ISSN: 2146 - 9466
www.ijtase.net
International Journal of New Trends in Arts, Sports & Science Education – 2025, volume 14, issue 4

ANALİZ DERSİ YAZILI SINAVLARININ ÜRETKEN YAPAY ZEKA İLE DEĞERLENDİRİLMESİ

EVALUATION OF ANALYSIS COURSE WRITTEN EXAMS USING GENERATIVE ARTIFICIAL INTELLIGENCE

Şevval YÜKSEL
Yüksek Lisans Öğrencisi, Dokuz Eylül Üniversitesi, İzmir, Türkiye
ORCİD: https://orcid.org/0009-0003-8441-8235
sevvalyuksel883@gmail.com

Cenk KEŞAN
Prof. Dr., Dokuz Eylül Üniversitesi, İzmir, Türkiye
ORCİD: https://orcid.org/0000-0003-2629-8119
cenkkesan@gmail.com

Received: June 16, 2025 **Accepted:** October 18, 2025 **Published:** October 31, 2025

Suggested Citation:
Yüksel, Ş., & Keşan, C. (2025). Analiz dersi yazılı sınavlarının üretken yapay zeka ile değerlendirilmesi. *International Journal of New Trends in Arts, Sports & Science Education (IJTASE), 14*(4), 133-147.

Öz

Bu araştırmanın amacı İlköğretim Matematik Öğretmenliği programında öğrenim gören öğrencilerin yazılı sınav kağıtlarının yapay zeka aracılığı ile notlandırılmasını incelemektir. Bu amaçla Ege Bölgesi'ndeki bir devlet üniversitesinde öğrenim görmekte olan birinci sınıf ilköğretim matematik öğretmenliği programındaki 2025-2026 güz yarıyılında "Analiz 1" derslerine katılım sağlayan 50 öğrencinin yazılı kağıtları incelenmiştir. Araştırmanın ilişkisel farklarını belirlemek amacıyla nicel araştırma yöntemlerinden tarama modeli; aynı zamanda bu notlandırmalarını detaylarını incelemek amacıyla nitel araştırma yöntemlerinden betimsel yöntem kullanılmıştır. Araştırmadaki veriler uzmanların vermiş olduğu puanlar, kendi oluşturmuş olduğu cevap anahtarına göre değerlendirme yapan üretken yapay zeka ve verilen cevap anahtarına göre değerlendirme yapan üretken yapay zekanın puanlarından elde edilmiştir. Bu puanlar SPSS programında ayrı ayrı analiz edilmiştir. Yapılan analizler sonucunda uzmanların aynı performansa farklı puanlar verebildiği tespit edilmiştir. Uzmanlar ve üretken yapay zeka arasındaki ilişki incelendiğinde de istatistiksel olarak anlamlı farklılık olduğu saptanmıştır. Üretken yapay zekanın ölçme değerlendirmede yapmış olduğu değerlendirmeler iyileştirilebilir.

Anahtar Terimler: Üretken yapay zeka, yazılı sınav kağıtları, ölçme değerlendirme.

Abstract

The purpose of this study is to examine the grading of written exam papers of students enrolled in the Elementary Mathematics Education program using artificial intelligence. To this end, the written papers of 50 students enrolled in the first year of the Primary Education Mathematics Teaching program at a state university in the Aegean Region who participated in the "Analysis 1" courses in the fall semester of 2025-2026 were examined. The survey model, one of the quantitative research methods, was used to determine the relational differences in the research; while the descriptive method, one of the qualitative research methods, was used to examine the details of these grading processes. The data in the research were obtained from the scores given by the experts, the scores given by the generative artificial intelligence that evaluated according to its own answer key, and the scores given by the generative artificial intelligence that evaluated according to the given answer key. These scores were analyzed separately using the SPSS program. The analyses revealed that experts could assign different scores for the same performance. When examining the relationship between experts and generative artificial intelligence, a statistically significant difference was also found. The assessments made by generative artificial intelligence in measurement and evaluation can be improved.

Keywords: Generative artificial intelligence, written exam papers, measurement and evaluation.

GİRİŞ

Son yıllarda yapay zekâ teknolojilerinin kullanımı hızla artmıştır (Adaş ve Erbay, 2022; Aşık vd., 2023). ChatGPT, Gemini, Stable Diffusion, Claude ve Deepseek gibi uygulamalar bu alana örnek olarak gösterilebilir. Söz konusu teknolojilerin kullanım alanları ise oldukça çeşitli ve geniştir (Özdal, 2023). Bu alanlardan birisi de eğitim alanıdır. Son yıllarda, yapay zeka uygulamalarının eğitim

alanında kayda değer bir ivme kazandığı, öğrenme-öğretme ve değerlendirme süreçlerini dönüştürme potansiyeliyle ön plana çıktığı görülmektedir (Xia vd., 2024). Özellikle ölçme ve değerlendirme süreçlerinde, yapay zeka destekli uygulamaların sunduğu olanaklar sayesinde geleneksel yöntemlerde köklü bir dönüşüm yaşanmaya başlamıştır (Akça ve Yılmaz Keleş, 2025). Günümüzde otomatik puanlama sistemleri, yapay zekanın yazılı belgeleri puanlamak için kullandığı Doğal Dil İşleme (NLP) alanının en önde gelen akademik uygulamalarından biri olarak kabul edilmektedir (Lagakis & Demetriadis, 2021).

Öğrenciler açısından yapay zeka kullanımı, bireysel ihtiyaca göre öğrenme planı sunarak öğrenmeyi daha verimli ve kişiselleştirilmiş hale getirilebilmektedir (Aşık vd., 2023). Öğretmenler açısından yapay zekâ tabanlı değerlendirmelerin etik kullanımını sağlama, sonuçları yorumlama, öğrencilere eyleme geçirebilir geri bildirim sağlama, öğrencilerin öğrenme ihtiyaçlarına göre eğitimi bireyselleştirme, eleştirel düşünmeyi teşvik etme gibi hususlarda kritik bir role sahip oldukları belirlenmiştir (Çavuş, 2024). Bunlar içinde özellikle açık uçlu soruların ve yazılı sınavların otomatik olarak notlandırılması, zaman tasarrufu sağlaması ve değerlendirmede tutarlılığı artırma potansiyeli nedeniyle öne çıkmaktadır. Ayrıca dil öğrenme uygulamaları ve soru-cevap sistemleri gibi alanlarda da yapay zekâdan yararlanılmaktadır.

Yapay zekânın değerlendirme süreçlerinde tutarlılık sağlayabileceğine dair bulgular yalnızca eğitim alanıyla sınırlı değildir. Örneğin Posner ve Saran'ın 2025 yılında yürüttükleri bir çalışmada, 31 ABD federal yargıcının kararları ile ChatGPT'nin karar verme performansı karşılaştırılmış; insan yargıçların duygusal faktörlerden etkilenerek tutarsız kararlar verebildiği, buna karşılık yapay zekâ modelinin duygusal unsurlardan bağımsız kalarak hukuki emsallere daha sıkı bağlı kaldığı tespit edilmiştir. Bu durum, yapay zekânın objektif ve tutarlı değerlendirme yapabilme potansiyelini ortaya koymaktadır. Bu bulgu, eğitimde objektif değerlendirme için yapay zekânın kullanılabileceğini düşündürmektedir.

Günümüzde yazılı sınav sonuçları liseye geçiş sürecinde yerel yerleştirmelerde okul başarı puanı (OBP) esas alınmakta (Yalçın, 2019), yükseköğretime geçişte ise bu puan hesaplamalara doğrudan etki etmektedir. Dolayısıyla yazılı sınav sonuçlarının doğru, objektif ve hatasız bir şekilde değerlendirilmesi büyük önem taşımaktadır (Bektaş, M., & Kudubeş, A. A., 2014).

Ancak yazılı sınavlarda, herhangi bir cevabın tamamen doğru veya tamamen yanlış olarak değerlendirilmesi çoğu zaman mümkün değildir (Bektaş, M., & Kudubeş, A. A., 2014). Bu durum, aynı cevabın farklı öğretmenlerce farklı puanlanmasına yol açabilmektedir. Yazılı sınavların en olumsuz yönü puanlama güvenirliğinin düşük olmasıdır. Bu nedenle puanlamaya başlamadan önce cevap anahtarı ve nereye ne kadar puan verileceğinin gösterildiği puanlama yönergesi hazır bulundurulmalıdır (Başol, 2019). Bu nedenle, verilen puanın doğruluğu sıklıkla tartışma konusu olur. Yapay zekâ tarafından desteklenen otomatik notlandırma sistemleri, sınav değerlendirme sürecinde devrim yaratarak doğruluğu ve verimliliği artırmıştır (Öksüz Gül,2024) . Bu tür değerlendirmeler tutarsızlıklarını azaltma ve ölçme sürecinde objektifliği artırma potansiyeli taşımaktadır. Bu araştırma, söz konusu potansiyelin birinci sınıf öğrencilerinin yazılı sınav kâğıtlarının değerlendirilmesine yansımalarını incelemeyi amaçlamaktadır.

Ölçmeye karışan hatalar nedeniyle aynı kişiler ya da farklı kişilerin tekrarlı ölçümleri sonucu farklı sonuçlar elde edilme ihtimali vardır. Ölçme aracının duyarlılığı ve ölçmeyi yapan bireyin dikkati, fiziksel ölçümlerin tekrarlı ölçmelerde değişmezliği üzerinde etkilidir (Uysal Saraç, 2025).

Bir ölçme aracında, ölçülmek istenen nitelikler dışında ölçmeye karışan etkenler ölçme hatasına sebep olur. Ölçülmek istenen nitelikler dışında önemli görülen hususlara ayrıca puan vermek bu hatayı ortadan kaldıracak, ölçmenin daha sağlıklı sonuçlar vermesini sağlayacaktır.

Zorbaz'ın 2005 yılındaki çalışmasında; literatürde eğitimcilerin önemli bir kısmının, ölçülmek istenen özellik dışında kalan özelliklere (anlatım güzelliği, dil bilgisi hataları, kâğıt düzeni, yazı güzelliği, noktalama ve imla kuralları) ayrı puan vermeyi tercih ettikleri; bir kısım eğitimcilerin ise (%34,1) ölçülmek istenen özelliklerin puanlamasına bu öğeleri de dâhil ettikleri belirtilmektedir. Ayrıca

literatürde, yazılı sınavların değerlendirilmesi için standart bir not baremi yönergesi oluşturulduğu ancak eğitimcilerin %31,9'unun bu yönergeyi fazla önemsemediklerini saptamıştır.

Bu araştırmada İlköğretim Matematik Öğretmenliği programında öğrenim gören öğrencilerin yazılı sınav kağıtlarının üretken yapay zeka aracılığı ile notlandırılmasını incelemek amaçlanmaktadır. Araştırmanın amacı doğrultusunda aşağıda verilen problemlere cevap aranmıştır.

1. Uzmanların aynı performansa vermiş oldukları farklı soru puanları arasında istatistiksel olarak anlamlı bir farklılık görülmekte midir?
2. Uzmanların puanlamaları toplam puanlamada istatistiksel olarak anlamlı bir fark göstermekte midir?
3. Üretken Yapay zekanın verilen cevap anahtarıyla yapmış olduğu puanlama ile kendi oluşturmuş olduğu cevap anahtarıyla yapmış olduğu puanlama arasında istatistiksel olarak anlamlı bir fark var mıdır?
4. Uzmanlar ile üretken yapay zekanın değerlendirmeleri arasında istatistiksel olarak anlamlı farklılık var mıdır?
5. Uzmanlar ve üretken yapay zekanın değerlendirme süreçlerinde gözlemlenen farklılıklar var mıdır?

YÖNTEM

Araştırmanın Deseni

İlköğretim Matematik Öğretmenliği yazılı sınavlarının yapay zekâ ile değerlendirilmesi inceleyen bu araştırmanın yöntemi, nicel ve nitel araştırma yöntemlerinin bir arada kullanıldığı karma bir araştırma yöntemi ile yürütülmüştür.

Araştırmanın nicel boyutunda tarama modeli kullanılmıştır. "Bir konuya ya da olaya ilişkin katılımcıların görüşlerinin ya da ilgi, beceri, yetenek, tutum vb. özelliklerinin belirlendiği genellikle diğer araştırmalara göre görece daha büyük örneklemler üzerinde yapılan araştırmalara tarama araştırmaları denir" (Büyüköztürk ve diğerleri, 2021, s. 184). Bu araştırma türünde asıl amaç bir özellik, eylem veya görüşü olduğu haliyle ortaya koymaktır (Karasar, 1984, s.79; Akt., Özdemir ve Doğruöz, 2020).

Araştırmanın nitel boyutunu ise betimsel çalışma oluşturmaktadır. Büyüköztürk ve diğerlerinin (2021) çalışmasında bilimsel araştırmalar şu şekilde tanımlanır: "Betimsel (descriptive) araştırmalar, verilen bir durumu olabildiğince tam ve dikkatli bir şekilde tanımlar"(s. 25). Betimsel analizde kullanılan gözlem, görüşme ve doküman gibi veri toplama araçlarında yer alan soru, konu ya da temalar temele alınarak analiz edilir (Ekiz, 2020). Çalışmamızda verilerimiz yazılı sınav kağıtlarında bulunan üç tane sorudan oluşmaktadır.

Çalışma Grubu

Araştırmanın çalışma grubunu, Ege Bölgesi'ndeki bir devlet üniversitesinde öğrenim görmekte olan birinci sınıf ilköğretim matematik öğretmenliği programındaki öğrenciler oluşturmaktadır. Çalışmanın katılımcılarını 2025-2026 güz yarıyılında "Analiz 1" derslerine katılım sağlayan 50 öğrenci oluşturmaktadır.

Veri Toplama Aracı

Aydın ve Önder'in 2010 yılında yürüttükleri bir çalışmada, klasik sorularla sınava hazırlanma yönteminin başarıyı artırmadaki etkisinin, çoktan seçmeli sorularla sınava hazırlanma yöntemine göre oldukça üstün olduğunu ortaya koymuşlardır.

McKeachie (1986) uzun cevap gerektiren açık uçlu soruların öğrencilerin sahip oldukları bilgileri kendi cümleleriyle düzenleme, birleştirme, yorumlama ve ifade etmeleri açısından önemli birer araç olduğunu belirtmekte; uzun cevap gerektiren ve açık uçlu soruların bulunduğu sınavlara, aynı derse ait fakat çoktan seçmeli olarak uygulanan bir sınava göre çok daha fazla ve kapsamlı çalıştıklarını dile getirmektedir.

Bu gerekçeler doğrultusunda, verilerimiz yazılı sınav kağıtlarından elde edilmiştir.

Veri Toplama Süreci

Araştırma için gerekli izinler alındıktan sonra 2025-2026 eğitim-öğretim yılının güz döneminde araştırmacılar tarafından belirlenen bir devlet üniversitesinde öğrenim gören ilköğretim matematik öğretmenliği bölümü öğrencilerinden toplanmıştır. Verilerin tamamı, çalışma kapsamında hazırlanan yazılı sınav soruları yoluyla toplanmıştır. Çalışma kapsamında hazırlanan yazılı sınav soruları "Analiz 1" dersinde uygulanmıştır. Sınav soruları tahtaya sırayla yazılıp öğrencilerden sıralı bir şekilde çözümlenmeleri istenmiştir. Sorular için her sınıfa 1 ders saati (40 dakika) verilmiştir.

Veri Analizi

Verilerin toplanması tamamlandıktan sonra 4 uzmanın değerlendirebilmesi için veriler çoğaltılıp 4 uzmana dağıtılmıştır. Daha sonra her kağıdın fotoğrafı çekilip Gemini Pro'ya her öğrenci için ayrı ayrı yüklenmiştir. Üretken yapay zekadan hem kendisinin bir cevap anahtarı oluşturması ve oluşturmuş olduğu cevap anahtarına göre değerlendirmesi istenmiştir. Sonrasında yeni bir sohbet başlatılarak bizim oluşturduğumuz cevap anahtarı ve sorular Gemini Pro'ya gönderilmiş ve bu çerçevede değerlendirilmesi istenmiştir. Uzmanlar ve üretken yapay zekanın değerlendirmesi bittikten sonra veriler önce Excel dosyasına aktarılmıştır.

Daha sonrasında her sorunun analizi için SPSS programı kullanılmıştır. Verilerin normal dağılıp dağılmadığının görülmesi için Shapiro-Wilk testi uygulanmıştır. Normal olmayan dağılım için Non-parametric testlerden Friedman testi ve Willcoxon testi uygulanmıştır. Her öğrencinin bir sorusuna verilen puanların değerlendirici uzmanlar arasındaki farklılığını ortaya koymak için tek yönlü varyans analizi (ANOVA) kullanılmıştır. Çalışmanın güvenirlik analizi için Cronbach alfa katsayısı incelenmiştir.

Geçerlik ve Güvenirlik

Araştırmanın güvenirliğini artırmak adına;

- İçerik bakımından homojen maddelerden oluşan sorular kullanılmıştır (Büyüköztürk vd., 2021).
- Öğrencilerin açık uçlu bir soruya verdikleri cevapların dört uzman tarafından belirlenmiş ölçütlere göre puanlandırılması sağlanmıştır (Büyüköztürk vd., 2021).

Araştırmanın Etik İzinleri

Etik değerlendirmeyi yapan kurulun adı: Dokuz Eylül Üniversitesi Eğitim Bilimleri Enstitüsü Müdürlüğü

Etik değerlendirme kararının tarihi: 05/11/2025
Etik değerlendirme belgesi sayı numarası: 20

BULGULAR

Bu bölümde Ege Bölgesi'ndeki bir devlet üniversitesinde öğrenim görmekte olan birinci sınıf ilköğretim matematik öğretmenliği programındaki öğrencilere uygulanan yazılı sınav kağıtlarından elde edilen veriler üzerinden yapılan istatistiksel analizler ve elde edilen bulgular sunulmuştur.

Araştırmanın birinci alt probleminde "Uzmanların aynı performansa vermiş oldukları farklı soru puanları arasında istatistiksel olarak anlamlı bir farklılık görülmekte midir?" sorusuna yanıt aranmıştır. Uzmanların birinci soru değerlendirmelerinden elde edilen puanların ortalamaları arasındaki farkı tespit edebilmek için tek yönlü varyans analizi (ANOVA) kullanılmıştır. Uzmanların birinci soruya vermiş oldukları puanların ortalamalarına ilişkin ANOVA testine ait bulgular Tablo 1'de verilmektedir.

Tablo 1 Birinci soruya ait değerlendirmelerin ortalamalarına ilişkin ANOVA testi sonuçları

Varyansın Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	F	P	Anlamlı Fark Tukey
Gruplar arası	109,720	3	36,573	5,101	,002	1-2, 1-3, 2-4, 3-4
Grup içi	1405,280	196	7,170			
TOPLAM	1515,000	199				

Elde edilen Tukey HSD testi sonuçları birinci soru için uzmanların değerlendirmelerine göre anlamlı bir farklılık olduğunu göstermektedir($p<.05$). ANOVA analizinde öncelikle grupların varyanslarının eşit olup olmadıklarını belirleyebilmek amacıyla Levene testi yapılmıştır. Levene testi sonucuna göre $p=.068$ ($p>.05$) olduğu için grupların varyansları eşittir. ANOVA testi sonrasında farkları belirlemek üzere tamamlayıcı post-hoc analizi olarak Tukey testi uygulanmıştır. Tukey test sonuçlarına göre birinci soru değerlendirmelerinde 1.uzmanın 2.uzmana karşı 2.uzmanın lehine anlamlı bir sonuç gösterdiği gözlenmiştir. Tukey test sonuçlarına göre birinci soru değerlendirmelerinde 1.uzmanın 3.uzmana karşı 3.uzmanın lehine anlamlı bir sonuç gösterdiği gözlenmiştir. Tukey test sonuçlarına göre birinci soru değerlendirmelerinde 2.uzmanın 4.uzmana karşı 2.uzmanın lehine anlamlı bir sonuç gösterdiği gözlenmiştir. Tukey test sonuçlarına göre birinci soru değerlendirmelerinde 3.uzmanın 4.uzmana karşı 3.uzmanın lehine anlamlı bir sonuç gösterdiği gözlenmiştir.

Uzmanların ikinci soru değerlendirmelerinden elde edilen puanların ortalamaları arasındaki farkı tespit edebilmek için tek yönlü varyans analizi (ANOVA) kullanılmıştır Uzmanların ikinci soruya vermiş oldukları puanların ortalamalarına ilişkin ANOVA testine ait bulgular Tablo 2'de verilmektedir.

Tablo 2 İkinci soruya ait değerlendirmelerin ortalamalarına ilişkin ANOVA testi sonuçları

Varyansın Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	F	P
Gruplar arası	78,535	3	26,178	2,527	,059
Grup içi	2030,660	196	10,361		
TOPLAM	2109,195	199			

ANOVA analizinde öncelikle grupların varyanslarının eşit olup olmadıklarını belirleyebilmek amacıyla levene testi yapılmıştır. Levene testi sonucuna göre $p=.001$ ($p<.05$) olduğu için grupların varyansları eşit değildir. Varyanslar eşit olmadığı için ANOVA testi sonrasında farkları belirlemek üzere Brown-Forstyhe testi kullanılmıştır. Elde edilen Brown-Forstyhe testi sonuçlarına göre ikinci soru için uzmanlar arasında anlamlı bir farklılık görülmemektedir ($p>.05$)

Uzmanların üçüncü soru değerlendirmelerinden elde edilen puanların ortalamaları arasındaki farkı tespit edebilmek için tek yönlü varyans analizi (ANOVA) kullanılmıştır. Uzmanların üçüncü soruya vermiş oldukları puanların ortalamalarına ilişkin ANOVA testine ait bulgular Tablo 3'de verilmektedir.

Tablo 3 Üçüncü soruya ait değerlendirmelerin ortalamalarına ilişkin ANOVA testi sonuçları

Varyansın Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	F	P
Gruplar arası	10,120	3	3,373	,561	,642
Grup içi	1179,160	196	6,016		
TOPLAM	1189,280	199			

ANOVA analizinde öncelikle grupların varyanslarının eşit olup olmadıklarını belirleyebilmek amacıyla levene testi yapılmıştır. Levene testi sonucuna göre $p=.118$ ($p>.05$) olduğu için grupların varyansları eşittir. ANOVA testi sonrasında farkları belirlemek üzere tamamlayıcı post-hoc analizi olarak Tukey testi kullanılmıştır. Elde edilen Tukey HSD testi sonuçlarına göre üçüncü soru için uzmanlar arasında istatistiksel olarak anlamlı bir farklılık olmadığı görülmemektedir.

Araştırmanın ikinci alt probleminde "Uzmanların puanlamaları toplam puanlamada istatistiksel olarak anlamlı bir fark göstermekte midir?" sorusuna yanıt aranmıştır. Uzmanların toplam puanlama değerlendirmelerinden elde edilen puanların ortalamaları arasındaki farkı tespit edebilmek için tek

yönlü varyans analizi (ANOVA) kullanılmıştır. Uzmanların toplam değerlendirmelerine ilişkin ANOVA testine ait bulgular Tablo 4'de verilmektedir.

Tablo 4. Toplam değerlendirmelerin ortalamalarına ilişkin ANOVA testi sonuçları

Varyansın Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	F	P	Anlamlı Fark Tukey
Gruplar arası	237,320	3	79,107	2,801	,041	1-3
Grup içi	5536,200	196	28,246			
TOPLAM	5773,520	199				

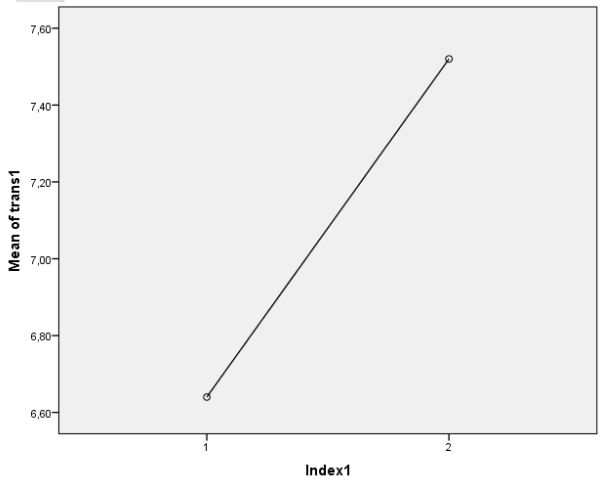
ANOVA analizinde öncelikle grupların varyanslarının eşit olup olmadıklarını belirleyebilmek amacıyla Levene testi yapılmıştır. Levene testi sonucuna göre p=.633 (p>.05) olduğu için grupların varyansları eşittir. ANOVA testi sonrasında farkları belirlemek üzere tamamlayıcı post-hoc analizi olarak Tukey testi kullanılmıştır. Tukey test sonuçlarına göre toplam değerlendirmelerinde 1.uzmanın 3.uzmana karşı 3.uzmanın lehine anlamlı bir sonuç gösterdiği gözlenmiştir.

Araştırmanın üçüncü alt probleminde "Üretken yapay zekanın verilen cevap anahtarıyla yapmış olduğu puanlama ile kendi oluşturmuş olduğu cevap anahtarıyla yapmış olduğu puanlama arasında istatistiksel olarak anlamlı bir fark var mıdır?" sorusuna yanıt aranmıştır. Üretken yapay zekanın verilen cevap anahtarıyla yapmış olduğu puanlama ile kendi oluşturmuş olduğu cevap anahtarıyla yapmış olduğu puanlamaların ortalamaları arasındaki farkı tespit edebilmek için tek yönlü varyans analizi (ANOVA) kullanılmıştır. Üretken yapay zekanın birinci soruya ait değerlendirmelerine ilişkin elde edilen ANOVA testine ait bulgular Tablo 5'te verilmektedir.

Tablo 5. Üretken yapay zekanın birinci soruya ait değerlendirmelerine ilişkin elde edilen ANOVA testi sonuçları

Varyansın Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	F	P
Gruplar arası	19,360	1	19,360	2,186	,142
Grup içi	868,000	98	8,857		
TOPLAM	887,360	99			

ANOVA analizinde öncelikle grupların varyanslarının eşit olup olmadıklarını belirleyebilmek amacıyla Levene testi yapılmıştır. Levene testi sonucuna göre p=.389 (p>.05) olduğu için grupların varyansları eşittir. ANOVA testi sonrasında farkları belirlemek üzere NonParametric test olarak Mann Whitney-U testi kullanılmıştır Elde edilen Mann Whitney U testi sonuçlarına göre üretken yapay zekanın verilen cevap anahtarıyla yapmış olduğu puanlama ile kendi oluşturmuş olduğu cevap anahtarıyla yapmış olduğu puanlama arasında arasında anlamlı bir farklılık görülmektedir (p=.016<.05).



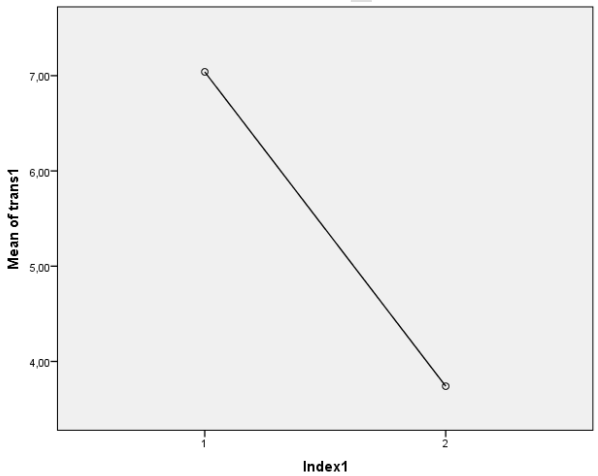
Görsel 1. Üretken yapay zekanın birinci soruya ilişkin değerlendirmelerinin ortalamaları

Üretken yapay zekanın ikinci soru değerlendirmelerinden elde edilen puanların ortalamaları arasındaki farkı tespit edebilmek için tek yönlü varyans analizi (ANOVA) kullanılmıştır. Üretken yapay zekanın ikinci soruya vermiş oldukları puanların ortalamalarına ilişkin ANOVA testine ait bulgular Tablo 6 'da verilmektedir.

Tablo 6. Üretken yapay zekanın ikinci soruya ait değerlendirmelerine ilişkin elde edilen ANOVA testi sonuçları

Varyansın Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	F	P
Gruplar arası	272,250	1	272,250	20,468	,000
Grup içi	1303,540	98	13,301		
TOPLAM	1575,790	99			

ANOVA analizinde öncelikle grupların varyanslarının eşit olup olmadıklarını belirleyebilmek amacıyla Levene testi yapılmıştır. Levene testi sonucuna göre p=.469 (p>.05) olduğu için grupların varyansları eşittir. ANOVA testi sonrasında farkları belirlemek üzere NonParametric test olarak Mann Whitney-U testi kullanılmıştır Elde edilen Mann Whitney U testi sonuçlarına göre üretken yapay zekanın verilen cevap anahtarıyla yapmış olduğu puanlama ile kendi oluşturmuş olduğu cevap anahtarıyla yapmış olduğu puanlama arasında arasında anlamlı bir farklılık görülmektedir (p=.000<.05).



Görsel 2. Üretken yapay zekanın ikinci soruya ilişkin değerlendirmelerinin ortalamaları

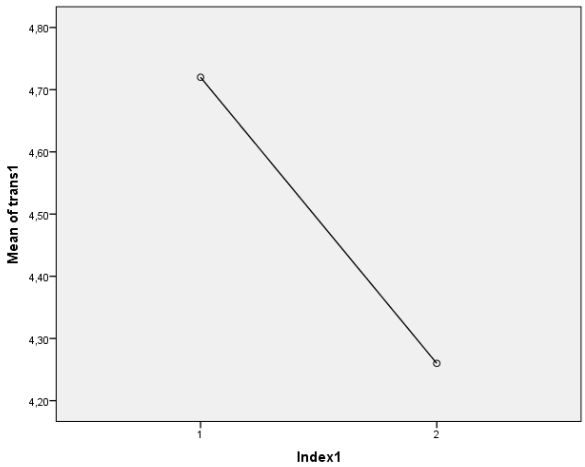
Üretken yapay zekanın üçüncü soru değerlendirmelerinden elde edilen puanların ortalamaları arasındaki farkı tespit edebilmek için tek yönlü varyans analizi (ANOVA) kullanılmıştır. Üretken yapay zekanın üçüncü soruya vermiş oldukları puanların ortalamalarına ilişkin ANOVA testine ait bulgular Tablo 7'de verilmektedir.

Tablo 7. Üretken yapay zekanın üçüncü soruya ait değerlendirmelerine ilişkin elde edilen ANOVA testi sonuçları

Varyansın Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	F	P
Gruplar arası	5,290	1	5,290	,536	,466
Grup içi	967,700	98	9,874		
TOPLAM	972,990	99			

ANOVA analizinde öncelikle grupların varyanslarının eşit olup olmadıklarını belirleyebilmek amacıyla Levene testi yapılmıştır. Levene testi sonucuna göre p=,788 (p>.05) olduğu için grupların varyansları eşittir. ANOVA testi sonrasında farkları belirlemek üzere NonParametric test olarak Mann Whitney-U testi kullanılmıştır Elde edilen Mann Whitney U testi sonuçlarına göre üretken yapay zekanın verilen cevap anahtarıyla yapmış olduğu puanlama ile kendi oluşturmuş olduğu cevap

anahtarıyla yapmış olduğu puanlama arasında arasında anlamlı bir farklılık görülmemektedir (p=.286>.05).

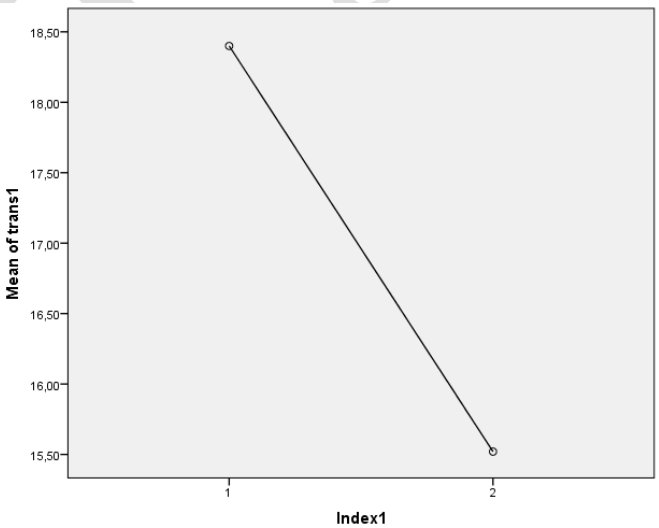


Görsel 3. Üretken yapay zekanın üçüncü soruya ilişkin değerlendirmelerinin ortalamaları

Üretken yapay zekanın toplam değerlendirmelerinden elde edilen puanların ortalamaları arasındaki farkı tespit edebilmek için tek yönlü varyans analizi (ANOVA) kullanılmıştır. Üretken yapay zekanın toplam değerlendirmeye vermiş oldukları puanların ortalamalarına ilişkin ANOVA testine ait bulgular Tablo 8'de verilmektedir

Tablo 8. Üretken yapay zekanın toplam değerlendirmelerine ilişkin elde edilen ANOVA testi sonuçları

Varyansın Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	F	P
Gruplar arası	207,360	1	207,360	5,248	,024
Grup içi	3872,480	98	39,515		
TOPLAM	4079,840	99			



Görsel 4. Üretken yapay zekanın toplam değerlendirmelerinin ortalamaları

ANOVA analizinde öncelikle grupların varyanslarının eşit olup olmadıklarını belirleyebilmek amacıyla Levene testi yapılmıştır. Levene testi sonucuna göre p=.409 (p>.05) olduğu için grupların varyansları eşittir. ANOVA testi sonrasında farkları belirlemek üzere NonParametric test olarak Mann Whitney-U testi kullanılmıştır Elde edilen Mann Whitney U testi sonuçlarına göre üretken yapay zekanın verilen cevap anahtarıyla yapmış olduğu puanlama ile kendi oluşturmuş olduğu cevap

anahtarıyla yapmış olduğu puanlama arasında arasında anlamlı bir farklılık görülmektedir (p=.016<.05).

Araştırmanın dördüncü alt probleminde "Uzmanlar ile üretken yapay zekanın değerlendirmeleri arasında istatistiksel olarak anlamlı farklılık var mıdır?" sorusuna yanıt aranmıştır. Uzmanlar ve üretken yapay zekanın değerlendirmelerinden elde edilen puanların ortalamaları arasındaki farkı tespit edebilmek için tek yönlü varyans analizi (ANOVA) kullanılmıştır. Birinci soruya ait değerlendirmelerin ortalamalarına ilişkin ANOVA testi sonuçları Tablo 9'da verilmektedir.

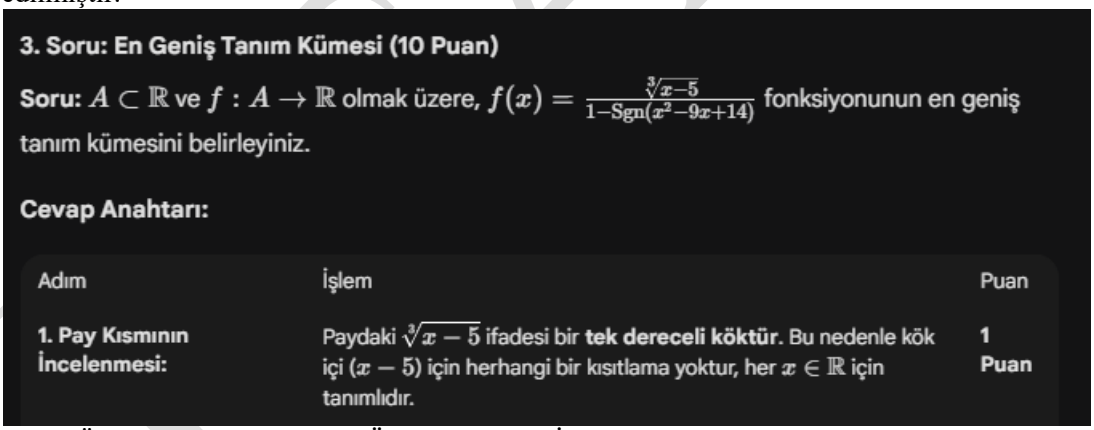
Tablo 9. Birinci soruya ait değerlendirmelerin ortalamalarına ilişkin ANOVA testi sonuçları

Varyansın Kaynağı	Kareler Toplamı	sd	Kareler Ortalaması	F	p
Gruplar arası	317, 240	5	63,448	8,206	,000
Grup içi	2273,280	294	7,732		
Toplam	2590,520	299			

ANOVA analizinde öncelikle grupların varyanslarının eşit olup olmadıklarını belirleyebilmek amacıyla levene testi yapılmıştır. Levene testi sonucuna göre p=.017 (p<.05) olduğu için grupların varyansları eşit değildir. ANOVA testi sonrasında farkları belirlemek üzere tamamlayıcı post-hoc analizi olarak Tukey testi kullanılmıştır.

Tablo 9 incelendiğinde uzmanlar ve üretken yapay zekanın değerlendirmelerine ilişkin istatistiksel olarak anlamlı bir fark tespit edilmiştir (F=8,206 ve p<.05).

Bu anlamlı farklılığın üretken yapay zekanın değerlendirmelerinden kaynaklandığı düşünülmektedir. Birinci soru olan ispat sorusunda üretken yapay zeka ispat adımlarını takip etmekte zorlanmamıştır. Fakat değerlendirme yaparken verilen cevap anahtarına göre uzmanlara daha yakın değerlendirme yaptığı görülürken; kendi oluşturmuş olduğu cevap anahtarına göre olan değerlendirmede kendi cevap anahtarına fazla sadık kalmadığı gözlenmiştir. Bu bulguya aşağıda verilmiş olan değerlendirmeden elde edilmiştir.


3. Soru: En Geniş Tanım Kümesi (10 Puan)

Soru: $A \subset \mathbb{R}$ ve $f : A \to \mathbb{R}$ olmak üzere, $f(x) = \frac{\sqrt[3]{x-5}}{1-\mathrm{Sgn}(x^2-9x+14)}$ fonksiyonunun en geniş tanım kümesini belirleyiniz.

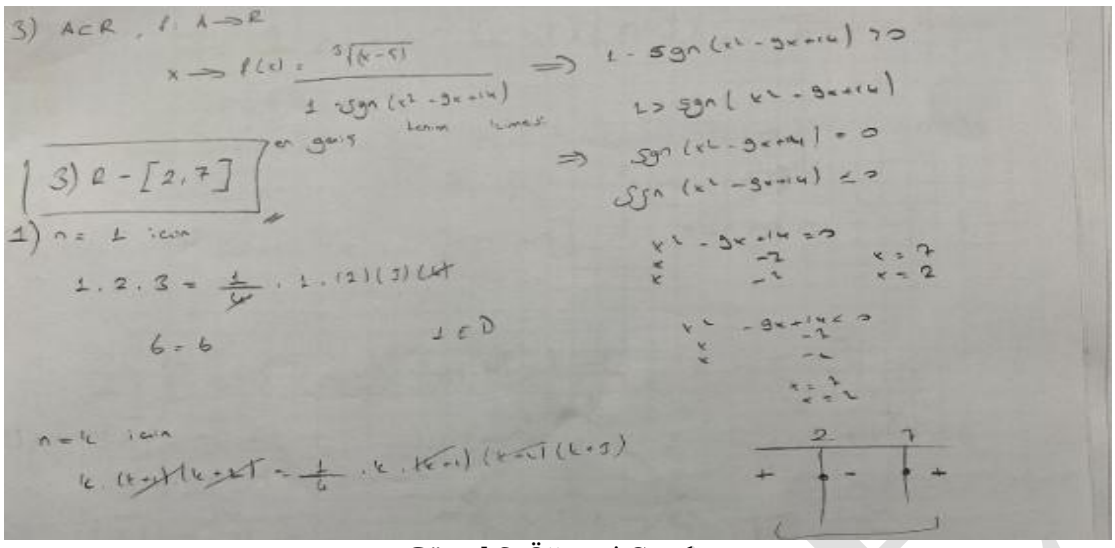
Cevap Anahtarı:

Adım	İşlem	Puan
1. Pay Kısmının İncelenmesi:	Paydaki $\sqrt[3]{x-5}$ ifadesi bir **tek dereceli köktür.** Bu nedenle kök içi $(x-5)$ için herhangi bir kısıtlama yoktur, her $x \in \mathbb{R}$ için tanımlıdır.	**1 Puan**

Görsel 5. Üretken Yapay Zekanın Üçüncü Soruya İlişkin Oluşturduğu Cevap Anahtarının Bir Kısmı

Bu görselde üretken yapay zekadan istenen cevap anahtarında 3. Sorunun cevap anahtarında bu şekilde puanlama yapılması gerektiğini söylemektedir.

ISSN: 2146 - 9466
www.ijtase.net
International Journal of New Trends in Arts, Sports & Science Education – 2025, volume 14, issue 4



Görsel 6. Öğrenci Cevabı

Bu görselde ise bir öğrencinin cevap kağıdına ilişkin yanıtı yer almaktadır. Burada öğrenci pay kısmını incelememiş, direkt çözüme geçmiştir. Bu yüzden üretken yapay zekanın kendi oluşturmuş olduğu cevap anahtarına göre bu kısımdan dolayı 1 puan vermemesi gerekmektedir. Ancak aşağıdaki görselde üretken yapay zekanın bu cevap kağıdına tam puan verdiği gözlemlenmektedir.

S3	En Geniş Tanım Kümesi	10	**10**	**Tam Puan. Payda Kısıtlaması (4 Puan):** Paydanın sıfır olmaması şartı $1 \neq \text{Sgn}(\dots)$ olarak doğru analiz edilmiş ve $x^2 - 9x + 14 \not> 0 \implies x^2 - 9x + 14 \leq 0$ eşitsizliğine doğru ulaşılmıştır. **Kökler (1 Puan):** Kökler $x = 2$ ve $x = 7$ doğru tespit edilmiştir. **Çözüm Kümesi (5 Puan):** $x^2 - 9x + 14 \leq 0$ eşitsizliğinin doğru çözümü olan $[2, 7]$ aralığına ulaşılmıştır. Tüm adımlar ve nihai sonuç doğrudur.

Görsel 7. Üretken Yapay Zekanın Kendi Oluşturmuş Olduğu Cevap Anahtarına İlişkin Değerlendiği Yazılı Kağıdının Açıklaması

Uzmanlar ve üretken yapay zekanın değerlendirmelerinden elde edilen puanların ortalamaları arasındaki farkı tespit edebilmek için tek yönlü varyans analizi (ANOVA) kullanılmıştır. İkinci soruya ait değerlendirmelerin ortalamalarına ilişkin ANOVA testi sonuçları Tablo 11'dw verilmektedir.

Tablo 11. İkinci soruya ait değerlendirmelerin ortalamalarına ilişkin ANOVA testi sonuçları

Varyansın Kaynağı	Kareler Toplamı	sd	Kareler Ortalaması	F	p
Gruplar arası	782,587	5	156,517	13,801	,000
Grup içi	3334,200	294	11,341		
Toplam	4116,787	299			

Uzmanlar ile üretken yapay zekanın değerlendirmelerine ilişkin ANOVA analizi sonucunda, uzmanlar ve üretken yapay zekanın ikinci soruya ilişkin değerlendirmelerinde okuyuculara göre istatistiksel olarak anlamlı bir fark gösterdiği bulunmuştur ($p<.05$)

Uzmanlar ve üretken yapay zekanın değerlendirmelerinden elde edilen puanların ortalamaları arasındaki farkı tespit edebilmek için tek yönlü varyans analizi (ANOVA) kullanılmıştır. Üçüncü soruya ait değerlendirmelerin ortalamalarına ilişkin ANOVA testi sonuçları Tablo 12'de verilmektedir.

Tablo 12. Üçüncü soruya ait değerlendirmelerin ortalamalarına ilişkin ANOVA testi sonuçları

Varyansın Kaynağı	Kareler Toplamı	sd	Kareler Ortalaması	F	p
Gruplar arası	782,587	5	156,517	13,801	,000
Grup içi	3334,200	294	11,341		
Toplam	4116,787	299			

Uzmanlar ile üretken yapay zekanın değerlendirmelerine ilişkin ANOVA analizi sonucunda, uzmanlar ve üretken yapay zekanın üçüncü soruya ilişkin değerlendirmelerinde okuyuculara göre istatistiksel olarak anlamlı bir fark gösterdiği bulunmuştur ($p<.05$)

Uzmanlar ve üretken yapay zekanın değerlendirmelerinden elde edilen puanların ortalamaları arasındaki farkı tespit edebilmek için tek yönlü varyans analizi (ANOVA) kullanılmıştır. Toplam değerlendirmelerin ortalamalarına ilişkin ANOVA testi sonuçları Tablo 9'da verilmektedir.

Tablo 13. Toplam değerlendirmelerin ortalamalarına ilişkin ANOVA testi sonuçları

Varyansın Kaynağı	Kareler Toplamı	sd	Kareler Ortalaması	F	p
Gruplar arası	782,587	5	156,517	13,801	,000
Grup içi	3334,200	294	11,341		
Toplam	4116,787	299			

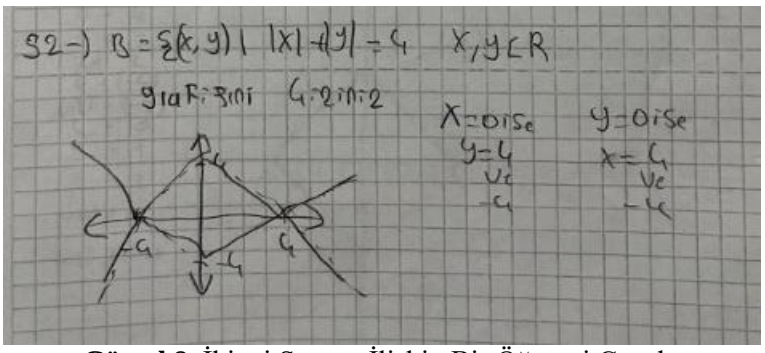
Uzmanlar ile üretken yapay zekanın değerlendirmelerine ilişkin ANOVA analizi sonucunda, uzmanlar ve üretken yapay zekanın genel puanlamasına ilişkin değerlendirmelerinde okuyuculara göre istatistiksel olarak anlamlı bir fark gösterdiği bulunmuştur ($p<.05$).

Uzmanlar ile üretken yapay zekanın değerlendirmelerinin tüm hepsinde anlamlı farklılıklar tespit edilmiştir. Bu anlamlı farklılıkların üretken yapay zekanın kendi oluşturmuş olduğu cevap anahtarına göre yapmış olduğu değerlendirmede uzmanlardan daha fazla puan vermesinden dolayı olabileceği düşünülmektedir. Aynı zamanda üretken yapay zekanın Analiz dersi kapsamındaki matematik sorularını tam olarak yorumlayamamasından kaynaklı olabileceği de düşünülmektedir. Üretken yapay zekanın kendi hazırlamış olduğu cevap anahtarı ile verilen cevap anahtarına göre yapmış olduğu değerlendirmeler kıyaslandığında da kendi hazırlamış olduğu cevap anahtarına göre yaptığı değerlendirmelerde daha fazla puan verdiği gözlenmiştir.

Bu araştırmanın beşinci alt probleminde "Uzmanlar ve üretken yapay zekanın değerlendirme süreçlerinde gözlemlenen farklılıklar var mıdır?" sorusuna yanıt aranmıştır. Bu süreçte gözlemlenen farklılıklardan biri zaman farkıdır. Üretken yapay zeka değerlendirme anlamında daha hızlı yanıt verse de genel olarak kağıtları üretken yapay zekaya değerlendirme yaptırmak için belli süreçlerden geçmesi gerekmektedir. Örneğin öncelikle tüm yazılı kağıtları ayrı ayrı fotoğraflanmalıdır. Bu işlem zaman alıcıdır. Daha sonrasında çekilen her fotoğrafı üretken yapay zeka ile olan sohbete yükleyebilmek için bilgisayara yüklemek gerekmektedir. Üretken yapay zeka ile olan sohbette her öğrenci için 2 ayrı promt girilmektedir. Bu promtları girerken bir promptta öğrenci kağıtları da sohbete yüklenmektedir. Tüm bu süreçler üretken yapay zekanın yaptığı değerlendirme sürecine dahil edildiğinde ortaya daha zaman alıcı bir değerlendirme çıkmaktadır. Uzmanlar için ayrı ayrı okuttuğumuz yazılı kağıtlarında uzmanların birbirlerinin puanlamalarını görmemeleri adına tüm yazılı kağıtlarını öncesinde 4 ayrı fotokopi aldık. Daha sonrasında her bir uzmana dağıttık. Bu süreç uzmanların değerlendirme zamanlarına eklendiğinde kağıtları okumadan önce uzmanlar için de üretken yapay zeka için de belli bir hazırlık süreci gerekmektedir.

Bir diğer farklılık grafikleri yorumlayabilmektedir. Üretken yapay zeka bazı grafikleri yanlış yorumlamıştır. Öğrencilerin el yazısı, resim çizme becerisi gibi farklılıkları yorumlayamadığı gözlenmiştir. Örneğin bir öğrenci cevabında grafik çiziminin düzgün olmaması sebebiyle yorum yapmakta güçlük çekmiştir.

ISSN: 2146 - 9466
www.ijtase.net
International Journal of New Trends in Arts, Sports & Science Education – 2025, volume 14, issue 4



Görsel 8. İkinci Soruya İlişkin Bir Öğrenci Cevabı

Üretken yapay zeka bu şekilde dışarıya doğru çizilen çizgileri hiperbol olarak yorumlamıştır. Bu da değerlendirme sürecine yansımıştır. Şekli doğru yorumlayabilseydi puanlaması daha farklı olabilirdi.

S2	Bağıntı Grafiği	10	4	**Eksen Kesim Noktaları (4 Puan):** $x = 0$ için $y = \pm 4$ ve $y = 0$ için $x = \pm 4$ noktaları doğru bulunmuştur. **Grafik Çizimi (0 Puan):** İstenen eşkenar dörtgen çizilmiş olmasına rağmen, kağıt üzerinde fazladan ve hatalı (hiperbol benzeri) eğriler çizildiği için puan kesintisi yapılmıştır. **Dört bölge denklemi** (kavramsal anlama) gösterilmemiştir.

Görsel 9. Üretken Yapay Zekanın Kendi Oluşturmuş Olduğu Cevap Anahtarına İlişkin Değerlendiği Yazılı Kağıdının Açıklaması

Ancak uzmanlar değerlendirme yaparken hiperbol veya başka şekilde düşünmeden çizimin becerisine odaklanmadan değerlendirme yapmışlardır.

Uzmanların değerlendirmelerinde hiçbir yazılı kağıdı tam puan almamıştır. Ancak üretken yapay zeka değerlendirmelerinde verilen cevap anahtarına göre değerlendirmede üç kişi tam puan alırken; kendi oluşturmuş olduğu cevap anahtarına göre olan değerlendirmede bir kişi tam puan almıştır.

TARTIŞMA, SONUÇ VE ÖNERİLER

Bu araştırmada İlköğretim Matematik Öğretmenliği programında öğrenim gören öğrencilerin yazılı sınav kağıtlarının üretken yapay zeka aracılığı ile notlandırılmasını incelemek amaçlanmıştır. Araştırmada hem nicel hem de nitel araştırma yöntemleri kullanılmıştır. Veri toplama aracı olarak yazılı sınav kağıtları kullanılmıştır. Verilerin analizinde SPSS programından yararlanılmıştır.

Üretken yapay zekanın son yıllardaki önemli gelişimi eğitim alanı için de oldukça önemlidir. Bu çalışmada ölçme değerlendirme alanında üretken yapay zekanın nasıl değerlendirme yaptığı incelenmiştir.

Üretken yapay zekadan tarafından hazırlanan cevap anahtarına göre yapmış olduğu değerlendirme araştırmacı tarafından hazırlanan cevap anahtarına göre yapmış olduğu değerlendirmeler arasında birinci, ikinci soru ve toplam değerlendirmelerde anlamlı farklılık elde edilmiştir. Yalnızca üçüncü soruya ait değerlendirmelerinde anlamlı bir farklılık gözlenmemiştir. Bu bulgulardan hareketle üretken yapay zekanın yapmış olduğu soru çözümlemeleri veya hazırladığı cevap anahtarının uzmanların hazırlamış olduğu cevap anahtarı ile farklılık gösterdiği sonucuna ulaşılmıştır. Bu da yapmış olduğu değerlendirmelerin doğruluğu ile ilgili kuşku uyandırmaktadır. Yapay zekanın değerlendirme yaparken soruyu okuma hızı bir uzmandan daha hızlıdır. Ancak vermiş olduğu yanıtlar mutlaka kontrol edilmelidir. Ve vermiş olduğu sonuçlarda karar öğretmende olmalıdır. Benzer şekilde Akça ve Yılmaz Keleş'in (2025) çalışmasında da nihai kararın mutlaka öğretmende olması gerektiği sonucu ile örtüşmektedir. Üretken yapay zekanın ölçme değerlendirme konusunda geliştirilmeye ihtiyacı vardır.

Üretken yapay zeka tarafında oluşturmuş cevap anahtarına göre ve araştırmacı tarafından hazırlanan cevap anahtarına göre yapmış olduğu değerlendirmeler sonucunda verilen puanların uzman değerlendirmelerinden fazla olduğu görülmüştür. Bu sonuç Flodén'in (2024) çalışmasındaki yapay zekanın genel olarak insanlardan daha fazla puan verdiği gözlenmiştir sonucuyla örtüşmektedir. Üretken yapay zekayı kendi içinde karşılaştıracak olursak kendi oluşturmuş olduğu cevap anahtarında daha fazla puan verdiği sonucuna ulaşılmıştır.

Uzmanların yapmış oldukları değerlendirmelerde ikinci ve üçüncü sorularda uzmanlar arasında anlamlı farklılıklar bulunmamasına rağmen birinci soru ve toplam değerlendirmelerde uzmanlar arasında anlamlı farklılıklar bulunmaktadır. Bu sonuçlardan hareketle aynı yazılı kağıdı için birbirinden bağımsız puanlar elde edildiği görülmektedir. Bu da ölçme değerlendirme anlamında istenen bir sonuç değildir. Okuyan kişinin yapmış olduğu kriter, yorgunluk durumu, yazı güzelliği vb. özelliklere puan vermiş olabileceği düşünülmektedir. Bu da puanlamadaki tutarsızlığı ortaya koymaktadır.

Ölçme ve değerlendirme alanında üretken yapay zekanın kullanımının hem olumlu hem de olumsuz yanları bulunduğu tespit edilmiştir. Üretken yapay zeka değerlendirmelerinde grafik sorusunu tam olarak yorumlayamadığı için hatalı değerlendirme yaptığı düşünülmektedir. Benzer şekilde Can'ın (2025) çalışmasında da üretken yapay zekanın bazen yanlış değerlendirmeler yapabilir sonucuyla örtüşmektedir. Bu konuda geliştirilmeye ihtiyacı vardır.

Uzmanlar ile üretken yapay zekanın değerlendirmeleri arasındaki ilişkiye bakıldığında arada anlamlı bir farklılık bulunmaktadır. Literatür incelendiğinde buna benzer çok fazla çalışma bulunamadığı için elde edilen bu bulgunun alanyazına katkıda bulunacağı aşikardır.

Öneriler

Bu çalışma kapsamında yalnızca Analiz dersindeki konuları içeren sorular sorulmuştur ve bunlar üzerinde değerlendirmeler yapılmıştır. Üretken yapay zekanın ölçme değerlendirme konusunda daha detaylı bir kanıya varmak için farklı dersler kapsamında veya farklı konular üzerindeki sorular ile bir çalışma yapılabilir.

Uygulanan sorularda grafik soruları veya cevabında şekil çizimi gerektiren sorular ile yapılan sınavlar uygulanabilir.

İlköğretim Matematik Öğretmenliği programında öğrenim gören öğrencileri ile çalışılan bu araştırma farklı bölümlerden öğrenciler ile de yürütülebilir.

Üniversitede öğrenim gören öğrencilerle yapılan bu çalışma ortaokul ve lise grubu öğrencilerinin yazılı sınav kağıtları için de uygulanabilir. Böylelikle hangi sınıf düzeyine daha tutarlı değerlendirmeler yapacağı incelenebilir.

GeminiPro ile yapılan bu değerlendirme farklı üretken yapay zeka araçları ile uygulanabilir.

Ölçme değerlendirme kapsamında birden fazla üretken yapay zeka uygulamasının birbiri ile karşılaştırılması incelenebilir.
Araştırmada kullanılacak cevapları üretken yapay zekanın okuyacağı şekilde soruların tasarlanarak en uygun şekilde öğrencilerden cevaplanmaları istenebilir. Böylelikle değerlendirmede hata payını azaltır ve ölçmenin de daha sağlıklı olmasını sağlar.

Uygulamada görülen cevabı farklı şekilde değerlendirilen cevapların doğru şekilde algılanması için üretken yapay zekaların bu kapsamda eğitilmesi sağlanabilir.

KAYNAKÇA

Adaş, E. ve Erbay, B. (2022). *Yapay zekâ sosyolojisi üzerine bir değerlendirme.* Gaziantep University Journal of Social Sciences, 21(1), 326-337.

Akça, G. ve Yılmaz Keleş, A. (2025). Yapay zeka ile ölçme değerlendirme. C. Şahin Kölemen (Ed.), *Eğitimde zekanın yeni yüzü yapay zeka* (s. 47-58). Pegem Akademi.

Aşık, F., Yıldız, A., Kılınç, S., Aytekin, N., Adalı, R., & Kurnaz, K. (2023). *Yapay zekânın eğitime etkileri.* International Journal of Social and Humanities Sciences Research (JSHSR), 10(98), 2100–2107. https://doi.org/10.5281/zenodo.8307107

Aydın, E. ve Önder, O. (2013). *Sınava hazırlık biçiminin farklı sınav türlerinde ölçülen matematik sınav başarı düzeylerine etkisi.* Marmara Üniversitesi Atatürk Eğitim Fakültesi Eğitim Bilimleri Dergisi, 31(31), 5-24.

Başol, G. (2019). *Eğitimde ölçme değerlendirme*. Pegem Akademi. (Orijinal yayın tarihi 2012).

Bektaş, M., & Kudubeş, A. A. (2014). *Bir Ölçme ve Değerlendirme Aracı Olarak: Yazılı Sınavlar*. Dokuz Eylül Üniversitesi Hemşirelik Fakültesi Elektronik Dergisi, 7(4), 330-336.

Büyüköztürk, Ş. ve Kılıç Çakmak, E. Ve Akgün, Ö. E. ve Karadeniz, Ş. Ve Demirel, F. (2021). *Eğitimde bilimsel araştırma yöntemleri*. Pegem Yayınevi.

Can, E. (2025). *Bilimsel yayınların yapay zeka programları tarafından değerlendirilmesine ilişkin literatür incelemesi*. Kapanaltı Dergisi(7), 61-77.

Çavuş, M. N. (2024). *Eğitimde Yapay Zekâ Tabanlı Ölçme ve Değerlendirme Üzerine Bir Derleme*. International Journal of English for Specific Purposes, 2(1), 39-54.

Ekiz, D. (2020). *Bilimsel araştırma yöntemleri.* Anı Yayıncılık.

Flodén, J. (2024). *Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT.* British Educational Research Journal, 51(1), 201-224. https://doi.org/10.1002%2Fberj.4069

Lagakis, P., & Demetriadis, S. (2021, November). Automated essay scoring: A review of the field. In 2021 International Conferrence on Computer, Information and Telecommunication Systems (CITS) (pp. 1-6). IEEE.

McKeachie, W.J. (1986). Teaching Tips (8.Baskı), Lexington, MA: Heath Inc.

Öksüz Gül, F. (2024). *Eğitimde yapay zeka: Eğitim fakültesi akademisyenleri için fırsatları ve riskleri*. Medeniyet Eğitim Araştırmaları Dergisi, 8(2), 71-97.

Özdal, M. A. (2023). *Görüntü içeriği sınıflandırmasında yapay zekanın rolü ve uygulamaları.* Van İnsani Ve Sosyal Bilimler Dergisi(6), 37-61. https://doi.org/10.62068/visbid.1352901

Özdemir, M. & Doğruöz, E. (2020). Bilimsel araştırma desenleri. N, Cemaloğlu (Ed.), *Bilimsel araştırma teknikleri ve etik. 65-98*. Pegem Akademi.

Posner, E. A. ve Saran, S. (2025). *Judge AI: Assessing large language models in judicial decision-making*. University of Chicago Coase-Sandor Institute for Law & Economics Research Paper, (2503).

Uysal Saraç, M. (2025). Geleneksel ve modern ölçme araçları. M. Şata vd (Ed.), *Eğitimde ölçme ve değerlendirme temel ilkelerden yenilikçi uygulamalara* (s. 13-35). Pegem Akademi.

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., &Chiu, T.K. (2024). *A scoping review on how generative artifical intelligence transforms assesment in higher education.* International Journey of Educational Technology in Higher Education, 21(1), 40. htthps://doi.org/10.1186/s41239-024-00468-z

Yalçın, E. (2019). *Liseye giriş sınavı (LGS)'nın yönetici, öğretmen, öğrenci ve veliye göre incelenmesi.* Eğitimde Ölçme ve Değerlendirme Tezli Yüksek Lisans Programı Yüksek Lisans Tezi, Akdeniz Üniversitesi Eğitim Bilimleri Enstitüsü, Antalya.

Zorbaz, K.Z. (2005*). İlköğretim okulları ikinci kademe Türkçe öğretmenlerinin ölçme değerlendirmeye ilişkin görüşleri ve yazılı sınavlarda sordukları sorular üzerine bir değerlendirme*. Eğitim Bilimleri Programı Yayınlanmamış Yüksek Lisans Tezi, Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü. Hatay, Türkiye.

EXTENDED ABSTRACT

This study aims to examine the grading of written examination papers of students enrolled in the Elementary Mathematics Education program through generative artificial intelligence. In line with the objective of the research, answers were sought for the following problems: 1. Is there a statistically significant difference between the different question scores given by experts for the same performance? 2. Do the experts' ratings show a statistically significant difference in the total scoring? 3. Is there a statistically significant difference between the scoring made by Generative Artificial Intelligence with the given answer key and the scoring made with the answer key it created? 4. Is there a statistically significant difference between the evaluations of experts and generative artificial intelligence? 5. Are there observed differences in the evaluation processes of experts and generative artificial intelligence? The method of this research, which examines the evaluation of Elementary

Mathematics Education written exams with artificial intelligence, was carried out with a mixed research method in which quantitative and qualitative research methods were used together. The screening model was used in the quantitative dimension of the research. The qualitative dimension of the research consists of a descriptive study. The study group of the research consists of students in the first-year elementary mathematics teaching program studying at a state university in the Aegean Region. The participants of the study consist of 50 students who participated in "Analysis 1" courses in the 2025-2026 fall semester. Written exam papers were used as a data collection tool in our research. After the necessary permissions were obtained for the research, the written exam papers prepared within the scope of the study were applied in the "Analysis 1" course. After the data collection was completed, the data were reproduced and distributed to 4 experts for evaluation. Then, a photograph of each paper was taken and uploaded to Gemini Pro separately for each student. Generative artificial intelligence was asked to both create an answer key itself and evaluate it according to the answer key it had created. SPSS program was used for the analysis of quantitative data. The Shapiro-Wilk test was applied to see whether the data were normally distributed. Friedman test and Wilcoxon test, which are non-parametric tests, were applied for non-normal distribution. One-way analysis of variance (ANOVA) was used to reveal the difference between the evaluating experts in the scores given to one question of each student. Cronbach's alpha coefficient was examined for the reliability analysis of the study. As a result of the findings obtained, a significant difference was obtained in the first, second questions, and total evaluations between the evaluation made by generative artificial intelligence according to the answer key prepared by itself and the evaluation made according to the answer key prepared by the researcher. A significant difference was not observed only in the evaluations of the third question. Based on these findings, it was concluded that the question analysis made by generative artificial intelligence or the answer key it prepared differed from the answer key prepared by the experts. This raises doubts about the accuracy of the evaluations it has made.