde verilen yanıtlar kaydedilmiştir. Her bir maddeye uygulamanın 2 kez yanıt vermesi istenmiştir. Cevaplar arasında tutarlılık varsa nihai cevap olarak kayıt altına alınmıştır. Aksi bir durum için tekrar bir cevap oluşturulması istenmiştir. Madde güçlük düzeyi sınıflaması için ChatGPT'ye "Bu maddeyi güçlük düzeyine göre Çok Kolay-Kolay-Orta-Zor-Çok Zor olarak gerekçesiyle birlikte sınıflandırır mısın?" istemi girilmiştir. Elde edilen sonuçlar tablolaştırılmıştır. Uzmanlardan da benzer şekilde bağımsız olarak güçlük sınıflaması alınmıştır. Uzmanlar arasındaki uyumu belirlemek için Fleiss Kappa istatistiği kullanılmıştır. Üç uzmanın yaptığı 1-5 arasındaki sınıflamaya göre ortalama puan değeri çıkarılmıştır. Gerçek madde güçlük düzeyleri, ChatGPT sınıflaması ve uzman görüşleri arasındaki güvenirliği belirlemek amacıyla SPSS paket programı ile sınıf içi korelasyon katsayısı hesaplanmıştır. Elde edilen sonuçlar 0,05 anlamlılık düzeyinde raporlaştırılmış ve veri analizi aşaması tamamlanmıştır.

BULGULAR

Araştırmanın ilk alt problemi için 9. sınıf matematik testinde yer alan 20 çoktan seçmeli maddeye ChatGPT tarafından verilen yanıtlar incelenmiştir. Her bir maddenin konu alanını, madde güçlük indeksini ve ChatGPT'nin verdiği yanıtların doğru ya da yanlış olma durumunu gösteren veriler Tablo 1'de sunulmuştur.

Tablo 1. Test Maddelerine Dair Konu Alanı, Güçlük İndeksi ve ChatGPT Yanıtları Tablosu

Madde No	Konu Alanı	Madde Güçlük İndeksi	ChatGPT Yanıtları
1	Mantık	0,81 (Çok Kolay)	Yanlış
2	Kümeler	0,49 (Orta)	Yanlış
3	Kümeler	0,50 (Orta)	Doğru
4	Bölünebilme	0,32 (Zor)	Yanlış
5	Denklemler	0,60 (Orta)	Doğru
6	Mutlak Değer	0,43 (Orta)	Doğru
7	Oran-Orantı	0,47 (Orta)	Doğru
8	Problemler	0,63 (Kolay)	Doğru
9	Problemler	0,33 (Zor)	Doğru
10	Problemler	0,41 (Orta)	Doğru
11	Mantık	0,64 (Kolay)	Yanlış
12	Kümeler	0,48 (Orta)	Yanlış
13	Kümeler	0,52 (Orta)	Doğru
14	Kümeler	0,52 (Orta)	Doğru
15	Kartezyen Çarpım	0,45 (Orta)	Yanlış
16	Bölünebilme	0,72 (Kolay)	Doğru
17	Denklemler	0,72 (Kolay)	Doğru
18	Mutlak Değer	0,31 (Zor)	Yanlış
19	Eşitsizlikler	0,19 (Çok Zor)	Yanlış
20	Bölünebilme	0,26 (Zor)	Yanlış

Tablo 1 incelendiğinde testteki 1 maddenin "Çok kolay", 4 maddenin "Kolay", 10 maddenin "Orta", 4 maddenin "Zor", 1 maddenin ise "Çok Zor" güçlük düzeyinde olduğu görülmektedir. ChatGPT'nin 20 çoktan seçmeli maddesine verdiği yanıtlardan 11 tanesi doğru iken 9 tanesi ise

yanlıştır. Tabloya göre madde güçlük indeksleri ile maddenin doğru cevaplanma durumu arasında net bir ilişki görülememiştir. Doğru ve yanlış yanıtlanan maddelerin madde güçlük indeksleri ortalamaları arasında fark olup olmadığı Mann-Whitney U testi ile sınanmıştır. Analiz sonucuna göre 0,05 anlamlılık düzeyinde gruplar arasında manidar bir farklılık görülmemiştir (U=31, p=0,160). Testin en kolay maddesi olan 1. madde ChatGPT tarafından yanlış cevaplanırken zor kategorisinde yer alan 9. madde doğru cevaplanmıştır. Yanlış cevaplara ilişkin uzman incelemeleri gerçekleştirilmiş ve bu duruma; maddedeki bilinmeyen sayısı, işlem adımları sayısı, çözüm için gereken muhakeme becerisi gibi çeşitli faktörlerin etki ettiği belirtilmiştir. Testte yer alan 20 madde için ortalama güçlük indeksi 0,49'dur. Bu indekse göre testin ortalama zorluk düzeyinde olduğunu söylemek mümkündür. ChatGPT ise test maddelerinin 11 tanesini doğru cevaplayarak %55 başarı oranı elde etmiştir. Bu verilere göre ChatGPT'nin yaklaşık olarak bu teste katılan ortalama başarıdaki bir öğrenci kadar doğru yanıtlama becerisine sahip olduğu söylenebilir.

Araştırmanın ikinci alt problemi için 3 alan uzmanı tarafından test maddeleri incelenmiş ve bağımsız olarak bu maddeler güçlük düzeylerine göre (Çok Kolay(5)-Kolay(4)-Orta(3)-Zor(2)-Çok Zor(1)) olarak sınıflandırılmıştır. Uzmanlar arasındaki uyumun derecesini belirlemek amacıyla Fleiss Kappa istatistiği hesaplanmış ve bu değer 0,86 olarak bulunmuştur. Landis ve Koch (1977)'a göre bu değerin çok yüksek düzeyde uyuma işaret ettiği ifade edilebilir. Elde edilen uzman görüşlerinin ortalama değeri alınarak araştırmada bu değerler üzerinden analizler yürütülmüştür. Daha sonra ChatGPT'den de test maddelerini benzer şekilde güçlük düzeylerine göre sınıflandırması istenmiştir. Tablo 2'de ChatGPT ve uzmanlar tarafından testteki bir maddeye dair güçlük düzeyi sınıflandırması ve bu sınıflandırmanın gerekçesini gösteren veriler sunulmuştur.

Tablo 2. ChatGPT'nin ve Uzmanların Madde Güçlük Sınıflandırmasını ve Gerekçesini Gösteren Bir Örnek

Madde 3:
Bilgi: Ortak elemanı olmayan kümelere ayrık kümeler denir. A ve B ayrık kümeler olmak üzere; s(AUB) = 13 s(A) = 3x-1 s(B) = x+2 ise **s(A) kaçtır?** A)3 B)5 C)8 D)10 E)12
<u>ChatGPT Sınıflandırması</u>: Bu soru orta düzeyde bir matematik sorusudur. Soru, küme teorisi temel prensiplerini anlama ve bu prensipleri kullanarak verilen koşullardan yararlanma becerisini ölçmektedir. Ayrıca soruda denklem çözme ve basit aritmetik işlemleri kullanma becerisi de gereklidir. Ancak soru oldukça açık ve net bir şekilde verildiği için zorlayıcı veya kafa karıştırıcı bir yanı yoktur. Dolayısıyla, bu soruyu orta düzeyde bir soru olarak sınıflandırabilirim.
<u>Uzman Sınıflandırması</u>: Soruda "Ayrık Küme" tanımı bilgi notu olarak verilmiştir. Dolayısıyla kümelerde temel işlem becerisine sahip olan bir öğrenci bu soruda isterse denklem kurarak isterse de Venn şeması çizerek çözüme ulaşabilir. Sorunun bilinmeyen bir ifade içermesi zorluk düzeyini bir miktar artırmaktadır. Ancak kümeler konusu ile ilgili temel bir soru kalıbı özelliği taşıması sebebiyle bu sorunun orta güçlükte olduğu düşünülmektedir.

Tablo 2 incelendiğinde testteki 3 numaralı maddeye ChatGPT'nin "orta" güçlük düzeyinde bir sınıflandırma yaptığı görülmektedir. ChatGPT'den maddeyi güçlük düzeyine göre sınıflandırma yapması istendiğinde sadece sınıflama yapmakla kalmamış aynı zamanda bu sınıflandırmanın gerekçesini de detaylı olarak ortaya koymuştur. Bu durum tüm test maddeleri için de benzer şekilde gerekçelendirilmiştir. Uzmanlar da bu madde için farklı değerlendirmeler yapmış ve maddeyi "orta" güçlük düzeyinde sınıflandırmıştır. Tablo 3'te 20 test maddesi için gerçek uygulamadan elde edilen madde güçlük indeksleri ile ChatGPT'nin ve uzmanların yaptığı güçlük düzeyi sınıflandırması verilmiştir.

Tablo 3. Test Maddelerinin Güçlük İndeksleri Sınıflandırması

Madde No	Madde Güçlük İndeksi	Madde Güçlük Düzeyleri	ChatGPT Değerlendirmesi	Uzman Görüşleri Ortalaması
1	0,81	Çok Kolay (5)	Çok Kolay (5)	5.00
2	0,49	Orta (3)	Kolay (4)	3.67
3	0,50	Orta (3)	Orta (3)	3.67
4	0,32	Zor (2)	Orta (4)	2.67
5	0,60	Orta (3)	Kolay (3)	4.33
6	0,43	Orta (3)	Kolay (4)	3.33
7	0,47	Orta (3)	Orta (3)	3.33
8	0,63	Kolay (4)	Kolay (4)	4.33
9	0,33	Zor (2)	Kolay (4)	2.67
10	0,41	Orta (3)	Kolay (4)	3.00
11	0,64	Kolay (4)	Çok Kolay (5)	4.67
12	0,48	Orta (3)	Orta (3)	2.67
13	0,52	Orta (3)	Orta (3)	3.67
14	0,52	Orta (3)	Orta (3)	2.67
15	0,45	Orta (3)	Orta (3)	3.00
16	0,72	Kolay (4)	Çok Kolay (5)	3.67
17	0,72	Kolay (4)	Çok Kolay (5)	4.00
18	0,31	Zor (2)	Kolay (4)	2.00
19	0,19	Çok Zor (1)	Orta (3)	3.00
20	0,26	Zor (2)	Orta (3)	2.67

Tablo 3 incelendiğinde bazı maddelerde hem uzman görüşleri hem ChatGPT değerlendirmesi madde güçlüklerini doğru sınıflamış olmasına rağmen göre iken bazı maddelerde ise sınıflandırmalar farklılık göstermektedir. Örneğin testteki 1. maddenin güçlük indeksi 0,81'dir ve bu madde "çok kolay" sınıfındadır. Hem ChatGPT hem de uzman görüşleri bu maddenin "çok kolay" sınıfında olduğu yönündedir. Testteki 18. maddenin ise güçlük indeksi 0,31'dir ve bu madde "zor" kategorisindedir. Ancak ChatGPT bu maddeyi "kolay" olarak sınıflandırmıştır. Gerçek madde güçlük düzeyleri, ChatGPT'nin belirlediği düzeyler ve uzman görüşlerinden elde edilen ortalama değerler arası güvenirliği belirlemek için sınıf içi korelasyon katsayıları hesaplanmıştır. İki yönlü karma yöntem ile ortalama ölçüm değerleri alınarak elde edilen sonuçlar Tablo 4'te verilmiştir.

Tablo 4. ChatGPT, Uzman Görüşü ve Madde Güçlük Düzeyleri Arasındaki Sınıf İçi Korelasyon Analizi

	Sınıf İçi	95% Güven Aralığı		F Testi Sonuçları			
	Korelasyon	Alt Sınır	Üst Sınır	Değer	df1	df2	Sig
Uzman - Madde Güçlüğü	,870	,673	,949	7,716	19	19	,000
ChatGPT- Madde Güçlüğü	,748	,363	,900	3,968	19	19	,002
ChatGPT- Uzman	,787	,462	,916	4,692	19	19	,001

Tablo 4 incelendiğinde, uzman görüşlerinden elde edilen madde güçlük düzeyi sınıflandırması ile gerçek test uygulamasından elde edilen madde güçlük düzeyleri arasındaki sınıf içi korelasyon katsayısı 0,870 olarak görülmektedir ($p<.05$). Bu değer uzman görüşlerine göre yapılan sınıflandırmanın gerçekte olan madde güçlük düzeyleri ile yüksek düzeyde benzerlik gösterdiği şeklinde yorumlanabilir. Benzer karşılaştırma ChatGPT tarafından yapılan madde güçlük düzeyi sınıflandırmasına göre incelendiğinde ise sınıf içi korelasyon değeri 0,748 olarak bulunmuştur ($p<.05$). ChatGPT sınıflama performansının uzman görüşlerine göre daha düşük düzeyde kaldığı ifade edilebilir. Ancak sınıf içi korelasyona dayalı puanlayıcı güvenirliğinde kabul edilebilir alt sınır 0,40'tır (Walter, Eliasziw ve Donner, 1998). Bu sebeple ChatGPT'nin yaptığı sınıflama performansının da yüksek düzeyde olduğunu belirtmek mümkündür. Son

olarak ChatGPT ve uzman görüşleri arasındaki güvenirlik incelenmiş ve bu değer 0,787 olarak bulunmuştur (p<.05).

SONUÇ, TARTIŞMA ve ÖNERİLER

Doğal dil işleme teknolojisindeki gelişmeler tüm alanlarda olduğu gibi eğitim alanında da çeşitli yenilikler ve fırsatlar sunmuştur. ChatGPT benzeri bu araçların eğitimin temel öğeleri olan öğretmen ve öğrencilere geri bildirim sunması, bireyselleştirilmiş öğrenmeye imkân tanıması, öğrenme ve öğretme sürecinde destekleyici bir mekanizma olarak kullanılabilmesi mümkündür. Son dönemlerde ChatGPT'nin farklı disiplinlerdeki yanıtlama becerisinin test edildiği birçok çalışmaya rastlanmaktadır. Literatürde hukuk, tıp, dil eğitimi gibi alanlarda eğitim sonundaki yeterlilik sınavlarında ChatGPT'nin gösterdiği başarılı performansı ortaya koyan araştırmalar mevcuttur (Choi vd., 2023; Kung vd., 2023; Ryznar, 2023).

Bu araştırmanın ilk aşamasında ChatGPT'nin 20 çoktan seçmeli matematik testi maddesine verdiği yanıtlar incelenmiştir. ChatGPT bu maddelerden 11 tanesine doğru yanıt verirken 9 maddeyi yanlış yanıtlamıştır ve 55 puan almıştır. ChatGPT, lise kademesinde bir dersten başarılı sayılabilmek için gerekli olan 50 puanın üzerinde bir başarı göstermiştir. Ancak ortalama madde güçlük indeksinin 0,49 olduğu bu test için yüksek düzeyde bir yanıtlama performansının elde edilemediği düşünülmektedir. Alanyazında bu araştırmadaki bulgulara benzer çalışmalar yer almaktadır. Frieder vd., (2023) çalışmalarında ChatGPT'nin matematiksel yeteneklerini test etmiştir. ChatGPT'nin soru yanıtlama, teorem arama gibi kullanım durumlarında matematikçiler için yararlı bir destekçi olup olmadığını ele almıştır. Ancak ChatGPT'nin matematiksel yeteneklerinin medyadaki olumlu haberlerin aksine ortalama bir yüksek lisans öğrencisinin bile altında kaldığı belirtilmiştir. ChatGPT'nin soruları genel olarak anladığı fakat bu sorulara doğru cevaplar üretmede yetersiz kaldığı ifade edilmiştir. Shakarian, Koyyalamudi, Ngu ve Mareedu (2023) araştırmasında ChatGPT'nin matematik problemlerine verdiği yanıtları farklı tekrarlar ile incelemiştir. ChatGPT'nin verdiği yanıtlara göre istem şekli değiştirilerek ve ek ifadeler eklenerek gerçekleştirilen sorgulamalarda ChatGPT'nin problemleri yanıtlamadaki başarısızlık oranının %84'ten %20'ye düştüğü ifade edilmiştir. Ayrıca problemlerdeki bilinmeyen sayısı ve işlem sayısı arttıkça ChatGPT'nin maddeleri yanıtlamadaki başarısızlık olasılığının da arttığı belirtilmiştir.

Araştırmanın ikinci aşamasında ChatGPT'nin ve uzmanların yaptığı madde güçlük düzeyleri sınıflandırma performansları incelenmiştir. Elde edilen sonuçlara göre ChatGPT, uzman sınıflamasının gerisinde kalsa da yine de gerçek madde güçlük sınıfları ile yüksek korelasyon gösteren bir sınıflandırma performansı ortaya koymuştur. Bu araştırmaya benzer bir çalışma Khademi (2023) tarafından yürütülmüştür. Araştırmada ChatGPT ve Google Bard araçlarının, yazma istemlerinin karmaşıklığını algılama ve derecelendirme konusunda uzman insanlara karşı güvenilirliği incelenmiştir. Hem ChatGPT hem de Google Bard'ın uzman puanlamasına kıyasla daha düşük güvenirlik düzeyinde olduğu sonucuna ulaşılmıştır.

Bu araştırmadan elde edilen sonuçlara göre ChatGPT veya benzer yapay zekâ teknolojilerinin eğitim alanında ve okul ortamında insan kontrolü altında destekleyici birer unsur olarak kullanılabileceği düşünülmektedir. Özellikle deneme uygulamasının yapılamadığı geniş ölçekli sınavlarda madde ve test hazırlama sürecinde ChatGPT'nin verdiği yanıtlardan faydalanılarak madde güçlükleri tahmin edilebilir. Ayrıca sınav güvenliği açısından uzman görüşlerine başvurulamadığı durumlarda da ChatGPT'den yararlanılabilir. Başka araştırmalarda madde sayısı ve uzman sayısı artırılarak farklı disiplinlerde benzer araştırmalar yürütülebilir. Yapay zekâ teknolojisindeki ilerlemeleri de gözlemlemek amacıyla benzer çalışmalar belirli aralıkta tekrarlanıp bu teknolojilerin maddelere verdikleri yanıtların ne ölçüde geliştiği incelenebilir. Ayrıca yapay zekâ teknolojisi ile istenilen konu alanı ve kazanıma göre otomatik madde üretimi yapılabilir.

KAYNAKÇA

Anıl, D. (2002). Deneme uygulamasının yapılamadığı durumlarda madde ve test parametrelerinin klasik ve örtük özellikler test teorilerine göre kestirilmesi. *Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Estitüsü, Ankara*.

Baykul, Y., & Sezer, S. (1993). Deneme yapılamayan durumlarda madde güçlük ve ayırıcılık gücü indekslerinin ve bunlara bağlı test istatiklerinin kestirilmesi. *Eğitim ve Bilim*, 17(83)

Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. Ankara: Pegem Akademi.

Bozkurt, A., Xiao, J., Lambert, S., Crompton, H., Koseoglu, S., Farrow, R., Bond, M., Nerantzi, C., Honeychurch, S., Bali, M., Dron, J., Mir, K., Stewart, B., Costello, E., Mason, J., Stracke, C., Romero-Hall, E., Koutropoulos, A., . . . Jandrić, P. (2023). Speculative futures on ChatGPT and Generative Artificial Intelligence (AI): A collective reflection Pazurek, A., from the educational landscape. *Asian Journal of Distance Education*, 18(1), 53-130. https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/709

Choi, J. H., Hickman, K. E., Monahan, A. B. & Schwarcz, D. (2023). ChatGPT Goes to Law School. Minnesota Legal Studies Research Paper No. 23-03.

CNN (2023). ChatGPT Passes Exams from Law and Business Schools. Available online: https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams (accessed on 10 March 2023).

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. USA:Harcourt Brace Javanovich College Publishers.

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

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (Vol. 7, p. 429). New York: McGraw-hill.

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., ... & Berner, J. (2023). Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*.

Güler, N., İlhan, M., & Taşdelen-Teker, G. (2021). Çoktan seçmeli maddelerde uzmanlarca öngörülen ve ampirik olarak hesaplanan güçlük indekslerinin karşılaştırılması. *Journal of Computer and Education Research*, 9(18), 1022-1036. DOI: 10.18009/jcer.1000934

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. Journal of Educational Measurement, 35(1), 69-81.

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.*

Khademi, A. (2023). Can ChatGPT and Bard Generate Aligned Assessment Items? A Reliability Analysis against Human Performance. *arXiv preprint arXiv:2304.05372*.

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.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics,* 159-174.

Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences, 13*(4), 410.

Lorge, I., & Diamon, L. K. (1954). The value of information to good and poor judges of item difficulty. Educational and Psychological Measurement, 14(1), 29–33. https://doi.org/10.1177/001316445401400103

OpenAI (2023). *Introducing OpenAI*. Erişim tarihi:08.05.2023. Erişim adresi: https://openai.com/blog/introducing-openai

Quereshi, M. Y., & Fisher, T. L. (1977). Logical versus empirical estimates of item difficulty. Educational and Psychologıcal Measurement, 37(1), 91–100. https://doi.org/10.1177/001316447703700110

Ryznar, M. (2023). Exams in the Time of ChatGPT. *Washington and Lee Law Review Online*, *80*(5), 305.

Sezer, S. (1992). Ön deneme yapılamayan durumlarda madde güçlük ve ayırıcılık gücü indekslerinin ve bunlara bağlı test istatistiklerinin kestirilmesi. *Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Estitüsü, Ankara*.

Shakarian, P., Koyyalamudi, A., Ngu, N., & Mareedu, L. (2023). An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP). *arXiv preprint arXiv:2302.13814*.

Sok, S., & Heng, K. (2023). *ChatGPT for education and research*: A review of benefits and risks. *Available at SSRN 4378735.*

Tinkelman, S. (1947). Difficulty prediction of test items. Teachers College Contributions to Education, 941, 55.

Urbina, S. (2014*). Essentials of psychological testing (2nd ed.).* Hoboken, New Jersey: Wiley

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in medicine*, *17*(1), 101-110.

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

The Journal of Interdisciplinary Educational Research

ISSN: 2602-2516
Journal of Interdisciplinary Educational Research
2023; 7(15);202-210

Comparison of Expert Opinions and ChatGPT Performance in Predicting Item Difficulties

Erdem BODUROĞLU[1] Oğuz KOÇ[2] Mahmut Sami YİĞİTER[3]

Cited:
Boduroğlu, E., Koç, O., Yiğiter, M. S.,(2023). Comparison of Expert Opinions and ChatGPT Performance in Predicting Item Difficulties. *Journal of Interdisciplinary Educational Research*. 7(15), 202-210, DOI: 10.57135/jier. 1296255

Abstract
In this study, ChatGPT's performance in answering multiple-choice test items and classifying the item difficulty levels of these items was examined. Item's actual difficulty levels were determined according to the responses of 4930 students to the five-choice multiple-choice test items consisting of 20 items. The relationships between these difficulty levels and the classifications made by ChatGPT and experts were tested. The findings demonsrated that ChatGPT's performance in giving correct answers to multiple-choice items was at moderate level (55%). However, in terms of classifying item difficulty levels, ChatGPT showed a correlation of 0.748 with actual item difficulty levels and 0.870 with expert opinions. According to these results, it is thought that ChatGPT can be used to support test development in cases where trial application cannot be conducted or expert opinions cannot be consulted. In large-scale exams, ChatGPT-like artificial intelligence technologies can be utilized under expert supervision.

Keywords: ChatGPT, artificial intelligence, item difficulties, expert opinion

INTRODUCTION

OpenAI (2023) is a non-profit artificial intelligence research company whose goal is to develop digital intelligence in a way that is most beneficial for humanity, without concern for financial return. In this context, the company's most famous product ChatGPT was launched on November 30, 2022. ChatGPT is a prototype chatbot trained in artificial intelligence and specializes in dialogue. ChatGPT is a large language model with the ability to generate contextually appropriate responses and converse naturally (Deng & Lin, 2022). The language model has been a huge success and has found applications in healthcare, education, and various other disciplines. For example, it has successfully completed four separate exams at the University of Minnesota Law School (CNN, 2023). Although the scores are not high for now, the results show that this artificial intelligence application is capable of obtaining a university degree (Choi, Hickman, Monahan, & Schwarcz, 2023).

Lo (2023) analyzed 50 studies conducted with ChatGPT and identified five main functions related to instructors in two themes. These were instructional preparation (developing materials, suggestions, translation, etc.) and assessment (creating tasks and evaluating performance). Although themes have been established in related studies, it is seen that there is still a limited number of studies on assisting both the teaching and assessment process. Zhai (2023) states that teachers can save time and effort by using ChatGPT for assessment. Since most teachers spend a lot of time preparing quizzes, monthly tests, and exams, educators can find the opportunity to

[1], Deputy Director, Niğde Measurement And Evaluation Center Republic Of Türkiye Ministry Of National Education, Niğde-Türkiye,erdemboduroglu@gmail.com, orcid.org/0000-0001-8318-4914

[2]Teacher, Niğde Measurement And Evaluation Center, Republic of Türkiye Ministry of National Education, Niğde-Türkiye, oguzkoc20@hotmail.com, orcid.org/0000-0002-8656-6069

[3] Lect., Social Sciences University Of Ankara, Distance Education Application And Research Center, Ankara-Türkiye, mahmutsami.yigiter@asbu.edu.tr, orcid.org/ 0000-0002-2896-0201

Received: 12/05/2023
Acceptance: 25/07/2023
Research Article

alleviate the pressure of assessment with the help of ChatGPT (Sok & Heng, 2023). ChatGPT can also offer an automated grading system with the necessary feedback to improve student's learning outcomes. ChatGPT can make grading semi-automated by identifying both weaknesses and strengths of students. In this way, it can help students work to be graded better (Kasneci et al., 2023). It can also help teachers save time by reducing their burden (Bozkurt et al., 2023).

Today, education is seen as a system that develops human behavior. Whether the components of this system are functioning or not, their deficiencies and strengths are revealed through measurement and evaluation (Baykul, 2015). Tests have an essential place in measurement and evaluation studies. The quality of a developed test is proportional to the quality of the prepared items. To reveal the quality of the items, various item indicators are calculated. Statistics such as item difficulty index, item discrimination index, and distractor analysis reveal the quality of the items. According to classical test theory, the item difficulty index, one of these indicators, is obtained by dividing the number of individuals who answered an item correctly by the number of all individuals in the group. In short, this index expresses the percentage of participants' correct answers to an item. It is expected that the average item difficulty index of the tests that will best distinguish between those who know and those who do not know should be around 0.50. If the difficulty index of an item is between 0.00-0.20, this item is interpreted as "very difficult", between 0.21-0.40 as "difficult", between 0.41-0.60 as "medium", between 0.61-0.80 as "easy" and between 0.81-1.00 as "very easy" (Baykul, 2015; Crocker & Algina, 1986; Urbina, 2014). When the literature is examined, it can be stated that there are many studies examining the compatibility between item difficulties determined based on expert opinions and actual item difficulty indices (Anıl, 2002; Baykul & Sezer, 1993; Güler, İlhan & Teker, 2021; Impara & Plake, 1998; Lorge & Diamon, 1954; Quereshi & Fisher, 1977; Tinkelman, 1947). However, it is seen that there are a limited number of studies in which artificial intelligence is used to determine item difficulty index. It can be said that studies on how ChatGPT will perform to develop qualified assessment and evaluation tools in cases where it is difficult to reach experts or where a trial application cannot be performed will contribute to the literature.

Purpose of the Study

In the test development process, the purpose and scope of the test are first determined and then the test items are prepared. Test items should be prepared for the general purpose of the test, be of a quality to reveal the observed behaviors, and be free from scientific and linguistic errors. To determine whether the items have these characteristics, expert opinions are consulted and trial applications are carried out (Baykul, 2015). In test development processes, some problems may be encountered in cases where trial applications cannot be conducted or expert opinions cannot be consulted due to exam security. In such cases, item and test statistics are tried to be estimated with the limited estimates of test developers or item writers and erroneous results may be obtained (Sezer, 1992). To minimize these problems, it was aimed to research whether ChatGPT technology can be used in a supportive role in the field of education. This study, it was aimed to examine the responses of ChatGPT to multiple-choice test items in the field of mathematics and to compare the performance of ChatGPT in classifying items according to item difficulty levels with expert opinions and actual item difficulty levels. For these purposes, answers to the following two sub-problems were sought in the study:

1- How is the performance of ChatGPT in answering multiple-choice test items?
2- In determining the item difficulty levels; what is the relationship between expert opinions, ChatGPT classification and actual item difficulty levels?

METHOD

Research Model

This research is a correlational research design, which is one of the quantitative research methods. Relational research aims to reveal the relationships between two or more variables as they are. Relational research is sometimes referred to as a type of descriptive research because it describes an existing relationship between variables (Fraenkel, Wallen, & Hyun, 2012). In this study, the relationships between the classification of item difficulty levels obtained according to the ChatGPT and expert opinions and the actual item difficulty indices were examined. For this reason, it can be stated that the study has a correlational research design.

Research Data

The item statistics used in the study were obtained from a 20-item, five-choice, multiple-choice mathematics test administered to 4930 9th-grade students in Niğde province. This test was developed by five domain experts, one language expert, and one measurement and evaluation expert. The test was piloted and the final items were created according to the test and item statistics obtained. The responses of the ChatGPT to these test items, which were developed by all the procedures of the test development process, and the item difficulty level classifications and item difficulty classifications made by three domain experts constitute the data of the study.

Data Collection Methods

To use the test items examined in the study and the item difficulty indices of these items, official permission was obtained from the Niğde Provincial Directorate of National Education and access to the data was provided. These data were obtained from a large-scale common exam administered to all 9th-grade students studying in Niğde province and serve as a criterion for other data to be used in the study. These multiple-choice items were given to the ChatGPT 3.5 application with their options and asked to answer these items. In addition, ChatGPT was asked to classify each answered item as Very Easy-Easy-Medium-Difficult-Very Difficult according to item difficulty levels and to justify. Afterward, expert opinions were consulted for item difficulty level classification. Three mathematics teachers working in public schools and competent in test development were selected as the expert team. These experts were asked to examine the 20 items in the test and classify the items according to their difficulty levels with their justifications. After this classification, the data collection process was completed.

Data Analysis

In the first sub-problem of the study, the answers given to 20 multiple-choice test items by ChatGPT 3.5 version on 01.03.2023 were recorded. The application was asked to respond to each item 2 times. If there was consistency between the answers, it was recorded as the final answer. In the opposite case, a new answer was requested. For item difficulty level classification, the prompt "Can you classify this item according to the difficulty level as Very Easy-Easy-Medium-Difficult-Very Difficult with the reasoning?" was entered into ChatGPT. The results obtained were tabulated. Similarly, difficulty classifications were obtained independently from the experts. Fleiss Kappa statistic was used to determine the agreement between the experts. The average score value was derived according to the classification between 1-5 made by the three experts. To determine the reliability between actual item difficulty levels, ChatGPT classification, and expert opinions, the intraclass correlation coefficient was calculated with the SPSS package program. The results obtained were reported at a 0.05 significance level and the data analysis phase was completed.

RESULTS

For the first sub-problem of the study, the answers given by ChatGPT to 20 multiple-choice items in the 9th-grade mathematics test were analyzed. The data showing the subject area of each item, the item difficulty index, and the correctness or incorrectness of the answers given by ChatGPT are presented in Table 1.

Table 1. Table of Subject Area, Difficulty Index and ChatGPT Responses for Test Items

Item Number	Subject Area	Item Difficulty Index	ChatGPT's Answers
1	Logic	0,81 (Very Easy)	False
2	Clusters	0,49 (Medium)	False
3	Clusters	0,50 (Medium)	Correct
4	Divisibility	0,32 (Difficult)	False
5	Equations	0,60 (Medium)	Correct
6	Absolute Value	0,43 (Medium)	Correct
7	Ratio and Proportion	0,47 (Medium)	Correct
8	Problems	0,63 (Easy)	Correct
9	Problems	0,33 (Difficult)	Correct
10	Problems	0,41 (Medium)	Correct
11	Logic	0,64 (Easy)	False
12	Clusters	0,48 (Medium)	False
13	Clusters	0,52 (Medium)	Correct
14	Clusters	0,52 (Medium)	Correct
15	Cartesian Product	0,45 (Medium)	False
16	Divisibility	0,72 (Easy)	Correct
17	Equations	0,72 (Easy)	Correct
18	Absolute Value	0,31 (Difficult)	False
19	Inequalities	0,19 (Very Difficult)	False
20	Divisibility	0,26 (Difficult)	False

When Table 1 is examined, it is seen that 1 item in the test is at "Very easy", 4 items are at "Easy", 10 items are at "Medium", 4 items are at "Difficult" and 1 item is at "Very Difficult" difficulty level. While 11 of the answers given to the 20 multiple-choice items of the ChatGPT were correct, 9 of them were incorrect. According to the table, there was no clear relationship between the item difficulty indices and the item being answered correctly. The Mann-Whitney U test was used to test whether there was a difference between the mean item difficulty indices of the items answered correctly and incorrectly. According to the results of the analysis, there was no significant difference between the groups at a 0.05 significance level (U=31, p=0.160). While item 1, the easiest item of the test, was answered incorrectly by ChatGPT, item 9 in the difficult category was answered correctly. Expert examinations were conducted regarding the incorrect answers and it was stated that various factors such as the number of unknowns in the item, the number of operation steps, and the reasoning skills required for the solution affected this situation. The average difficulty index for the 20 items in the test was 0.49. According to this index, it is possible to say that the test is at an average difficulty level. ChatGPT achieved a 55% success rate by answering 11 of the test items correctly. According to these data, it can be said that ChatGPT has approximately as many correct answering skills as an average achievement student participating in this test.

For the second sub-problem of the study, the test items were examined by 3 domain experts, and these items were independently classified according to their difficulty levels (Very Easy (5)-Easy (4)-Medium (3)-Difficult (2)-Very Difficult (1)). Fleiss Kappa statistic was calculated to determine the degree of agreement between the experts and this value was found to be 0.86. According to Landis and Koch (1977), this value indicates a very high level of agreement. The average value of the expert opinions obtained was taken and analyzes were conducted on these values in the study. Then, ChatGPT was asked to classify the test items according to their difficulty levels in a similar way. Table 2 presents the difficulty level classification of an item in the test by ChatGPT and experts and the data showing the rationale for this classification.

Table 2. An Example of ChatGPT's and Experts' Item Difficulty Classification and Rationale

Item 3:
Info: Sets with no elements in common are called discrete sets. Let A and B be discrete sets; If s(AUB) = 13 s(A) = 3x-1 s(B) = x+2 **What is s(A)?** A)3 B)5 C)8 D)10 E)12 **ChatGPT's Classification:** This is an intermediate-level math question. The question measures the ability to understand the basic principles of set theory and to use these principles to make use of given conditions. It also requires the ability to solve equations and use simple arithmetic operations. However, since the question is given in a very clear and straightforward way, nothing is challenging or confusing about it. Therefore, I would classify this question as an intermediate-level question. **Expert's Classification:** In the question, the definition of "Discrete Set" is given as an information note. Therefore, a student who has basic operation skills in sets can solve this question either by constructing an equation or by drawing a Venn diagram. The fact that the question contains an unknown expression increases the difficulty level to some extent. However, this question is considered to be of medium difficulty since it is a basic question pattern related to the subject of sets.

When Table 2 is examined, it is seen that ChatGPT made a classification at the "medium" difficulty level for item number 3 in the test. When ChatGPT was asked to classify the item according to the difficulty level, he not only made a classification but also provided a detailed justification for this classification. This situation was similarly justified for all test items. The experts also made different evaluations for this item and classified it at the "medium" difficulty level. Table 3 shows the item difficulty indices obtained from the actual administration and the difficulty level classifications made by ChatGPT and experts for 20 test items.

Table 3. Classification of Difficulty Indices of Test Items

Item Number	Item Difficulty Index	Item Difficulty Levels	ChatGPT's Review	Average of Expert Opinions
1	0,81	Very Easy (5)	Very Easy (5)	5.00
2	0,49	Medium (3)	Easy (4)	3.67
3	0,50	Medium (3)	Medium (3)	3.67
4	0,32	Difficult (2)	Medium (4)	2.67
5	0,60	Medium (3)	Easy (3)	4.33
6	0,43	Medium (3)	Easy (4)	3.33
7	0,47	Medium (3)	Medium (3)	3.33
8	0,63	Easy (4)	Easy (4)	4.33
9	0,33	Difficult (2)	Easy (4)	2.67
10	0,41	Medium (3)	Easy (4)	3.00
11	0,64	Easy (4)	Very Easy (5)	4.67
12	0,48	Medium (3)	Medium (3)	2.67
13	0,52	Medium (3)	Medium (3)	3.67
14	0,52	Medium (3)	Medium (3)	2.67
15	0,45	Medium (3)	Medium (3)	3.00
16	0,72	Easy (4)	Very Easy (5)	3.67
17	0,72	Easy (4)	Very Easy (5)	4.00
18	0,31	Difficult (2)	Easy (4)	2.00
19	0,19	Very Difficult (1)	Medium (3)	3.00
20	0,26	Difficult (2)	Medium (3)	2.67

When Table 3 is examined, it is seen that although both expert opinions and ChatGPT assessment correctly classified item difficulties in some items, the classifications differ in some items. For example, the difficulty index of item 1 in the test is 0.81 and this item is in the "very easy" class. Both ChatGPT and expert opinions suggest that this item is in the "very easy" class. The difficulty index of item 18 in the test is 0.31 and this item is in the "difficult" category. However, ChatGPT categorized this item as "easy". Intraclass correlation coefficients were calculated to determine

the reliability between the actual item difficulty levels, the levels determined by ChatGPT, and the mean values obtained from expert opinions. The results obtained by taking average measurement values with the two-way mixed method are given in Table 4.

Table 4. Intraclass Correlation Analysis between ChatGPT, Expert Opinion and Item Difficulty Levels

	Intraclass	95% Confidence Interval		F Test Results			
	Correlation	Lower	Upper	Estimate	df1	df2	Sig
Expert - Item Difficulty	,870	,673	,949	7,716	19	19	,000
ChatGPT- Item Difficulty	,748	,363	,900	3,968	19	19	,002
ChatGPT- Expert	,787	,462	,916	4,692	19	19	,001

When Table 4 is examined, the intraclass correlation coefficient between the item difficulty level classification obtained from expert opinions and the item difficulty levels obtained from the actual test administration is 0.870 ($p<.05$). This value can be interpreted as a high level of similarity between the classification made according to expert opinions and the actual item difficulty levels. When a similar comparison was analyzed according to the item difficulty level classification made by ChatGPT, the intraclass correlation value was found to be 0.748 ($p<.05$). It can be stated that ChatGPT classification performance was lower than the expert opinions. However, the acceptable lower limit for rater reliability based on intraclass correlation is 0.40 (Walter, Eliasziw, & Donner, 1998). For this reason, it is possible to state that the classification performance of ChatGPT is at a high level. Finally, the reliability between ChatGPT and expert opinions was examined and found to be 0.787 ($p<.05$).

CONCLUSION, DISCUSSION and RECOMMENDATIONS

Developments in natural language processing technology have provided various innovations and opportunities in the field of education as in all other fields. It is possible that these ChatGPT-like tools can be used as a supportive mechanism in the learning and teaching process by providing feedback to teachers and students, which are the basic elements of education, enabling individualized learning. Recently, there have been many studies testing ChatGPT's responsiveness in different disciplines. In the literature, some studies reveal the successful performance of ChatGPT in proficiency exams at the end of education in fields such as law, medicine, and language education (Choi et al., 2023; Kung et al., 2023; Ryznar, 2023).

In the first phase of this study, ChatGPT's responses to 20 multiple-choice math test items were analyzed. ChatGPT answered 11 of these items correctly, and 9 items incorrectly, and scored 55 points. ChatGPT scored above the 50 points required to be considered successful in a high school course. However, it is thought that a high level of answering performance could not be achieved for this test with an average item difficulty index of 0.49. There are studies in the literature similar to the findings of this study. Frieder et al. (2023) tested the mathematical abilities of ChatGPT in their study. They addressed whether ChatGPT is a useful supporter for mathematicians in use cases such as answering questions and searching theorems. However, contrary to the positive media reports, ChatGPT's mathematical abilities were reported to be below that of an average graduate student. It was stated that ChatGPT understood the questions in general but failed to produce correct answers to these questions. Shakarian, Koyyalamudi, Ngu, and Mareedu (2023) examined ChatGPT's responses to math problems with different repetitions. According to the answers given by ChatGPT, it was stated that the failure rate of ChatGPT in answering the problems decreased from 84% to 20% in the queries performed by changing the prompt form and adding additional expressions. It was also stated that as the number of unknowns and the number of operations in the problems increased, the probability of ChatGPT's failure in answering the items also increased.

In the second stage of the study, the item difficulty level classification performances of ChatGPT and experts were analyzed. According to the results obtained, although ChatGPT was behind the expert classification, it still showed a classification performance that was highly correlated with the actual item difficulty classes. A similar study to this research was conducted by Khademi (2023). In the study, the reliability of ChatGPT and Google Bard tools against people who are experts in perceiving and rating the complexity of writing prompts was examined. It was concluded that both ChatGPT and Google Bard had lower reliability levels compared to expert ratings.

According to the results obtained from this study, it is thought that ChatGPT or similar artificial intelligence technologies can be used as a supportive element under human control in the field of education and school environment. Item difficulties can be estimated by utilizing the answers given by ChatGPT in the item and test preparation process, especially in large-scale exams where the trial application cannot be done. In addition, ChatGPT can also be utilized in cases where expert opinions cannot be consulted in terms of exam security. Similar studies can be conducted in different disciplines by increasing the number of items and experts in other studies. To observe the advances in artificial intelligence technology, similar studies can be repeated at certain intervals and the extent to which these technologies improve the responses to the items can be examined. In addition, artificial intelligence technology can be used to automatically generate items according to the desired subject area and outcome.

REFERENCES

Anıl, D. (2002). Deneme uygulamasının yapılamadığı durumlarda madde ve test parametrelerinin klasik ve örtük özellikler test teorilerine göre kestirilmesi. *Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Estitüsü, Ankara*.

Baykul, Y., & Sezer, S. (1993). Deneme yapılamayan durumlarda madde güçlük ve ayırıcılık gücü indekslerinin ve bunlara bağlı test istatiklerinin kestirilmesi. *Eğitim ve Bilim*, 17(83)

Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. Ankara: Pegem Akademi.

Bozkurt, A., Xiao, J., Lambert, S., Crompton, H., Koseoglu, S., Farrow, R., Bond, M., Nerantzi, C., Honeychurch, S., Bali, M., Dron, J., Mir, K., Stewart, B., Costello, E., Mason, J., Stracke, C., Romero-Hall, E., Koutropoulos, A., . . . Jandrić, P. (2023). Speculative futures on ChatGPT and Generative Artificial Intelligence (AI): A collective reflection Pazurek, A., from the educational landscape. *Asian Journal of Distance Education*, 18(1), 53-130. https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/709

Choi, J. H., Hickman, K. E., Monahan, A. B. & Schwarcz, D. (2023). ChatGPT Goes to Law School. Minnesota Legal Studies Research Paper No. 23-03.

CNN (2023). ChatGPT Passes Exams from Law and Business Schools. Available online: https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams (accessed on 10 March 2023).

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. USA:Harcourt Brace Javanovich College Publishers.

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

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (Vol. 7, p. 429). New York: McGraw-hill.

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., ... & Berner, J. (2023). Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*.

Güler, N., İlhan, M., & Taşdelen-Teker, G. (2021). Çoktan seçmeli maddelerde uzmanlarca öngörülen ve ampirik olarak hesaplanan güçlük indekslerinin karşılaştırılması. *Journal of Computer and Education Research*, 9(18), 1022-1036. DOI: 10.18009/jcer.1000934

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. Journal of Educational Measurement, 35(1), 69-81.

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.*

Khademi, A. (2023). Can ChatGPT and Bard Generate Aligned Assessment Items? A Reliability Analysis against Human Performance. *arXiv preprint arXiv:2304.05372*.

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.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics,* 159-174.

Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences, 13*(4), 410.

Lorge, I., & Diamon, L. K. (1954). The value of information to good and poor judges of item difficulty. Educational and Psychological Measurement, 14(1), 29–33. https://doi.org/10.1177/001316445401400103

OpenAI (2023). *Introducing OpenAI*. Erişim tarihi:08.05.2023. Erişim adresi: https://openai.com/blog/introducing-openai

Quereshi, M. Y., & Fisher, T. L. (1977). Logical versus empirical estimates of item difficulty. Educational and Psychologıcal Measurement, 37(1), 91–100. https://doi.org/10.1177/001316447703700110

Ryznar, M. (2023). Exams in the Time of ChatGPT. *Washington and Lee Law Review Online*, *80*(5), 305.

Sezer, S. (1992). Ön deneme yapılamayan durumlarda madde güçlük ve ayırıcılık gücü indekslerinin ve bunlara bağlı test istatistiklerinin kestirilmesi. *Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Estitüsü, Ankara*.

Shakarian, P., Koyyalamudi, A., Ngu, N., & Mareedu, L. (2023). An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP). *arXiv preprint arXiv:2302.13814*.

Sok, S., & Heng, K. (2023). *ChatGPT for education and research*: A review of benefits and risks. *Available at SSRN 4378735.*

Tinkelman, S. (1947). Difficulty prediction of test items. Teachers College Contributions to Education, 941, 55.

Urbina, S. (2014*). Essentials of psychological testing (2nd ed.).* Hoboken, New Jersey: Wiley

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in medicine*, *17*(1), 101-110.

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