Eğitim ve Bilim

2025 Ek Sayı 1 1-18

Orijinal Makale

Öğrencilerin El Yazısıyla Yanıtladığı Açık Uçlu Maddelerin Puanlanmasında Yapay Zekâ Performansının İncelenmesi *

Mahmut Sami Yiğiter [1], Erdem Boduroğlu [2]

Öz

Öğrenci başarılarının değerlendirilmesinde yüzyıllardır bir ölçme yöntemi olarak kullanılan açık uçlu maddeler, üst düzey becerilerin ölçülmesi, öğrenci hakkında zengin tanısal bilgi sağlaması, şans başarısının olmaması gibi pek çok avantaja sahiptir. Fakat günümüzde açık uçlu maddeler, puanlama işlemine hata karışabilmesi ve emek, zaman ve para açılarından dezavantajlı olması sebebiyle fazla sayıda öğrencinin katıldığı sınavlarda kullanılamamaktadır. Bu noktada Yapay Zekâ (YZ) açık uçlu maddelerin puanlanmasında önemli bir potansiyel içermektedir. Bu çalışmanın amacı, öğrencilerin açık uçlu maddelere el yazısıyla verdiği yanıtların puanlanmasında YZ'nin puanlama performansını incelemektir. Araştırmada bir devlet üniversitesinde Eğitimde Ölçme ve Değerlendirme dersi kapsamında 3 açık uçlu ve 10 çoktan seçmeli maddeden oluşan bir başarı testi geliştirilmiştir. Açık uçlu maddeler yanıtı yapılandırılmış biçimde (0-1-2) puanlanırken, çoktan seçmeli maddeler doğru-yanlış (0-1) şeklinde puanlanmıştır. 84 katılımcının yer aldığı çalışmada yer alan açık uçlu maddeler uzman grubu ve YZ aracı (ChatGPT-4o) tarafından puanlanmıştır. YZ aracına öğrencilerin el yazıları ile yazdıkları görsel yanıtlar iki farklı senaryoda puanlatılmıştır. Birinci senaryoda YZ'ye herhangi bir puanlama ölçütü verilmeden YZ aracının puanlama yapması istenirken, ikinci senaryoda standart puanlama ölçütlerine göre YZ'den puanlama yapması istenmiştir. Araştırmanın bulguları, YZ ile ölçütsüz puanlar ile uzman puanları arasında düşük uyum ve ilişki katsayıları olduğunu gösterirken, YZ ile standart ölçütle puanlama ve uzman puanlamaları arasında yüksek uyum ve ilişki katsayıları olduğu görülmüştür. Bu bulgulara benzer şekilde, YZ ile ölçütsüz puanlamanın madde ayırt edicilikleri oldukça düşük iken, YZ ile standart ölçütle puanlamanın madde ayırt edicilikleri yüksektir. Araştırmada ayrıca uzman puanları ve YZ ile standart ölçütlü puanları arasındaki uyumsuzlukların nedenleri de araştırılmış ve raporlanmıştır. Sonuçlar, YZ'nin standart puanlama ölçütleriyle el yazısıyla yanıtlanmış açık uçlu maddeleri iyi düzeyde puanlayabildiğini göstermektedir. Gelecekte YZ'nin gelişim ve dönüşümüyle birlikte tutarlılık açısından uzman puanlayıcılarla karşılaştırılabilir puanlama doğruluğuna ulaşabileceği düşünülmektedir.

Anahtar Kelimeler

Açık uçlu madde
Yapay zekâ
YZ
ChatGPT
Otomatik puanlama
El yazısı yanıtlar
Yapılandırılmış yanıtlı madde

Makale Hakkında

Gönderim Tarihi: 16.10.2024
Kabul Tarihi: 07.01.2025
Elektronik Yayın Tarihi: 03.03.2025

DOI: 10.15390/EB.2025.14119

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

[1] Ankara Sosyal Bilimler Üniversitesi, Uzaktan Eğitim Uygulama ve Araştırma Merkezi, Türkiye, mahmutsamiyigiter@gmail.com

[2] Millî Eğitim Bakanlığı, Niğde Ölçme Değerlendirme Merkezi, Türkiye, erdemboduroglu@gmail.com

Giriş

Eğitimde ölçme ve değerlendirme, eğitim sistemlerinde öğrenci öğrenmeleri hakkında dönütler sağlayan, öğrencilerin üst öğrenime yerleşmesi için kanıt sunan ve eğitim politikalarına rehberlik eden kritik görevler üstlenmektedir (Lohman, 1993). Öğrencilerin başarılarını belirlemede çeşitli değerlendirme yöntemleri bulunmaktadır. Öğrencilerin bağımsız yanıt üretmelerini sağlayan açık uçlu maddeler, uzun yıllardır kullanılan ve önemli bir yere sahip olan ölçme aracıdır (Freedman, 1994). Önceden belirlenmiş yanıtlar sunan çoktan seçmeli maddelerin aksine, açık uçlu maddeler öğrencilerin kendi eşsiz yanıtlarını üretmelerini gerektirir ve bu da öğrencilerin bilgi, bilişsel süreç ve yeteneklerinin daha derinlemesine anlaşılmasını sağlar (Agustianingsih ve Mahmudi, 2019; Doğan, 2019). Fakat, açık uçlu maddelerin adil bir şekilde değerlendirilmesi oldukça zor olduğundan ve puanlama için zaman ve emek gerektirmesinden dolayı ülke genelinde yapılan sınavlarda ve geniş ölçekli değerlendirmelerde kullanılamamaktadır (Karadag, Boz Yuksekdag, Akyildiz ve Ibileme, 2020). Sınıf içi ölçmelerde kullanılan açık uçlu maddelerde ise öğretmenler zaman ve emek harcamakta, zaman zaman da öznel puanlama yapıldığından dolayı eleştirilmektedir (Baykul ve Turgut, 2012). Yapay Zeka'nın (YZ) sunduğu puanlama desteği sayesinde bu iki sorunun üstesinden geleceği düşünülmektedir (Gao, Merzdorf, Anwar, Hipwell ve Srinivasa, 2024).

Günümüzde dijital teknolojiler, eğitim süreçlerinin birçok alanında köklü değişikliklere yol açmakta ve bu süreçlerin yeniden yapılandırılmasını gerektirmektedir (Beksultanova, Vatyukova ve Yalmaeva, 2020; Senkivska, 2022). Bu değişim, değerlendirme yöntemleri üzerinde de etkisini göstermekte olup, özellikle YZ tabanlı araçların eğitimde kullanım alanları genişlemektedir (Owan, Abang, Idika, Etta ve Bassey, 2023). Öğrenci performanslarının hızlı ve güvenilir bir şekilde değerlendirilmesine katkı sağlayan YZ, eğitimde ölçme ve değerlendirme yöntemlerini dönüştürme potansiyeline sahiptir. YZ'nin bu alandaki potansiyelinin yeterince anlaşılabilmesi için uzman puanlayıcılarla karşılaştırılması önem arz etmektedir (Chen, Chen ve Lin, 2020).

Açık uçlu maddeler, öğrencilerin sadece bilgiyi hatırlama değil, aynı zamanda bu bilgiyi uygulama, analiz etme ve kendi ifadeleriyle aktarma becerilerini ortaya koymalarına olanak tanıyan önemli bir ölçme aracıdır (Badger ve Thomas, 2019; Geer, 1988). Bu tür soruların değerlendirilmesi, çoktan seçmeli maddelere kıyasla daha fazla zaman ve uzmanlık gerektirdiğinden, öğretmenler açısından büyük bir iş yükü yaratmaktadır. Öğretmenlerin bu iş yükünü hafifletmek amacıyla YZ tabanlı araçların kullanımı, özellikle büyük öğrenci gruplarının olduğu ortamlarda giderek yaygınlaşmaktadır. Ancak, açık uçlu maddelerin puanlanmasında YZ'nin performansı ve güvenilirliği, hâlâ literatürde tartışılmaya devam eden bir konudur (Fernandez vd., 2022; Yaneva, Baldwin, Jurich, Swygert ve Clauser, 2023).

YZ'nin, çoktan seçmeli maddelerin puanlanmasında etkinliği geniş çapta kabul edilse de, açık uçlu maddelerin puanlanmasında karşılaşılan zorluklar daha karmaşık bir yapıya sahiptir. Açık uçlu maddeler, öğrencilerin düşüncelerini ifade etme biçimleri açısından çeşitlilik gösterdiği için, bu tür soruların değerlendirilmesinde standart bir kriterin geliştirilmesi zor olabilir (Sychev, Anikin ve Prokudin, 2020). Bu noktada, YZ'nin özellikle el yazısıyla yanıtlanan açık uçlu maddeleri ne ölçüde doğru ve tutarlı bir şekilde değerlendirebildiği kritik bir sorudur. YZ'nin bu soruları değerlendirirken uzman puanlayıcılar kadar tutarlı sonuçlar verip veremeyeceği önemli bir araştırma problemidir (Lin vd., 2020).

Açık Uçlu Maddeler

Açık uçlu maddeler, katılımcıların önceden tanımlanmış seçeneklerle veya belirli bir formatla sınırlandırılmadan kendi fikirleri ve düşünceleri doğrultusunda kendi sözcükleriyle yanıt vermelerine olanak tanıyan bir soru türüdür (Karadag vd., 2020). Katılımcıların düşüncelerini, görüşlerini veya deneyimlerini derinlemesine ifade etmelerini sağlayarak detaylandırmayı, keşfetmeyi ve düşünmeyi teşvik etmek açık uçlu maddelerin öne çıkaran özellikleridir (Sarwanto, Fajari ve Chumdari, 2021; Suherman ve Vidákovich, 2022). Açık uçlu maddeler, öğrencilerin öğrenmelerini tespit etmede diğer ölçme değerlendirme yöntemlerine göre pek çok avantaja sahiptir ve değerlendirme ve yaratma gibi üst düzey düşünme becerilerinin ölçülmesini sağlar (Brookhart, 2010). Öğrencileri, düşüncelerini ifade etmeye, akıl yürütmelerini gerekçelendirmeye ve kavrayışlarının kapsamını göstermeye zorlar. Açık

uçlu maddelerin sağladığı bu anlayış derinliği, öğrencilerin doğru cevabı gerçek bir anlayıştan ziyade tahmin veya eleme yoluyla seçebildiği çoktan seçmeli formatlarda genellikle çok azdır. Diğer taraftan, açık uçlu maddelerin yapısı yaratıcılığı ve eleştirel düşünmeyi teşvik eder (Fitriyah, Wahyudin, Suhendra, Nurhayati ve Febrianti, 2024; Monrat, Phaksunchai ve Chonchaiya, 2022; Winarso ve Hardyanti 2019). Ayrıca açık uçlu maddeler, öğrencilere soruya çeşitli açılardan yaklaşma, özgün yanıtlar geliştirme ve yenilikçi çözümler sunma özgürlüğü verir ve öğrencilerin entelektüel keşfini destekler (Kartikasari, Usodo ve Riyadi, 2022; Septiani, Retnawati ve Arliani, 2022).

Açık uçlu maddeler, öğrenci hakkında zengin tanısal bilgiler sağlar. Öğretmenler, öğrencilerin kavram yanılgıları, problem çözme stratejileri ve düşüncelerinin mantıksal akışı hakkında bilgi edinebilirler ve ayrıntılı geri bildirim sağlayabilirler (Karakaya, 2022). Buna karşılık, çoktan seçmeli maddeler genellikle öğrencilerin seçimlerinin altında yatan mantığı ortaya çıkaramaz ve tanısal faydalarını sınırlar. Ayrıca, açık uçlu maddeler otantik değerlendirme uygulamalarıyla daha uyumludur. Bireyler, önceden tanımlanmış seçeneklerle kısıtlanmadan yanıtlar, çözümler veya açıklamalar ürettirmeleri gereken gerçek dünya görevlerini simüle ederler. Bu da açık uçlu maddeleri, öğrencilerin gerçek hayattaki zorluklara karşı hazırlıklı olma ve pratik durumlarda bilgiyi uygulama becerilerini değerlendirmede faydalı kılar. Açık uçlu maddeler, kısmi puanlanabildiğinden (örn. 0-1-2) öğrenci başarısının değerlendirilmesinde çoktan seçmeli maddelere (0-1) göre daha fazla bilgi sağlar ve sınavın geçerliğini arttırır (Karimi, 2014). Açık uçlu maddelerde şans başarısı olmadığından öğrenciler haksız puanlar elde edemezler. Ayrıca açık uçlu maddelerde kopya oranı daha düşüktür (Abdolreza Gharehbagh, Mansourzadeh, Montazeri Khadem ve Saeidi, 2022). Açık uçlu maddelerin seçenek içermemesi hatırlama etkisini ortadan kaldırır. Seçenek içeren çoktan seçmeli maddelerde ise öğrenciler seçeneklerdeki çağrışımlar ile doğru cevabı hatırlayabilir veya seçeneklerden yardım alarak deneme-yanılma yolu ile doğru cevaba erişebilir. Açık uçlu maddelerin burada sıralanan önemli avantajlarının ötesinde öğrencilerin gerçek performanslarını keşfetmede birçok avantajı vardır.

Açık Uçlu Maddelerin Dezavantajları ve Puanlama Zorlukları

Açık uçlu maddelerin puanlanması, değerlendirmelerin güvenirliğini ve verimliliğini önemli ölçüde etkileyebilecek çeşitli zorluklara ve dezavantajlara sahiptir. Bu zorluklar, açık uçlu yanıtların doğasında var olan karmaşıklıktan ve uzman puanlayıcıların öznel doğasından kaynaklanabilmektedir. Bu zorluk ve dezavantajlardan birincisi, öznellik ve puanlayıcı yanlılığıdır (Hogan ve Murphy, 2007; Karakaya, 2022). Açık uçlu maddelerin yanıtlarını puanlamak için genellikle insan değerlendiricilere ihtiyaç duyulur ve bu değerlendiriciler kişisel yanlılıklarından etkilenerek puanlamada tutarsızlıklara yol açabilir. Bu öznellik sebebiyle, farklı puanlayıcılar tarafından verilen puanlar farklılaşabilir ve tutarsızlıklar oluşabilir. Bu durum puanlamanın güvenirliğini azalabilir (Güler, 2014; Maris ve Bechger, 2006). Puanlayıcıların uzmanlıkları ve öznel yargıları da öğrenci yanıtlarına verilen puanları etkileyebilir (Baburajan, de Abreu e Silva ve Pereira, 2022). İkinci olarak ise açık uçlu maddelerin puanlanması; emek, zaman ve maliyet açısından ekonomik değildir. Uzman değerlendiricilerin her yanıtı dikkatle okuyup puanlaması gerekir ki bu da özellikle büyük ölçekli değerlendirmelerde hem maliyetli hem de zahmetli olabilir. Bu durum sadece eğitimcilerin iş yükünü artırmakla kalmaz, aynı zamanda geri bildirim sürecini de geciktirebilir (Aznar-Mas, Atarés Huerta ve Marin-Garcia, 2023; Sychev vd., 2020). Üçüncüsü, açık uçlu maddelerin yanıtlarının puanlanmasının karmaşıklığıdır. Öğrencilerin açık uçlu maddelere verdiği yanıtlar, öğrencinin soruları kendi düşünce ve fikirleriyle yanıtlamasını gerektirir. Bu nedenle öğrenciler, soruları bilişsel şemalarında anlamlandırdığı düşünce ve fikirleri doğrultusunda yanıtlar ve bu nedenle bu tür yanıtlar daha karmaşık olabilir. Otomatik sistemlerle puanlanan çoktan seçmeli maddelerin aksine, açık uçlu maddeler öğrencilerin cevaplarındaki anlamı, nüansı ve bağlamı yakalamak için daha detaylı analizler gerektirir. Bu durumda yanıtların doğruluğunu ve kalitesini doğru bir şekilde puanlamak için standart puanlama formlarının hazırlanması gerekir. Standart puanlama formlarının hazırlanmasındaki zorluklar da açık uçlu maddelerin değerlendirilmesinde yaşanan başka bir güçlüktür. Özetle, açık uçlu maddeler öğrencilerin anlama ve bilişsel becerileri hakkında daha derin bir bilgi sağlarken, puanlayıcı değişkenliği, zaman ve emek açısından potansiyel dezavantajlara sahiptir. Açık uçlu maddelerin bu zorlukları ve dezavantajları ise gelişmekte olan üretken YZ sistemleri ile en aza indirilebilir (Mizumoto ve Eguchi, 2023; Pinto vd., 2023).

Açık Uçlu Maddelerin YZ ile Otomatik Puanlanması

YZ, gelişmiş doğal dil işleme (NLP) ve makine öğrenmesi (ML) teknikleriyle açık uçlu maddeleri puanlamaktadır (Beiting-Parrish ve Whitmer, 2023). Puanlama süreci, görüntü işlemeyi ve tahmin analitiğini birleştiren birkaç aşamayı içermektedir. Yanıtlar el yazısıyla yazılmışsa, bunları metne dönüştürmek için optik karakter tanıma (OCR) teknolojisi kullanılmaktadır. OCR sistemleri, özellikle karmaşık el yazısı stilleriyle uğraşırken yüksek doğruluk elde etmek için derin öğrenme algoritmaları kullanır. Karakter tanımadaki hatalar, doğru ve okunaklı metin çıktısı sağlamak için bu aşamada düzeltilir. Metin, okunduktan sonra gelişmiş NLP teknikleri aracılığıyla anlamsal analize tabi tutulur (Jescovitch vd., 2021). Word2Vec, GloVe ve BERT gibi dil modelleri, yanıtları hem sözcük düzeyinde hem de bağlamsal çerçeveleri içinde analiz eder. Bu modeller, yapay zekâ sistemlerinin yalnızca yüzeysel içeriği değil, aynı zamanda yanıtların daha derin anlamlarını da değerlendirmesini sağlar (Zhang ve Yuan, 2022). Ardından, YZ, puanlama kriterlerini ana hatlarıyla belirten önceden tanımlanmış rubrik veya puanlama araçlarına göre yanıtları puanlar. Bu aşamada sınıflandırma veya regresyon modelleri kullanılır (Jamil ve Hameed, 2023).

Literatür Taraması

Bu bölümde YZ ile puanlama üzerine literatürde yer alan bazı çalışmaların dikkat çekici bulgularına yer verilmiştir. Alers, Malinowska, Meghoe ve Apfel (2024), öğrenci yanıtlarını puanlamak için GPT-4 modelinin performansını incelediği çalışmasında 105 öğrencinin yanıtları ile YZ ve insan puanlarını karşılaştırdı. Bazı uyumsuzluklar olmasına karşın iki puanlama arasında güçlü bir korelasyon olduğunu belirtmektedir. Araştırmacılar ayrıca YZ teknolojilerinin puanlamayı önemli derecede hızlandırabileceğini raporlamaktadır. Jukiewicz (2024), öğrencilerin programlama görevlerini puanlamada ChatGPT ve insan puanlamasını karşılaştırmıştır. Bu çalışma, insan puanlarının ChatGPT puanlarına göre daha yüksek olduğunu, fakat notlar arasında yüksek bir korelasyon olduğunu bildirmektedir. Poole ve Coss (2024), ChatGPT'nin ikinci dilde yazılan kompozisyonların değerlendirilmesindeki etkinliğini uzman ve YZ aracını karşılaştırarak araştırmıştır. Araştırmacılar ayrıca farklı promptların etkililiğini de araştırmıştır. Bu araştırmanın bulguları, ChatGPT'ye puanlama kriterleri ve örnek yanıtlar sunuldukça puanlama kalitesinin iyileştiğini göstermiştir. Demir (2023), öğrencilerin açık uçlu maddelere verdiği yanıtların puanlanmasında uzman puanları ve YZ puanları arasındaki tutarlılığı incelemiştir. Bu araştırmanın bulguları, açık uçlu maddelerin puanlanmasında ChatGPT ve uzman puanları arasında yüksek düzeyde uyuşma ve korelasyon değerlerin olduğunu ifade etmektedir. Quah, Zheng, Sng, Yong ve Islam (2024), diş hekimliği lisans öğrencilerinin sınavlarını üç farklı uzmana ve YZ aracına (ChatGPT) değerlendirmiştir. Bu araştırmanın bulguları YZ aracının uzmanlara göre orta düzeyde korelasyona sahip olduğunu, ayrıca YZ'nin daha katı puanlama eğiliminde olduğunu ve konuyla ilgisiz veya yanlış içeriklere sıfır puan verme yeteneğine sahip olmadığını belirtmektedir (Quah vd., 2024). Başka bir araştırma ise el yazısı stillerinin değişkenliğinin ve standartlaştırılmış cevap formatlarının eksikliğinin puanlama sürecini daha da karmaşık hale getirdiğini ve potansiyel hatalara yol açtığını belirtmektedir (Lu, Zhou ve Ji, 2021). Sonuç olarak, el yazısı ile yanıtlanan açık uçlu maddelerin puanlanmasında YZ'nin performansının incelenmesi üzerine az sayıda çalışma olduğu ve halen konunun tam olarak açıklığa kavuşturulamadığı görülmektedir.

Araştırmanın Amacı ve Önemi

Bu çalışmanın amacı, öğrencilerin açık uçlu maddelere el yazısıyla verdikleri yanıtların puanlanmasında YZ'nin etkinliğini araştırmaktır. Bu kapsamda YZ'nin el yazısıyla yanıtlanan açık uçlu maddeleri puanlama performansını iki farklı senaryo altında incelenmiştir. İlk senaryoda, YZ'ye herhangi bir puanlama ölçütü verilmeden, yalnızca öğrenci yanıtlarını kendi algoritmasına göre puanlaması istenmiştir. İkinci senaryoda ise, YZ'ye standart puanlama ölçütleri sunulmuş ve bu kriterlere dayanarak puanlama yapması talep edilmiştir. Her iki durumda da YZ'nin verdiği puanlar, uzman puanlayıcılar tarafından verilen puanlarla karşılaştırılmıştır. Mevcut araştırmanın bulguları, YZ'nin özellikle standartlaştırılmış puanlama ölçütleri kullanıldığında ne derece tutarlı sonuçlar verebildiğini ortaya koyacaktır. Aynı zamanda, uzman puanlayıcılarla karşılaştırıldığında YZ'nin puanlamadaki farklarını ve bu farklılıkların anlamlı olup olmadığını analiz ederek, YZ'nin eğitimde daha yaygın ve güvenilir bir şekilde kullanılması için gereken iyileştirme alanlarına ışık tutacaktır. YZ'nin bu alanda sağlayacağı katkılar, eğitimde değerlendirme süreçlerinin daha hızlı ve daha etkili hale getirilmesine olanak tanıyabilir.

Açık uçlu maddeler, üst düzey bilişsel becerileri değerlendirme ve kapsamlı tanısal geri bildirim sağlama gibi avantajlarından dolayı uzun yıllardır kullanılmaktadır. Bu avantajlarına rağmen, puanlamaya öznel yargıların karışabilmesi ve puanlama sürecinin emek, zaman ve para gerektirmesi gibi dezavantajlara sahip olması nedeniyle büyük ölçekli değerlendirmelerde kullanılmamaktadır. Bu araştırma, YZ'nin verimli ve tutarlı puanlama sunma potansiyelini araştırarak, açık uçlu maddelerin büyük ölçekli değerlendirmelerde daha yaygın bir şekilde uygulanmasına ışık tutabilecektir. Ayrıca bu araştırmada öğrencilerin el yazısıyla yazarak sorulara verdiği yanıtlara odaklanılmıştır. Türkiye'de büyük örneklemler üzerinde uygulanan sınavların çoğunluğu bu araştırmada olduğu gibi halen kağıt-kalem formatında uygulanmaktadır (örn. LGS, ALES, YKS vb.). Dolayısıyla bu araştırmanın bulguları Türkiye'nin gelecekte açık uçlu maddeleri büyük ölçekli değerlendirmelere entegre etmesi için yol gösterici olabilir.

Araştırma Soruları

Araştırmada yanıt aranan sorular aşağıda sunulmuştur:

1. YZ aracına puanlama ölçütleri verilmediğinde, YZ ile uzman puanlayıcılar arasındaki puanlama uyumu ne düzeydedir?
2. YZ aracına standart puanlama ölçütleri verildiğinde, YZ ile uzman puanlayıcılar arasındaki puanlama uyumu ne düzeydedir?
3. Uzman puanlayıcılarla YZ'nin ölçütsüz puanlamaları arasında istatistiksel olarak anlamlı bir fark var mıdır?
4. Uzman puanlayıcılarla YZ'nin standart puanlama ölçütlerine göre yaptığı puanlamalar arasında anlamlı bir fark var mıdır?
5. Uzman puanlayıcılardan, YZ'nin ölçütsüz puanlamasından ve YZ'nin standart puanlama ölçütlerine göre yaptığı puanlamadan elde edilen madde istatistikleri nasıldır?
6. YZ'nin uzman puanlarına göre hatalı puanlama nedenleri nelerdir?

Yöntem

Araştırmanın Türü

Bu çalışma, YZ'nin el yazısıyla verilen açık uçlu maddelerin puanlanmasındaki performansını incelemeye yönelik betimsel bir araştırma olarak yapılandırılmıştır. Betimsel araştırmalar, bir durum veya olayı herhangi bir müdahale olmadan olduğu gibi tanımlamayı amaçlar (Karasar, 2012).

Örneklem

Araştırmanın örneklemi, bir devlet üniversitesinde Eğitimde Ölçme ve Değerlendirme dersi alan lisans düzeyindeki 84 öğrenciden oluşmaktadır. Örneklem, araştırmanın amacına uygun bilgiye sahip bireylerin seçildiği bir yöntem olan uygun örnekleme yöntemi ile belirlenmiştir (Patton, 2002). Bu örneklemde yer alan katılımcı öğrencilerin açık uçlu maddelere verdikleri el yazısı yanıtlar, analizler için veri kaynağı olarak kullanılmıştır.

Veri Toplama Aracı

Araştırma verilerinin elde edilmesi amacıyla üç açık uçlu ve on çoktan seçmeli olmak üzere toplamda 13 maddeden oluşan bir başarı testi geliştirilmiştir. Başarı testinde yer alan açık uçlu maddeler yapılandırılmış yanıt (0=Yanlış, 1=Kısmi doğru, 2=Tam doğru) puanlama sistemiyle oluşturulurken, çoktan seçmeli maddeler iki kategorili (0=Yanlış, 1=Doğru) puanlama sistemi ile puanlanacak şekilde geliştirilmiştir. Başarı testinin kapsam geçerliğinin sağlanması için öncelikle belirtke tablosu oluşturulmuş ve belirlenen yedi konunun ağırlıklarına göre soruların dağılımı sağlanmıştır. Açık uçlu maddeler farklı konularda yer alacak şekilde hazırlanmıştır. Ayrıca açık uçlu maddelerin hazırlanmasında madde türü hesaplamalı veya açıklamalı madde olacak şekilde yapılandırılmıştır. Hesaplamalı madde, sorunun çözümünde öğrencinin matematiksel işlemler

yapmasını gerektiren sorular olarak; açıklamalı madde ise sorunun çözümünde öğrencinin sorunun çözümüne yönelik sözel ifadelerle açıklama yapması beklenen soru olarak tanımlanabilir. Açık uçlu maddelerin birincisi hesaplamalı, ikincisi hesaplamalı + açıklamalı (sorunun iki seçeneğinin birinde hesaplama, diğerinde ise açıklama isteniyor), üçüncüsü ise açıklamalı madde türünde hazırlanmıştır. Açık uçlu maddelerin yenilenmiş Bloom taksonomisine göre bilişsel düzeyleri ise Soru 1,2 ve 3 için sırasıyla uygulama, uygulama ve analiz etmeydi. Ayrıca açık uçlu maddeler için standart puanlama yönergesi hazırlanmıştır. Hazırlanan başarı testi ve standart puanlama yönergeleri soruların bilimsel açıdan doğruluğu, okunabilirliği ve kapsam geçerliği açısından değerlendirilmek üzere üç uzmana gönderilmiş ve uzman görüşü alınmıştır. Uzman görüşleri doğrultusunda sorular ve standart puanlama yönergeleri üzerinde revizyonlar gerçekleştirilmiş ve Başarı Testi'nin nihai formu oluşturulmuştur. Formun geliştirme sürecinin ardından Başarı Testi'nin nihai formu, katılımcı öğrencilere kağıt-kalem formatında uygulanmıştır. Uzman puanlayıcılar, ölçme ve değerlendirme alanında doktora derecesine sahip iki alan uzmanından oluşmaktadır. Başarı testinin uygulanmasından sonra iki uzman puanlayıcının yaptığı puanlamaların güvenirliği, uyum ve tutarlılık analizleriyle test edilmiş ve yüksek düzeyde uyum olduğu görülmüştür (bkz. Tablo 1), bu bulgu çalışmanın güvenirliğini desteklemektedir.

Verilerin Analizi

Araştırmada iki farklı puanlama senaryosu uygulanmıştır. Birinci senaryoda (Senaryo-1: YZ ile Ölçütsüz Puanlama), YZ aracına herhangi bir puanlama kriteri sunulmaksızın öğrenci yanıtlarını puanlaması istenmiştir. İkinci senaryoda (Senaryo-2: YZ ile Standart Ölçütlerle Puanlama) ise YZ'ye önceden hazırlanan standart puanlama yönergeleri sunularak bu yönergedeki kriterler doğrultusunda yanıtları puanlaması istenmiştir. Tüm YZ puanlamaları ChatGPT-4o modelinin "gpt-4o-2024-08-06" sürümü ile 17 Eylül 2024 - 20 Eylül 2024 tarihleri arasında gerçekleştirilmiştir. Bu iki senaryodaki YZ puanları, uzmanlar tarafından verilen puanlarla karşılaştırılmıştır. Uzmanlar arası ve uzmanlar ile YZ puanları arasındaki uyum ve tutarlılıkları değerlendirmek için Uyuşma Yüzdesi, Cohen's Kappa, Karesel Ağırlıklandırılmış Kappa (Quadratic Weighted Kappa), Gwet'in AC1 istatistiği ve korelasyon katsayısı kullanılmıştır. Ayrıca, iki senaryodan elde edilen YZ puanları ve uzman puanları arasındaki farklılığı belirlemek için bağımsız örneklem t testi uygulanmıştır. Eğer t testinde varyansların homojenliği varsayımı ihlal ediliyorsa varyans eşdeğerliği varsayımı olmayan Welch t testi uygulanmıştır (Aydın, Algina, Leite ve Atılgan, 2018). Ayrıca t testi sonuçlarındaki farklılığın önemini belirlemek için Cohen's d etki büyüklükleri hesaplanmış ve raporlanmıştır. Cohen's d etki büyüklüğü küçük etki (d = .20), orta etki (d = .50) ve büyük etki (d = .80) olarak yorumlanmıştır (Cohen, 1992). Ardından, uzman puanları ve iki farklı senaryoyla YZ'den elde edilen puanların madde ayırt edicilik ve madde güçlüğü katsayıları hesaplanmıştır. 0-1 şeklinde kodlanan çoktan seçmeli maddelerin madde ayırt edicilikleri Nokta Çift Serili Korelasyon katsayısı ile hesaplanırken, 0-1-2 şeklinde puanlanan açık uçlu maddelerin madde ayırt edicilikleri düzeltilmiş madde toplam korelasyonu ile hesaplanmıştır. Madde güçlüğü ise ortalama doğru yanıt oranı ile hesaplanmıştır. Son olarak, uzman puanları ile Senaryo-2'den elde edilen YZ puanları arasındaki puanlama farklılıkları tespit edilmiş ve bu farklılıkların nedenini araştırmak için YZ aracına bu puanı neden atadığının açıklanması istenmiştir. Uzman ve Senaryo-2 puanları arasındaki farklılıkların sebebi kategoriler haline getirilerek raporlanmıştır.

Bulgular

Bu bölümde araştırmanın bulguları yer almaktadır. Öncelikle açık uçlu maddeleri puanlayan iki uzman arasındaki uyumlar incelenmiş, ardından uzman puanları ve YZ puanları arasındaki ilişkiler uyum katsayıları, t testi ve madde parametreleri ile analiz edilmiştir. Bu bölümde son olarak uzman ve YZ puanları arasındaki farklılıkların temel nedenleri belirlenmiştir.

Uzmanlar Arası Uyum

Araştırmada iki uzman puanlayıcı açık uçlu maddeleri birbirinden bağımsız olarak puanlamıştır. Tablo 1'de iki uzman arasındaki uyum katsayıları yer almaktadır.

Tablo 1. Uzmanlar Arası Uyum İndeksleri

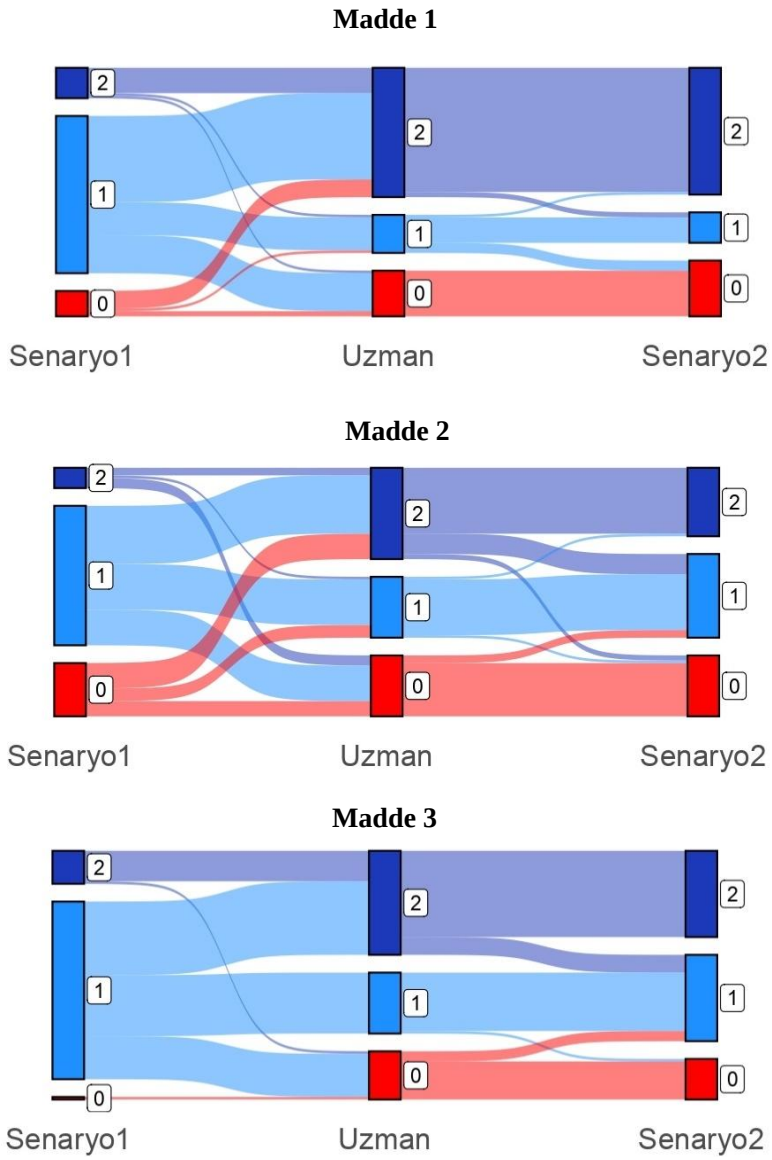
Madde	Uyum	Cohen's Kappa	Karesel Ağırlıklandırılmış Kappa	Gwet AC1	Correlation
Madde 1	0.98	0.98	0.99	0.98	0.99
Madde 2	0.95	0.92	0.93	0.93	0.97
Madde 3	0.94	0.91	0.91	0.91	0.93
Ortalama	0.96	0.94	0.94	0.94	0.96

Madde 1: Hesaplamalı, Madde 2: Hesaplamalı + Açıklamalı, Madde 3: Açıklamalı

Tablo 1 incelendiğinde hem uyum hem de Cohen's kappa katsayılarına göre iki uzman arasında yüksek düzeyde uyum görünmektedir (Landis ve Koch, 1977). Yüksek uyum katsayıları, hem soruların iki farklı uzman tarafından benzer şekilde puanlanabildiğini hem de puanlama kriterlerinin oldukça iyi tasarlanmış olduğunu göstermektedir. Uzmanlar arasında uyumsuzluk gösteren her yanıt, iki uzman tarafından düzenlenen toplantıda tek tek incelenmiş ve ortak karara varılmıştır. İki uzmanın ortak kararı olan puanlara araştırmanın devamında "Uzman Puanları" olarak değinilecektir.

Uzman Puanları ve YZ Arasındaki Uyum

Araştırmada yer alan üç açık uçlu maddeyi yanıtlayan 84 öğrenciden toplamda 252 öğrenci yanıtı elde edilmiştir. 252 öğrenci yanıtı YZ aracı (ChatGPT-4o modeli) tarafından 2 farklı senaryo altında puanlanmıştır (252*2= 504 resimli yanıt analiz edildi). Bu bölümde uzman puanları ve iki farklı senaryoda YZ puanları arasındaki ilişkiler incelenmiştir. İlk olarak Uzman ve iki farklı senaryoda YZ puanları arasındaki ilişkiler Sankey diyagramı ile Şekil 1'de sunulmuştur.



Madde 1: Hesaplamalı, Madde 2: Hesaplamalı + Açıklamalı, Madde 3: Açıklamalı, Senaryo-1: YZ ile Ölçütsüz puanlama, Senaryo-2: YZ ile Standart Ölçütlerle Puanlama

Şekil 1. Uzman ve YZ Puanları Arasındaki Uyum (Sankey Diyagramı)

Şekil 1'de açık uçlu maddeye puanlayıcının verdiği puanlar gösterilmektedir. Kırmızı renk puanlayıcının "0 = Yanlış"; açık mavi renk puanlayıcının "1 = Kısmi Doğru"; koyu mavi renk ise puanlayıcının "2 = Tam Doğru" puan verdiğini göstermektedir. Grafiklerin alt kısımlarında ise Puanlayıcıların isimlerinin kısaltmaları yer almaktadır. Madde 1 olarak isimlendirilen hesaplamalı madde türüne dair puanlar incelendiğinde Uzman ve Senaryo-2 puanları, puanlama kategorileri arasında daha benzer bir dağılım sergilerken, Senaryo-1 puanları daha farklı bir dağılım izlemiştir. Bu bulgular Uzman ve Senaryo-2 puanlarının daha uyumlu olduğunu gösterirken, Senaryo1 puanlarının farklılaştığını belirtmektedir. Bu soruya benzer şekilde Madde 2 olarak isimlendirilen hesaplamalı ve açıklamalı maddede ve Madde 3 olarak isimlendirilen açıklamalı maddede de Uzman ve Senaryo-2 puanları benzerlik gösterirken Senaryo-1 ayrışmıştır. Bulgular genel olarak incelendiğinde, her üç maddede de uzman puanlarında 2 (Tam doğru) puanları kategorisi Senaryo-1 ve Senaryo-2 puanlarına göre daha fazla yer almaktadır. Bu bulgu Uzmanların iki farklı YZ senaryosuna göre daha fazla tam puan verdiğini göstermektedir. Diğer taraftan, YZ ile ölçütsüz puanlama olarak isimlendirilen Senaryo-1 puanlarında diğer iki puanlayıcıya göre 1 (Kısmi doğru) puanının daha fazla ve 0 (Yanlış)

kategorisinin daha az yer aldığı görülmektedir. Bu durum YZ ile ölçütsüz puanlamanın diğer iki puanlayıcıdan oldukça farklılaştığını göstermektedir.

Tablo 2'de uzmanların ve iki farklı YZ'nin açık uçlu maddeleri puanlamasıyla elde edilen puanların uyum indekslerine yer verilmiştir.

Tablo 2. YZ ve Uzman Puanları Arasındaki Uyum İndeksleri

Madde	Gruplar	Uyum	Cohen's Kappa	Karesel Ağırlıklandırılmış Kappa	Gwet AC1	Correlation
Madde 1	Uzman-Senaryo-1	0.30	0.07	0.07	-0.02	0.09
	Uzman-Senaryo-2	0.92	0.85	0.94	0.88	0.95
Madde 2	Uzman-Senaryo-1	0.32	0.03	-0.06	0.00	-0.08
	Uzman-Senaryo-2	0.82	0.73	0.81	0.73	0.82
Madde 3	Uzman-Senaryo-1	0.44	0.18	0.26	0.22	0.34
	Uzman-Senaryo-2	0.86	0.78	0.88	0.79	0.89
Ortalama	Uzman-Senaryo-1	0.35	0.09	0.09	0.07	0.12
	Uzman-Senaryo-2	0.87	0.79	0.88	0.80	0.89

Madde 1: Hesaplamalı, Madde 2: Hesaplamalı + Açıklamalı, Madde 3: Açıklamalı, Senaryo-1: YZ ile Ölçütsüz puanlama, Senaryo-2: YZ ile Standart Ölçütlerle Puanlama

Tablo 2 incelendiğinde her üç madde için de Uzman ve Senaryo-1 arasındaki puanların uyum indekslerinin oldukça düşük olduğu görülürken Uzman ve Senaryo-2 arasındaki uyum indekslerinin orta veya yüksek düzeyde olduğu görülmektedir. Bu bulgu, Senaryo-1 olarak isimlendirilen YZ ile ölçütsüz yapılan puanlamanın uzman puanlarına göre oldukça farklılaştığını ve çok düşük uyum gösterdiğini belirtmektedir. Diğer bir deyişle YZ ile ölçütsüz yapılan puanların güvenirliğinin oldukça düşük olduğu söylenebilir. Diğer taraftan, Senaryo-2 olarak isimlendirilen YZ ile Standart Ölçütlerle yapılan puanlamanın uzman puanları ile orta veya yüksek düzeyde uyum gösterdiği görülmektedir. Diğer bir deyişle, YZ ile standart ölçütlerle yapılan puanların güvenirliğinin orta veya yüksek düzeyde olduğu söylenebilir.

Senaryo-1, Senaryo-2 ve Uzman Puanları Arasındaki Farklılıklar

Bir diğer araştırma sorusu ile ilgili olarak uzman puanları ve iki farklı senaryo ile YZ puanları arasındaki farklılığın anlamlılığı bağımsız örneklemler t testi ile incelenmiştir. t testi uygulanmadan önce verilerin normalliği incelenmiştir. Hair, Black, Babin ve Anderson (2010) ve Byrne (2010), çarpıklığın -2 ile +2 arasında ve basıklığın -7 ile +7 arasında olması durumunda verilerin normal kabul edilebileceğini belirtmişlerdir. Araştırma verisinde yer alan tüm değişkenlerin çarpıklık ve basıklık değerleri belirtilen aralıklarda olduğundan tüm verilerin normal dağıldığına karar verilmiştir (Hair vd., 2010; Byrne, 2010). Varyansların homojenliği varsayımı incelendiğinde tüm maddelerde uzman ve Senaryo-1 puanları arasında varyansların homojenliği varsayımının ihlal edildiği görülmüştür (Levene Test, $p<.05$). Uzman ve Senaryo-2 puanları arasında ise varyanslar homojen dağılmaktadır (Levene Test, $p>.05$). Varyansların homojenliğinin ihlal edildiği durumlarda varyansların homojenliği varsayımı bulunmayan Welch t testi kullanılmıştır. t testi bulguları Tablo 3'te sunulmuştur.

Tablo 3. Uzman ve YZ Puanları Arasındaki Farklılıklar

Madde	Grup	Mean	SS	t	p	Etki Büyüklüğü
Madde 1	Uzman	1.393	0.822	3.48	0.00*	0.54
	Senaryo-1	1.024	0.514			
	Uzman	1.393	0.822	0.45	0.65	0.07
	Senaryo-2	1.333	0.869			
Madde 2	Uzman	1.143	0.838	2.69	0.00*	0.41
	Senaryo-1	0.845	0.57			
	Uzman	1.143	0.838	0.85	0.39	0.13
	Senaryo-2	1.036	0.783			
Madde 3	Uzman	1.262	0.808	1.22	0.23	0.19
	Senaryo-1	1.143	0.385			
	Uzman	1.262	0.808	0.39	0.69	0.06
	Senaryo-2	1.214	0.746			

Madde 1: Hesaplamalı, Madde 2: Hesaplamalı+Açıklamalı, Madde 3: Açıklamalı, Senaryo-1: YZ ile Ölçütsüz puanlama, Senaryo-2: YZ ile Standart Ölçütlerle Puanlama

Tablo 3 incelendiğinde uzman ve Senaryo-1 puanları arasında iki soruda anlamlı farklılık bulunmakta iken bir maddede anlamlı farklılık bulunmamaktadır. Üç açık uçlu maddenin tamamında Senaryo-1'den elde edilen puanların ortalamasının uzman puanları ortalamasına göre daha düşüktür. İki açık uçlu maddede ise anlamlı farklılık vardır. Etki büyüklükleri incelendiğinde Senaryo-1 ile uzman puanları arasında sırasıyla orta, orta ve küçük etki düzeyinde etki büyüklükleri bulunmaktadır. Bu bulgular, YZ'nin ölçütsüz puanlarının uzman puanlarına göre farklılaştığını göstermektedir. Diğer taraftan, Senaryo-2 ve uzman puanları karşılaştırıldığında iki puan grubu arasında her üç maddede de anlamlı farklılığın olmadığı görülmektedir. Ayrıca Senaryo-2 ve uzman puanları arasındaki etki büyüklükleri sıfıra oldukça yakındır. Bu bulgular, YZ ile Standart ölçütlerle yapılan puanlamanın uzman puanlarına göre anlamlı biçimde farklılaşmadığını, dolayısıyla benzerlik gösterdiğini ifade etmektedir.

YZ ve Uzman Puanlarından Elde Edilen Madde Parametreleri

İki farklı YZ senaryosundan ve uzmanlardan elde edilen puanların geçerliği örneklem sayısı sebebiyle Klasik Test Kuramı parametreleriyle incelenmiştir. Senaryo-1, Senaryo-2 ve uzman puanlarının madde ayırt ediciliği ve madde güçlüğü parametreleri Tablo 4'te sunulmuştur.

Tablo 4. YZ ve Uzman Puanları Arasındaki Farklılıklar

Madde		Madde Ayırt Ediciliği	Madde Güçlüğü
Madde 1	Uzman	0.37	0.69
	Senaryo-1	-0.01	0.51
	Senaryo-2	0.39	0.67
Madde 2	Uzman	0.47	0.57
	Senaryo-1	0.02	0.43
	Senaryo-2	0.43	0.52
Madde 3	Uzman	0.34	0.63
	Senaryo-1	0.31	0.57
	Senaryo-2	0.42	0.61
Madde 4		0.47	0.68
Madde 5		0.46	0.80
Madde 6		0.61	0.83
Madde 7		0.45	0.85
Madde 8		0.25	0.88
Madde 9		0.45	0.76
Madde 10		0.48	0.73
Madde 11		0.47	0.67
Madde 12		0.49	0.86
Madde 13		0.57	0.57

Madde 1: Hesaplamalı, Madde 2: Hesaplamalı+Açıklamalı, Madde 3: Açıklamalı, Senaryo-1: YZ ile Ölçütsüz puanlama, Senaryo-2: YZ ile Standart Ölçütlerle Puanlama

Tablo 4 incelendiğinde Madde 1, Madde 2 ve Madde 3'te Senaryo 1'in madde ayırt edicilikleri sırasıyla -0.01, 0.02, 0.31 ve Senaryo-2'nin sırasıyla 0.39, 0.43, 0.42 iken uzman puanlarının sırasıyla 0.37, 0.47 ve 0.43'tür. Bu bulgular, Senaryo-1'in madde ayırt edicilik katsayılarının Senaryo-2 ve uzmanlara göre oldukça düşük olduğunu göstermektedir. Dolayısıyla YZ ile ölçütsüz yapılan puanlama işlemlerinin geçerliğinin oldukça düşük olduğu görülmektedir. Diğer taraftan, Senaryo-2 ile uzman puanlarının madde ayırt edicilikleri karşılaştırıldığında Madde 1 ve Madde 3'te Senaryo-2'nin ayırt ediciliği az bir fark ile yüksek iken, Madde 2'de uzman puanlarının ayırt ediciliği az bir fark ile daha yüksektir. Bu bulgu, Senaryo-2 ve uzman puanlarının ayırt ediciliklerinin birbirine benzer olduğunu ve madde geçerliğinin yüksek olduğunu göstermektedir.

Madde güçlükleri incelendiğinde her üç maddede de Senaryo-1'in madde güçlüğü en düşük iken, Senaryo-1'i sırasıyla Senaryo-2 ve uzman puanları takip etmektedir. Bu bulgu, Senaryo-1'in Senaryo-2 ve uzmanlara göre daha düşük puanlar verdiğini gösterirken, Senaryo-2 ve uzmanların daha yüksek puanlar verme eğiliminde olduğunu ifade etmektedir.

YZ ve Uzman Puanları Arasındaki Farklılıkların Nedenleri

Araştırmanın bu bölümünde, Senaryo-2 ve uzman puanları arasındaki farklılığın nedenleri araştırılmıştır. Daha önceki bölümlerde YZ'nin ölçütsüz puanlama (Senaryo-1) performansının uzman puanlarına göre oldukça uyumsuz olduğundan ve düşük güvenirlik-geçerliğe sahip olduğundan bahsedilmişti. Dolayısıyla bu bölümde sadece YZ ile Standart Ölçütlerle puanlama (Senaryo-2) ve uzman puanları arasındaki farklılıkların temel sebebini öğrenmek için Senaryo-2 ile uzman puanları arasındaki 32 uyumsuz yanıt incelenmiş ve bu yanıtlarda YZ aracına bu puanı neden atadığının açıklanması istenmiştir. Elde edilen açıklamalar doğrultusunda puanlama farklılıkları oluşan kategoriler altında birleştirilerek analiz edilmiştir. Bulgular Tablo 5'te sunulmuştur.

Tablo 5. YZ ve Uzman Puanları Arasındaki Farklılıkların Nedenleri

S.N.	YZ ve Uzmanların Puanlama Farklılıkları	f	%
1	Kötü yazılmış el yazısını tam olarak okuyamama	4	12.5
2	Cümlenin bağlamını tam olarak anlayamama	4	12.5
3	Şişirme yanıtlardaki doğruluğu tam kestirememe	8	25.0
4	Kurşun kalemin silik yazdığı durumlarda yanıtın tam olarak okunamaması	2	6.3
5	Basit hataları YZ'nin daha net puanlaması	8	25.0
6	Kısa yazılan yanıtlarda anlamı yakalayıp tam olarak puanlayamama	5	15.6
7	Ok veya sembol ile işaret edilen yanıtın anlaşılamaması	1	3.1

YZ ile uzman puanları arasındaki farklılığın en önemli sebeplerinden ikisi "Şişirme yanıtlardaki doğruluğu tam kestirememe" (f = 8, % = 25.0) ve "Basit hataları YZ'nin daha net puanlaması" (f = 8, % = 25.0) seçenekleridir. Öğrencilerin açık uçlu maddelere verdiği şişirme (doğru olan fakat konu ile ilgisiz argümanlar) yanıtların puanlanmasında YZ ile uzman puanları arasında fark olduğu görülmektedir. Örneğin Madde 3'ün birinci seçeneğinde öğrencilerden çoktan seçmeli maddelerin güvenirlik açısından avantajları istenmiş ve Öğrenci_62 bu seçeneğe "Çoktan seçmeli maddelerden oluşan çoktan seçmeli testleri katılımcılara tekrar tekrar uyguladığımız zaman aynı sonuçları elde etme olasılığımız yüksek olduğu için güvenirlik yüksektir." şeklinde cevap vermiştir. Bu yanıt, test-tekrar test güvenirlik türünün tanımına benzer olup soru ile alakası bulunmayan bir argümandır. Bu yanıta YZ tam puan verirken, uzmanlar yanlış olarak değerlendirmişlerdir. Bir diğer seçenek olan "Basit hataları YZ'nin daha net puanlaması" seçeneğinde ise öğrencilerin matematiksel işlemlerinde yaptığı basit hatalara uzmanlar daha hoşgörülü iken, YZ daha net ve katı davranmaktadır. Örneğin Madde 1'de Z puanı hesaplarken doğru cevap 0/10=0 olması gerekirken; Öğrenci_23, 0/10= 0.10 olarak hesaplama yapmıştır. Uzmanlar bu yanıtta, öğrencinin yaptığı işlem hatasının bilişsel olarak ölçülen özellik için kritik olmadığını düşünerek tam puan verirken, YZ bu yanıtı yanlış olarak değerlendirmiştir. Bir başka seçenek olan "Kısa yazılan yanıtlarda anlamı yakalayıp tam olarak puanlayamama" seçeneğinde ise YZ, öğrencinin açık uçlu maddeye kısa bir şekilde verdiği yanıttaki anlamı tam olarak yakalayamamakta ve dolayısıyla eksik puan verebilmektedir. Örneğin Madde 2'nin bir seçeneğinde Öğrenci_23, madde güçlüğünü 0.24

olarak doğru bir şekilde hesaplamış ve yorumlarken "Madde güçlüğü yüksektir" şeklinde yorumlamıştır. Bu yanıtta öğrencinin aslında belirttiği şey madde güçlüğü, sayısal olarak düşük olsa da maddenin zor olduğu anlamındadır. Ancak bu yanıtı YZ yanlış olarak değerlendirirken, uzmanlar doğru olarak değerlendirmiştir. Bu seçenekleri sırasıyla "Kötü yazılmış el yazısını tam olarak okuyamama" (f = 4, % = 12.5) ve "Cümlenin bağlamını tam olarak anlayamama" (f = 4, % = 12.5) seçenekleri izlemektedir. Bu seçeneklerde ise YZ'nin bazı kötü el yazısına sahip yanıtları anlayamadığı ve düşük ifade içeren veya anlamsal olarak iyi ifade edilmemiş cümlelerin de YZ tarafından iyi bir şekilde anlaşılmadığı görülmüştür. Ayrıca, "Kurşun kalemin silik yazdığı durumlarda yanıtın tam olarak okunamaması" seçeneğinde ise kurşun kalemi yeterince bastırmadan yazan iki yanıtın YZ tarafından okunamadığı görülmüştür (f = 2, % = 6.3). Son olarak, bir yanıtta ise öğrenci yanıtını ok çıkararak göstermiştir, bu yanıt ile öğrencinin kısmi doğru puanı alması gerekirken YZ aracı ok işaretini yorumlamamış ve bu yanıtı yanlış olarak değerlendirmiştir.

Tartışma, Sonuç ve Öneriler

Son dönemlerde, YZ'nin açık uçlu maddelerin puanlanmasındaki performansına yönelik çalışmaların sayısının arttığı görülmektedir. Özellikle geniş ölçekli değerlendirmeler kapsamında, iş yükünü azaltma ve puanlama tutarlılığını artırmadaki potansiyeli nedeniyle büyük ilgi görmektedir. Literatür, uzman puanlayıcılara kıyasla YZ tabanlı puanlama sistemlerinin güvenilirliği, doğruluğu ve sınırlılıkları ile ilgili önemli bulgular ortaya koymaktadır.

Literatürdeki çalışmalar, ChatGPT gibi YZ araçlarının açık uçlu maddeleri puanlarken insan değerlendiricilerle orta veya yüksek düzeyde korelasyon ve uyum sergilediğini göstermiştir. Bu durum YZ'nin özellikle insan kaynaklarının sınırlı olduğu büyük ölçekli değerlendirmelerde puanlama için güvenilir bir araç olma potansiyeli taşıdığını göstermektedir (Demir, 2023; Uysal ve Doğan, 2021). Xiao ve diğerleri (2024) kompozisyonları puanlamak için GPT-4 ve GPT-3.5-turbo'yu çeşitli yaklaşımlar altında incelemişlerdir. Farklı konfigürasyonlar için ağırlıklandırılmış kappa değerlerinin 0.67 ile 0.80 arasında olduğu görülmektedir ve bu da YZ'nin uzman puanlayıcılarla uyumlu olduğunu göstermektedir. von Davier, Tyack ve Khorramdel (2022), TIMSS 2019'daki grafiksel açık uçlu madde yanıtlarının yapay sinir ağları kullanılarak otomatik puanlandığı araştırmasında, bu araçların karmaşık yapılandırılmış madde yanıtlarını etkili bir şekilde işleyebileceğini ve potansiyel olarak ikinci insan değerlendiricilere olan ihtiyacı ortadan kaldırabileceğini ifade etmiştir. Bu araştırmada da literatürle uyumlu olarak, standart puanlama ölçütleri ve ayrıntılı puanlama anahtarının kullanıldığı durumlarda ChatGPT tarafından yapılan puanlamanın insan değerlendiricilerle yüksek düzeyde uyum sergilediğini görülmüştür. Ancak herhangi bir ölçüt kullanılmadan ChatGPT tarafından yapılan puanlamada ise uyum düzeyi son derece düşüktür. Bu durum, YZ'nin yararlı bir araç olabileceğini ancak dikkatli ve insan gözetimi ile birlikte eğitilerek kullanılması gerektiğini göstermektedir.

Uzman ve YZ puanları arasındaki farklılıklar t testi ile incelenmiştir. Uzman puanları ile YZ ile ölçütsüz puanlama arasında üç maddenin ikisinde anlamlı farklılık olduğu görülmüştür. Diğer taraftan, uzman puanları ile YZ ile standart ölçütlerle puanlama puanları arasında ise üç maddenin hiçbirinde anlamlı farklılık bulunmadığı görülmüştür. Bu bulgu, YZ ile ölçütsüz puanlamanın uzman puanlarına göre farklılaştığını göstermektedir. Diğer taraftan uzman puanları ile YZ ile standart ölçütlü puanları arasında anlamlı farklılaşma bulunmamaktadır. Ortalamalar incelendiğinde uzman puanlarının YZ'nin standart ölçütlü ve ölçütsüz puanlarına göre daha yüksek olduğu görülmektedir. Jukiewicz (2024), literatürde yer alan pek çok çalışma mevcut çalışmanın bulgularına benzer şekilde, ChatGPT'nin puanlarının insan puanlarından daha düşük olduğunu belirtmiştir (Almusharraf ve Alotaibi, 2023; Bui ve Barrot, 2024; Jukiewicz, 2024).

Uzman ve iki senaryoda YZ puanlarının madde parametreleri incelendiğinde YZ ile ölçütsüz puanlamadan elde edilen madde ayırt edicilik parametrelerinin düşük olmasından dolayı bu puanlamanın geçerliliğinin olmadığı görülmüştür. Bir diğer ifade ile YZ araçları standart puanlama araçları ile eğitilmeden herhangi bir puanlama yapılmasının geçerli olmayan sonuçlar üreteceği sonucuna ulaşılmıştır.

Bu çalışmada ayrıca uzmanlar puanları ve YZ ile standart ölçütle puanlama puanları arasındaki farklılıkların nedenleri araştırılmıştır. Bu analizde yedi farklı spesifik neden olduğu görülmüştür. Bu nedenlerden en yüksek frekansa sahip olan ikisi, şişirme yanıtlara YZ'nin puan vermesi ve basit hatalarda YZ'nin daha kesin ve net puanlama yapmasıdır. Literatürde yer alan çalışmalarda, açık uçlu maddelerin YZ ile puanlanması sürecinde karşılaşılabilecek bazı sorunlardan bahsedilmiştir. Bunlardan birisi "ölçeklenebilirlik" problemidir. Mevcut YZ modelleri genellikle her bir madde için ölçeklenebilir olmayan ayrı bir modelin eğitilmesini gerektirir. Bu durum YZ'nin, madde sayısının çok olduğu ve geniş çapta hesaplama gereken durumlarda pratik bir yöntem olarak kullanılmasını engelleyebilmektedir. Ayrıca YZ modelleri, birden fazla maddenin bir ortak okuma metni ile ilişkilendirildiği bağlamlı sorularda bağlamsal ilişkileri anlamada başarısız olabilmektedir. Bu durum puanlama hatalarına yol açmaktadır. Puanlama sürecine karışan hata türleri ve yanlılık ise bir diğer sorundur. YZ modelleri, puanlamanın güvenilirliğini etkileyen hata türleri ve yanlılık sergileyebilir. Bu durum, eğitim verileri kaynaklı olabilir veya modelin yanıtların bağlamını tam olarak anlayamamasından kaynaklanabilir (Fernandez vd., 2022). Bir başka sınırlılık ise farklı YZ algoritmalarının, otomatik puanlamada farklı performans düzeyleri gösteriyor olmasıdır (Uysal ve Doğan, 2021). Yaneva ve diğerleri (2023), çalışmasında büyük dil modellerinin, aynı maddelere verdiği tekrarlı yanıtların, anlamlı ölçüde farklılıklar gösterdiğini belirtilmiştir. Kullanılan YZ araçları için uzman doğrulamasına ihtiyaç duyulduğunu ifade etmiştir. Diğer taraftan YZ'nin puanlama performansına ilişkin araştırma yapan pek çok yazar, mevcut çalışmanın yazarları ile benzer şekilde YZ'nin puanlama performansında daha katı puanlama eğiliminde olduğunu ve alakasız, şişirme veya yanlış içerikleri cezalandırma (düşük not ile puanlama) yeteneğine sahip olmadığını ifade etmektedir (Bui ve Barrot, 2024; Parker, Becker ve Carroca, 2023; Quah vd., 2024).

YZ araçlarının puanlama performansının maddelerin zorluk düzeyine ve belirli bilgi alanına bağlı olarak değişebileceği ifade edilmektedir (Zesch, Horbach ve Zehner, 2023). ChatGPT'nin, sınav katılımcıları tarafından daha kolay bulunan maddelere doğru yanıt verme olasılığının daha yüksek olduğu ve uygulamaya dayalı maddelerde önemli ölçüde daha kötü performans gösterdiği belirtilmiştir (Yaneva vd., 2023). Demir (2023) araştırmasında YZ araçlarının genellikle insan değerlendiricilerle yüksek korelasyon ve uyum gösterdiğini belirtmiş ancak genellenebilirlik kuramı gibi daha hassas yöntemlerle hesaplanan güvenilirlik katsayıları daha düşük bulunmuştur. YZ'nin her zaman uzman puanlayıcı standartlarıyla eşleşmeyebileceğini ifade etmiştir. Bu araştırmada da YZ'nin uzman değerlendiricilerle uyum yüzdesinin maddelerin özelliklerine göre değiştiği gözlenmiştir. Özellikle doğrudan hesaplama gerektiren madde için uyum yüzdesi ve korelasyon değerleri daha yüksek bulunmuştur. Açıklama gerektiren maddelerde ise uyum yüzdesi göreceli olarak daha düşüktür.

Mevcut araştırmanın dört sınırlılığı sunulmaktadır. Birincisi, araştırmacılar, makine öğrenmesi algoritmaları ve doğal dil işleme yöntemlerini kullanarak puanlama sistematiği geliştirmek yerine GPT modelini kullanması bu araştırmanın bir sınırlılığıdır. Araştırmada öğrencilerin el yazısı yanıtlarını puanlamak için iki temel teknolojik gereksinim bulunmaktadır. Bunlar optik karakter okuma ve doğal dil işleme modeli ile yanıtları puanlamadır. Bu iki işlevi aynı anda gerçekleştirdiğinden GPT modeli kullanılmıştır. Mevcut araştırmanın ikinci sınırlılığı, araştırmada yer alan sorular geliştirilirken 0-1-2 şeklinde üç kategorili puanlanacak şekilde geliştirilmiştir. Dolayısıyla YZ'nin dört ve daha fazla kategorili puanlama performansı değişiklik gösterebilir. YZ'nin dört ve daha fazla kategorili puanlama performansı hakkında bir fikir sunmaması mevcut araştırmanın bir diğer sınırlılığıdır. Üçüncüsü, mevcut araştırmada GPT-4o modeli kullanılması bu araştırmanın bir diğer sınırlılığıdır. Gelecekte yeni GPT modellerinin geliştirilmesi ile açık uçlu maddelerin puanlama performansının da değişebileceği düşünülmektedir. Dördüncü sınırlılık, mevcut araştırmanın örnekleminin küçük olmasıdır. Nokta çift serili korelasyon katsayısı ve düzeltilmiş madde toplam korelasyonu ile hesaplanan madde ayırt edicilik katsayıları başarı testinin geçerliliğine dair bilgi sunmakta olsa da faktör analitik yöntemler ile inceleme yapılamamış olması mevcut araştırmanın bir sınırlılığıdır.

Açık uçlu maddelere yönelik YZ tabanlı puanlama sistemleri, verimlilik ve insan çabasının azaltılması açısından umut vaat etmektedir. Ancak algoritma performansındaki değişkenlik, ölçeklenebilirlik sorunları, yanlılık ve insan değerlendiricilere kıyasla daha düşük güvenilirlik gibi önemli sınırlılıklardır. Gelecekte, YZ tabanlı puanlama sistemlerinin performansını ve uygulanabilirliğini daha da artırabilecek çalışmalara ihtiyaç duyulmaktadır. YZ performansının optimize edilmesinin, eğitimsel ve psikolojik ölçümlerdeki kullanım alanlarının genişletilmesinin faydalı olacağı düşünülmektedir. Sonuç olarak, bu araştırmanın yazarları, önceki yazarlar ile benzer şekilde (Poole ve Coss, 2024; Ramineni ve Williamson, 2018) YZ araçlarının eğitsel değerlendirmelerde insan değerlendiricilerin yanında "ikinci puanlayıcı" olarak yer alabileceğini düşünmektedirler.

Uygulayıcılara Yönelik Öneriler

Araştırmanın sonuçları doğrultusunda uygulayıcılara yönelik altı öneri sunulmuştur. Birincisi, YZ ile puanlama yapılırken, puanlamada YZ'nin hangi kriterlerin kullanılacağına dair net bir çerçeve oluşturulması oldukça önemlidir. Bu araştırma sonucunda da görüldüğü üzere, standart ölçütlerle yapılan puanlamada (Senaryo-2) uzman puanlaması ile yakın sonuçlar elde edilmiştir. İkinci olarak, ChatGPT sorunun bağlamıyla ilişkili olmayan veya şişirme olan bazı yanıtlara puan vermiştir. Bu durumda öğrenciler YZ aracının bu özelliğinden faydalanabilir. Dolayısıyla şişirme yanıtlara yönelik uygulayıcıların önlem alması ve dikkatli olması önerilmektedir. Üçüncüsü, ChatGPT, özellikle matematiksel ifadelerde küçük hatalar içeren bazı yanıtları daha keskin ve net biçimde puanlamış ve öğrenciye düşük puan vermiştir. Dolayısıyla uygulayıcılar, YZ araçlarının küçük hatalarda notu azalttığını göz önüne almalıdırlar. Dördüncüsü, ChatGPT öğrencilerin el yazılarının kötü veya silik olması durumunda net biçimde anlayamamakta, dolayısıyla puanlamada sorunlar oluşabilmektedir. Bu konuda uygulayıcılar el yazısı silik ise yazıların kontrastını arttırabilirler, el yazısı kötü ise öğrencileri uyarabilirler veya yazı değerlendirilmeden önce YZ araçlarından yazıyı anlamlı biçimde yeniden düzenlemesini talep edebilirler. Beşinci olarak, YZ'nin verdiği puanların insan değerlendiriciler tarafından gözden geçirilmesinin faydalı olabileceği düşünülmektedir. Bu durum YZ'nin kaçırabileceği ince ayrıntıları gözden geçirme fırsatı sunacaktır. Bu araştırmada da gözlemlendiği gibi; şişirme yanıtlar, silik ve okunaksız el yazıları vb. faktörlerden kaynaklanan yanıt uyumsuzluklarının giderilmesi gerekmektedir. Puanlama süreçlerinde hibrit bir sistemin (insan + yapay zekâ) kullanılmasının puanlama doğruluğunu artıracağı ifade edilebilir. Altıncısı, açık uçlu maddelerin YZ ile nihai puanlamasını yapmadan önce, küçük pilot çalışmalar yaparak sistemin doğruluğunun ve güvenirliğinin test edilmesi, sistemin güçlü ve zayıf yanlarını belirlenmesi faydalı olacaktır. Pilot uygulamadan sonra puanlama anahtarı üzerinde iyileştirmeler yapılabilir.

Araştırmacılara Yönelik Öneriler

Bu bölümde bu araştırma konusu ile ilgili olarak araştırmacıların gelecekte çalışabilecekleri konular üzerine araştırmacılara sekiz öneri sunulmuştur. Birincisi, bu araştırmada yalnızca ChatGPT kullanılmıştır. Araştırmacılar, açık uçlu maddelerin puanlanmasında Google Bard, Microsoft Copilot, Gemini gibi farklı YZ araçlarını karşılaştırabilirler. İkinci olarak, araştırmacılar aynı YZ aracına aynı yanıtı tekrar puanlatarak puanlama güvenirliği üzerine çalışmalar gerçekleştirebilirler. Üçüncüsü, bu araştırmada, YZ'nin bir final sınavında elde edilen açık uçlu yanıtları puanlama performansı incelenmiştir. Araştırmacılar gelecekte YZ'nin öğrencilere kişiselleştirilmiş geri bildirimler vermesini sağlayarak süreç değerlendirmeye yönelik yaklaşımları ele alabilirler. Dördüncüsü, YZ'nin zamanla gelişen ve öğrenen bir yapıya sahip olması sebebiyle aynı yanıtların belirli periyodlarda tekrarlı puanlandığı ve bunlar arasındaki tutarlılığın incelendiği çalışmalar yürütülebilir. Beşincisi, YZ'nin insan değerlendiriciler ile uyumsuzluk gösterdiği yanıtlar kapsamlı incelenerek bu uyumsuzlukların nedenleri ve çözüm yolları üzerine araştırmalar gerçekleştirilebilir. Altıncı olarak, açık uçlu maddelerin kullanıldığı, yüksek riskli sınav kapsamında olmayan, geniş ölçekli ulusal ve uluslararası eğitim araştırmalarında (PISA, TIMSS, ABİDE vb.) YZ'nin puanlama performansına yönelik araştırmalar yürütülebilir. Yedincisi, bu araştırmada açık uçlu maddeler 0-1-2 olmak üzere üç kategori üzerinden değerlendirilmiştir. Gelecekte araştırmacılar daha fazla kategori ile benzer araştırmalar gerçekleştirebilirler. Sekizinci olarak, bu araştırmada el yazısı ile yanıtların YZ aracı tarafından okunmasında ChatGPT'nin optik karakter tanıma teknolojisi kullanılmıştır. Gelecek araştırmalarda farklı OCR araçlarının puanlama açısından etkililiğini karşılaştıran araştırmalar tasarlanabilir.

Kaynakça

Abdolreza Gharehbagh, Z., Mansourzadeh, A., Montazeri Khadem, A. ve Saeidi, M. (2022). Reflections on using open-ended questions. *Medical Education Bulletin, 3*(2), 475-482.

Agustianingsih, R. ve Mahmudi, A. (2019). How to design open-ended questions?: Literature review. *Journal of Physics: Conference Series, 1320*(1). doi:10.1088/1742-6596/1320/1/012003

Alers, H., Malinowska, A., Meghoe, G. ve Apfel, E. (2024). Using ChatGPT-4 to grade open question exams. K. Arai (Ed.), *Advances in information and communication* içinde (s. 1-9). Switzerland: Springer Nature. doi:10.1007/978-3-031-53960-2_1.

Almusharraf, N. ve Alotaibi, H. (2023). An error-analysis study from an EFL writing context: Human and automated essay scoring approaches. *Technology Knowledge and Learning, 28*(3), 1015-1031.

Aydın, B., Algina, J., Leite, W. L. ve Atılgan, H. (2018). *Sosyal bilimler için r'a giriş.* Ankara: Anı Yayıncılık.

Aznar-Mas, L. E., Atarés Huerta, L. ve Marin-Garcia, J. A. (2023). Effectiveness of the use of open-ended questions in student evaluation of teaching in an engineering degree. *Journal of Industrial Engineering and Management, 16*(3), 521. doi:10.3926/jiem.5620

Baburajan, V., de Abreu e Silva, J. ve Pereira, F. C. (2022). Open vs closed-ended questions in attitudinal surveys-comparing, combining, and interpreting using natural language processing. *Transportation Research. Part C, Emerging Technologies, 137*(12), 103589. doi:10.1016/j.trc.2022.103589

Badger, E. ve Thomas, B. (2019). Open-ended questions in reading. *Practical Assessment, Research, and Evaluation, 3*(1), 4.

Baykul, Y. ve Turgut, M. F. (2012). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayıncılık.

Beiting-Parrish, M. ve Whitmer, J. (2023). Lessons learned about evaluating fairness from a data challenge to automatically score NAEP reading items. *Chinese/English Journal of Educational Measurement and Evaluation*, *4*(3). doi:10.59863/nkcj9608

Beksultanova, A. I., Vatyukova, O. Y. ve Yalmaeva, M. A. (2020). *Application of digital technologies in the educational process*. Proceedings of the 2nd International Scientific and Practical Conference on Digital Economy (ISCDE 2020).

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

Bui, N. M. ve Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies*. doi:10.1007/s10639-024-12891-w

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New York: Routledge.

Chen, L., Chen, P. ve Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access*, 8, 75264-75278.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101. doi:10.1111/1467-8721.ep10768783

Demir, S. (2023). Investigation of ChatGPT and real raters in scoring open-ended items in terms of inter-rater reliability. *Uluslararası Türk Eğitim Bilimleri Dergisi, 2023*(21), 1072-1099. doi:10.46778/goputeb.1345752

Doğan, N. (2019). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayınevi.

Fernandez, N., Ghosh, A., Liu, N., Wang, Z., Choffin, B., Baraniuk, R. ve Lan, A. (2022). Automated scoring for reading comprehension via in-context bert tuning. M. M. Rodrigo, N. Matsuda, A. I. Cristea ve V. Dimitrova (Ed.), *International conference on artificial intelligence in education* içinde (s. 691-697). Cham: Springer International Publishing.

Fitriyah, Y., Wahyudin, Suhendra, Nurhayati, H. ve Febrianti, T. S. (2024). Open-ended approach for critical thinking skills in mathematics education: A meta-analysis. *EduMatSains: Jurnal Pendidikan, Matematika Dan Sains, 9*(1), 156-174. doi:10.33541/edumatsains.v9i1.5975

Freedman, R. L. H. (1994). *Open-ended questioning: A handbook for educators*. Boston: Addison-Wesley.

Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C. ve Srinivasa, A. R. (2024). Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, 6, 100206. doi:10.1016/j.caeai.2024.100206

Geer, J. G. (1988). What do open-ended questions measure?. *Public Opinion Quarterly, 52*(3), 365-367.

Güler, N. (2014). Analysis of open-ended statistics questions with many facet Rasch model. *Eurasian Journal of Educational Research,* 55, 73-90. doi:10.14689/ejer.2014.55.5

Hair, J., Black, W. C., Babin, B. J. ve Anderson, R. E. (2010). *Multivariate data analysis* (7. bs.). Upper Saddle River, NJ: Pearson Educational International.

Hogan, T. P. ve Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education, 20*(4), 427-441. doi:10.1080/08957340701580736

Jamil, F. ve Hameed, I. A. (2023). Toward intelligent open-ended questions evaluation based on predictive optimization. *Expert Systems with Applications*, 231, 120640. doi:10.1016/j.eswa.2023.120640

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J., Urban-Lurain, M., Doherty, J. H. ve Haudek, K. C. (2021). Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. *Journal of Science Education and Technology*, *30*(2), 150-167. doi:10.1007/s10956-020-09858-0

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. doi:10.1016/j.tsc.2024.101522

Karadag, N., Boz Yuksekdag, B., Akyildiz, M. ve Ibileme, A. I. (2020). Assessment and evaluation in open education system: Students' opinions about Open-Ended Question (OEQ) practice. *Turkish Online Journal of Distance Education*, *22*(1), 179-193. doi:10.17718/tojde.849903

Karakaya, İ. (2022). *Açık uçlu soruların hazırlanması, uygulanması ve değerlendirilmesi*. Ankara: Pegem Yayınları.

Karasar, N. (2012). *Bilimsel araştırma yöntemi* (24. bs.). Ankara: Nobel Yayın Dağıtım.

Karimi, L. (2014). The effect of constructed-responses and multiple-choice tests on students' course content mastery. *Southern African Linguistics and Applied Language Studies, 32*(3), 365-372. doi:10.2989/16073614.2014.997067

Kartikasari, S. A., Usodo, B. ve Riyadi (2022). The effectiveness open-ended learning and creative problem solving models to teach creative thinking skills. *Pegem Journal of Education and Instruction, 12*(4), 29-38. doi:10.47750/pegegog.12.04.04

Landis, J. R. ve Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lin, Y., Zheng, L., Chen, F., Sun, S., Lin, Z. ve Chen, P. (2020). *Design and implementation of intelligent scoring system for handwritten short answer based on deep learning*. IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), Dalian, China. doi:10.1109/ICAIIS49377.2020.9194943

Lohman, D. F. (1993). Learning and the nature of educational measurement. *NASSP Bulletin*, *77*(555), 41-53. doi:10.1177/019263659307755506

Lu, M., Zhou, W. ve Ji, R. (2021). Automatic scoring system for handwritten examination papers based on YOLO algorithm. *Journal of Physics: Conference Series*, 2026. doi:10.1088/1742-6596/2026/1/012030.

Maris, G. ve Bechger, T. (2006). Scoring open ended questions. *Handbook of statistics* içinde (s. 663-681). Hollanda: Elsevier.

Mizumoto, A. ve Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. doi:10.1016/j.rmal.2023.100050

Monrat, N., Phaksunchai, M. ve Chonchaiya, R. (2022). Developing students' mathematical critical thinking skills using open-ended questions and activities based on student learning preferences. *Education Research International,* 2022, 1-11. doi:10.1155/2022/3300363

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O. ve Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(8), em2307. doi:10.29333/ejmste/13428

Parker, J. L., Becker, K. ve Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *Journal of Nursing Education, 62*(12), 721-727.

Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Thousand Oaks: CA: Sage Publications.

Pinto, G., Cardoso-Pereira, I., Ribeiro, D. M., Lucena, D., de Souza, A. ve Gama, K. (2023). *Large language models for rducation: Grading open-ended questions using ChatGPT.* arXiv. doi:10.48550/ARXIV.2307.16696

Poole, F. J. ve Coss, M. D. (2024). Can ChatGPT reliably and accurately apply a rubric to L2 writing assessments? The devil is in the prompt(s). *Journal of Technology & Chinese Language Teaching*, *15*(1).

Ramineni, C. ve Williamson, D. (2018). Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the GRE® General Test. *ETS Research Report Series, 2018*(1), 1-31. doi:10.1002/ets2.12192

Quah, B., Zheng, L., Sng, T. J. H., Yong, C. W. ve Islam, I. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Medical Education*, *24*(1). doi:10.1186/s12909-024-05881-6

Sarwanto, Fajari, L. E. W. ve Chumdari. (2021). Open-ended questions to assess critical thinking skills in Indonesian elementary school. *International Journal of Instruction, 14*(1), 615-630. doi:10.29333/iji.2021.14137a

Senkivska, L. (2022). The role of digital technologies in education. *Journal of Education, Health and Sport*, *12*(1), 419-423. doi:10.12775/jehs.2022.12.01.036

Septiani, S., Retnawati, H. ve Arliani, E. (2022). Designing closed-ended questions into open-ended questions to support student's creative thinking skills and mathematical communication skills. *JTAM (Jurnal Teori Dan Aplikasi Matematika), 6*(3), 616. doi:10.31764/jtam.v6i3.8517

Suherman, S. ve Vidákovich, T. (2022). Assessment of mathematical creative thinking: A systematic review. *Thinking Skills and Creativity*, 44, 101019. doi:10.1016/j.tsc.2022.101019

Sychev, O., Anikin, A. ve Prokudin, A. (2020). Automatic grading and hinting in open-ended text questions. *Cognitive Systems Research, 59,* 264-272. doi:10.1016/j.cogsys.2019.09.025

Uysal, İ. ve Doğan, N. (2021). How reliable is it to automatically score open-ended items? An application in the Turkish language. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 12*(1), 28-53. doi:10.21031/epod.817396

von Davier, M., Tyack, L. ve Khorramdel, L. (2022). *Automated scoring of graphical open-ended responses using artificial neural networks.* arXiv. doi:10.48550/arXiv.2201.01783

Winarso, W. ve Hardyanti, P. (2019). Using the learning of reciprocal teaching based on open ended to improve mathematical critical thinking ability. *EduMa: Mathematics Education Learning and Teaching, 8*(1). doi:10.24235/eduma.v8i1.4632

Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y. ve Fu, Q. (2024). *Human-AI collaborative essay scoring: A dual-process framework with LLMs*. arXiv. doi:10.48550/arXiv.2401.06431

Yaneva, V., Baldwin, P., Jurich, D. P., Swygert, K. ve Clauser, B. E. (2023). Examining ChatGPT performance on USMLE sample items and implications for assessment. *Academic Medicine, 99*(2), 192-197.

Zesch, T., Horbach, A. ve Zehner, F. (2023). To score or not to score: Factors influencing performance and feasibility of automatic content scoring of text responses. *Educational Measurement Issues and Practice, 42*(1), 44-58. doi:10.1111/emip.12544

Zhang, D. ve Yuan, X. (2022). Intelligent scoring of English composition by machine learning from the perspective of natural language processing. *Mathematical Problems in Engineering*, 2022, 1-9. doi:10.1155/2022/9070272