
GAZİ ÜNİVERSİTESİ
Gazi
1926

İNGİLİZCE AKADEMİK YAZMA BECERİLERİNİN DEĞERLENDİRİLMESİNDE PUANLAYICI EĞİTİMİNİN AKRAN PUANLAYICI DAVRANIŞLARI ÜZERİNE ETKİSİNİN ÇOK YÜZEYLİ RASCH MODELİ İLE BELİRLENMESİ

Fatıma Nur Fişne

YÜKSEK LİSANS TEZİ

EĞİTİM BİLİMLERİ ANA BİLİM DALI

EĞİTİMDE ÖLÇME VE DEĞERLENDİRME BİLİM DALI

GAZİ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

ARALIK, 2025

TELİF HAKKI VE TEZ FOTOKOPİ İZİN FORMU

Bu tezin tüm hakları saklıdır. Kaynak göstermek koşuluyla tezin teslim tarihinden itibaren 12 ay sonra tezden fotokopi çekilebilir.

YAZARIN

Adı	: Fatıma Nur
Soyadı	: Fişne
Bölümü	: Eğitim Bilimleri/Eğitimde Ölçme ve Değerlendirme
İmza	:
Teslim tarihi	:

TEZİN

Türkçe Adı : İngilizce Akademik Yazma Becerilerinin Değerlendirilmesinde Puanlayıcı Eğitiminin Akran Puanlayıcı Davranışları Üzerine Etkisinin Çok Yüzeyli Rasch Modeli ile Belirlenmesi

İngilizce Adı : Examining the Effect of Rater Training on Peer Rater Behaviors in the Assessment of Academic Writing Skills in English Through the Many-Facet Rasch Model

ETİK İLKELERE UYGUNLUK BEYANI

Tez yazma sürecinde bilimsel ve etik ilkelere uyduğumu, yararlandığım tüm kaynakları kaynak gösterme ilkelerine uygun olarak kaynakçada belirttiğimi ve bu bölümler dışındaki tüm ifadelerin şahsıma ait olduğunu beyan ederim.

Yazar Adı Soyadı: Fatıma Nur FİŞNE

İmza:

JÜRİ ONAY SAYFASI

Fatıma Nur Fişne tarafından hazırlanan "İngilizce Akademik Yazma Becerilerinin Değerlendirilmesinde Puanlayıcı Eğitiminin Akran Puanlayıcı Davranışları Üzerine Etkisinin Çok Yüzeyli Rasch Modeli ile Belirlenmesi" adlı tez çalışması aşağıdaki jüri tarafından oy birliği/oy çokluğu ile Gazi Üniversitesi Eğitimde Ölçme ve Değerlendirme Bilim Dalı'nda Yüksek Lisans tezi olarak kabul edilmiştir.

Danışman: Prof. Dr. İsmail KARAKAYA
Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, ……
Gazi Üniversitesi

Başkan: Prof. Dr. Hakan Yavuz ATAR
Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, ……
Gazi Üniversitesi

Üye: Doç. Dr. Mahmut Sami KOYUNCU
Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, ……
Afyon Kocatepe Üniversitesi

Tez Savunma Tarihi: …./…/……..

Bu tezin Eğitimde Ölçme ve Değerlendirme Bilim Dalında Yüksek Lisans tezi olması için şartları yerine getirdiğini onaylıyorum.

Prof. Dr. Osman ÇİMEN
Eğitim Bilimleri Enstitüsü Müdürü ……....

Akademik hayatta her an yanımda olan güzel aileme ithafen…

TEŞEKKÜR

Hepimizin hayatında bazı dönüm noktaları vardır. Benim dönüm noktalarından biri de Gazi Üniversitesi Eğitimde Ölçme ve Değerlendirme Tezli Yüksek Lisans programına kabul aldığım andır. Mutlulukla aldığım bu haber ile içimde barındırdığım sayıların gücünü akademik çalışmalara yansıtabilme şansını yakaladığımı düşünmüştüm ve öyle de oldu. Her ne kadar uzun bir zaman alsa da bu tezi burada sonlandırabilme hissiyatının sevincini yaşamak tez danışmanım Prof. Dr. İsmail KARAKAYA ve ailemin bana verdiği destek sayesindedir. Beni destekleyen saygıdeğer danışmanıma ve sevgili aileme teşekkür etmek isterim. Ölçme ve değerlendirme alanında yüksek lisans yapmak bana sadece akademik anlamda katkı sağlamadı; aynı zamanda pek çok değerli öğretim üyesi ile tanışma ve iş birliği yapma fırsatı da vuldum. Öncelikle tezimin jüri üyeleri Prof. Dr. Hakan Yavuz ATAR ve Doç. Dr. Mahmut Sami Koyuncu'ya tezime sundukları katkılarından dolayı çok teşekkür ederim. Benzer şekilde, Rasch analizi ile aklıma takılan her soruda beni yanıtsız bırakmayan ve desteğini her an hissettiğim saygıdeğer Dr. Umur ÖÇ'e içten teşekkürlerimi sunmak isterim. Hem meslektaşım hem de uzun seneler aynı ofisi paylaştığım değerli arkadaşım Dr. Merve ÖKSÜZ ZEREY'e tez sürecim boyunca yanımda oluşu ve akademik yolculuğumda gösterdiği içtenliği için ayrıca teşekkür ederin. Benim için ölçme ve değerlendirme alanında yüksek lisans yapmak nispeten uzun sürdü. Bu yolculukta fikir alışverişi yaptığım çok değerli araştırmacılar ile tanıştım. Aynı şekilde öğrencilerim ile uzun uzun ölçme ve değerlendirme konularını tartışma olanağım oldu. Çalışmalarıma ilham olan tüm araştırmacı arkadaşlarıma ve öğrencilerilerime teşekkürü bir borç bilirim.

İNGİLİZCE AKADEMİK YAZMA BECERİLERİNİN DEĞERLENDİRİLMESİNDE PUANLAYICI EĞİTİMİNİN AKRAN PUANLAYICI DAVRANIŞLARI ÜZERİNE ETKİSİNİN ÇOK YÜZEYLİ RASCH MODELİ İLE BELİRLENMESİ

(Yüksek Lisans Tezi)

Fatıma Nur Fişne
GAZİ ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Aralık 2025

ÖZET

İngilizce akademik yazma becerileri, öğrencilerin aktif olarak katılım göstererek farklı türlerde performans sergilemelerini gerektiren bir beceri türüdür. Öğrencinin akademik yazma performansı ise belirli ölçütler ve düzey tanımlamaları doğrultusunda puanlayıcılar tarafından değerlendirilmektedir. Puanlayıcı öznelliğinden kaynaklanan hataları en aza indirgemek için farklı puanlayıcıların aynı performansı benzer şartlarda ve aynı ölçme aracı kullanarak puanlaması gerekmektedir. Diğer taraftan bu gereklilik, puanlayıcı güvenirliliği tartışmalarını ve puanlayıcı davranışlarını gündeme getirmektedir. Bu davranışlardan kaynaklanan hatayı minimize etmek ve ölçme güvenirliğini sağlayarak daha geçerli kararlara ulaşabilmek için puanlayıcı eğitimi başvurulan yöntemlerden biridir. Bu çalışmada bir yazma türü olan kompozisyon görevlerinden reaksiyon türünde kompozisyonun akran değerlendirme kapsamında incelenmesi hedeflenmiştir. İngilizce öğretmenliği öğrencileri olan ve akademik yazma becerileri dersini alan 51 akran, öncelikle önceki sene aynı eğitimi alan akranlarının yazdığı 5 adet kompozisyon görevini puanlamıştır. Ardından akranlar "Puanlayıcı Hatası Eğitimi" ve "Performans Boyutları Eğitimi" çerçevesinde altı haftalık bir puanlayıcı eğitim almış ve uygulama sonrasında da aynı görevleri tekrar puanlamışlardır.

Madde tepki kuramına dayalı olan Çok Yüzeyli Rasch Ölçme Modeli kullanarak akran puanlamaları birey, zaman, kriter ve kompozisyon yüzeylerine göre analiz edilmiştir. Her ne kadar akranlar, eğitim sonrasında katılık ve cömertlik gibi puanlama davranışlarında az düzeyde farklılık sergileseler ve yanlılık davranışlarında azalma olsa da eğitim öncesi ve sonrası puanlamalarda anlamlı bir fark bulunamamıştır. Değişim olmayışının nedenleri arasında eğitim süresinin kısa tutulması, puanlama zamanı ve puanlayıcıların uzmanlık alanları düşünülmektedir. Model ölçüm değerleri incelendiğinde ise akademik yazma becerilerinin Çok Yüzeyli Rasch Ölçme Modeli ile güvenilir düzeyde analiz edilebileceği ve performansın uygun biçimde yorumlanabileceği tespit edilmiştir.

Anahtar Kelimeler: Yazma becerisini ölçme ve değerlendirme; akran değerlendirme, Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM)

Sayfa Sayısı: xiv + 85

Tez Danışmanı: Prof. Dr. İsmail KARAKAYA

EXAMINING THE EFFECT OF RATER TRAINING ON PEER RATER BEHAVIORS IN THE ASSESSMENT OF ACADEMIC WRITING SKILLS IN ENGLISH THROUGH THE MANY-FACET RASCH MODEL

(Yüksek Lisans Tezi)

Fatıma Nur Fişne
GAZİ UNIVERSITY
INSTITUTE OF EDUCATIONAL SCIENCES

December 2025

ABSTRACT

English academic writing skills are a type of skill that requires students' active participation and performance in various genres. Raters assess the students' academic writing performance in accordance with specific criteria and proficiency levels. To minimize errors arising from rater subjectivity, it is essential for raters to assess the same performance under similar conditions and using the same measurement tool. On the other hand, this requirement brings discussions on rater reliability and rater behaviors into play. One of the methods used to consider these behaviours and ensure measurement reliability is obviously rater training. In this study, the aim was to examine reaction essays within the scope of peer assessment. 51 participants, who are pre-service English teachers taking the Academic Writing Skills course, first scored five essays written by others who had taken the course in the previous year. Afterward, within the framework of "Rater Error Training, and "Performance Dimensions Training", peers received six weeks of rater training and then re-scored the same tasks after training. Based on Item Response Theory, the Many-Facet Rasch Measurement Model was used to analyze peer ratings in terms of the facets of person, time, criteria, and student/essay. Although peers slightly differed in scoring behaviors such as severity and leniency, and scoring bias was reduced after the training, no significant difference was found

between the pre- and post-training scores. This may be due to the short duration of the training and the ineffectiveness of the training type, considering the raters' proficiency levels, educational backgrounds, and rating/scoring times. Regarding the model measurement indices, it can be inferred that academic writing skills can be analyzed reliably using the Many-Facet Rasch Measurement Model, and that performance can be interpreted appropriately.

Key words: Assessing writing skills; peer assessment; Multi-facet Rasch Measurement Model (MFRM)

Page number: xiv + 85

Thesis supervisor: Prof. Dr. İsmail KARAKAYA

İÇİNDEKİLER

İÇİNDEKİLER x
TABLOLAR LİSTESİ xiii
ŞEKİLLER LİSTESİ xiv
SİMGELER VE KISALTMALAR LİSTESİ xiv
BÖLÜM I 1
GİRİŞ 1
Araştırmanın Problemi 1
Araştırmanın Amacı ve Soruları 4
Araştırmanın Önemi 5
Araştırmanın Varsayımları 5
Araştırmanın Sınırlıkları 6
Tanımlamalar 6
BÖLÜM II 8
ALAN YAZIN TARAMASI 8
İngilizce Akademik Yazma Becerileri 8
Akran Değerlendirme 11
Puanlayıcı Davranışları ve Puanlayıcı Eğitimi 13
Çok Yüzeyli Rasch Ölçme Modeli 17
İlgili Araştırmalar 21

BÖLÜM III 27
YÖNTEM 27
Araştırma Deseni 27
Evren ve Örneklem 28
Verilerin Toplanması 29
Reaksiyon Kompozisyonu Görevi (Reaction Essay) 30
Yazma Becerilerine Yönelik Analitik Dereceli Puanlama Aracı 31
Puanlayıcı Eğitimi 32
Veri Analizi 32
BÖLÜM IV 36
BULGULAR 36
Birinci Araştırma Sorusuna Yönelik Bulgular 36
Ikinci Araştırma Sorusuna Yönelik Bulgular 43
Üçüncü Araştırma Sorusuna Yönelik Bulgular 50
BÖLÜM V 57
TARTIŞMA 57
Birinci Araştırma Sorusuna Yönelik Bulgular ve Tartışma 57
İkinci Araştırma Sorusuna Yönelik Bulgular ve Tartışma 58
Üçüncü Araştırma Sorusuna Yönelik Bulgular ve Tartışma 59
BÖLÜM VI 61
SONUÇ VE ÖNERİLER 61
KAYNAKLAR 63
EKLER 76
Ek 1. Kriter/Ölçüt Yüzeyi Ön Uygulama ve Son Uygulama Logit Değerleri ... 77
Ek 2. Yanlılık Test Sonuçları – Puanlayıcı Sırası ve Tartışma 78

Ek 3. Veri Toplama Aracı (Kompozisyon Ödevi) 81

Ek 4. Veri Toplama Aracı (Dereceli Puanlama Anahtarı)..... 82

Ek 5. Etik Kurul Onayı 85

TABLOLAR LİSTESİ

Tablo 1. *Yaygın Puanlayıcı Hataları*……...…..…..…14

Tablo 2. *Katılımcıların Demografik Bilgileri*…….……..……..…..…28

Tablo 3. *Uygulama Kapsamında Puanlanması Beklenilen 5 Kompozisyon Görevi*……35

Tablo 4. *Uygulama Öncesi Puanlayıcı Yüzeyine Ait Ölçüm Raporu*……...37

Tablo 5. *Uygulama Öncesi Öğrenci/Kompozisyon Yüzeyine Ait Ölçüm Raporu*…….....39

Tablo 6. *Uygulama Öncesi Kriter Yüzeyine Ait Ölçüm Raporu*…….……...40

Tablo 7. *Uygulama Öncesi Puanlayıcı/Kompozisyon Etkileşim Raporu*……..……41

Tablo 8. *Uygulama Sonrası Puanlayıcı Yüzeyine Ait Ölçüm Raporu* ……..…....44

Tablo 9. *Uygulama Sonrası Öğrenci/Kompozisyon Yüzeyine Ait Ölçüm Raporu* ……...46

Tablo 10. *Uygulama Sonrası Kriter Yüzeyine Ait Ölçüm Raporu* …….……47

Tablo 11. *Uygulama Sonrası Puanlayıcı/Kompozisyon Etkileşim Raporu* ……48

Tablo 12. *Uygulama Kapsamında Puanlayıcı Yüzeyine Ait Ölçüm Raporu*……50

Tablo 13. *Uygulama Kapsamında Kompozisyon Yüzeyine Ait Ölçüm Raporu*…….52

Tablo 14. *Uygulama Kapsamında Kriter Yüzeyine Ait Ölçüm Raporu* ……..…….53

Tablo 15. *Uygulama Kapsamında Zaman Yüzeyine Ait Ölçüm Raporu* …….……...….54

Tablo 16. *Ön ve Son Puanlama Logit Değerleri T Testi Sonuçları* ……...….55

ŞEKİLLER LİSTESİ

Şekil 1. Akademik yazma süreci……..…..……...…9

Şekil 2. Veri toplama süreci…….……...…...……29

Şekil 3. Uygulama kapsamında elde edilen logit cetveli…….…….…..……56

SİMGELER VE KISALTMALAR LİSTESİ

AFA: Açımlayıcı Faktör Analizi

ASWRAE: Analitik Dereceli Puanlama Anahtarı

ÇYRM: Çok Yüzeyli Rasch Modeli

ÇYRÖM: Çok Yüzeyli Rasch Ölçme Modeli

DFA: Doğrulayıcı Faktör Analizi

G-Kuramı: Genellenebilirlik Kuramı

KTK: Klasik Test Kuramı

MTK: Madde Tepki Kuramı

BÖLÜM 1

GİRİŞ

Bu bölüm araştırmanın problem durumunu, araştırma amacını ve sorularını, çalışmanın önemini ve mevcut koşullar dahilinde varsayımlarını ve sınırlılıklarını betimlemektedir. Ayrıca bu bölümde alana ve çalışmaya özgü bazı tanımlamalar sergilenerek bir sonraki bölümde yer verilen alan yazın taramasının daha açık bir şekilde anlamlandırılması hedeflenmektedir.

Araştırmanın Problemi

Yabancı dilde ölçme ve değerlendirme hem ulusal hem de uluslararası düzeyde araştırmaların yoğun olarak gerçekleştiği alanlardan biridir. Yabancı dilin gerçek hayatta eş zamanlı karşılığının olması ve dilin farklı beceri ve ögelerden oluşması, ölçme ve değerlendirme süreçlerini de aktif yönde etkilemektedir. Yabancı dilde ölçme ve değerlendirmenin, pek çok birey için farklı akademik ve sosyal alanlarda (örn. eğitim veya yurtdışı yaşam) belirleyici bir ölçüt niteliği taşıdığı bilinmektedir (Bachman & Purpura, 2008; Hamid vd., 2019). Norton (1997) dil ile ilgili ölçme ve değerlendirme süreçlerinin sadece ölçülen birey için değil aynı zamanda politika geliştiricilerin ve uygulayıcı paydaşlar içinde önemli sonuçlar ortaya çıkardığını ifade etmektedir (s. 313). Benzer şekilde Bachman ve Palmer (2010) yabancı dilde ölçme ve değerlendirme faaliyetlerinin iç öğrenme halkasını oluşturan öğrencilerden dış halkaları kapsayan toplum ve kurumlar için hayati kararlara zemin hazırladığını vurgulamaktadır. Bu kadar geniş etkileri olan yabancı dilde ölçme ve değerlendirme süreci hakkında doğru çıkarımlar yapabilmek adına güvenilir ve geçerli ölçümler elde etmek önem taşımaktadır.

Toplumda ve akademik alanlarda geniş yankı uyandıran bir yabancı dil olarak İngilizcenin öğretilmesi ve ölçülmesi de benzer derecede geniş etkili ve çok yönlüdür. Özellikle İngilizce dil becerilerinin ölçülmesinde belirlenen amaca ve hedeflenilen düzeye göre geleneksel ve alternatif yaklaşımı içinde barındıran çeşitli yöntem ve teknik kullanılmaktadır (Brown, 2004; Carr, 2011; Douglas, 2010; Hughes, 2013; Madsen, 1980). Geleneksel yöntemler ile daha çok dil yapısı ve özelliklerine odaklanan ölçme süreci 1980'li yıllarda dil ölçme alanına eklenen "iletişim yetkinliği" (communicative competence) tartışmaları (Canale & Swain, 1980; Canale, 1987) ile daha fazla iletişimsel amacı gerçekleştirmeye yönelik anlam kazanmıştır. Bu bağlamda öğrenci performansının ön plana çıkması ve performans değerlendirmeyi odak alan yaklaşımların yaygınlaşması söz konusu olmuştur.

Performans değerlendirmenin yaygın olarak yapıldığı alanlardan biri de İngilizce dil ölçme ve değerlendirmedir. Özellikle yazma ve konuşma gibi üretici dil becerilerinde bireylerin ürün sergilemeleri, bu ürünün yargılayıcılar (rater, judge) tarafından puanlanmasını zorunlu kılmaktadır. Akademik bağlamlarda yapılan yazma pratikleri, akademik yazma olarak adlandırılmaktadır (Oshima & Hogue, 2007) ve akademik yazma sürecinde de bireylerin performans göstermesi gereken farklı yazım türleri vardır. Bu türlere derlemeler (review documents) kompozisyon/deneme çalışmaları (compositions/essays) ve araştırma raporları (research papers) örnek olarak gösterilebilir. Üniversite düzeyinde bilişsel ve dil düzeyine uygunluk açısından en yaygın kullanılan yazım türlerinden biri kompozisyondur.[1] Komposizyon/deneme türlerinin ölçme görevi olarak kullanılmasında da puanlama güveniriliği sıklıkla gündeme gelmektedir. Sudweeks vd. (2004) öğrencilerin kompozisyon puanlarının bazı faktörlerden etkilendiğine dikkat çekmiştir (s. 240). Bu faktörler sırasıyla (a) yazma görevine ait özellikler, (b) ürünü değerlendiren puanlayıcılar, (c) puanlama durumuna bağlı unsurlar, (d) öğrencinin konuya dair bilgi ve tutumu ve (e) belirtilen faktörlerin etkileşimidir. Görüldüğü üzere, yazma becerilerinin puanlanmasında puanlayıcılar ve diğer değişkenler, ölçme ve değerlendirme sürecinin tarafsız olmasını, geçerli ve güvenilir tamanlanmasını etkileyebilmektedir.

Öğrenci performansını öne çıkaran ve öğrencilerin puanlayıcı olarak aktif katılımı gerektiren yaklaşımlardan birisi de akran değerlendirmedir. Akran değerlendirme, öğrenciye ait bir performansın aynı öğretim ortamındaki diğer bir öğrenci tarafından belirli ölçütler

[1] Kompozisyon, yabancı alan yazında deneme (essay) olarak da kullanılabilmektedir. Kramer (2022) ise denemenin bir komposizyon türü olduğunu ifade etmiştir.

doğrultusunda değerlendirilmesidir (Kutlu vd., 2014; Spiller, 2012). Akran değerlendirmenin, geleneksel yaklaşımlar ile kıyaslandığında farklı avantajlara sahip olduğu görülmektedir (Brown, 2013; Douglas, 2010). Bu avantajlardan başlıcaları öğrencilerin sürece aktif katılımı ve performans ölçütleri ile ilgili deneyim kazanarak kendi çalışmalarına eleştirel gözle bakabilmesidir. Benzer şekilde, öğrenciler akran değerlendirme yaparken hem öğrenme süreçlerini hem de akranların ürünlerini izleme, analiz etme ve değerlendirme sorumluluğu taşıdığı bilinmektedir (Cheng & Warren, 2005, s. 94). Öte yandan bu süreçte yaşanan en büyük kaygılardan biri akranlardan kaynaklanan davranışlar ve bu davranışların geçerlik ve güvenirlilik süreçlerini etkilemesi olmuştur (Marsaulina, 2025; Zhang vd., 2020). Li vd. (2022) akran değerlendirmenin İngilizce ölçme güvenirliği için neden tehdit oluşturabileceğini dört madde ile açıklamıştır (s. 2):

1. Akranların belirli dil yetkinliğine sahip olamaması
2. Rekabetçi bir öğrenme ortamının olması
3. Akranların puanlama deneyiminin veya eğitiminin yetersiz olması
4. Akran değerlendirmenin uygun yönetilememesi

Yukarıda tartışıldığı üzere puanlayıcı davranışları ve ölçme ortamı, akran değerlendirmenin güvenirlik niteliğini şekillendiren etmenlerdir. Akranların, birbirlerinden gelen dönütleri risk olarak değil de öğrenmeyi destekleyen bir unsur olarak görmeleri ile akran değerlendirme amacına daha çok hizmet edebilmektedir (Brookhart, 2013). Puanlama pratiğinin ve alan bilgisi yeterliğinin eksik olması da puanlayıcılar olarak akranlarda ölçme güvenirliğini tehdit eden davranışlara neden olmaktadır.

Puanlayıcı davranışları Myford ve Wolfe (2004) ve Royal ve Hecker (2016) tarafından daha geniş bir çerçevede kategorize edilmiştir. Burada öne çıkan temaların puanlayıcılar arası güvenirlik, yanlılık, katılık, cömertlik ve hale etkisi olduğu görülmektedir. Daha spesifik bir bakış açısı ile puanlayıcılar arası güvenirlik, puanlayıcıların aynı performans hakkında benzer yargılara varması; yanlılık, puanlayıcıların belirli niteliklere göre performansa katı veya cömert puan verme eğilimi; katılık ve cömertlik, puanlayıcının genel anlamda az veya fazla puanları tercih etmesi; hale etkisi ise performans ile ilgili bir niteliğin etkisinin puanlamada diğer değişkenlere yansımasıdır. İngilizce dil becerilerinin ölçülmesinde, özellikle yazma ve konuşma gibi üretici dil becerilerinde puanlama davranışları yaygın bir şekilde araştırılmıştır (Farrokhi & Esfandiari, 2011; Kennedy vd., 2015; Kim, 2015; McNamara & Adam, 1991; O'Grady, 2023; Phung, 2013; Polat, 2020). Bu çalışmalarda

puanlayıcıların farklı bağlamlarda katılık, cömertlik, yanlılık ve hale etkisi gibi farklılaşan davranışları incelenmiştir. Öte yandan davranışlardan kaynaklanan tutarsızlıkları azaltmak amacıyla sunulan önerilerden biri de genel anlamda puanlayıcı eğitimi olmuştur.

Puanlayıcı eğitimi, puanlama davranışından kaynaklanan uyuşmazlıkları minimize etmek ve puanlayıcılar arasında güvenirliliği sağlamak için uygulanabilmektedir (Razı, 2015; Weigle, 1994). Bu kapsamda puanlayıcılara, puanlama eğitimi verilmesi ile ölçme ve değerlendirme sonuçlarının daha tutarlı ve amaca uygun hale getirileceği düşünülmektedir. Mevcut çalışmada, bir performans değerlendirme yaklaşımı olarak öğrencinin aktif katılımını teşvik eden akran değerlendirmenin, İngilizce akademik yazma becerilerinde puanlayıcı eğitimi ile uygulanması planlanmakta ve uygulanan eğitimin akran puanlayıcı davranışlarını nasıl etkilediği incelenmek istenmektedir. Böylece hem akran değerlendirmenin İngilizce yazma becerilerini ölçme ve değerlendirmede kullanılmasının etkililiği hakkında hem de puanlayıcı eğitiminin akran davranışlarını ne yönde etkilediği ile ilgili yorumlamalar yapılmaktadır.

Araştırmanın Amacı ve Soruları

Bu araştırmanın amacı, akran değerlendirmesi kapsamında puanlayıcı eğitimi uygulamak ve eğitim öncesinde, sonrasında ve araştırma kapsamında İngilizce akademik yazma becerilerini puanlamada akran davranışlarını incelemektir. Bu bağlamda araştırma soruları aşağıdaki gibidir:

1. Puanlayıcı eğitimi öncesinde, akademik yazma becerilerinin puanlanması sürecinde;
2. Puanlayıcı eğitimi sonrasında, akademik yazma becerilerinin puanlanması sürecinde;
3. Puanlayıcı eğitimi kapsamında yapılan uygulama dikkate alındığında;

    a. İngilizce öğretmenliği öğrencilerinden oluşan puanlayıcı grubunun puanlamalarında öğrenci, ölçüt, zaman ve puanlayıcı yüzeylerinden elde edilen verilerin uygunluk değerleri ve model ölçümleri nasıldır?

    b. İngilizce öğretmenliği öğrencilerinden oluşan puanlayıcı grubunun puanlama davranışları nasıldır (katılık, cömertlik, yanlılık)?

c. İngilizce öğretmenliği öğrencilerinden oluşan puanlayıcı grubunun puanlama davranışlarında ön ve son uygulama arasında herhangi bir farklılık var mıdır?

Araştırmanın Önemi

İngilizce akademik yazma becerisi, üretime dayalı bir dil becerisidir (productive skill). Öğrencilerin bu beceriye yönelik ölçme ve değerlendirme sürecinde belirlenen akademik bir türde performans sergilemesi ve bu performansın puanlayıcılar tarafından analitik veya bütüncül dereceli puanlama anahtarı kullanarak puanlanması gerekmektedir (Brown, 2004). Performansın alternatif değerlendirmesinde, öz-değerlendirme ve akran değerlendirme gibi öğrenci katılımını vurgulayan yaklaşımlar yaygın tercih edilmeye başlanmıştır. Akran değerlendirmede farklı puanlayıcılar aynı performansı belirli kriterlere göre puanlandırmaktadır (Topping, 2009). Farklı puanlayıcıların ölçme sürecinde birlikte yer alması, performans değerlendirme bağlamında puanlayıcı güvenirliği konusunu gündeme getirmektedir (Alderson vd., 1995; Zhang vd., 2020). Puanlayıcı güvenirliğine ek olarak yine puanlayıcılardan kaynaklı temel davranışların (yanlılık, katılık/cömertlik ve merkezi yönelim vb.) puanlama davranışını etkilediği alan yazında özellikle ön plana çıkmaktadır. Puanlayıcı davranışlarının ölçme güvenirliğine tehdit oluşturmasının önüne geçme yöntemlerinden biri de puanlayıcı eğitimidir (Weigle, 1994). Bu araştırmada üniversitede İngilizce akademik yazma becerileri dersi alan İngilizce öğretmenliği öğrencilerine puanlama eğitimi verilerek puanlama davranışlarından kaynaklanan ölçme hatalarını minimize etmek istenmektedir. Bu sayede akran değerlendirmenin daha güvenilir ve uygulanabilir hale getirilebileceğine inanılmaktadır. Aynı zamanda akran değerlendirme, öğretmenlik öğrencilerinin puanlama deneyimi kazanmasına olanak sağlamakta ve öğrenciler ölçme süreçlerine aktif katılım göstermektedir. Bu sayede öğrenciler, mezun olmadan önce de puanlama deneyimi kazanmış olmaktadır.

Araştırmanın Varsayımları

Bu araştırma için 51 İngilizce öğretmenliği öğrencisinden "reaksiyon" türünde yazılmış 5 adet kompozisyonu (reaction essay) puanlandırması istenmiştir. İki temel akran değerlendirme türünden, belirli bir grup öğrenci ürününü diğer tüm akranların puanlandırması yöntemi seçilmiştir (Kutlu vd., 2014). Bu kapsamda puanlayıcıların,

üniversite eğitiminin birinci sınıfında verilen "Yazma Becerileri I ve II" derslerini aldığı, deneme veya kompozisyon türünde pratikler yaptığı ve kompozisyon yazma ve geri bildirim verme süreci ile ilgili yeterli bilgiye sahibi olduğu varsayılmaktadır. 51 puanlayıcının kendilerine verilen 5 adet kompozisyonu gerçek düşüncesini yansıtarak puanladığı düşünülmektedir. Araştırma da "Akademik Yazma Becerisine Yönelik Analitik Puanlama Anahtarı" kullanılmıştır. Bu puanlama anahtarının reaksiyon kompozisyon türü ve içeriğine uygun olduğu düşünülmektedir.

Araştırmanın Sınırlıkları

Bu araştırmanın en temel sınırlılığı puanlayıcı eğitimine ayrılan süredir. Puanlayıcı eğitimi pratik nedenlerden dolayı toplam altı haftada gerçekleşmiştir. Diğer bir sınırlılık ise örneklem grubunun İngilizce öğretmen adayları olması nedeni ile akademik yazma deneyimlerinin oldukça iyi düzeyde olması ve puanlayıcı eğitiminde sunulan çoğu içeriği akademik yazma derslerinde de pekiştirmiş olmalarıdır. Bu durum eğitim neticesinde ortaya çıkan bazı kazanımları sınırlandırmıştır.

Tanımlamalar

Akran Değerlendirme: Performansa yönelik bir davranışın sergilendiği bağlamda, davranışı sergileyen katılımcıların bir diğer katılımcıyı puanlamasıdır (Spiller, 2012).

Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM ya da ÇYRM): Madde tepki kuramına bağlı bir model olarak ölçüt, puanlayıcı ve aday gibi farklı boyutların tek bir "digit" ölçeğinde sergilendiği modeldir (Linacre, 2017).

Katılık/Cömertlik: Birey yeteneklerin madde güçlülüğü ile etkileşimi katılık ve cömertlik olarak betimlenir (Linacre, 2002).

Madde Karakteristik Eğrisi: Bir maddeye ait doğru cevap verme olasılığı ile yetenek düzeyi arasındaki ilişki, madde karakteristik eğrisi olarak tanımlanır (Baker, 2001).

Performans Değerlendirme: Gerçek hayata yönelik görevler doğrultusunda ortaya konan eylemin belirli ölçütlere göre puanlayıcılar tarafından değerlendirilmesidir (Brown vd., 2002).

Puanlayıcı Hatası: Ölçme puanlarına, ölçülen özellik dışında yansıyan puanlayıcı davranışıdır (İlhan vd., 2024).

Yüzey: ÇYRÖM'de temsil edilen değişkenlik kaynağıdır (İlhan vd., 2024). Örneğin, ölçütler, zaman veya puanlayıcı birer yüzey olarak analize dahil edilebilir.

BÖLÜM II

ALAN YAZIN TARAMASI

Bu bölümde sırasıyla akademik yazma becerileri, akran değerlendirme ve puanlayıcı davranışları ilgili alan yazın çalışmaları incelenmiştir. Öncelikle hedef konular ile ilgili teorik düzeyde kavramsal içerik sunulmuştur. Ardırdan bu konular ile ilgili yapılan çalışmaların amaçları, bağlamı ve bulguları tartışılmıştır.

İngilizce Akademik Yazma Becerileri

Akademik yazma becerileri, bireylerin herhangi bir konu ile ilgili bilgi edinmelerini, edindikleri bilgileri yorumlayıp ve sentezlemelerini ve bu çıkarımlar ile söz konusu yazım biçimine göre üretim yapmalarını gerektiren bir beceri türüdür. Günümüz bilgi, eğitim ve yaygınlaştırma çağında başlıca iletişim dilinin İngilizce olması nedeni ile yabancı dil olarak İngilizce öğretiminde akademik yazma becerilerinin öğretimi önemli bir yer tutar. Bireyler veya yaygın kullanıcılar olan öğrenciler, akademik yazma pratiklerini genellikle üniversite dersleri veya akademik çalışma bağlamında gerçekleştirir (Bailey, 2003). Bu pratikler neticesinde kendi kazanımlarını ifade eden ve belirli bir amaca hizmet eden ürünler ortaya koyarlar.

Akademik yazma sonucunda her ne kadar yazımsal bir ürün ortaya çıksa da akademik yazma süreci, ortaya çıkan ürünün niteliğini şekillendiren temel unsurlardan biridir. Şekil 1'de de görüldüğü üzere, Oshima ve Hogue (2007) akademik yazmanın dört temel aşamadan oluştuğuna dikkat çekmektedir. Bu kapsamda, akademik yazı belirlenen konu ile ilgili ön yazma uygulamaları gerektirir. Bu aşama, öğrencilerin veya bireylerin konuyu belirledikleri, konunun kapsamını daralttığı, daraltılan konu ile ilgili beyin fırtınasının yapıldığı yerdir. Bu

aşama neticesinde ortaya anlamlı veya anlamsız düşünce birikimleri çıkacaktır. Yazarın bir sonraki aşamada yapması gereken ise konu ile anlamlı bütün oluşturan düşünceleri seçerek bunları belirli temalara göre gruplandırmaktır. Gruplanma sonucunda, hedeflenen türün organizasyon yapısı doğrultusunda fikirlerin düzenlenmesi gerekecektir. Organizasyonu takip eden bir diğer adım asıl yazma pratiğinin yapıldığı yerdir. Bu aşamada yazar, kendinden beklenen ölçütler doğrultusunda organize ettiği fikirlerden anlamlı bir eser oluşturur. Ancak, oluşan bu eser birinci taslak görevi görür. Yapılandırılan taslak öğretmenler, akranlar veya öğrencinin kendi tarafından ön değerlendirilmeye alınır. Bu ön değerlendirme kapsamında eser veya ürün ile ilgili geri bildirimler, bir başka deyiş ile, dönüt yazara iletilir.



Şekil 1. Akademik yazma süreci

Son aşamada yazardan beklenilen, geri bildirim doğrultusunda düzenlemeler yapmaktır. Görüldüğü üzere akademik yazma, çıkan ürünün yanı sıra süregelen bir çalışmanın, sistematik bir sürecin sonucudur. Bu bağlamda akademik yazma becerilerinin ölçülmesi ve değerlendirilmesi, yazma becerisinin doğası gereği hem sürece yönelik hem de ürüne yöneliktir. Önceden de belirtildiği üzere üretici bir beceri olan yazma eyleminin ölçülmesi dolaylı araçlar olan dereceli puanlama anahtarları ile nicelleştirilir. Ancak, akademik yazma süreçlerini görünür kılmak ve geri bildirim ile ilgili düzenlemeler sunmak için ise öğretmen veya akranların ölçme sürecine aktif katılımı gerekmektedir. Geri bildirim bir diğer ifade ile dönüt, öğretmen dönütü, akran dönütü ve öz dönüt olarak genel anlamda üç temel kategoride

ele alınır. Yazma becerilerinin gelişmesinde önemli bir aşama olan geri bildirim sunma, ortaya çıkan ürünün değerlendirilmesinde de önemli bir role sahiptir.

Akademik yazma süreçlerine ve geri bildirim sağlayan bir mekanizmaya ek olarak etkili bir akademik ürün için yazarların kendilerinden beklenilen yazım türüne hizmet eden çalışmalar ortaya koyması da akademik yazma sürecinde önemli bir paya sahiptir. Alan yazında akademik yazıların kaleme alındığı pek çok yazım türü vurgulanmaktadır (Fang, 2021; Bailey, 2003). Paragraf, deneme (kompozisyon), derleme, teknik veya bilimsel rapor, araştırma makaleleri, kitap eleştirilerileri, bloglar, epostalar, eğitim veya kurumsal politikalar ve konferans bildirleri bu akademik yazım türlerinin başlıcalarıdır. Akademi olarak bilinen üniversitelerde çalışmalar yapan öğrencilerin de yukarıda belirtilen türlerde yazılar kaleme aldığı yaygın olarak bilinmektedir. Bu türlerden deneme (kompozisyon), öğrencilerin sıklıkla başvurduğu akademik yazma türlerinden biridir (Hani & Perlina, 2022; West vd., 2019).

Hedef konu ile ilgili yazarın sahip olduğu argümana göre ise deneme türleri de çeşitlilik göstermektedir. Folse ve Pugh (2019) deneme türlerini dört temel çerçevede ele almaktadır. Bunlar:

1. Neden-Sonuç Denemesi (Cause and Effect)
2. Karşılaştırma Denemesi (Comparison)
3. Reaksiyon Denemesi (Reaction)
4. Argüman/Tartışma Denemesi (Argument)

Tüm bu deneme türleri göz önünde bulundurulduğunda öğrenci performanslarından beklenilen temel kriterler şu şekildedir (Brown, 2004; Brookhart, 2013; Folse & Pugh, 2019; Harrington vd., 2006; Saddler & Andrade; 2004): Tez cümlesinin deneme türüne uygun yazılması, tez cümlesinde yer alan kontrol ifadelerine göre gelişme bölümünün organize edilmesi, her bir gelişme bölümünün ilişkili konu cümlesi ile başlaması, gerekli kanıtlar ile argümanların desteklenmesi ve tüm tartışmayı özetleyen bir kapanış cümlesinin var olması, doğru dil bilgisi kullanımı ve yazım kuralları, kelime çeşitliliği ve yeterlik düzeyi, formal ifadeler ile anlaşılabilirliği ön plana çıkarılması, cümle geçişlerin bağlantılı olmasıdır. Belirtilen bu temel kriterler ve performans düzeyleri dikkate alınarak öğrenci çalışmalarına puanlayıcılar tarafından niceliksel bir değer atfedilir.

Diğer taraftan yapılan bazı çalışmalarda öğrencilerin akademik yazma sürecinde hangi alanlarda zorlandıklarına dikkat çekilmiştir. (Aldabbus & Almansouri, 2022; Mustafa vd., 2022; Yağız & Yiğiter, 2012). Weir'in(2016) ele aldığı bir eserde öğrencilerin akademik yazmada en çok zorlandıkları alanlar, özetle şöyle ifade edilmektedir:

- ❖ Dil bilgisi yapıları ve yazım kuralları
- ❖ Uygun ve çeşitli kelime seçimi
- ❖ Konu bilgisi ve organizasyon
- ❖ Açıklık ve düzen

Akademik yazma becerilerinde öğrenci performansını beklenilen düzeye çıkarabilmek için mevcut olan problemlerin tespit edilmesi ve bunlar ile ilgili gerekli eğitimlerin ve çalışmaların yapılması gerekmektedir. Bu durum da akademik yazmayı bir süreç olarak ele alınmasının önemini vurgulamaktadır. Dolayısıyla temel ölçütlerin sınıfta tartışılması, yazma sürecine dönütleri ile öğretmenlerin ve akranların da katılması ortaya çıkacak ürünün niteliğini artıracaktır.

Akran Değerlendirme

Akran, bir diğer adıyla, öğrencinin, kendisi ile aynı öğrenme yaşantılarını paylaşan diğer arkadaşları ile beraber ölçme sürecine katılarak bir çalışma veya performans hakkında yargıya varması akran değerlendirme olarak kabul edilmektedir. Alan yazında aynı ana temaya vurgu yapan farklı tanımlamalar söz konusudur:

- Akran değerlendirme, öğrenen katılımını gerektiren bir değerlendirmedir (Kutlu vd., 2014).
- Öğrencilerin akran ürünlerini nitelik, düzey ve değer açısından incelemesidir (Topping, 2009).
- Akran ürünleri hakkında öğrencilerin değer biçtiği etkinlikler bütünüdür (Reinholz, 2016).
- Öğrencilerin belli ölçütler çerçevesinde birbirilerini değerlendirmeleridir (Brown, 2005)

Bu tanımlamalardan yola çıkarak akran değerlendirme de öne çıkan unsurlar, öğrencilerin aktif katılımı ve ölçütler doğrultusunda diğer öğrencilerin yazma performansı hakkında

yargıya varmasıdır. Akran değerlendirmenin öğrencilere sunduğu pek çok avantaj vardır (Brown, 2013; Douglas; 2010; Kutlu vd., 2014; Tan, 2014: Topping, 2009):

- Öğrenciler, birbirlerine ürünleri hakkında detaylı geri bildirim verir.
- Öğrencilerin eleştirel düşünme becerilerini zenginleştirir.
- Öğrencilere sorumluluk hissi verir.
- Öğrencinin kendi ürününü daha iyi değerlendirmesine olanak tanır.
- Öğrencilerin ders hedeflerine yönelik farkındalığı artar.
- Öğrenci motivasyonunu ve özerkliğini teşvik eder.
- Yaşam boyu öğrenmeyi destekler.

Akran değerlendirme iki şekilde yapılabilmektedir. Bunlardan ilkinde her bir öğrenci bir diğer akranının çalışmasını değerlendirir. Diğerinde ise seçilen bazı çalışmalar sınıftaki tüm akranlar tarafından değerlendirilir. Diğer değerlendirme türlerinde olduğu gibi burada da ölçme aracı olarak "kontrol listeleri, açık uçlu sorular ve dereceli puanlama anahtarları" kullanılabilir (Kutlu vd., 2014, p. 104). Öğrencilerden performans göstermesi beklenilen ilgili yazma türüne ait özelliklere göre değerlendirme aracına karar verilebilir.

Her ne kadar aktif öğrenci katılımını teşvik etse de akran değerlendirme yaparken dikkat edilmesi gereken bazı önemli noktalar vardır. Ölçülen becerinin düzeyi arttıkça ve performans çıktıları daha komplike hale geldikçe, akran değerlendirmenin yapılması zorlaşabilir (Douglas, 2010). Öğrenciler, birbirlerini değerlendirirken yanlı davranma davranışları da gösterebilir (Stonewall, 2018). Bu durumda, ölçme ve değerlendirmenin güvenirliği ve geçerliğin tehdit edilebilir. Bu hata kaynaklarını minimize etmek için öğrencilere puanlayıcı eğitiminin verilmesi, değerlendirme bağlamında güven ortamının oluşturulması ve yapılan akran değerlendirme çalışmalarının sınıfta etkileşimli tartışılması tavsiye edilmektedir (Saka, 2023; Weigle, 1994). Böylece öğrenme ortamı rekabetten ziyade öğrenme odaklı olacağından ve puanlayıcılar olarak akranlara puanlama deneyimi ve bilgisi kazandırılacağından, ölçme ve değerlendirmenin de daha tutarlı ve geçerli sonuçlar sunacağı düşünülmektedir.

İngilizce yazma becerileri özelinde bakıldığında ise Shohamy ve diğerleri (1992) yazma becerisinin ölçülmesine dair eğitim alan puanlayıcıların almayanlara göre daha yüksek puanlayıcı güvenirliğine sahip olduğunu açıklamıştır. Bu durum, puanlayıcı eğitiminin dil

becerileri kapsamında uygulanmasının, akranlar gibi puanlayıcılar arasında güvenirliği artırmanın bir yolu olarak kullanılabileceğine işaret etmektedir.

Puanlayıcı Davranışları ve Puanlayıcı Eğitimi

Performansa dayalı değerlendirmelerde puanlama güvenirliliğini etkileyen bazı faktörler söz konusudur. Puanlamayı etkileyen bu faktörler, hatalı puanlamaya sebebiyet vererek performans hakkında doğru olmayan çıkarımlar yapmaya neden olur. Verilen puanlardan elde edilen çıkarımların, puanlanan bireylerin akademik ve sosyal alanlardaki hayati kararlarını önemli düzeyde etkilediği düşünüldüğünde (Eckes, 2017) bu faktörlerin tespit edilmesi ve bunları engellemek için önlem alınması da önem arz etmektedir (Johnson & Lim, 2009). Bu faktörlerden başlıcaları puanlayıcı kaynaklı unsurlardır. Puanlayıcı etkileri, hataları veya puanlayıcı davranışları olarak bilinen bu unsurlar (Anthony vd., 2022; Engelhard, 1994; Myford & Wolfe, 2004), puanlamayı ve dolayısıyla puanlama sonucunda yapılan çıkarımları da büyük ölçüde etkilemektedir.

Kavramsal bakış açısı ile bir performansın, puanlayıcılar tarafından beklenilen başarı düzeyinden farklı olarak puanlanması, puanlayıcı hatalarına neden olmaktadır (Wind, 2020). Eckes (2017) hatalı puanlayıcı davranışlarını temel olarak dört başlık altında sıralamıştır. Bunlar aşağıda belirtildiği gibi katılık/cömertlik, merkezi eğilim, hale etkisi ve yanlılıktır. Temel tanımlamalara aşağıda yer verilmektedir (444-445):

1. *Puanlayıcı katılığı:* Çok düşük puanlama yapma anlamına gelir. En yaygın hatalardan biridir.
2. *Merkezi eğilim:* Verilen puanların orta puan çevresinde toplanmasıdır.
3. *Hale etkisi*: Bir ölçüte ait performans etkisinin diğerlerine genellenmesidir.
4. *Puanlayıcı yanlılığı:* Farklı özellikleri çok katı veya çok cömert puanlama anlamına gelmektedir.

Performans değerlendirmeye yönelik puanlayıcı hataları performans değerlendirmeyi gerektiren farklı disiplinlerde de ele alınmıştır. Örneğin, Royal ve Hecker (2016) klinik performansa yönelik ölçme ve değerlendirme süreçlerini öne çıkartıkları çalışmalarında yaygın olarak karşılaşılan puanlayıcı hatalarını listeledikleri tablo aşağıdaki gibidir (s. 6):

Tablo 1

*Yaygın Puanlayıcı Hataları**

Asimilasyon Etkisi:	Aşırı farklı görünmemek için diğer puanlayıcılara yakın puan verme
Aktarma Etkisi:	Bir öğrencinin sahip olduğu performansın, puanlamasını olumlu veya olumsuz etkilemesi
Merkeze Eğilim:	Puanlama ölçeğinde çoğunlukla orta kategorileri tercih etme
Amigo Etkisi:	Öğrencileri desteklemek adına performansa ne olursa olsun yüksek puan verme
Uyuşmayan Standartlar:	Puanlayıcının kişisel ölçütleri ile ölçme aracının ölçütleri arasındaki uyuşmazlıktan dolayı herkese düşük puan verme
Uyuşmayan Değerler:	Performansta gözlenen değerler ile puanlayıcının sahip olduğu değerler arasında uyum olmamasından dolayı düşük puan verme
Çekimserlik:	Öğrencinin itiraz etme veya karşı gelme ihtimali nedeni ile yüksek puan verme
Kayma Etkisi:	Zaman zaman artan şekilde çok katı veya çok cömert puanlar verme
Aşırılık:	Ölçme aracının sadece iki uç/karşıt değerini kullanma
Yorgunluk:	Yorgunluk nedeni ile puanların etkilenmesi
Taraf Tutma:	Tanınan bir öğrencinin bazı hatalarını görmezden gelme
İlk İzlenim:	Bir öğrencin sonraki performansını göz ardı ederek bıraktığı ilk izlenime göre değerlendirme
Hale Etkisi:	Öğrencide şüphe uyandırarak yüksek puan puanlar verme
Gestalt Olgusu:	Performansın farklı yönleri ayırt edilemediğinden öğrenciye genel izlenimlerine dayalı puan verme
Bağlantı Etkisi:	Öğrencinin başka biri ile olan bağlantısının puanı etkilemesi
Güvensizlik:	Puanlayıcının daha düşük bir puanı açıklamada yetersiz kalmasından dolayı yüksek puan vermesi
Uzunluk:	Daha uzun süreli oluşturulan performansı daha yüksek puanlama
Cömertlik:	Sürekli olarak yüksek puan verme
Kişilik Çatışması:	Öğrencilere karşı sahip olunan olumsuz yargılardan dolayı düşük puan verme
Puanlayıcı Yeterliği:	Gözlemlenmeyen veya değerlendirilemeyen durumlarda yine de puan/derece verilmesi
Tekrar Etkisi:	Aynı performansın tekrarlanması nedeni ile düşük puan verme
Ölçme Aracını Yorumlama:	Ölçme aracının yanlış anlaşılması nedeni ile doğru olmayan puanlar verme

Puan-Ranj Sınırlaması:	Ölçülmesi hedeflenen gizli değişkenleri performansta ayırt edememe
Öz-Puanlama:	Performansı gerektiğinden fazla inceleme ve değerlendirme
Benzerlik Etkisi:	Puanlayıcıya yakın performans gösteren öğrencilere yüksek puan verme
Yüzeysel Puanlama:	Performansın çok az bir bölümünü inceleyerek puanlama yapma
Ani Kayıp:	Puanlayıcıda olumsuz bir tepki oluşturacak performansı daha düşük puanlama
Sempati Puanı:	Puanlayıcın duygularına yönelik veya ödüllendirme amacı ile ortaya konan bir performansı değerlendirme
Katılık:	Sürekli olarak daha düşük puanlar verme
Kişisel Özellik Etkisi:	Öğrencinin kişisel bir özelliğini (zamanlama, iletişim veya görünüm) dikkate alarak performansını puanlama

*Bu liste Royal ve Hecker (2016) "Rater errors in clinical performance assessments" çalışmasında yer alan terminolojiyi doğrudan çevirerek oluşturulmuştur. Çevirilerde Şata'nın (2019) "Performans değerlendirme sürecinde puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisinin incelenmesi" isimli doktora tezindeki tanımlamalarından da yararlanılmıştır.

Görüldüğü üzere puanlayıcı hataları, puanlayıcının kendisinden, ölçme aracından ve ölçme bağlamından kaynaklanmak üzere farklı şekillerde incelenebilir. İlgili araştırmalar göz önünde bulundurulduğunda (Anthony vd., 2022; Eckes, 2017; Myford & Wolfe, 2004; Styck vd., 2021; Wind, 2020) yapılan çalışmaların spesifik olarak odaklandığı temel hata kaynaklarının cömertlik, katılık, hale etkisi, merkezi eğilim ve yanlılık olduğu görülmektedir.

Puanlayıcıların dahil olduğu her ölçme ve değerlendirme sürecinde olduğu gibi akran değerlendirme de puanlamaya karışan puanlayıcı kaynaklı bazı hatalı davranışlar olabilir. Akranlar bir diğerinin performansına beklenenden fazla veya az puan verebilir (Esfandiari, 2015). Akran değerlendirmeyi konu edinen bir diğer çalışmada Rotsaert ve diğerleri (2017) akran değerlendirme kapsamında ölçme güvenirliliği etkileyebilecek altı önemli kişisel değişkene alan yazından alıntılar yaparak dikkat çekmektedir (s. 30):

1. *Arkadaşlık etkisi:* Arkadaşlık ilişkilerinden dolayı puanlamada yanlılık olmasıdır.
2. *Onaylanmama korkusu:* Olumsuz bir değerlendirmede akran tarafından olumsuz yargılanma korkusudur.
3. *Psikolojik güvenlik:* Farklı görüşleri tehdit değil fırsat olarak algılamak anlamına gelmektedir.

4. *Değer uyumu:* Değerlendirme amaçlarında ve ölçütlerinde aynı görüşte olunmasının uygunluğudur.
5. *Akranın kendine güvenmesi:* Değerlendirme yaparken akranın kendisine duyduğu güveni ifade etmektedir.
6. *Akrana duyulan güven:* Akranın yaptığı değerlendirmeye diğer arkadaşları tarafından duyulan güvendir.

Bunlara ek olarak puanlayıcının geçmiş eğitim yaşantıları ve puanlamaya getirdikleri bakış açıları da ölçme güvenirliğini ve sonrasında kararların geçerliliğini etkileyen önemli hususlardan bir diğeridir (He vd., 2013; Johnson & Lim, 2009). Anlaşıldığı üzere performans temelli ölçme ve değerlendirme sürecine puanlayıcı ile ilgili farklı hata kaynakları karışmaktadır. İlhan vd. (2024) puanlayıcı davranışları hata kaynaklarını azaltmak amacıyla aşağıdaki yöntemleri önermiştir:

1. Birden fazla puanlayıcının puanlamaya dahil edilmesi
2. Puanlayıcı eğitimleri
3. Puanlama ölçeklerinin kullanılması

Bu yöntemlerden biri olan puanlayıcı eğitimi (rater training), ölçme ve değerlendirme sonuçlarını daha tutarlı ve geçerli kılabilmek için önerilen yöntemlerden biridir (Barrett, 2001; Plake vd., 1991; Schaefer, 2008; Weigle, 1998). Wang (2010) puanlayıcı eğitimin özellikle yanlılık bağlamında daha istikrarlı sonuçlara zemin hazırladığını bildirmektedir.

Woehr ve Huffcutt (1994) puanlayıcı eğitimlerini dört temel kategoride sunmuştur: a) puanlayıcı hatası eğitimi, b) performans boyutları eğitimi, c) referans çerçevesi eğitimi ve d) davranış gözlem eğitimidir. İlhan vd. (2024) ise bunlara ek olarak "Puanlayıcı Değişkenliği Eğitimi"ni de öne sürerel puanlayıcı eğitimlerini beş ayrı bölümde incelemiştir. Bu eğitimlerin kısa tanımlamaları aşağıda verilmiştir (Woehr & Huffcutt, 1994; İlhan vd., 2024).

1. *Puanlayıcı Hataları Eğitimi:* Bu eğitimin ana amacı puanlayıcıları temel hata türleri (katılık/cömertlik, merkeze yönelme, yanlılık vb.) hakkında bilgilendirmek ve bu hata türlerinden nasıl uzak kalınabileceğine dair yönlendirmelerde bulunmaktır.
2. *Performans Boyutları Eğitimi:* Burada amaç puanlama hatalarından ziyade puanlanacak performans ve alt özellikleri hakkında puanlayıcıya bilgi vermektir.

Böylece puanlayıcılar ölçülecek performansı ve alt boyutlarını tanıyarak puanlama davranışlarında daha ayırt edici olabilecektir.

3. *Referans Çerçevesi Eğitimi:* Bu eğitimde amaç puanlayıcıların aktif olarak deneyim kazanması ile ilgilidir. Puanlayıcı hata eğitiminin aksine, burada doğru puanlamanın nasıl yapılacağı uygulamalı olarak ele alınır.
4. *Davranış Gözlem Eğitimi:* Bu eğitim, performansı puanlayabilmek için öncelikle gözlem becerilerinin yeterli düzeyde gelişmiş olması gerektiğine vurgu yapar. Bu dayanaktan yola çıkarak puanlayıcıya temel gözlem yeterliklerinin kazandırılması amaçlanır.
5. *Puanlayıcı Değişkenliği Eğitimi:* Puanlayıcı hataları eğitimi ve referans çerçeve eğitimi modellerinin zayıf ve güçlü yanlarının göz önüne alınması ile öne sürülen bu modelde puanlayıcıların farklı yeterlik düzeyinde yer alan performanslar için nasıl bir puanlama dağılımına sahip olması gerektiği vurgulanmaktadır.

Görüldüğü üzere akran değerlendirmeyi de ele alan performans temelli değerlendirmelerde ölçme sonuçlarına puanlayıcı kaynaklı hatalar yansıyabilmektedir. Bu hata kaynaklarına en aza indirgemenin yollarından biri de puanlayıcılara eğitim vermektedir. Eğitimde odaklanılan özellik ve eğitimden beklenilen çıktı dikkate alındığında verilecek eğitimin türü de çeşitlilik gösterebilir.

Çok Yüzeyli Rasch Ölçme Modeli

Ölçme süreçlerine puanlayıcı davranışlarından ve akran değerlendirme dinamiğini etkileyen farklı değişkenler sebebi ile farklı hata kaynakları karışmaktadır. Ölçülen yapıyla ilgisi olmayan hata kaynaklarını, yapı ile ilgili olmayan varyans (contruct-irrelavance) olarak tanımlayan Messick (1996) bu kaynakların geçerliliği tehdit ettiğini ileri sürmüştür. Diğer bir ifade ile, belirtilen puanlayıcı hataları ölçme yapısı (measurement construct) ile ilgili olmadığı için değerlendirme sürecini olumsuz yönde etkileyebilmektedir (Eckes, 2019; Anthony vd., 2022). Bu nedenden dolayı, hata kaynaklarının tespit edilmesi ve ölçmenin bu hata kaynaklarından bağımsız olarak gerçekleştirilmesi gerekmektedir.

Test kuramları dikkate alındığında, klasik tesk kuramında (KTK), genellenebilirlik kuramında (G-Kuramı) ve madde tepki kuramında (MTK) hata kaynaklarını incelemek için

farklı yöntemler geliştirilmiştir. KTK, hata kaynaklarını tek bir bileşen olarak analiz ederken, G-Kuramı hata kaynaklarının çeşitliliğini açıklamaktadır (Akiyama, 2001; İlhan vd., 2024). Daha net bir biçimde ifade etmek gerekirse, KTK ölçme sonucunu etkileyen madde, birey ve bağlam kaynaklı hata kaynaklarını çeşitlilik düzeyinde açıklayamazken (Ntumi & Bulala, 2025), G-kuramı farklı hata kaynaklarını ve bu hata kaynakları arasındaki etkileşimi de gözler önüne serer (Atılgan, 2005). Anlaşıldığı üzere G-kuramı hata kaynalarının detaylı analizini sunar. MTK'ya bakıldığında ise hata kaynaklarının madde ve birey düzeyinde incelenme olanağının olduğu gözlemlenir.

Örtük özellikler kuramı veya *madde karakteristik eğrisi* kuramı olarak da bilenen MTK'nın (Weiss & Yoes, 1991), KTK ve G-kuramına göre farklı avantajları söz konusudur. G-kuramı ölçmeye ait hata kaynaklarını detaylı biçimde incelese de bu hataların madde veya birey düzeyinde incelenmesi yönünden sınırlıdır (Linacre, 1993; Öç, 2024). MTK'da ölçme hatası her bir birey için ayrı bir şekilde kestirilebilmektedir (Çelen, 2008). Bu açıdan MTK, bireylerin madde düzeyinde gösterdiği performanslar ile ilgilenir (Baker, 2001). KTK'da madde parametreleri, ölçme yapılan gruba bağlı iken MTK'da madde parametreleri gruptan/bireylerden bağımsız olarak kestirilir; bu da MTK'nın değişmez madde parametreleri sağladığının kanıtı olarak değerlendirilebilir (Weiss & Yoes, 1991). MTK'nin en ayırıcı özelliklerinden biri ise "değişmezlik ilkesi"dir; madde parametreleri bireylerin örtük özelliklerinden ve birey parametreleri ise maddelerden bağımsızdır (Bortolotti vd., 2013). Bu çerçevede MTK'nın temel amacı, değişmez madde istatistikleri ve yetenek kestirimleri vermektir (Hamlbeton & Swaminathan, 2013).

Hambleton ve diğerleri (1991) MTK'nın iki temel prensipini şöyle açıklamaktadır: (1) bireye ait madde performansı, örtük özellikler seti tarafından kestirilir, (2) madde performansı ve örtük özellikleri seti arasındaki ilişki madde karakteristik eğrisi tarafından tanımlanır. Madde karakteristik eğrisi, MTK modellerine dayanak oluşturmaktadır. İlhan ve diğerleri (2024) MTK modellerinin iki kategorili puanlanan maddeler ve çoklu puanlanan maddelere göre çeşitlilik gösterdiğini ifade etmiştir. İki kategorili (doğru ve yanlış veya çoktan seçmeli vb.) puanlanan testler için geliştirilen modeller aşağıdaki gibidir (İlhan vd., 2024):

1. Bir parametreli lojistik model (Madde güçlük indeksi, madde düzeyinde değişkenlik gösterir; madde ayırt ediciliğin 1'e sabitlenmesi ile Rasch modeline karşılık gelir.)
2. İki parametreli lojistik model (Ayırt edicilik ve güçlük parametreleri)

3. Üç parametreli lojistik model (Şans, ayırt edicilik ve güçlük parametreleri)
4. Dört parametreli lojistik model (Dikkatsizlik, şans, ayırt edicilik ve güçlük parametreleri)

Çok kategorili MTK modellerine ise "Sıralama Ölçekli Model", "Kısmi Puanlama Modeli", "Geliştirilmiş Kısmi Puanlama Modeli" ve "Aşamalı Tepki Modeli" örnek olarak gösterilebilir (Büyükkıdık & Atar, 2018; İlhan vd., 2024). Bu modellerden bir bölümü Rasch modelleri başlığı altında toplanmaktadır. Bu modeller aşağıdaki gibidir (Eckes, 2019; Knoch & McNamara, 2015):

1. *Basit Rasch Modeli (The Simple Rasch Model):* Georg Rasch tarafından 1960 senesinde öne sürülmüştür. Yanlış ve doğru gibi iki kategorili puanlamalarda/değerlendirmelerde kullanılmaktadır.
2. *Dereceleme Ölçek Modeli (The Rating Scale Model):* İki kategorili değerlendirmeler dışında sıralama/dereceleme ölçeklerinden elde edilen puanlarda kullanılmaktadır.
3. *Kısmi Kredi Modeli (The Partial Credit Model):* Farklı aşama/adım gerektiren maddeler için tasarlanmıştır.
4. *Çok Yüzeyli Rasch Modeli (ÇYRÖM) (The Many-Facet Rasch Model):* Analize ölçme sonuçlarını etkileyebilecek farklı yüzeylerin de eklenmesine olanak sağlar.

Rasch modelleri ölçümün yapıldığı örneklemden ve sınav maddelerinden bağımsız sonuçlar verir (Yu, 2020). Rasch modelinin bir uzantısı olarak geliştirilen ÇYRÖM'de ise yukarıda da ifade edildiği gibi analize yetenek düzeyi ve madde güçlüğü dışında daha farklı değişkenler yani yüzeyler (facets) eklenir (Aryadoust vd., 2021; Lynch & McNamara, 1998). Diğer bir ifade ile ÇYRÖM, tek parametreli Rasch modelinin gelişmiş bir sürümüdür ve geliştirilen bu model birey yetenekleri hakkında puanlayıcı, görev ve diğer etkenlere karşı istatiksel olarak değişkenlik göstermeyen bağımsız ölçümler sunar (Park, 2024, s. 2). Güler ve Gelbal (2010), ÇYRÖM'ün birey yetenek düzeyi, madde güçlü ve puanlayıcı katılık/cömertliğini kestirmek için istatiksel bir teknik olduğunu ifade etmektedir.

Bu ifadede de belirtildiği üzere üç yüzeye sahip olan bir rasch modeli için aşağıdaki formül sunulabilir (Linacre, 2023, s. 17).

$$\log\left(\frac{\mathrm{P_{nijk}}}{\mathrm{P_{nij(k-1)}}}\right) = B_n - D_{gi} - C_j - F_{gk}$$

Formülde;

$\mathrm{P_{nijk}}$: j numaralı puanlayıcının n numaralı bireyin i numaralı maddesine k puan verme olasılığı

$\mathrm{P_{nij(k-1)}}$: j numaralı puanlayıcının n numaralı bireyin i numaralı maddesine k-1 puan verme olasılığı

B_n : n numaralı bireyin yetenek düzeyi

D_{gi} : g isimli ölçeğin i maddesine ait güçlük düzeyi

C_j : j numaralı puanlayıcının cömertliği veya katılığı

F_{gk} :g numaralı ölçekte k-1 puana göre k puanı alma güçlüğü

Formülde yer verilen yüzeyler ÇYRÖM ile birbirinden bağımsız bir şekilde logit cetveli olarak adlandırılan doğrusal bir ölçek üzerine yerleştirilerek kestirilir (Öç, 2024). Yabancı dilde ölçme ve değerlendirme alanına bakıldığında da ÇYRÖM'e sıklıkla başvurulduğu gözlenmektedir. Knoch ve McNamara (2015), özellikle yazma ve konuşma becerilerinin ölçülmesinde puanlayıcı davranışları hakkında güçlü kestirimler sunması ile Rasch modellerinin dil ölçme de avantajlı bir yöntem olduğunu ileri sürmüştür. Park (2004), ÇYRÖM'un, yazma sınavlarının yapı geçerliliği hakkında kanıt sunmasını açıklamaktadır. Performans değerlendirme kapsamında değerlendirildiğinde ise, ölçme ve değerlendirme güvenirlik ve geçerliliğini tehdit eden puanlayıcı hatalarının incelenmesinde ÇYRÖM'un etkili olabilieceği Engelhard (1994) tarafından ifade edilmektedir. Anlaşıldığı üzere, KTK ve G-Kuramına göre hata kaynaklarını belirlenen değişkenlerden bağımsız analiz etmede avantaja sahip olan MTK'ya dayalı olarak geliştirilen ÇYRÖM'un yabancı dilde performans gerektiren ölçme ve değerlendirmelerde puanlayıcı kaynaklı davranışları analiz etmek açısından oldukça elverişli olduğu anlaşılmaktadır.

İlgili Araştırmalar

Yazma performansının değerlendirilmesinde akran puanlayıcılarını konu alan çalışmalarda öğrenci performansı, farklı analiz yöntemleri ile incelemiştir. Bu çalışmalarda yaygın olarak kullanılan yöntemlerden biri de Rasch Analizidir. Madde tepki kuramına dayanan bu modelde performans ölçüt, birey ve puan gibi farklı yüzeylerde örneklemden bağımsız olacak şekilde değerlendirilebilmektedir. Farklı yüzeylerin ve etkileşime giren değişkenlerin de analizi dahil edilmesi ile ÇYRÖM/ÇYRM, özellikle üretime dönük yabancı dil becerilerinin ölçülmesinde hem ulusal hem de uluslararası araştırmalarda önemli bir yer tutmaktadır.

Kuramsal çalışmalara bakıldığında, Sudweeks ve diğerlerinin (2004) akademik yazmanın ölçülmesi kapsamında G-Kuramı ve ÇYRM'u karşılaştırdığı görülmektedir. Çalışma, G-Kuramının performansı analiz etmede daha bütüncül bir bakış açısı sunarken, ÇYRM'in yüzey bazında daha detaylı ve spesifik bilgiye ulaşma olanağı sağladığını ifade etmektedir. Yabancı dil öğretim bağlamında çalışma yapan Polat ve diğerleri (2022) yazma becerisinin ölçülmesinde Klasik Test Kuramı (KTK) ve Çok Yüzeyli Rasch Modeli (ÇYRM) karşılaştırmasına odaklanmıştır. Çalışmanın bulguları, ÇYRM'nin, KTK'ya göre ölçüt geçerliliğinin daha yüksek olduğunu göstermektedir. Myford ve Wolfe (2004) çalışmalarında ÇYRM'nin nasıl kullanıldığını tanıtmış ve çalışmanın ana amacı olarak puanlayıcı davranışlarını katılık/cömertlik, merkezi eğilim, tesadüfilik, hale etkisi ve farklılaşan katılık/cömertlik olmak üzere beş kategoride değerlendirmiştir. Eckes (2017) Rasch modellerinin puanlama kalitesi, puanlama eğitimi ve izlenmesi ile ilgili bilgi sağlayabileceğini iletmiştir; ancak, puanlayıcı eğitiminin, puanlayıcı değişkenini azaltmada beklenilden daha az etkili olduğunu ileri sürmüştür (Eckes, 2008).

Yazma performansın değerlendirilmesine yönelik yapılan diğer çalışmalarda ise geri bildirim/dönüt türleri incelenmiş ve bu türlerin yazma performansı ile ilişkisi ortaya koyulmaya çalışılmıştır. Akran dönütü ile ilgili yapılan çalışmalara bakıldığında, akran dönütünün performansa olan etkisi, diğer türler ile kıyaslandığında göreceli etkililiği, puanlayıcı davranışlarının farklı analiz yöntemleri ile incelenmesi öne çıkan araştırma konularıdır. Yu ve Lee (2016) akran geri bildirimi üzerine oluşturdukları derleme baş yapıt (state of art) makalesinde ilgili araştırmaları toplamda yedi ana konu altında tartışmıştır: (1) akran geri bildiriminin, öz ve öğretmen geri bidildiriminine göre etkiliği, (2) akran geri

bildiriminin, bireylere sağladığı katkılar, (3) bilgisayar destekli akran geri bildirimi, (4) akran geri bildirimi eğitimi, (5) öğrenci duruşları ve motivasyonları, (6) akran etkileşimi ve grup dinamikleri ve (7) akran geri bildiriminde kültürel unsurlar (s. 461). Bu çalışma akran geri bildirimi konusunda araştırmacılar için genel bir çerçeve özelliği sunmaktadır. Mumpuni ve diğerleri (2022) Endenezyo'da akran değerlendirme üzerine ÇYRM kullanarak bir çalışma yapmış ve analiz sonuçlarına göre akranların nesnel biçimde değerlendirme yapılabileceğine işaret etmiştir. Ravand ve diğerleri (2016) zaman içerisinde yabancı dilde yazma becerilerinin ölçülmesini akran, öz ve öğretmen değerlendirme kapsamında ÇYRM ile incellemiş ve zaman içerisinde, eğitim ve deneyim ile, puanlayıcı yanlılıklarında azalma meydana geldiğini ifade etmiştir.

Göreceli etkililik ve performansa olan katkısı bağlamında, Nelson ve Schun'un (2009) geri bildirim modeli üzerine yaptıkları çalışmada etkili geri bildirim ile ilgili öneriler de bulunmuş, performans hakkında açıklamalar yapma yönünden öğretmen ve akran dönürlerindeki farklılaşmaya da ışık tutmuştur. Bir diğer çalışmada ise Huisman ve diğerleri (2018) akran geri bildiriminin, performans değerlendirme ile ilgili süreçlerde bireylerin aktif rol almasını gerektirmesi nedeni ile yazma davranışlarına katkı sağladığı ifade etmiştir. Benzer şekilde Carrió-Pastor (2016, s. 63) akran değerlendirmenin öğrenci katılımını teşvik etmesi, bu süreçte öğrenciye yansıtım becerileri kazandırması ve diğer öğrencilere de geri bildirim vermesi yönünden bazı yararlara sahip olduğunu sergilemektedir. Sun ve diğerleri (2023) İngilizce yazma becerilerinin değerlendirmesinde akran geri bildirimini incelemiş ve akran geri bildirimin, öğretmen geri bildirimine göre sınırlı kalmasına rağmen yazma becerilerinin gelişmesinde önemli rol oynadığını ifade etmiştir. Birjandi ve Siyyari (2010) yazma becerilerinin ölçülmesinde öz ve akran değerlendirmeyi incelemiş ve iki değerlendirme yönteminin de yazma becerilerin gelismesi ve puanlama doğruluğu açısından verimli olduğunu; hatta, akran değerlendirmenin yazma performansı açısından daha etkili sonuçlar verdiğini tartışmışlardır.

Aryadoust ve diğerleri (2021) dil ölçme ve değerlendirme alanında Rasch temelli çalışmalar ile ilgili 215 yayını kapsayan derleme araştırması yapmıştır. Araştırma bulguları çalışmaların 63 tanesinin (% 29,3) yazma becerileri alanında yapıldığını göstermiştir. Rentiana ve diğerleri (2024) İngilizce akademik yazma performansını Rasch modeli kullanarak değerlendirmiş ve "İçerik", "Yapı", "Yazım ve İmla" ve "Söyleyiş" ölçütlerinin, hedef yapının geçerli bir şekilde ölçülmesine önemli katkı sağladığını tespit etmiştir. Engelhard

(1994) yazma becerileri kapsamında kompozisyonların ölçülmesinde ÇYRM ile puanlayıcı etkilerini incelemiştir. Puanlayıcı katılığında ve cömertliğinde anlamlı farklılıklar, hale etkisi ve merkezi eğilim gibi etkilerin mevcut olduğu anlaşılmıştır.

Goodwin (2016) Rasch temelli yaptıkları çalışmada, bulguları yorumlarken yazma performansının puanlanmasını etkileyebilecek puanlayıcı kaynaklı faktörlerin ve diğer farklı faktörlerin de araştırılmasının önemine vurgu yapmıştır. Eckes (2008), yazma ile ilgili performans değerlendirmede puanlayıcı değişkenini ÇYRM ile ele almış ve analiz sonucunda altı farklı türde puanlama yapan puanlayıcıya dikkat çekmiştir. Yazma performansı ölçütleri ile bu puanlayıcı türleri arasındaki ilişkileri incelemiştir. Eckes (2005) yaptığı bir diğer çalışmada üretici beceriler olarak yazma ve konuşma performans değerlendirmede puanlayıcı etkilerini araştırmış; puanlayıcıların katılık ve cömertlikte güçlü farklılaşma eğiliminden bahsetmiş ve puanlamaların cinsiyete göre yanlılık göstermediğini ifade etmiştir.

Engelhard (1994) yazma becerileri kapsamında kompozisyonların ölçülmesinde ÇYRM ile puanlayıcı etkilerini incelemiştir. Puanlayıcı katılığında ve cömertliğinde anlamlı farklılıklar, hale etkisi ve merkezi eğilim gibi etkilerin mevcut olduğu anlaşılmıştır. Kim (2020) yabancı dilde yazma performansını ölçmede hale etkisini ÇYRM ile incelemiştir. Bu çalışmada 11 puanlayıcı toplam 30 deneme/kompozisyonu "standart", "ters" ve "tesadüfi" olarak üç farklı türde sıralanan kriter listesine göre değerlendirmiştir. Araştırma sonuçları, tesadüfi kriter sırasının hale etkisini azaltmada yardımcı olduğunu açıklamıştır. Yazma becerilerinin ölçmede ÇYRM'i kullanan ilk çalışmalardan biri olan ve Gyagenda ve Engelhard (1998) tarafından yapılan araştırmada da puanlayıcılar arasında önemli değişiklerin olduğu ileri sürülmüş, kriterler/yazma boyutları ve puanlayıcılar arasında etkileşim olduğu açıklanmıştır.

Ayoobiyan ve Ahmadi (2023) yabancı dilde yazma performansının ölçülmesinde puanlayıcılara hale etkisi kapsamında eğitim vermiş ve bu eğitimin hale etkisini azaltmada önemli bir rol üstlendiğinin altını çizmiştir. Farrokhi ve Esfandiari (2011) de yabancı dilde yazılan kompozisyonların puanlanmasında hale etkisi tespit etmiştir. Araştırmalarında öz, akran ve öğretmen puanlamaya odaklanmışlar ve özellikle öz puanlamada daha fazla hale etkisinin ortaya çıktığını açıklamışlardır. Esfandiari (2015), yazma becerilerini ölçen bir diğer çalışmasında puanlayıcı olarak akranların performanslarını incelemiştir ve bireysel

akran puanlamalarının merkezi eğilim ve hale etkisi içerdiğini raporlamıştır. Farrokhi (2012) ve diğerleri benzer şekilde ÇYRM kullanarak öz, akran ve öğretmen puanlamalarını incelemiş ve puanlama davranışları arasındaki farklılığa işaret etmiştir.

Trace ve diğerleri (2017) yazma performansının değerlendirilmesi kapsamında puanlayıcı uzlaşması/görüşmesinin puanlayıcı davranışlarını nasıl etkilediğini araştırmıştır. Bu uzlaşmanın katılık/cömertlik davranışını değiştirmese de yanlılık etkilerinde azalma meydana getirdiğini açıklamıştır. Wiseman (2012) çalışmasında deneme/kompozisyon değerlendirmede etkili olan puanlayıcı davranışlarını incelemiş ve puanlayıcı özelliklerinin puanlama sürecinin etkileyebileceğini göstermiştir. Khamboonruang (2023) Taylant'ta üniversite düzeyinde yazma becerilerinin değerlendirilmesine odaklanmış, öz puanlama ve akran puanlama üzerine araştırma gerçekleştirmiştir. Öz puanlama daha çok cömert davranışlar içerirken, akran puanlamanın öğretmen puanlamasına daha yakın olduğunu bulgusu paylaşılmıştır.

Yan ve Chuang (2023) puanlayıcı davranışını ÇYRM ile modellemeye çalışması yapmış ve puanlayıcı davranışlarını ve gelişim sürecini yıllar içerisinde takip ederek anlamaya çalışmıştır. Araştırmacılara göre puanlama davranışının gelişimine dair bir anlayış geliştirmek puanlama eğitiminin etkililiğini anlayabilmek için dayanak oluşturmaktadır. Wu ve Tan (2016) yaptıkları çalışmada büyük ölçekli önemli kararları olan bir sınav bağlamında puanlayıcı davranışlarını incelemiş ve ÇYRM'in bu davranışları analiz etmede nasıl veri sağlayabileceğini göstermiştir. Benzer şekilde Schaefer (2008) de yazma becerilerinin değerlendirilmesinde yanlı puanlayıcı davranışlarını ÇYRM ile analiz etmiş ve bu analizin puanlayıcı etkilerine dair bilgi sunabileceğini ifade etmiştir. Yin ve Eckes (2024), ÇYRM'nin grup değerlendirmeler için geliştirilmiş versiyonunu akran etkileşiminin olduğu konuşma performansını değerlendirmek için kullanmıştır.

Şahin ve diğerleri (2016) ÇYRM ile akran değerlendirmesini incelemiştir. Lisans düzeyinde gerçekleşen bu çalışmada altı akran ve dersin öğretim öğretim elemanı, diğer öğrencilerin performanslarını holsitik rubrik ile değerlendirmiştir. Sonuç olarak, ÇYRM'in akran değerlendirme kapsamında öğrenci performansını güvenilir bir şekilde ayırt etmede başarılı olduğu ankaşılmıştır. Ayrıca akranların başarı düzeyinin puanlamalarında diğerlerinden anlamlı şekilde farklılalşmasında etkili olduğu bulunmuştur. Yeşilçınar ve Şata (2021) yaptıkları çalışmada kimlik bilgileri gizlenerek yapılan akran değerlendirmenin güvenirlik

ve geçerliliğinin daha yüksek olduğu sonucuna varmıştır. Tursynbayeva ve diğerleri (2023) tarafından yapılan çalışmada ortaokul düzeyinde akran değerlendirme kapsamında puanlayıcı eğitimi vermiş ve bu eğitimin katılık ve cömertlik davranışlarında değişimlere neden olduğu sonucuna varmıştır. Köse ve diğerleri (2016) yüksek lisans programına kayıtlı öğrencilerin sunum yapma becerilerinde akran değerlendirme uygulamış ve puanları ÇYRM ile analiz etmiştir. Sarıtaş-Akyol ve Karakaya (2021) problem çözme becerileri üzerine yaptıkları çalışmada farklı zamanlarda verilen öz, akran ve öğretmen puanlarını ÇYRM ile analiz etmiştir. Araştırma çıktıları akran puanlarının katılık ve cömertlik açısından öğretmen puanlarına yakınlık gösterdiği ve akran değerlendirmenin verilen eğitim ve kazanılan deneyim ile güvenilir bir şekilde kullanabileceğini göstermektedir. Araştırma çıktılarında ÇYRM'in performans değerlendirmede alternatif bir yol olarak kullanılabileceğine dikkat çekilmektedir. Farklı bir disiplin alanında araştırma yapan Yüzüak ve diğerleri (2015), ÇYRM'nin performans ve akran değerlendirmelerinde kullanım uygunluğuna dikkat çekmiştir.

Şata ve Karakaya (2022) yazma becerilerinin değerlendirmesinde puanlayıcılara referans çerçeve eğitimi ve puanlayıcı hataları eğitimi kapsamında dört haftalık bir eğitim sunmuştur. Deney ve kontrol grubunun kullanıldığı çalışmada, gruplar arasında deney sonrası farklılığın oluştuğu tespit edildi. Ayrıca, eğitimin puanlayıcı hatalarını azaltmada etkili olabileceğinin altı çizildi. Sayın ve Şata (2022), yazma becerilerinin puanlanmasında sahip olunan uzmanlık ile puanlama güvenirliği arasında bir ilişkinin olduğu ifade edilmektedir. Yazma becerisinin incelendiği bir diğer çalışmada Fişne ve diğerleri (2022) ÇYRM ile standart belirleme yöntemi ile yükseköğretim öğrencilerinin akademik yazma becerisi ile ilgili yeterlik düzeylerini belirlemişlerdir.

Demir (2023) sözlü sunum becerileri kapsamında akran ve öz değerlendirmeleri ÇYRM ile analiz etmiştir. Araştırma sonucu, puanlayıcı katılık ve cömertliklerinde ve kriterlerin zorluk düzeylerinde farklılıkların olduğunu ve ÇYRM'nin puanlayıcı yanlılıklarının kaynak ve yönünü incelemede kullanabileceğini ifade etmektedir. Ayçiçek (2025) drama becerileri dersi kapsamında akran değerlendirme kullanmış ve puanlayıcı davranışlarında istatistiksel anlamlı farklılıklar tespit etmiştir. Akran değerlendirmeyi etkileyen faktörlerin araştırılması tavsiye edilmiştir. Aryadoust (2015) sözlü sunum becerilerinin ölçülmesinde öz ve akran değerlendirmelerini ve puanlayıcı yanlılıklarını incelemiş, yanlılık davranışının öz

değerlendirme de negatif, akran ve öğretmen değerlendirmesinde karışık yapıda olduğunu belirtmiştir.

Özetlemek gerekirse ÇYRM, yazma becerileri ve akran değerlendirme kapsamında çalışmalar genel anlamda puanlayıcı davranışlarına, ÇYRM'nin etkiliğine, puanlayıcı eğitimine, geri bildirim ve öğrenmeye etkisine ve son olarak psikometrik/kuramsal araştırmalara odaklanmaktadır. Hem yabancı dil öğretimi alanında hem farklı disiplinlerde ÇYRM'in ölçme ve değerlendirme süreçlerinde etkili olarak kullanılabileceğine dikkat çekilmiştir.

BÖLÜM III

YÖNTEM

Bu bölüm, mevcut araştırmanın yöntemsel süreçlerini kapsamaktadır. Öncelikle araştırma deseni ve örnekleme yöntemi teorik ve pratik olarak açıklanacaktır. Daha sonra veri toplama süreçleri ve araçları detaylı olarak ele alınacaktır. Son olarak nicel verilerin nasıl analiz edildiği ile ilgili bilgiler tartışılacaktır.

Araştırma Deseni

Bu araştırmada nicel araştırma yöntemleri kapsamında zayıf deneysel araştırma deseni kullanılmaktadır. Deneysel çalışmaları, diğer araştırma desenlerinden ayıran en önemli özellik bağımsız değişkenin manipüleye açık olarak araştırmaya dahil edilmesidir (Fraenkel vd., 2012). Deneysel çalışmalar, deney ve kontrol grubun varlığına göre veya uygulama yapılacak deneklerin tesadüfi seçilip seçilmemesine göre farklı türlere ayrılmaktadır. Bu çalışmada yapılan uygulamada kontrol grubunun olmaması ve deneklerin evrenden tesadüfi olarak seçilememesinden dolayı zayıf deneysel desen tercih edilmiştir (Büyüköztürk et al., 2014). Zayıf deneysel desen çerçevesinde "Puanlayıcı Eğitimi" ana uygulama olarak düşünülmüş; akran puanlama davranışları bağımlı değişken olarak tanımlanmıştır. Akran puanlama davranışları uygulama öncesi ve sonrası izlenmiş ve değişimler raporlanmıştır. Deneysel çalışma kapsamında verilerin toplanması için dereceli puanlama anahtarları kullanılmış ve yine veriler nicel araştırma yaklaşımına uygun olarak çok yüzeyli Rasch yöntemi ile analiz edilmiştir.

Evren ve Örneklem

Çalışmanın evrenini, İngilizce akademik yazma becerileri dersinde akran geri bildirimini veren tüm puanlayıcılar oluşturmaktadır. Bu evrenden uygun örneklem yoluyla toplam 51 puanlayıcı seçilmiştir. Uygun örneklem ya da kazara örneklem olarak bilinen bu yöntem araştırmacıların pratik yollar ile ulaşabileceği katılımcılardan araştırma kapsamında veri toplayabilmesi anlamına gelmektedir (Büyüköztürk et al., 2014). Araştırma örneklemini oluşturan katılımcılar Gazi Üniversitesi İngilizce Öğretmenliği lisans programında 1. sınıfta öğrenimine devam eden İngilizce öğretmeni adaylarıdır. Sorumlu oldukları lisans programında "Yazma Becerileri I" ve "Yazma Becerileri II" başlıklı derslerde akademik yazma davranışlarını pekiştiren öğretmen adaylarının çalışma esnasında araştırmaya katılım gönüllülüğü esas alınmıştır. Yaş aralığı 19-22 arasında olan öğretmen adayları sırasıyla güz ve bahar döneminde aldıkları dersler ile akademik kompozisyon yapısı ve yazımı ile ilgili pratikler yapmıştır.

Tablo 2

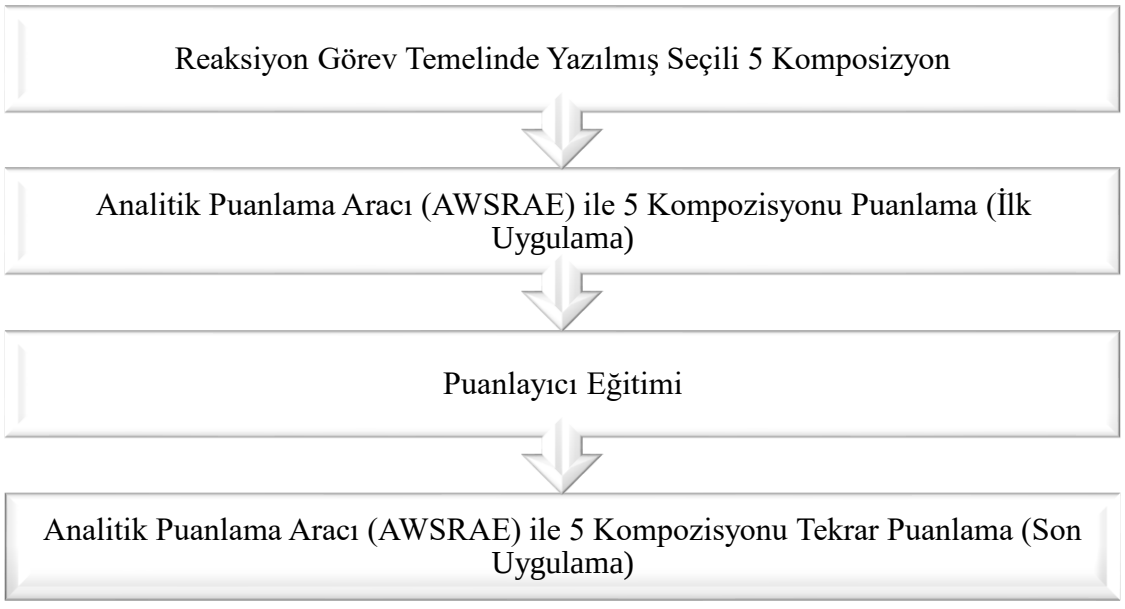
Katılımcıların Demografik Bilgileri

		f	%
Cinsiyet			
	Kadın	16	31,37
	Erkek	35	68,63
	Toplam	51	100
Üniversite Hazırlık Eğitimi			
	Evet	36	70,59
	Hayır	15	29,41
	Toplam	51	100
Yazma Becerileri I - Güz Harf Notu			
	AA	40	78,43
	BA	3	5,88
	BB	4	7,84
	Belirsiz	4	7,84
	Toplam	51	100

Bir akademik yıl boyunca akademik yazma alanında deneyim kazanan öğretmen adaylarından bir önceki sene kaleme alınan ve kompozisyon kriterlerinin farklı boyutlarında farklı nitelik özellikleri sergileyen (Tablo 3) toplam 5 kompozisyonu uygulama öncesi ve sonrası puanlandırmaları istenmiştir. Katılımcılara ait demografik bilgiler aşağıdaki tabloda sunulmaktadır. Tablo 2'de özetlendiği üzere 51 öğretmen adayının akademik yazma becerileri dersinde önemli düzeyde başarılı olduğu anlaşılmaktadır. Bu başarılı performans göstergesi adayların puanlayıcı olarak görev almasının uygunluğuna dair de kanıt oluşturmaktadır.

Verilerin Toplanması

Araştırma için veri toplama sürecinde 4 aşama vardır. Öncelikle uygulama yapılmadan bir önceki sene akademik yazma dersini alan İngilizce öğretmeni adaylarının reaksiyon türünde yazdığı kompozisyonlar değerlendirilmiştir. Bu kompozisyonlar arasında en iyi düzeyde, iyi düzeyde, orta düzeyde, zayıf düzeyde ve en zayıf düzeyde olan toplam 5 ürün seçilmiştir. Puanlama eğitimi verilecek olan araştırma örnekleminde yer alan toplam 51 öğrenciden eğitim öncesi ve sonrası 5 kompozisyonu puanlandırması istenmiştir. Şekil 2'de veri toplama süreci özetlenmiştir.



Şekil 2. Veri toplama süreci

Ardından puanlama grubunda yer alan öğrencilere hem ölçme aracı hem de puanlayıcı hatalarına dair eğitimler verilmiş ve eğitim kapsamında öğrencilerin öz yansıtma yapmaları

istenmiştir. Bu eğitimler özellikle "Puanlayıcı Hataları Eğitimi" ve "Performans Boyutları Eğitimi" kapsamında sunulmuştur. Ön uygulama sonrası öğrencilere temel puanlayıcı hatalarının neler olduğu aktarılmıştır ve öğrenciler kendi pratiklerini gözden geçirerek hata türlerine karşı farkındalık kazanmıştır. Ardından ilgili reaksiyon kompozisyon görevinde beklenilen performansın ne olduğuna dair tartışmalar yapılmış, performans için gerekli olan ölçütler ve boyutlar derinlemesine incelenmiştir. Son veri toplama aşamasında puanlama grubunda yer alan öğrencilerden aynı 5 kompozisyonu 6 hafta ara ile tekrar puanlaması istenmiştir.

Reaksiyon Kompozisyonu Görevi (Reaction Essay)

Reaksiyon kompozisyonu, öğrencilerin belirli bir konuda tepkilerini veya görüşlerini ortaya koydukları ve kompozisyon boyunca bu tepki veya görüşü destekleyecek argümanlar geliştirdikleri bir görev türüdür (Folse & Pugh, 2019). Bu kompozisyon türü giriş, gelişme ve sonuç olmak üzre toplam üç ana bölümden oluşmaktadır.

Giriş bölümü reaksiyonu özetleyen bir başlık ile başlamaktadır. Başlık sonrası, okurların dikkatini çeken bir giriş cümlesi (hook sentence) ile "Giriş" bölümü başlar, arka plan bilgisi verilir ve konu ile ilgili bir görüş ve tepki ile tez ifadesi (thesis statement) bu bölümün son cümlesi olarak sunulur. "Gelişme" bölümünde konu ile ilgili tepkinin veya görüşün ne açıdan oluştuğuna yönelik tartışmalar bütün olarak ele alınır ve kompozisyon organizasyon yapısına göre gelişme bölümü şekillendir. Son olarak, "Sonuç" bölümünde tez cümlesi tekrar sunulur, tepki veya görüşü destekleyen argümanlar tekrar özetlenir ve kapanış cümlesi ile kompozisyon bitirilir.

Bu araştırmada bir önceki seneden seçilen farkı kriter nitelik düzeylerine sahip beş öğrenci kompozisyonu, 51 İngilizce öğretmenliği öğrencisine puanlamak için sunulmuştur. Kompozisyonun ana konusu Daniel Keyes tarafından kaleme alınan "Algernona Çiçekler" (Flowers for Algernon) başlıklı hikayedir. Önceki sene dersi alan öğrencilerden hikâyeyi okuduktan sonra tepkilerini anlatan bir kompozisyon yazmaları istenmiştir. Tablo 3'te kompozisyonların temel nitelikleri, kriterlere uygunluk seviyeleri ve ders öğretmeni tarafından verilen puanlar izah edilmiştir.

Yazma Becerilerine Yönelik Analitik Dereceli Puanlama Aracı (AWSRAE)

İngilizce karşılığı "Analytical Writing Scoring Rubric for Academic Essays" (ASWRAE) olan yazma becerilerine yönelik analitik dereceli puanlama anahtarı, veri toplama aracı olarak (2019) tarafından geliştirilmiş ve geliştirilen bu dereceli puanlama anahtarı, aynı araştırmacılar tarafından Şata ve Karakaya (2020) ve Fişne ve diğerleri (2022) çalışmalarında da kullanılmıştır. Bu veri toplama aracının toplam 16 boyutu/ölçütü vardır. Bu boyutlar sırasıyla aşağıda listelenmiştir.

1. Başlık (Title of Essay)
2. Organizasyon/ Giriş–Gelişme–Sonuç (Introduction–Body–Conclusion)
3. Tez Cümlesi (Thesis Statement)
4. Konu Cümlesi (Topic Sentence)
5. Destekleyici Cümleler (Supporting Sentences)
6. Uygun Uzunluk (Appropriate Length)
7. Konuya Uygunluk (Topic Relevance)
8. Fikir Geliştirme (Idea Development)
9. Tutarlılık (Coherence)
10. Bağlantı Unsurları (Linking)
11. Dilbilgisi (Accuracy of Grammatical Forms)
12. Sözdizimsel Karmaşıklık (Syntactic Complexity)
13. Kelime Seçimi (Word Choice)
14. Sözcük Çeşitliliği (Lexical Range)
15. Yazım Kuralları/İmla (Spelling)
16. Noktalama (Punctuation)

Şata ve Karakaya (2020)'de de belirtildiği üzere bu boyut/ölçütlerin akademik yazma performansını ne kadar kapsadığına dair uzman görüşü alınmış, ve görüşler doğrultuunda yapılan düzenlemeler ile kapsam geçerliliği sağlanmıştır. Veri toplama aracının yapı geçerliliğini test etmek için, açımlayıcı faktör analizi (AFA) yapılmıştır. AFA yapmak için elde edilen örneklem incelendiğinde çoğu katılımcının komposizyon başlığı yazmadığı veya özensiz yazdığı gözlemlenmiştir. Bu davranışın ölçme yapısına zarar vermesinin önüne

geçmek için analiz 15 alt boyutla gerçekleşmiştir. Mevcut çalışmada 16 alt boyutta kullanılacağı için öncelikle kullanılan dereceli puanlama anahtarına doğrulayıcı faktör analizi (DFA) yapılmıştır. Ancak 16 ölçüte sahip olan model DFA tarafından doğrulanamadığından (RMSE>0.08), aynı ölçütlere AFA yapılmıştır. Analiz sonucunda tek faktörlü yapıda ölçüt varyanslarının yazma becerileri yapısını açıklamakta yeterli olduğu görülmektedir. Açıklanan toplam varyansın %38 olduğunu tespit edilmiştir. Tek faktörlü ölçeklerde açıklanan varyansın %30'dan büyük olması beklenmektedir (Çokluk vd., 2014). Bu açıdan ölçeğin tek föktörlü yapıya sahip olduğu görülmektedir. Güvenirlik için Cronbach Alpha hesaplanmış ve güvenirlik katsayısının ,88 olduğu anlaşılmıştır.

Puanlayıcı Eğitimi

Bu çalışma kapsamında verilen puanlayıcı eğitiminde "Puanlayıcı Hataları Eğitimi" ve "Performans Boyutları Eğitimi" modelleri çerçevesinde geliştirilmiştir. Toplam 6 hafta boyunca süren bu eğitimde öncelikle beklenilen performansın boyutları, dereceli puanlama anahtarı kapsamında incelenmiştir. Katılımcıların aşina olmadıkları performans boyutları eğitimde tartışılmış ve boyutlara ait alt özellikler açıklanmıştır. Bu içeriğin ardından katılımcılar ile hatalı puanlayıcı davranışları incelenmiştir. Her bir hata türünün ne anlama geldiği ve neden güvenirliği ve geçerliliği tehdit ettiği açıklanmıştır. Puanlayıcılar hata eğitimi sonrası, ön uygulama da yapmış olabilecekleri potansiyel hata türlerine dair farkındalık geliştirmiştir.

Veri Analizi

Bu çalışmada veri analizi ÇYRÖM kullanılarak yapılmıştır. ÇYRÖM, madde tepki kuramına dayalı bir analiz yöntemidir. Analiz her ne kadar varyans kaynaklarını bağımsız olarak inceleme fırsatı sunsa da Hamp-Lyons (1989) Rasch ölçme modelinin yüzeylerin birbirleri ile nasıl etkileşimde bulunduğunu gösterdiğini belirtmektedir. Bu kapsamda ÇYRÖM, yüzeyleri hem bireysel hem de bütüncül inceleme fırsatı verecektir. Rasch modelinin bir uzantısı olarak geliştiren bu modelde ön ve son uygulama ayrı ayrı ele alındığında toplam üç, uygulama bütünü ele alındığında ise toplam dört yüzey analize dahil edilmiştir. Dört yüzeyli kurulan modelde ön ve son uygulamanın da bir değişken olarak ele alındığı zaman yüzeyi mevcuttur. Zamandan bağımsız birey, yetenek ve madde

parametreleri incelenmiş ve zaman dilimleri arasında anlamlı bir fark olup olmadığına bakılmıştır. Dört yüzeyli Rasch analizi için temel alınan formül aşağıda sunulmaktadır.

$$\log\left(\frac{\mathrm{P_{nijlk}}}{\mathrm{P_{nijl(k-1)}}}\right) = B_n - D_{gi} - C_j - E_l - F_{gk}$$

Formülde;

$\mathrm{P_{nijlk}}$: j numaralı puanlayıcının n numaralı bireyin l numaralı zamanda i numaralı ölçütüne k puan verme olasılığı

$\mathrm{P_{nijl\,(k-1)}}$: j numaralı puanlayıcının n numaralı bireyin l numaralı zamanda i numaralı ölçütüne k-1 puan verme olasılığı

B_n : n numaralı bireyin akademik yazma puanı

D_{gi} : g isimli dereceli puanlama anahtarının i ölçütüne ait güçlük düzeyi

C_j : j numaralı puanlayıcının cömertliği veya katılığı

E_l : l numaralı zaman etkisi

F_{gk} :g isimli dereceli puanlama anahtarında k-1 puana göre k puanı alma güçlüğü

Diğer taraftan ÇYRÖM kullanabilmek için bazı varsayımların da karşılanması gerekmektedir: Bu varsayımlar Yu (2020) tarafından şu şekilde belirtilmiştir: (a) tek boyutluluk (unidimensionality), (b) yerel bağımsızlık, (c) madde kalibrasyonu ve yetenek kestirimidir. Tek boyutluluk varsayımı dereceli puanlama anahtarının yapı geçerliliğinin sağlanması ile karşılanmıştır. Hambleton ve diğerleri (1991), tek boyutluluk sağlandığında yerel bağımsızlık, bir diğer deyiş ile puanlayıcıların bir ölçüte verdiği puanın, bir diğer ölçütten bağımsız olması, varsayımı da karşılanmış olmaktadır. Madde kalibrasyonu için ise madde veri uyumu incelenmiştir.

Madde analizlerinde -3 ve +3 arasında kalan standartlaştırılmamış artık değerlerin, toplam veri sayısının %1'ini geçmemesi model verim uyumuna işaret etmektedir (Linacre, 2023). Mevcut veri setinde toplam 72 standartlaştırılmış artık veri tespit edilmiştir. Toplam gözlem

sayısının (51x5x2x16) 8160'ın %1'i 81.6'ya kaşılık gelmektedir. Bu durumda model veri uyumu varsayımının da karşılandığı görülmektedir.

ÇYRÖM'den elde edile ön ve son puanlamaya ait logit değerlerinin arasında anlamlı fark olup olmadığı eşleştirilmiş örneklem t testi (paired samples t test) ile inclenmiştir. Bir karşılaştırma testi olan t testinde varsayımlar kontrol edilmiş ve veri setlerinin varsayımı sağladığı anlaşılmıştır. Analizlerin yapıldığı paket programlar incelendiğinde ise araştırmada toplam üç paket programın kullanıldığı ifade edilebilir. ÇYRÖM analizleri, FACETs programı kullanılarak yapılmıştır. Eşleştirilmiş örneklem t-testi, AFA ve güvenirlik analizi SPSS 26 paket programı aracılığıyla gerçekleştirilmiştir. Son olarak DFA, JASP paket programından yararlanılmıştır.

Tablo 3

*Uygulama Kapsamında Puanlanması Beklenilen 5 Kompozisyon Görevi**

	Temel Özellikler	Seçilme Niteliği	Ürün Niteliği	Puan**
Kompozisyon 1	Etkili bir giriş cümlesi ile başlamaktadır. Büyük ölçüde akademik bir dil kullanılmaktadır. Ürün organizasyonu tez ifadesine göre şekillenmektedir. Sonuç bölümünde, kontrol ifadelerinin daha kapsamlı özetlenmesine ihtiyaç vardır.	Etkili bir kompozisyonun sahip olması gereken ölçütleri çoğunlukla yerine getirmektedir.	İleri Düzey	92
Kompozisyon 2	Etkili bir giriş cümlesi ile başlayan üründe altyapı bilgisi sunulmamaktadır. Reaksiyon olmasına rağmen kontrol ifadeleri açık değildir. İlk gelişme paragrafının detaylandırılmaya ihtiyacı vardır. Paragrafa ait giriş ve sonuç cümleleri açıktır.	Eksik ve güçlü yanları olan bir ürünün, puanlayıcılar tarafından nasıl puanlandırıldığı anlaşılmak istenmektedir.	Orta Düzey	82
Kompozisyon 3	Giriş cümlesi eksiktir. Tez ifadesi mevcuttur ama yanlış yerde verilmiştir. Sonuç bölümü, tez ifadesi ile uyumludur. Gelişme bölümüne ait ilk iki paragrafın detaylandırılmaya ihtiyacı vardır. Akademik bir dil kullanılmıştır.	Eksik ve güçlü yanları olan bir ürünün, puanlayıcılar tarafından nasıl puanlandırıldığı anlaşılmak istenmektedir.	Orta Düzey	75
Kompozisyon 4	Ürün, organizasyon yapısından eksiktir. Etkili bir giriş cümlesi ile başlamamaktadır. Tez ifadesi açık değildir. Kişisel dil ifadeleri içerir. Tepki kompozisyonu yerine özet niteliği taşır.	Puanlayıcıların, eksik yönleri olan bir ürünü nasıl puanladıklarını anlayabilmek için seçilmiştir.	Geliştirilmeli	38
Kompozisyon 5	Etkili bir giriş cümlesi ile başlar. Detaylı altyapı bilgisi verir. Tez ifadesi karışıktır. Fazla kişisel dil kullanımı içerir. Gelişme bölümünün daha fazla detaya ihtiyacı vardır. Ürün, tez ifadesi çerçevesinde gerektiği gibi şekillenmemiştir.	Bu ürün, etkili bir kompozisyon ölçütlerinin bir bölümünü karşılamaktadır. Puanlayıcıların, bu ölçütlere nasıl yaklaştığı öğrenilmek istenmiştir.	Başlangıç Düzeyinde	53

*Kompozisyonlar bulgular bölümünde “Kompozisyon” “Öğrenci” etiketi ile sunulmaktadır.

**Puanlar derse giren öğretmenin/araştırmacının performanslara verdiği niceliksel değerlerdir.

BÖLÜM IV

BULGULAR

Bu bölümde çalışmaya ait bulgular araştırma soruları doğrultusunda sunulacaktır. Öncelikle puanlayıcı eğitimi öncesinde ve sonrasında gerçekleştirilen kompozisyon görevi puanlamalarına ait ölçüm raporları sunulacaktır. Sonrasında ise puanlayıcı davranışlarının eğitim kapsamında nasıl değiştiği incelenecek ve iki ölçüm arasında herhangi bir fark olup olmadığı tartışılacaktır.

Birinci Araştırma Sorusuna Yönelik Bulgular

Araştırma Sorusu 1: Puanlayıcı eğitimi öncesinde akademik yazma becerilerinin puanlanması sürecinde;

a. İngilizce öğretmen adaylarından oluşan puanlayıcı grubunun puanlamalarında öğrenci, ölçüt ve puanlayıcı yüzeylerinden elde edilen verilerin uygunluk değerleri ve model ölçümleri nasıldır?
b. İngilizce öğretmen adaylarından oluşan puanlayıcı grubunun puanlama davranışları (katılık, cömertlik ve yanlılık) nasıldır?

Puanlama eğitimi öncesinde puanlayıcı, öğrenci performansı (kompozisyon) ve kriterler yüzeylerine ait ölçüm raporları sırasıyla tablolaştırılmıştır. Toplam 51 puanlayıcının 5 kompozisyon görevine verdikleri puanların ÇYRM'e göre uygunluk değerleri ve model ölçüm raporları analiz edilmiş ve bulgular Tablo 4'te sunulmuştur. Tabloda yer verilen logit değerinin büyüklüğü bireylerin yüksek puanlama (cömertlik davranışı) yaptığı küçüklüğü ise düşük düzeyde puanlama (katılık davranışı) sergilediği anlamına gelmektedir.

Tablo 4 incelendiğinde puanlama düzeylerinin 0,29 ile 2,14 logit arasında değiştiği gözlemlenmektedir. Bireylerin logit değerlerinin ortalaması 1,18 ve standart sapması 0,14'tür. Bu değerler ışığında en cömert puanlamaya sahip olan öğretmen adayı 44 (logit: 2,03) numaralı puanlayıcı; en katı puanlama davranışa sahip olan birey ise 30 (logit: 0,29) numaralı puanlayıcıdır. Birey yüzeyi için uygunluk içi değerleri 0,48 ile 1,94 arasında değişirken ortalaması 0,99'dur. Birey yüzeyi için uygunluk dışı değerleri 0,47 ile 2,43 arasında değişirken ortalaması 1,03'tür. Uygunluk değerlerinin 0,5 ile 1,5 arasında olması verinin modele iyi uyum gösterdiği anlamına gelmektedir (Linacre, 2002). Bu kapsamda model veri uyumunun iyi düzeyde olduğu söylenebilir. Birey yüzeyine ait ayırma katsayısı 4,44 ve güvenirlik indeksi ise 0,90'dır. Bu birey puanlamaların güvenilir şekilde birbirinden ayrıldığı anlamına gelmektedir.

Tablo 4

Uygulama Öncesi Puanlayıcı Yüzeyine Ait Ölçüm Raporu

Puanlayıcı	Gözlenen ortalama	Düzelt. Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	3,98	4,08	1,11	0,13	0,85	0,80
2	4,46	4,56	1,92	0,16	0,92	0,91
3	4,00	4,10	1,14	0,13	0,55	0,54
4	3,92	4,03	1,04	0,13	0,48	0,47
5	4,36	4,46	1,72	0,15	0,88	0,77
6	4,34	4,44	1,67	0,15	0,74	0,73
7	4,08	4,18	1,25	0,13	0,90	0,97
8	4,35	4,45	1,70	0,15	1,28	1,29
9	3,79	3,88	0,86	0,12	1,68	1,85
10	3,99	4,09	1,12	0,13	0,73	0,76
11	4,33	4,43	1,65	0,15	1,14	1,03
12	3,81	3,91	0,89	0,13	0,82	0,79
13	3,63	3,71	0,67	0,12	0,87	0,88
14	4,15	4,26	1,36	0,14	0,83	0,88
15	4,13	4,23	1,32	0,14	1,26	1,14
16	4,28	4,38	1,56	0,15	0,88	1,32
17	3,90	4,00	1,01	0,13	0,71	0,77
18	3,51	3,59	0,54	0,12	0,92	1,21
19	3,74	3,83	0,8	0,12	0,84	0,8
20	3,94	4,04	1,06	0,13	0,98	1,1
21	3,73	3,82	0,79	0,12	0,99	0,97
22	4,00	4,10	1,14	0,13	1,13	1,23

23	3,92	4,03	1,04	0,13	0,79	0,74
24	3,65	3,74	0,70	0,12	1,03	0,99
25	4,22	4,33	1,48	0,14	1,10	1,27
26	4,04	4,14	1,19	0,13	0,70	0,91
27	4,44	4,53	1,87	0,16	0,81	0,83
28	4,20	4,31	1,44	0,14	0,90	1,01
29	3,71	3,80	0,77	0,12	1,20	1,19
30	3,29	3,34	0,29	0,12	1,36	1,33
31	3,85	3,95	0,94	0,13	1,66	1,65
32	4,29	4,39	1,58	0,15	1,16	1,04
33	3,42	3,49	0,44	0,12	1,10	1,11
34	3,44	3,51	0,45	0,12	0,73	0,78
35	4,34	4,44	1,67	0,15	1,94	2,43
36	3,75	3,84	0,82	0,12	1,21	1,35
37	4,14	4,24	1,34	0,14	1,01	1,02
38	3,92	4,03	1,04	0,13	1,24	1,31
39	3,84	3,94	0,93	0,13	0,49	0,55
40	4,16	4,27	1,38	0,14	0,82	0,88
41	3,95	4,05	1,07	0,13	0,54	0,54
42	4,45	4,54	1,89	0,16	0,74	0,71
43	4,26	4,37	1,54	0,15	1,18	1,07
44	4,51	4,60	2,03	0,17	1,10	1,21
45	3,81	3,91	0,89	0,13	1,68	1,67
46	4,16	4,27	1,38	0,14	0,70	0,69
47	3,49	3,56	0,51	0,12	1,49	1,43
48	4,47	4,57	1,95	0,17	0,97	0,84
49	3,65	3,74	0,7	0,12	0,81	0,78
50	3,83	3,92	0,91	0,13	0,92	0,9
51	4,44	4,53	1,87	0,16	0,85	0,87
Ortalama	4	4,10	1,18	0,14	0,99	1,03
Evren	0,31	0,32	0,44	0,01	0,31	0,35
Örneklem	0,31	0,32	0,45	0,01	0,31	0,36

Model Evren: RMSE ,14 Düzelt. S.H. ,42 Ayırma İndeksi 3,08 Ayırma Katsayısı 4,44 Güvenirlik ,90
Model Örneklem: RMSE ,14 Düzelt. S.H. ,43 Ayırma İndeksi 3,12 Ayırma Katsayısı 4,49 Güvenirlik ,91
Model (Sabit etkili) ki-kare: 519,2 s.d.: 50 anlamlılık (olasılık): ,00
Model (Normal) ki-kare: 45,5 s.d..: 49 anlamlılık (olasılık): ,62

Tablo 4'te model örnekleminin ki-kare değeri incelendiğinde bu değerin istatistiksel olarak anlamlı olduğu belirlenmiştir. ($\chi 2(sd)$ =519,2(50), p= 0,00 < 0,01). Bu bulgu bireylerin puanlama davranışları arasında istatistiksel olarak anlamlı bir fark olduğunu işaret etmektedir.

Birey yüzeyine ilişkin ölçümlerin incelenmesinin ardından öğrenci veya kompozisyon yüzeyine ait ölçüm değerleri de Tablo 5'te raporlanmaktadır. Öğrenci performanslarına ait ölçüm değerlerine bakıldığında logit değerlerinin -0,71 ile 0,76 arasında olduğu görülmektedir ve bu model veri uyumuna işaret etmektedir. Logit değeri artıkça kompozisyonların kriterlere uygun yazılma davranışının, başka bir deyişle öğrenci performanslarının başarılı olma düzeylerinin artığı gözlenmektedir. Bu sonuçlar, kompozisyon seçilme kriterleri ile uygunluk göstermektedir. Seçilme kriterlerindeki düzeyler ile logit değerleri uyum göstermektedir. Kompozisyon yüzeyine ait uygunluk düzeylerine bakıldığında ise en düşük uygunluk içi değerinin 0,90 ve en yüksek uygunluk içi değerinin 1,16 olduğu görülmektedir. Benzer şekilde, en düşük uygunluk dışı değerinin 0,90 ve en yüksek uygunluk dışı değerin ise 1,18 olduğu anlaşılmaktadır. Değerler, verinin modele iyi uyum sergilediğini işaret etmektedir. Kompozisyon yüzeyine ait ayırma katsayısı 16,40 ve güvenirlik indeksi ise 0,99 olarak belirlenmiştir. Bu kompozisyonların güvenilir şekilde birbirinden ayrıldığı anlamına gelmektedir.

Tablo 5

Uygulama Öncesi Öğrenci/Komposizyon Yüzeyine Ait Ölçüm Raporu

Öğrenci/ Kompozisyon	Gözlenen ortalama	Düzelt. Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	4,26	4,35	0,33	0,05	1,10	1,10
2	4,49	4,57	0,76	0,05	1,16	1,18
3	3,47	3,53	-0,71	0,04	100	1,01
4	3,72	3,8	-0,41	0,04	0,91	0,94
5	4,06	4,16	0,03	0,04	0,90	0,90
Ortalama	4	4,08	0	0,04	1,01	1,03
Evren	0,37	0,37	0,52	0,01	0,10	0,10
Örneklem	0,41	0,42	0,58	0,01	0,12	0,11

Model Evren: RMSE ,04 Düzelt. S.H. ,52 Ayırma İndeksi 12,05 Ayırma Katsayısı 16,40 Güvenirlik ,99
Model Örneklem: RMSE ,04 Düzelt. S.H. ,58 Ayırma İnd.13,48 Ayırma Katsayısı 18,31 Güvenirlik ,99
Model (Sabit etkili) ki-kare: 712,8 s.d: 4 anlamlılık (olasılık): ,00
Model (Normal) ki-kare: 4.0 s.d: 3 anlamlılık (olasılık): ,26

Tablo 5'te verilen kompozisyon yüzeyine ait ki-kare değeri incelendiğinde ise bu değerin istatistiksel olarak anlamlı olduğu görülmektedir. ($\chi 2(sd)$ =712,8(4), p= 0,00 < 0,01). Bu bulgu bireylerin kompozisyon düzeyleri arasında istatistiksel olarak anlamlı bir fark

olduğunu işaret etmektedir. Tablo 6 ise puanlayıcı eğitimi öncesinde kompozisyon kriterlerine verilen puanlara ait ölçüm raporunu göstermektedir. Toplam 16 kriterin logit değeri incelendiğinde en katı puanlanan ölçütün -0.83 ile "Tez Cümlesi" (Thesis Statement) olduğu görülmektedir.

Tablo 6

Uygulama Öncesi Kriter Yüzeyine Ait Ölçüm Raporu

Kriterler/Ölçütler	Gözlenen ortalama	Düzelt. Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1. Başlık	3,89	3,97	-0,22	0,07	1,09	1,03
2. Organizasyon	3,78	3,86	-0,35	0,07	0,68	0,70
3. Tez Cümlesi	3,36	3,41	-0,83	0,07	1,38	1,36
4. Konu Cümlesi	3,60	3,67	-0,56	0,07	1,10	1,07
5. Destekleyici Cümleler	3,88	3,97	-0,22	0,07	0,68	0,71
6. Uygun Uzunluk	4,25	4,34	0,32	0,08	1,31	1,46
7. Konuya Uygunluk	4,35	4,44	0,49	0,09	0,90	0,83
8. Fikir Gelişimi	4,03	4,12	-0,02	0,07	0,70	0,73
9. Tutarlık	4,26	4,35	0,33	0,08	0,85	0,85
10. Bağlantı Unsurları	3,8	3,88	-0,33	0,07	1,12	1,19
11. Dil Bilgisi	4,33	4,42	0,45	0,09-8	1,00	1,02
12. Sözdizimsel Karmaşıklık	3,73	3,80	-0,41	0,07	0,87	0,88
13. Kelime Seçimi	3,61	3,68	-0,55	0,07	0,92	0,94
14. Kelime Çeşitliliği	3,97	4,06	-0,11	0,07	0,94	1,00
15. Yazım Kuralları/İmla	4,67	4,73	1,26	0,11	1,36	1,29
16. Noktalama	4,48	4,56	0,75	0,09	1,33	1,36
Ortalama	4,00	4,08	0,00	0,08	1,01	1,03
Evren	0,35	0,36	0,54	0,01	0,23	0,24
Örneklem	0,36	0,37	0,56	0,01	0,24	0,25

Model Evren: RMSE ,08 Düzelt. S.H. .53 Ayırma İnd. 6,82 Ayırma Katsayısı 9,42 Güvenirlik ,98
Model Örneklem: RMSE ,08 Düzelt. S.H. .55 Ayırma İnd. 7,04 Ayırma Katsayısı 9,72 Güvenirlik ,98
Model (Sabit etkili) ki-kare: 640,5 s.d.: 15 anlamlılık (olasılık): ,00
Model (Normal) ki-kare: 14,6 s.d..: 14 anlamlılık (olasılık): ,41

Diğer tarafta 1,26 logit değeri ile "Yazım Kuralları/İmla" ölçütünün en cömert puanlana ölçüt olduğu görülmektedir. Bu yüzeye ait logit değrlerin ortalaması 0'dır. Uygunluk içi değerleri 1,38 ile 0,68 arasında değişiklik gösterirken ortalaması 1,01'dir. Uygunluk dışı değerlerin 0,70 ve 1,46 arasında farklılık gösterdiği anlaşılmaktadır ve bu değerlerin ortalaması 1,03'tür. Bu değerlerin 1'e yakınlık göstermesi verinin modele iyi uyum gösterdiğini işaret etmektedir.

Kriter/ölçüt yüzeyine ait ki-kare değeri incelendiğinde ise bu değerin istatistiksel olarak anlamlı olduğu görülmektedir. ($\chi 2(sd)$ =640,5 (15), p= 0,00 < 0,01). Bu bulgu bireylerin kompozisyon ölçütleri/kriterleri arasında istatistiksel olarak anlamlı bir fark olduğunu işaret etmektedir.

Model ölçüm raporları incelendiğinde veri setlerinin modele iyi uyum gösterdiği, puanlayıcıların puanlama davranışlarının birbirinden farklılaştığı, bazı puanlayıcıların oldukça cömert veya katı davrandığı, kompozisyon/öğrenci performanslarının seçim kriterine uygun puanlandığı, en katı puanlanan ölçeğin tez cümlesi olduğu anlaşılmaktadır. Model veri uygunluğunun ve ölçüm raporlarının ardından uygulama öncesinde puanlayıcıların yanlılık davranışını incelenmiştir. Puanlayıcı x birey etkileşimini anlayabilmek için t değerlerine bakılmıştır. Linacre (2023, s. 190), her bir puanlayıcı için ±2 aralığı dışında kalan t değerlerinin istatistiksel olarak anlamlı olduğunu ifade etmektedir. Bireysel istatistikler incelenmeden önce grup istatistikleri ($\chi^2(255)=680,6$, $p<0,001$) incelenmiştir ve bu değerin istatisiksel olarak anlamlı olduğu bulunmuştur. Bu durum grup düzeyinde puanlayıcılarda yanlılık davranışının olduğunun göstergesidir. Tablo 7, yanlılık davranışı gösteren puanlayıcılara ait ölçüm raporlarını ve t değerlerini sunmaktadır.

Tablo 7

Uygulama Öncesi Puanlayıcı/Kompozisyon Etkileşim Raporu

Puanlayıcı	**Birey/ Kompozisyon**	**Gözlenen Puan**	**Beklenen Puan**	**Yanlılık (logit)**	**Standart hata**	**t**
P30	Kompozisyon 3	46,00	76,82	-0,87	0,17	-5,01
P17	Kompozisyon 3	72,00	96,83	-0,75	0,17	-4,47
P31	Kompozisyon 1	89,00	107,85	-0,67	0,18	-3,79
P36	Kompozisyon 1	89,00	106,68	-0,62	0,18	-3,51
P32	Kompozisyon 1	112,00	123,66	-0,69	0,22	-3,20
P39	Kompozisyon 2	82,00	98,88	-0,53	0,17	-3,12

P40	Kompozisyon 5	75,00	92,17	-0,51	0,17	-3,04
P40	Kompozisyon 3	66,00	83,51	-0,50	0,17	-2,98
P34	Kompozisyon 1	106,00	118,39	-0,60	0,20	-2,97
P39	Kompozisyon 3	71,00	88,12	-0,50	0,17	-2,97
P8	Kompozisyon 1	108,00	119,77	-0,60	0,21	-2,91
P39	Kompozisyon 4	66,00	82,59	-0,47	0,17	-2,81
P26	Kompozisyon 4	75,00	90,24	-0,45	0,17	-2,67
P15	Kompozisyon 2	96,00	108,84	-0,48	0,18	-2,64
P22	Kompozisyon 4	70,00	84,88	-0,42	0,17	-2,55
P14	Kompozisyon 3	93,00	105,75	-0,45	0,18	-2,53
P33	Kompozisyon 1	107,00	117,46	-0,50	0,20	-2,47
P34	Kompozisyon 2	76,00	89,44	-0,40	0,17	-2,35
P7	Kompozisyon 3	89,00	101,15	-0,41	0,18	-2,30
P7	Kompozisyon 2	100,00	110,69	-0,43	0,19	-2,26
P35	Kompozisyon 4	94,00	105,07	-0,39	0,18	-2,18
P5	Kompozisyon 2	97,00	107,54	-0,39	0,18	-2,14
P35	Kompozisyon 3	70,00	82,42	-0,35	0,17	-2,11
P17	Kompozisyon 2	97,00	107,17	-0,38	0,18	-2,05
P25	Kompozisyon 1	111,00	119,31	-0,44	0,21	-2,05
P3	Kompozisyon 5	99,00	108,83	-0,38	0,19	-2,03
P38	Kompozisyon 2	106,00	114,89	-0,40	0,20	-2,01
P35	Kompozisyon 2	106,00	94,87	0,40	0,20	2,01
P32	Kompozisyon 5	128,00	119,91	0,66	0,33	2,02
P37	Kompozisyon 4	117,00	106,83	0,48	0,24	2,04
P30	Kompozisyon 5	101,00	89,04	0,40	0,19	2,09
P9	Kompozisyon 3	112,00	100,70	0,46	0,22	2,13
P20	Kompozisyon 2	122,00	111,97	0,57	0,27	2,14
P39	Kompozisyon 5	110,00	98,18	0,46	0,21	2,18
P3	Kompozisyon 2	120,00	109,40	0,55	0,25	2,19
P7	Kompozisyon 5	121,00	110,14	0,59	0,26	2,26
P24	Kompozisyon 3	102,00	88,12	0,46	0,19	2,41
P14	Kompozisyon 2	126,00	114,35	0,78	0,30	2,57
P12	Kompozisyon 5	121,00	107,51	0,70	0,26	2,70
P31	Kompozisyon 3	106,00	90,34	0,55	0,20	2,74
P23	Kompozisyon 5	121,00	107,14	0,71	0,26	2,76
P25	Kompozisyon 5	128,00	114,60	0,96	0,33	2,94
P39	Kompozisyon 1	121,00	105,34	0,79	0,26	3,04
P34	Kompozisyon 3	121,00	105,29	0,79	0,26	3,04
P17	Kompozisyon 1	127,00	112,58	0,96	0,31	3,05

P18	Kompozisyon 2	117,00	100,03	0,74	0,24	3,13
P17	Kompozisyon 4	111,00	92,84	0,68	0,21	3,20
P33	Kompozisyon 2	106,00	87,30	0,64	0,20	3,21
P22	Kompozisyon 5	118,00	100,10	0,80	0,24	3,29
P36	Kompozisyon 3	108,00	88,78	0,68	0,21	3,31
P38	Kompozisyon 3	108,00	87,68	0,71	0,21	3,47
P8	Kompozisyon 4	123,00	104,07	0,98	0,27	3,57
P34	Kompozisyon 1	118,00	96,84	0,91	0,24	3,76
P30	Kompozisyon 1	121,00	97,20	1,09	0,26	4,18
P15	Kompozisyon 1	132,00	114,01	1,51	0,39	3,82

Tablo 7 incelendiğinde toplam 55 yanlı puanlama tespit edilmiştir. 27 katı puanlama davranışı söz konusu iken 28 adet cömert puanlama davranışı tespit edilmiştir. En çok yanlı davranış gösteren puanlayıcının P39, P34, P17, P34 ve P30 olduğu görülmektedir. Aynı zamanda yüksek oranlarda hem katı ve hem cömert davranış sergileyen puanlayıcıların P 15, P17, P30 ve P34 olduğu anlaşılmaktadır. Öte yandan Öğrenci/Komposizyon 4 dengeli bir şekilde puanlanırken, Öğrenci/Kompozisyon 1, 2 ve 3'ün en çok katı puanlanan performans olduğu görülmektedir. Öğrenci/Komposizyon 5 ise en çok cömert puanlanan performanstır. Puanlayıcı sırasına göre yanlılık davranışları detaylı olarak EK 2'de sunulmaktadır.

İkinci Araştırma Sorusuna Yönelik Bulgular

Araştırma Sorusu 2: Puanlayıcı eğitimi sonrasında akademik yazma becerilerinin puanlanması sürecinde;

a. İngilizce öğretmen adaylarından oluşan puanlayıcı grubunun puanlamalarında öğrenci, ölçüt ve puanlayıcı yüzeylerinden elde edilen verilerin uygunluk değerleri ve model ölçümleri nasıldır?
b. İngilizce öğretmen adaylarından oluşan puanlayıcı grubunun puanlama davranışları nasıldır (katılık, cömertlik ve yanlılık)?

Puanlayıcı eğitimi sonrasında birey, kriter ve kompozisyon yüzeyine ait ölçümler tekrar raporlanmıştır. Tablo 8'de puanlayıcı öğretmen adaylarına ait yüzeyin ölçümleri yer almaktadır. Logit değerleri incelendiğinde ölçümlerin 2,55 ile 0,02 arasında değiştiği anlaşılmaktadır. En cömert puanlama davranışı 2 numaralı öğrenci tarafından gösterilirken (2,58 logit değeri); en katı puanlama yapan katılımcının 40 numaralı puanlayıcı olduğu

anlaşılmaktadır (0,02 logit değeri). Logit değer ortalaması 1,24 ve standart hata ortalaması 0,14'tür.

Uygunluk içi ve uygunluk dışı değerlerin 1 değeri çevresinde konumlanması verilerin modele uyum gösterdiğini sergilemektedir. Bu bağlamda en yüksek uygunluk içi değeri 1,96 iken en düşük uygunluk içi değeri 0.43'tür. Uygunluk dışı değerlerine bakıldığında en yüksek değerin 2,07 ve en düşük değerin 0,45 olduğu anlaşılmaktadır.

Tablo 8

Uygulama Sonrası Puanlayıcı Yüzeyine Ait Ölçüm Raporu

Puanlayıcı	Gözlenen ortalama	Düzelt. Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	3,89	3,97	1,07	0,13	0,84	0,80
2	4,66	4,72	2,55	0,20	1,10	0,79
3	4,21	4,30	1,55	0,14	0,97	0,98
4	3,54	3,59	0,63	0,12	1,06	1,03
5	3,71	3,78	0,84	0,13	0,93	0,94
6	4,64	4,70	2,47	0,19	0,90	0,63
7	4,21	4,30	1,55	0,14	0,84	0,80
8	4,26	4,35	1,64	0,15	1,00	1,25
9	4,47	4,55	2,05	0,17	1,09	1,27
10	3,81	3,89	0,97	0,13	0,54	0,53
11	3,46	3,51	0,54	0,12	1,75	2,07
12	4,29	4,37	1,68	0,15	0,65	0,64
13	4,16	4,25	1,47	0,14	0,67	0,65
14	4,39	4,47	1,87	0,16	1,11	0,93
15	4,04	4,12	1,28	0,14	0,99	0,98
16	3,94	4,02	1,14	0,13	1,32	1,60
17	4,15	4,24	1,45	0,14	1,32	1,34
18	4,10	4,19	1,37	0,14	0,68	0,76
19	3,08	3,09	0,09	0,12	0,51	0,52
20	4,44	4,52	1,97	0,16	1,02	0,95
21	3,71	3,78	0,84	0,13	0,87	0,84
22	3,66	3,75	0,81	0,13	1,96	1,99
23	4,18	4,26	1,49	0,14	0,43	0,45
24	3,95	4,03	1,16	0,13	1,07	1,18
25	4,35	4,43	1,80	0,15	1,54	1,91
26	3,89	3,97	1,07	0,13	0,97	0,94
27	3,75	3,82	0,89	0,13	1,09	1,10
28	4,18	4,26	1,49	0,14	0,96	1,00
29	4,22	4,31	1,57	0,14	0,94	1,06

30	3,73	3,79	0,86	0,13	0,96	0,96
31	4,26	4,35	1,64	0,15	1,28	1,07
32	4,30	4,39	1,70	0,15	1,31	1,27
33	3,41	3,46	0,48	0,12	0,93	0,90
34	3,54	3,59	0,63	0,12	0,69	0,68
35	4,31	4,40	1,73	0,15	1,66	1,49
36	3,88	3,95	1,05	0,13	0,72	0,79
37	4,59	4,65	2,33	0,18	0,63	0,81
38	3,67	3,74	0,79	0,12	0,55	0,70
39	3,65	3,71	0,76	0,12	1,31	1,31
40	3,01	3,02	0,02	0,12	0,57	0,58
41	3,76	3,83	0,91	0,13	1,13	1,10
42	4,46	4,54	2,03	0,17	0,90	1,05
43	4,16	4,25	1,47	0,14	0,88	0,89
44	3,96	4,04	1,17	0,13	1,69	1,67
45	3,29	3,32	0,33	0,12	0,84	0,83
46	3,89	3,97	1,07	0,13	0,83	0,84
47	3,31	3,35	0,36	0,12	1,22	1,19
48	4,08	4,16	1,34	0,14	0,86	0,91
49	3,90	3,98	1,09	0,13	1,27	1,26
50	3,55	3,61	0,64	0,12	0,82	0,88
51	4,18	4,26	1,49	0,14	0,74	0,75
Ortalama	3,97	4,04	1,24	0,14	1,00	1,02
Evren	0,39	0,40	0,57	0,02	0,33	0,36
Örneklem	0,39	0,40	0,58	0,02	0,33	0,36

Model Evren: RMSE ,14 Düzelt. S.H. ,56 Ayırma İnd. 3,96 Ayırma Katsayısı 5,61 Güvenirlik ,94
Model Örneklem: RMSE ,14 Düzelt. S.H. ,56 Ayırma İnd. 4,00 Ayırma Katsayısı 5,67 Güvenirlik,94
Model (Sabit etkili) ki-kare: 803,1 s.d.: 50 anlamlılık (olasılık): ,00
Model (Normal) ki-kare: 46,8 s.d..: 49 anlamlılık (olasılık): ,56

Uygulama sonrası birey yüzeyine ait ayırma katsayısı 5,61 iken ve güvenirlik ise 0,94 olarak belirlenmiştir. Uygulama öncesi ölçümüne göre daha yüksek ayırma katsayısına sahip olan bu ölçüm, bireylerin puanlayıcı eğitimi sonrası daha fazla ayrıştığına işaret etmektedir. Ki-kare değeri incelendiğinde ise bu değerin istatistiksel olarak anlamlı olduğu görülmektedir. ($\chi 2(sd)$ =803,1 (50), p= 0,00 < 0,01). Bu bulgu bireylerin puanlama davranışları arasında istatistiksel olarak anlamlı bir fark olduğunu işaret etmektedir.

Uygulama sonrasında ölçümleri incelenen bir diğer yüzey kompozisyon (öğrenci performansı) yüzeyidir. Tablo 9 dikkate alındığında, logit değerleri -0,47 ile 0,46 arasında değişen bu yüzeyde uygulama öncesine göre ölçüm ranjının küçüldüğü görülmektedir. Bu

bulguyla orantılı olarak ayırma katsayısında da düşüklük yaşandığı tespit edilmiştir.Bu durum öğrenci performanslarının daha benzer puanlandığını göstermektedir. Ayırma katsayısı 11,56 iken güvenirlik katsayısı ise 0,99'dur. Tüm uygunluk içi ve uygunluk dışı değerlerin 0,90 ile 1,16 arasında değiştiği görülmektedir. Uygunluk değerlerinin 0,5 ile 1,5 arasında olması verinin modele yüksek düzeyde uyum gösterdiğine işaret etmektedir (Linacre, 2002).

Tablo 9

Uygulama Sonrası Öğrenci/Komposizyon Yüzeyine Ait Ölçüm Raporu

Öğrenci/ Kompozisyon	Gözlenen ortalama	Düzelt. Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	4,27	4,38	0,46	0,05	1,13	1,16
2	3,70	3,79	-0,39	0,04	0,92	0,92
3	4,13	4,24	0,22	0,04	1,08	1,06
4	3,63	3,72	-0,47	0,04	0,99	0,98
5	4,10	4,21	0,17	0,04	0,90	0,96
Ortalama	3,97	4,07	0	0,04	1,00	1,02
Evren	0,25	0,26	0,36	0	0,09	0,08
Örneklem	0,28	0,30	0,41	0	0,10	0,09

Model Evren: RMSE ,04 Düzelt. S.H. ,36 Ayırma İnd. 8,42 Ayırma Katsayısı 11,56 Güvenirlik ,99
Model Örneklem: RMSE ,04 Düzelt. S.H. ,41 Ayırma İnd. 9,43 Ayırma Katsayısı 12,91 Güvenirlik ,99
Model (Sabit etkili) ki-kare: 368,1 s.d.: 4 anlamlılık (olasılık): ,00
Model (Normal) ki-kare: 4,00 s.d..: 3 anlamlılık (olasılık): ,27

Tblo 9'da yer alan anlamlılık değerleri incelendiğinde, ki-kare değerleri ($\chi 2(sd)$ =368,1 (4), p= 0,00 < 0,01) bu yüzeyde kompozisyonun başarılı olma düzeyleri arasında istatistiksel olarak anlamlı bir farkın olduğunu açıklamaktadır. Diğer bir ifadeyle, kompozisyonlar uygulama sonrası yapılan puanlamada da başarı düzeyi açısından anlamlı düzeyde farklılaşmaktadır. Yapılan puanlamalar Öğrenci/Kompozisyon 2 dışında seçim kriterlerine uygunluk göstermektedir.

Uygulama sonrasında kriter yüzeyi incelendiğinde, 51 akranın kriterlere verdiği puanlara ait logit değerleri Tablo 10'da sunulmuştur. Logit değerlerinin -0,97 ile 1,50 arasında değiştiği tespit edilmektedir. En düşük logit değerine sahip kriter, "Tez Cümlesi", uygulama öncesi ve sonrasında değişlik göstermemiştir. Benzer şekilde "Yazım Kuralları/İmla", en cömert

puanlanan, en başarılı bulunan kriter olmuştur. Uygunluk içi değerleri 0,63 ile 1,77 arasında değişen yüzeyde, uygunluk dışı değerler 0,64 ile 1,79 arasındadır. Değerlerin 1'den uzaklaşmaması verilerin, modele uyum sağladığını işaret etmektedir. Kriter yüzeyine ait ayırma katsayısı 10,42 ve güvenirlik indeksi ise 0,98 olarak belirlenmiştir. Bu kompozisyonların güvenilir şekilde birbirinden ayrıldığı anlamına gelmektedir.

Tablo 10

Uygulama Sonrası Kriter Yüzeyine Ait Ölçüm Raporu

Öğrenci/ Kompozisyon	Gözlenen ortalama	Düzelt. Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1. Başlık	3,80	3,87	-0,29	0,07	1,10	1,06
2. Organizasyon	3,69	3,75	-0,43	0,07	0,79	0,78
3. Tez Cümlesi	3,24	3,27	-0,97	0,07	1,35	1,34
4. Konu Cümlesi	3,52	3,56	-0,64	0,07	1,05	1,08
5. Destekleyici Cümleler	3,78	3,84	-0,32	0,07	0,63	0,64
6. Uygun Uzunluk	4,29	4,37	0,44	0,08	1,77	1,79
7. Konuya Uygunluk	4,21	4,29	0,30	0,08	1,01	1,01
8. Fikir Gelişimi	3,88	3,95	-0,19	0,07	0,81	0,81
9. Tutarlık	4,18	4,26	0,25	0,08	0,72	0,72
10. Bağlantı Unsurları	3,82	3,88	-0,28	0,07	0,89	0,88
11. Dil Bilgisi	4,36	4,44	0,56	0,09	1,08	1,08
12. Sözdizimsel Karmaşıklık	3,76	3,82	-0,35	0,07	0,81	0,84
13. Kelime Seçimi	3,69	3,75	-0,43	0,07	1,06	1,10
14. Kelime Çeşitliliği	3,96	4,04	-0,08	0,07	0,83	0,87
15. Yazım Kuralları/İmla	4,72	4,77	1,50	0,12	1,04	0,95
16. Noktalama	4,53	4,60	0,93	0,10	1,26	1,29
Ortalama	3,97	4,03	0	0,08	1,01	1,02
Evren	0,37	0,38	0,61	0,01	0,27	0,27
Örneklem	0,39	0,10	0,63	0,01	0,28	0,28

Model Evren: RMSE ,08 Düzelt. S.H. ,60 Ayırma İnd. 7,56 Ayırma Katsayısı 10,42 Güvenirlik ,98
Model Örneklem: RMSE ,08 Düzelt. S.H. ,62 Ayırma İnd. 7,86 Ayırma Katsayısı 10,76 Güvenirlik ,98
Model (Sabit etkili) ki-kare: 735,2 s.d.: 15 anlamlılık (olasılık): ,00
Model (Normal) ki-kare: 14,6 s.d..: 14 anlamlılık (olasılık): ,40

Öte yandan, ayırma katsayısı uygulama öncesine göre düşüklük göstermiştir. Bu durum puanlayıcıların, uygulama sonrasında kriterler için daha benzer puanlamalar yaptığına kanıt

olarak sunulabilir. Tablo 10'da yer verilen, ki-kare değerleri incelendiğinde ($\chi 2(sd)$ =735,2 (15), p= 0,00 < 0,01) ile bu yüzeyde kriterlerin zorluğu ve kolaylığı açısından istatiksel olarak anlamlı bir farkın olduğu anlaşılmaktadır.

Uygulama sonrası ait son olarak yanlılık analizleri yapılmıştır. Öncelikle grup istatistikleri incelenmiş ve ki-kare değerinin ($\chi 2(255)=692,3$, p<.001) istatiksel olarak anlamlı olduğu tespit edilmiştir. Bu değer puanlamada yanlı davranış olduğunu göstermektedir. Tablo 11 incelendiğinde t değerinin +2 ve -2 aralığı dışında yer alan puanlayacıların olduğu görülmektedir ve böylece uygulama sonrası yanlı davranışların olduğuna karar verilmiştir (Eckes, 2019).

Tablo 11

Uygulama Sonrasında Puanlayıcı/Kompozisyon Etkileşim Raporu

Puanlayıcı	Birey/ Kompozisyon	Gözlenen Puan	Beklenen Puan	Yanlılık (logit)	Standart hata	t
P17	Kompozisyon 3	53,00	73,26	-1,61	0,26	-6,21
P26	Kompozisyon 4	39,00	60,08	-1,45	0,27	-5,33
P39	Kompozisyon 3	47,00	65,51	-1,30	0,26	-5,02
P31	Kompozisyon 1	54,00	69,93	-1,21	0,26	-4,65
P40	Kompozisyon 3	48,00	63,60	-1,08	0,26	-4,19
P39	Kompozisyon 2	47,00	61,51	-0,99	0,26	-3,84
P34	Kompozisyon 1	61,00	72,93	-1,02	0,27	-3,77
P32	Kompozisyon 3	65,00	75,55	-1,02	0,28	-3,64
P39	Kompozisyon 4	47,00	60,31	-0,91	0,26	-3,50
P33	Kompozisyon 1	66,00	75,46	-0,93	0,28	-3,27
P22	Kompozisyon 3	54,00	65,70	-0,85	0,26	-3,26
P32	Kompozisyon 1	72,00	79,29	-0,99	0,32	-3,09
P31	Kompozisyon 4	57,00	67,54	-0,79	0,26	-3,02
P30	Kompozisyon 3	56,00	66,46	-0,77	0,26	-2,95
P20	Kompozisyon 1	72,00	78,87	-0,91	0,32	-2,84
P38	Kompozisyon 2	55,00	64,98	-0,72	0,26	-2,77
P36	Kompozisyon 1	63,00	71,61	-0,73	0,27	-2,67
P35	Kompozisyon 4	60,00	68,81	-0,69	0,27	-2,58
P40	Kompozisyon 5	45,00	54,08	-0,61	0,26	-2,34
P27	Kompozisyon 3	59,00	67,02	-0,61	0,27	-2,29
P7	Kompozisyon 2	60,00	67,84	-0,61	0,27	-2,27

P8	Kompozisyon 2	61,00	68,56	-0,60	0,27	-2,22
P3	Kompozisyon 5	67,00	73,63	-0,63	0,29	-2,18
P33	Kompozisyon 1	57,00	64,63	-0,56	0,26	-2,12
P24	Kompozisyon 1	66,00	72,60	-0,60	0,28	-2,11
P18	Kompozisyon 3	66,00	72,37	-0,58	0,28	-2,03
P14	Kompozisyon 2	78,00	71,26	0,83	0,40	2,05
P11	Kompozisyon 1	73,00	65,54	0,68	0,33	2,08
P31	Kompozisyon 1	83,00	76,68	1,54	0,72	2,15
P39	Kompozisyon 4	64,00	55,53	0,61	0,28	2,19
P34	Kompozisyon 3	78,00	70,55	0,90	0,40	2,22
P31	Kompozisyon 3	82,00	74,85	1,36	0,59	2,30
P24	Kompozisyon 4	70,00	61,24	0,71	0,30	2,33
P17	Kompozisyon 1	83,00	75,31	1,73	0,72	2,41
P18	Kompozisyon 2	74,00	65,22	0,82	0,34	2,42
P32	Kompozisyon 4	77,00	68,55	0,93	0,38	2,42
P31	Kompozisyon 4	67,00	57,55	0,71	0,29	2,46
P8	Kompozisyon 3	83,00	74,85	1,78	0,72	2,48
P7	Kompozisyon 5	82,00	73,81	1,48	0,59	2,51
P15	Kompozisyon 1	84,00	73,89	2,60	1,01	2,59
P33	Kompozisyon 2	63,00	52,26	0,75	0,27	2,74
P38	Kompozisyon 3	82,00	72,19	1,66	0,59	2,80
P39	Kompozisyon 1	83,00	71,94	2,11	0,72	2,93
P39	Kompozisyon 3	80,00	69,45	1,38	0,47	2,94
P27	Kompozisyon 2	70,00	58,58	0,90	0,30	2,96
P26	Kompozisyon 3	80,00	69,26	1,39	0,47	2,98
P30	Kompozisyon 5	78,00	65,73	1,32	0,40	3,26
P31	Kompozisyon 3	80,00	67,21	1,57	0,47	3,36
P34	Kompozisyon 1	80,00	66,62	1,62	0,47	3,46
P17	Kompozisyon 4	81,00	65,31	1,97	0,52	3,80

Uygulama sonrası yanlı puanlama davranışı 50'ye düşmüştür. 26 katı puanlama davranışı varken 24 cömert puanlama davranışı tespit edilmiştir. Özellikle cömert puanlama davranışında düşüş olduğu görülmektedir. En katı puanlama yapan puanlayıcılar P17, P26, P39, P31ve P40 olarak görülürken iken en cömert yanlılık gösterenlerin yine P17, P26 ve P31, ek olarak P34 ve P30 olduğu anlaşılmıştır. En fazla yanlılıkta bulunan kişi P31 ve P39'dur. Puanlayıcıların, özellikle P34, P30, P17 ve P39'un puanlama öncesinde yanlı

davranışlarda bulunduğu gözlemlenmektedir. Bu durum puanlama eğitimi dışında puanlayıcı yanlılığına neden olabilecek farklı değişkenlerin olabileceğine işaret etmektedir.

Üçüncü Araştırma Sorusuna Yönelik Bulgular

Araştırma Sorusu 3: Puanlayıcı eğitimi kapsamında yapılan uygulama dikkate alındığında;

a. İngilizce öğretmen adaylarından oluşan puanlayıcı grubunun puanlamalarında öğrenci, ölçüt, zaman ve puanlayıcı yüzeylerinden elde edilen verilerin uygunluk değerleri ve model ölçümleri nasıldır?
b. İngilizce öğretmen adaylarından oluşan puanlayıcı grubunun puanlama davranışları nasıldır (katılık, cömertlik)?
c. İngilizce öğretmen adaylarından oluşan puanlayıcı grubunun puanlama davranışlarında ön ve son uygulama arasında herhangi bir farklılık var mıdır?

Mevcut çalışmada üçüncü araştırma sorusuna cevap verebilmek için analizlere zaman yüzeyi eklenmiş ve ölçümler uygulama kapsamını ifade edecek şekilde yorumlanmıştır. Tablo 12 detaylı olarak incelendiğinde puanlayıcı yüzeyine ait ölçüm değerleri görülmektedir.

Tablo 12

Uygulama Kapsamında Puanlayıcı Yüzeyine Ait Ölçüm Raporu

Puanlayıcı	Gözlenen ortalama	Düzelt. Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	3,93	4,01	0,95	0,09	0,83	0,78
2	4,56	4,62	1,96	0,12	1,03	0,93
3	4,11	4,18	1,17	0,09	0,77	0,74
4	3,73	3,80	0,72	0,08	0,88	0,87
5	4,04	4,11	1,08	0,09	0,96	0,95
6	4,49	4,55	1,79	0,11	0,84	0,75
7	4,14	4,22	1,22	0,09	0,91	0,90
8	4,31	4,38	1,47	0,10	1,25	1,35
9	4,13	4,21	1,21	0,09	1,29	1,50
10	3,90	3,97	0,91	0,09	0,65	0,65
11	3,89	3,97	0,91	0,09	1,50	1,81
12	4,05	4,13	1,10	0,09	0,97	0,93
13	3,89	3,97	0,91	0,09	0,88	0,91
14	4,27	4,34	1,41	0,10	0,93	0,96
15	4,08	4,16	1,14	0,09	1,03	0,97

16	4,11	4,18	1,17	0,09	0,98	1,23
17	4,03	4,10	1,07	0,09	0,81	0,79
18	3,81	3,88	0,81	0,08	0,83	1,08
19	3,41	3,46	0,39	0,08	0,75	0,79
20	4,19	4,26	1,28	0,09	1,15	1,16
21	3,72	3,79	0,71	0,08	0,90	0,88
22	3,84	3,92	0,85	0,08	1,42	1,43
23	4,05	4,13	1,10	0,09	0,84	0,78
24	3,80	3,87	0,80	0,08	0,94	0,95
25	4,29	4,36	1,44	0,10	1,28	1,44
26	3,96	4,04	0,99	0,09	0,95	0,94
27	4,09	4,17	1,16	0,09	0,92	0,90
28	4,19	4,26	1,28	0,09	0,89	0,93
29	3,97	4,04	1,00	0,09	1,22	1,26
30	3,51	3,56	0,49	0,08	1,18	1,18
31	4,06	4,13	1,11	0,09	1,49	1,43
32	4,29	4,37	1,45	0,10	1,07	0,97
33	3,42	3,47	0,40	0,08	0,94	0,95
34	3,49	3,54	0,47	0,08	0,55	0,55
35	4,33	4,40	1,50	0,10	1,72	1,78
36	3,81	3,88	0,81	0,08	0,77	0,85
37	4,36	4,43	1,56	0,10	1,02	1,02
38	3,80	3,87	0,80	0,08	0,79	0,90
39	3,74	3,81	0,74	0,08	0,73	0,75
40	3,59	3,65	0,57	0,08	0,90	0,89
41	3,86	3,93	0,86	0,08	0,94	0,91
42	4,46	4,52	1,73	0,11	0,80	0,86
43	4,21	4,29	1,32	0,10	1,00	0,98
44	4,24	4,31	1,36	0,10	1,45	1,51
45	3,55	3,61	0,53	0,08	1,34	1,39
46	4,03	1,10	1,07	0,09	0,80	0,80
47	3,40	3,45	0,38	0,08	1,19	1,17
48	4,28	4,35	1,42	0,10	1,06	1,11
49	3,78	3,85	0,77	0,08	1,14	1,09
50	3,69	3,75	0,68	0,08	0,76	0,76
51	4,31	4,38	1,47	0,10	0,84	0,84
Ortalama	3,98	4,05	1,05	0,09	1.00	1,02
Evren	0,29	0,29	0,37	0,01	0,24	0,27
Örneklem	0,29	0,30	0,37	0,01	0,24	0,27

Model Evren: RMSE ,09 Düzelt. S.H. ,36 Ayırma İnd. 3,95 Ayırma Katsayısı 5,60 Güvenirlik ,94
Model Örneklem: RMSE ,09 Düzelt. S.H. ,36 Ayırma İnd. 3,99 Ayırma Katsayısı 5,66 Güvenirlik ,94
Model (Sabit etkili) ki-kare: 811,7 s.d.: 50 anlamlılık (olasılık): ,00
Model (Normal) ki-kare: 46,9 s.d..: 49 anlamlılık (olasılık): .,54

51 puanlayıcının logit değerlerinin 0,38 ile 1,96 logit arasında değiştiği anlaşılmaktadır. Logit değeri ortalaması 1,05'tir. Uygunluk değerlerine bakıldığında ise uygunluk içi değerlerin 0,55 ile 1,72 arasında değiştiği anlaşılmaktadır. Uygunluk dışı değerleri ise 0,65 ile 1,81 arasında değişmektedir. Bu ölçümler, verinin modele uyum gösterdiğini kanıtlamaktadır (Linacre, 2002). Puanlayıcı yüzeyine ait ayırma katsayısı 5,60, güvenirlik katsayısı ile 0,94 olarak belirlenmiştir. Ki-kare değerleri ($\chi 2(sd)$ =811,7 (50), p= 0,00< 0,01) incelendiğinde ise bu yüzeyde 51 akranın puanlama davranışları arasında istatiksel açıdan anlamlı bir farkın olduğu anlaşılmaktadır.

Diğer bir yüzeyde uygulama bütününde Öğrenci/Kompozisyon yüzeyi ölçümleri incelendinmiştir. Tablo 13'te yer alan logit değerlerin -0,24 ike 0,37 arasında değiştiği görülmektedir.

Tablo 13

Uygulama Kapsamında Öğrenci/Komposizyon Yüzeyine Ait Ölçüm Raporu

Öğrenci/ Kompozisyon	Gözlenen ortalama	Düzelt. Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	4,26	4,35	0,37	0,03	1,04	1,09
2	4,09	4,18	0,13	0,03	1,12	1,12
3	3,80	3,88	-0,24	0,03	1,10	1,11
4	3,68	3,76	-0,37	0,03	0,88	0,91
5	4,08	4,17	0,11	0,03	0,84	0,88
Ortalama	3,98	4,07	0	0,03	1,00	1,02
Evren	0,21	0,22	0,27	0	0,11	0,10
Örneklem	0,24	0,24	0,30	0	0,13	0,12

Model Evren: RMSE ,03 Düzelt. S.H. ,27 Ayırma İnd. 9,48 Ayırma Katsayısı 12,97 Güvenirlik ,99
Model Örneklem: RMSE ,03 Düzelt. S.H. ,06 Ayırma İnd. 10,61 Ayırma Katsayısı 14,47 Güvenirlik ,99
Model (Sabit etkili) ki-kare: 458,00 s.d.: 4 anlamlılık (olasılık): ,00
Model (Normal) ki-kare: 4 s.d..: 3 anlamlılık (olasılık): ,26

Logit değerlerin ortalaması 0'a eşittir. En başarısız görülen kompozisyon 4 ve 3 iken, puanlayıcılar 1 numaralı kompozisyonu oldukça başarılı bulmuşlardır. Uygunluk değerleri incelendiğinde ise uygunluk içi değerlerinin 0,84 ile 1,12 arasında; uygunluk dışı değerlerinin ise 0,88 ile 1,12 arasında değişti anlaşılmaktadır. Diğer yüzey ölçümlerine benzer şekilde, bu değerlerin de 1 çevresinde konumlanması verinin modele iyi uyum gösterdiğine işaret etmektedir. Öğrenci perfomansının başarı yeterlik düzeyine göre

incelenmesi sağlandığında 3 numaralı performans dışındakilerin seçim kriterleri ile uyuştuğu anlaşılmaktadır. Ayırma katsayısının 12,97 ve güvenirlik katsayısının 0,99 olduğu bu düzeyde ki-kare değerleri ($\chi 2(sd)$ =458,00 (4), p= 0,00 < 0,01) kompozisyonların başarı düzeyleri arasında istatiksel olarak anlamlı bir farkın olduğunu göstermektedir.

Tablo 14'te yer verilen Kriter/Ölçüt yüzeyi ölçümlerine ait logit değerlerin -0,78 ile 1,24 arasında değer aldığı anlaşılmaktadır. Bu durum, öğrenci performansların bazı kriterlerde diğerlerine göre daha az başarılı bulunduğuna işaret etmektedir.

Tablo 14

Uygulama Kapsamında Kriter Yüzeyine Ait Ölçüm Raporu

Öğrenci/ Kompozisyon	Gözlenen Ortalama	Düzelt. Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1. Başlık	3,85	3,89	-0,23	0,05	1,20	1,17
2. Organizasyon	3,74	3,78	-0,35	0,05	0,80	0,81
3. Tez Cümlesi	3,30	3,32	-0,78	0,04	1,35	1,34
4. Konu Cümlesi	3,56	3,60	-0,53	0,04	1,12	1,12
5. Destekleyici Cümleler	3,83	3,88	-0,24	0,05	0,72	0,73
6. Uygun Uzunluk	4,27	4,32	0,33	0,06	1,32	1,34
7. Konuya Uygunluk	4,28	4,33	0,34	0,06	0,99	0,96
8. Fikir Gelişimi	3,96	4,01	-0,10	0,05	0,82	0,84
9. Tutarlık	4,22	4,27	0,25	0,05	0,79	0,80
10. Bağlantı Unsurları	3,81	3,85	-0,27	0,05	0,86	0,87
11. Dil Bilgisi	4,35	4,39	0,44	0,06	1,04	1,09
12. Sözdizimsel Karmaşıklık	3,75	3,79	-0,34	0,05	0,84	0,88
13. Kelime Seçimi	3,65	3,69	-0,43	0,05	0,94	0,97
14. Kelime Çeşitliliği	3,97	4,02	-0,09	0,05	0,85	0,91
15. Yazım Kuralları/İmla	4,70	4,73	1,24	0,08	1,24	1,24
16. Noktalama	4,51	4,55	0,74	0,06	1,29	1,33
Ortalama	3,98	4,03	0	0,05	1.01	1,02
Evren	0,36	0,36	0,51	0,01	0,21	0,20
Örneklem	0,37	0,38	0,52	0,01	0,22	0,21

Model Evren: RMSE ,05 Düzelt. S.H. ,50 Ayırma İnd. 9,56 Ayırma Katsayısı 13,08 Güvenirlik ,99
Model Örneklem: RMSE ,05 Düzelt. S.H. ,52 Ayırma İnd. 9,88 Ayırma Katsayısı 13,51 Güvenirlik ,99
Model (Sabit etkili) ki-kare: 1182,7 s.d.: 15 anlamlılık (olasılık): ,00
Model (Normal) ki-kare: 14,8 s.d..: 14 anlamlılık (olasılık): ,39

Hem ön hem de son puanlamada olduğu gibi en az başarısız bulunan “Tez Cümlesi” genel anlamda da en az puan verilen ölçüt olmuştur. Bunu takiben “Konu Cümlesi”, "Kelime Seçimi” ve “Organizasyon” en az başarılı bulunan kriterler olmuştur. Diğer taraftan “Yazım Kuralları/İmla” ve “Noktalama” en cömert puanlanan, en başarılı bulunan ölçütler arasında değerlendirilmiştir. Uygunluk değerleri, verilerin model ile iyi uyum gösterdiğini kanıtlamaktadır. Daha spesifik bir bakış açısı ile uygunluk içi değerleri 0,72 ile 1,35 arasında değer alırken; uygunluk dışı değerleri 0,73 ile 1,34 arasında değişiklik göstermektedir. Ayırma katsayısı 13,08ve güvenirlik katsayısı 0,99’dur. Bu değerlerin kriterlerin başarılı bir şekilde birbirinden ayrıştığı anlamına gelmektedir. Bu durumu kanıtlar şekilde ki-kare değeri ($\chi 2(sd)$ =1182,7 (15), p= 0,00< 0,01) bu farklılığı istatitiksel olarak doğrulamaktadır.

Uygulama öncesi ve sonrasında yapılan ölçümlerden farklı olarak, uygulama bütününde ele alınan bir diğer yüzey zaman yüzeyidir. Tablo 15, zaman yüzeyi kapsamında ön puanlama ve son puanlamaya dair ölçüm değerlerini sergilemektedir. Logit değerleri sırasıyla 0,02 ve -0,02’dir. Uygunluk değerlerinin 1’e çok yakın olmasının verinin modele uyum gösterdiğini ifade etmektedir. Ancak logit değerlerden anlaşıldığı üzere ön ve son uygulama arasında dikkate değer bir değişimin olmadığı görülmektedir.

Tablo 15

Uygulama Kapsamında Zaman Yüzeyine Ait Ölçüm Raporu

Uygulama	Gözlenen ortalama	Düzelt. Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
ÖN	4,00	4,10	0,02	0,02	1,02	1,05
SON	3,97	4,07	-0,02	0,02	0,97	1,00
Ortalama	3,98	4,09	0,00	0,02	0,99	1,02
Evren	0,02	0,02	0,02	0,00	0,03	0,03
Örneklem	0,02	0,03	0,03	0,00	0,04	0,04

Model Evren: RMSE ,02 Düzelt. S.H. ,01 Ayırma İnd. ,77 Ayırma katsayısı 1,36 Güvenirlik ,37
Model Örneklem: RMSE ,02 Düzelt. S.H. ,03 Ayırma İnd. 1,48 Ayırma katsayısı 2,30 Güvenirlik ,69
Model (Sabit etkili) ki-kare: 3,2 s.d.: 1 anlamlılık (olasılık): ,07

Ayırma katsayısın düşük olması (0,77) ve güvenirlik katsayısının (0,37) düşük çıkması da bu bulguyu desteklemektedir. Benzer şekilde, ön ve son uygulama puanlamaları arasında

istaitksel olarak bir farklılığın olmadığı da ki-kare değerlerinden de ($\chi 2(sd)$ =3,2 (1), p= 0,07 > 0,01) anlaşılmaktadır.

ÇYRM kullanılarak ön ve son uygulama puanları arasında herhangi bir istatiksel farklılığım olmadığı görülmüştür. İki uygulama arasında farkın olup olmadığına dair farklı bir kanıt sunmak için eşleştirilmiş örneklem t testi (paired sample t test) yapılmıştır.

Tablo 16

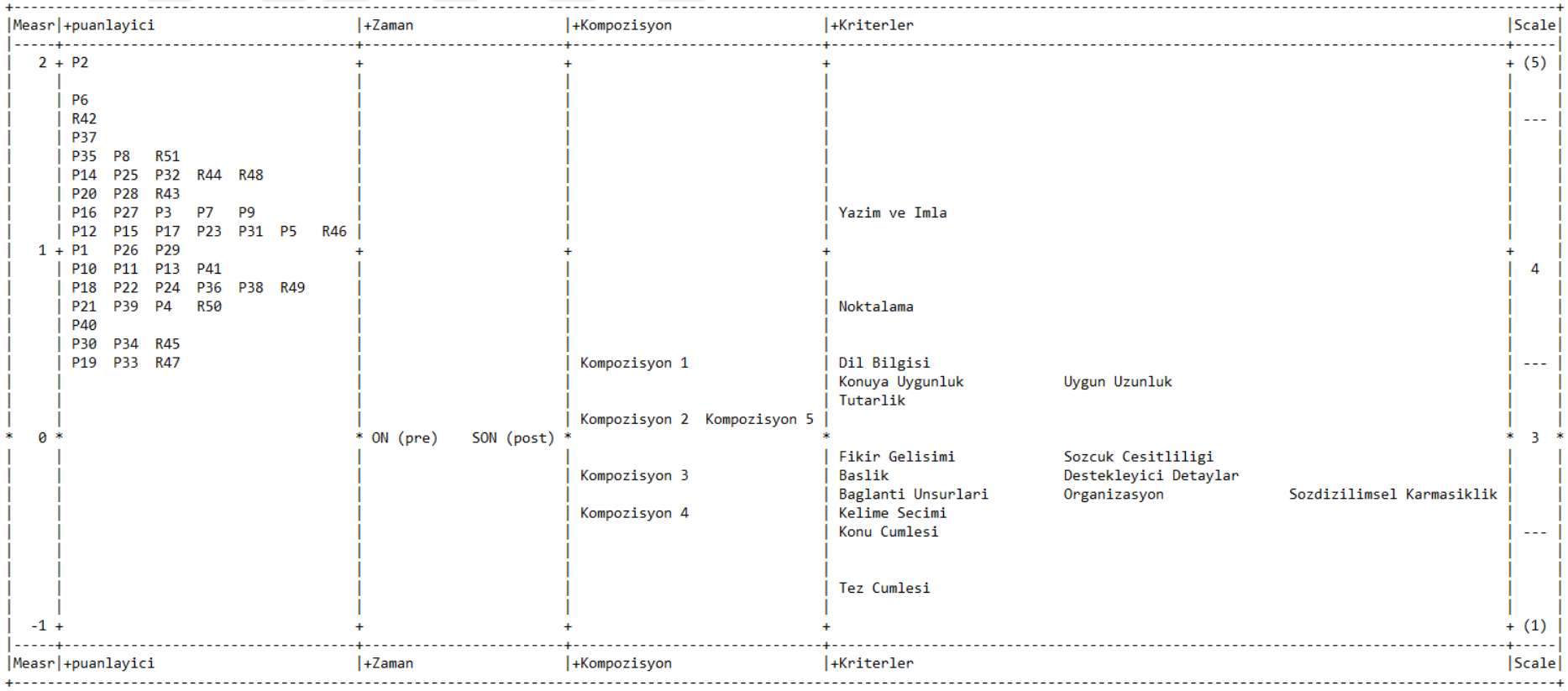
Ön ve Son Puanlama Logit Değerleri Eşleştirilmiş Örneklem T Testi Sonuçları

Uygulama	**N**	$\bar{X}$	S_x	**sd**	**t**	**p**
ÖN	51	1,19	0,45	50	-0,67	0,51
SON	51	1,24	0,58			

p>0.05 olması sebebi ile istatiksel olarak anlamlı bir farlılık bulunmamıştır.

Tablo 16'da görüldüğü üzere ön uygulama logit değerleri ve son uygulama değerleri logit değerleri arasında anlamlı olarak bir fark tespit edilememiştir ($t(50)$=-0,67, p=0,51>0,05). İki uygulama arasında herhangi bir farkın olmayışı puanlayıcı eğitimi, puanlama koşulları ve puanlayıcıların özellikleri kapsamında yorumlanacaktır.

Ölçüm raporlarına ek olarak Şekil 3'te zaman yüzeyinin de ele alındığı uygulama kapsamının logit cetvel görseli sunulmuştur. Şekil 3'te de görüldüğü üzere uygulama kapsamında en cömert davranan akran 2 numaralı puanlayıcıdır. Sonrasında 6 ve 42 numaralı puanlayıcılar da cömert puanlama davranışında bulunmuştur. Öte taraftan, 19, 30, 34, 33, 47 ve 45 numaralı puanlayıcılar daha katı puanlama yapmıştır. Uygulama kapsamında ön ve son puanlama arasında logit cetvelinde herhangi bir farklılık görülmemektedir. Kompozisyon yüzeyine bakıldığında en başarılı bulunan kompozisyonun 1 numaralı kompozisyon olduğu anlaşılmaktadır. Kompozisyon 1, seçme kriterlerinin arasında "İleri Düzey" etiketi ile araştırmaya dahil edilmiştir. Öte yandan, Kompozisyon 4'ün puanlayıcılar tarafından kriterleri yeterli karşılamadığı düşünülmüştür. Bu bulgu da seçme kriterini yansıtmaktadır. Kriter yüzeyine gelince en başarılı bulunan kriter yazım kuralları/imla iken en düşük puanları tez cümlesi olmuştur.



Şekil 3. Uygulama kapsamında elde edilen logit cetvel

BÖLÜM V

TARTIŞMA

Bu araştırmada yabancı dil olarak öğretilen İngilizcede yazma becerilerinin ölçülmesinde ve değerlendirilmesinde akran değerlendirme kapsamında puanlayıcı davranışları incelenmiştir. Bölüm IV'te elde edilen temel bulgular araştırma soruları başlığı altında sunulmuştur. Bu bölümde de aynı başlıklarla bulgular özetlenecek ve ilgili alan yazında yapılan benzer çalışmalar ile mevcut bulguların benzerlik ve farklılıklarına değinilecektir.

Birinci Araştırma Sorusuna Yönelik Bulgular ve Tartışma

Birinci araştırma sorununda puanlayıcı eğitimi vermeden önce katılımcılardan beş farklı akran kompozisyonu puanlandırmaları istenmiştir. ÇYRM sonucunda model veri uyumu, ölçüm raporları ve puanlayıcı davranışları incelenmiştir. Puanlayıcı, kompozisyon ve ölçüt yüzeylerinde model veri uyumunun sağlandığı anlaşılmaktadır. Bu bulgu, akademik yazma performansının ÇYRÖM ile uygun bir şekilde analiz edilebileceğine işaret etmektedir. Alan yazında da ÇYRM'in performans temelli değerlendirmelerde analitik incelemelere olanak sağladığı ve geçerli çıktılar verdiği belirtilmiştir (Sudweeks vd., 2004; Polat, 2022). Tüm yüzeylerde ayırma katsayısı güvenilir bir şekilde birimlerin birbirinden ayrıldığına işaret etmektedir. Yani puanlayıcıların puanlama davranışı, kompozisyonların yeterlik düzeyi ve ölçütlerin farklı performans boyutları arasında farklılaşma söz konudur ve bu farklılık istatiksel olarak anlamlı bulunmaktadır. Engelherd (1994) ve Eckes (2005) tarafından yapılan çalışmalarda da bu farklılığa dikkat çekilmiştir. Benzer şekilde, Demir (2023) ve Farrokhi (2012) puanlayıcı farklılıklarına dair bulgular sunmuştur. Uygulama öncesinde en cömert davranan puanlayıcılar 2, 44 ve 48 numaralı katılımcılar iken, en katı davranan puanlayıcılar 30, 33 ve 34 numaralı akranlardır.

Katı ve cömert puanlama yapan bazı akranların yanlı davrandığı tespit edilmiştir. Ön uygulamada toplamda 55 adet yanlılık gösteren davranış bulunmuştur. Yanlılık davranışı, özellikle öz, akran ve öğretmen değerlendirmelerinde odak noktası olmuştur (Aryadoust, 2015; Schaefer, 2008; Trace vd., 2017) ve yanlılık davranışını azaltmak için farklı değişkenler incelenmiştir. Bu çalışmada da puanlayıcı eğitiminin yanlılık davranışlarına etkisine bakılmak istenmiştir. Özellikle bazı akranların, örneğin P17, P30 ve P34, hem çok cömert hem de çok katı puanlama yaptıkları anlaşılmıştır. Bu durum ölçme sürecinden veya puanlayıcı özelliklerinden kaynaklı bazı hatalara işaret etmektedir. Bir kompozisyonu puanlandırmak nispeten uzun zaman alacağı için beş farklı kompozisyon puanlama sürecinde yorgunluk etkisi gözlenebilir. Ya da puanlayıcıların geçmiş yaşantıları veya kişilik özellikleri de puanlama davranışlarını değiştirebilir.

Performans boyutlarına/ölçütlerine bakıldığında yazım kuralları/imla ve noktalamanın en başarılı bulunan, tez cümlesi ve kelime cümlesinin en başarısız bulunan ölçüt olduğu anlaşılmaktadır. Kompozisyon yüzeyi incelendiğinde beş farklı eserin yetenek düzeyi olarak birbirinden farklılaşacak şekilde puanlandığı anlaşılmaktadır. Şahin ve diğerleri (2016) ÇYRM'in öğrenci performansını ayırt etmede önemli bir analiz olduğunu ifade etmiştir. Mevcut çalışmada da performans düzeyleri arasındaki fark puanlayıcılar tarafından fark edilmiş ve bu da analiz ile detaylıca açıklanmıştır.

İkinci Araştırma Sorusuna Yönelik Bulgular ve Tartışma

Altı hafta süren puanlayıcı eğitimi sonrasında katılımcılardan aynı 5 kompozisyonu tekrar puanlandırmaları istenmiştir. Tablo 8, 9 ve 10'daki ölçüm raporları incelendiğine veri setinin bazı istisnalar hariç ÇYRM analize uygun olduğu tespit edilmiştir. Rentiana ve diğerleri (2024) yazma becerilerinin ölçülmesinde Rasch modelinin önemli bir yeri olduğunun altını çizmiştir. İlk uygulamadan farklı olarak puanlayıcılar arasında ayırma katsayısında artış gözlenmiştir. Bu durumda puanlama davranışlarının ilk uygulamaya göre biraz daha farklılaştığı anlaşılmaktadır. Puanlayıcı 2, 6 ve 37'nin en cömert puanlar veren akranlar, 19 ve 40 numaralı puanlayıcının en katı puanlar veren katılımcılar olduğu anlaşılmıştır. Bu puanlayıcılar arasında 40 numaralı katılımcının yanlılık gösterdiği tespit edilmiştir. İlk uygulamaya nazaran yanlılık davranışında azalma söz konusu olmuştur. Ravand ve diğerleri

de (2016) eğitim ve pratik ile zaman içerisinde puanlayıcıların yanlılık davranışlarında azalma olduğunun altını çizmiştir.

İkinci puanlamada 50 adet yanlılık gösteren etkileşim tespit edilmiştir. Genel anlamda cömert puanlama davranışında düşüş gözlemlenmiştir. Ancak katılık ve cömertlik davranışlarında büyük değişimler yaşanmamıştır. Bu bulgu Trace ve diğerleri (2017) ile doğru orantılıdır. Eckes (2008) de puanlayıcı eğitiminin değişkenliği azaltmak açısından diğer faktörlere göre daha az etkisi olduğunu savunmuştur. Diğer taraftan, Tursynbayeva ve diğerleri (2023) tarafından yapılan çalışmada puanlayıcı eğitimi sonrası katılık ve cömertlik davranışlarında değişimler oldupu rapor edilmiştir. Uygulama sonrasında seçim kriterlerine uygun olarak en başarılı bulunan kompozisyon 1 numaralı performanstır. Benzer şekilde en başarısız puanlanan performans ise 4 numaralı performanstır. İlk uygulamadan farklı biçimde eğitim sonrası 2 numaralı kompozisyon performansının başarılı bulunma sırası düşüş göstermiştir. Eğitim sonrasında başarılı bulunan performans ölçütleri benzerlik göstermektedir. Sadece "Konuya Uygunluk", "Fikir Gelişimi" ve "Tez Cümlesi" ölçütlerinde puanlama cömertliğinde azalma söz konusu olmuştur.

Üçüncü Araştırma Sorusuna Yönelik Bulgular ve Tartışma

Üçüncü araştırma sorununa yönelik bulgular, zaman yüzeyinin de eklendiği analizde uygulama bütünlüğü düşünülerek sergilenmiştir. Puanlayıcılar uygulama genelinde de anlamlı olarak birbirinden farklı puanlama davranışı göstermektedir. En cömert davranan puanlayıcılar 2 ve 6 numaralı katılımcılar iken, en katı davranan puanlayıcılar 19, 33 ve 47 numaralı akranlardır. Uygunluk içi ve uygunluk dışı değerler, akademik yazmada akran puanlanmasın ÇYRM analizine uygun olduğuna işaret etmektedir. Öğrenci/kompozisyon yüzeyinin analizleri, uygulama öncesi yeterlik düzeylerine 3 ve 5 numaralı eser dışında benzerlik göstermektedir. Performans boyutları/ölçütleri incelendiğinde "Uygun Uzunluk", "Konuya Uygunluk", "Tutarlılık", "Dil Bilgisi", "Yazım Kuralları/İmla" ve "Noktalama" kriterlerinde performansların başarılı bulunduğu anlaşılmaktadır. Kriterlerin ayırma katsayısının güvenilir bir şekilde yüksek olduğu görülmektedir. ÇYRM'nin ayırt etme gücünün yüksek olmasına Şahin ve diğerleri (2016) ve Demir (2023) çalışmaları da değinmiştir.

Zaman yüzeyine bakıldığında ön ve son uygulama arasında herhangi bir farkın olmadığı anlaşılmaktadır. Ki-kare analizinin ve eşleştirilmiş örneklem t testinin anlamlı çıkmaması da bu bulguyu desteklemektedir. Uygunluk değerleri veri setinin modele bir kez daha uyum sağladığını göstermektedir.

Genel bir bakış açısıyla, altı haftalık puanlayıcı eğitimi puanlayıcı davranışlarında istatiksel olarak anlamlı bir değişiklik meydana getirmese de öğrencilerde farkındalık artırıcı etkide bulunmuştur. Alan yazında yapılan çalışmalarnda (Ayoobiyan & Ahmadi, 2023; Ravand vd., 2016; Şata & Karakaya, 2022; Tursynbayeva vd., 2023) puanlayıcıya verilen eğitimin olumlu sonuçlarına değinilmiştir. Burada dikkat edilmesi gereken nokta verilen eğitimin modeli ve puanlayıcıların alt yapı özellikleridir. Wiseman (2012) puanlayıcı özelliklerinin davranışları şekillendirebileceğini iletmiştir.

Mevcut çalaışma bulgularına göre puanlayıcıların yanlılık davranışında azalma meydana gelmiştir. Kompozisyonlar yeterlilik düzeylerine daha uygun bir şekilde değerlendirilmiştir. Ölçütlerde ön uygulamaya göre daha katı bir tutum sergilenmiştir. Her iki uygulamada da en katı ve en cömert puanlanan ölçütler değişmemiştir. Eğitimin belirgin bir farklılık çıkarmamasının farklı sebepleri olabilir. Öncelikle puanlayıcılar yaklaşık bir sene boyunca akademik yazma eğitimi almıştır. Bu da puanlama davranışlarını ön uygulamada olumlu etkilemiş olabilir. Puanlayıcı geçmiş özellikleri puanlamada etkili olabilir (Goodwin, 2016; Wiseman, 2012). Eğitim süresi daha geniş ve içerik daha kapsamlı tutulabilir. “Referans Çerçevesi Eğitimi” gibi puanlama yaşantılarının zengin olduğu modeller de mevcut eğitime eklenebilir (Şata & Karakaya, 2012). Bazı öğrencilerin yanlılık davranışlarının değişmemesi puanlayıcı etkilerinde farklı değişkenlerin de etkili olabileceği izlenimini vermiştir.

BÖLÜM VI

SONUÇ VE ÖNERİLER

Bu araştırmada yabancı dil olarak öğretilen İngilizcede yazma becerilerinin ölçülmesinde ve değerlendirilmesinde akran değerlendirme kapsamında puanlayıcı davranışları incelenmek istenmiştir. Puanlayıcılara altı hafta süreyle eğitim verilmiş ve akranların eğitim öncesinde ve sonrasında puanladıkları reaksiyon türünde beş adet kompozisyon görevi ÇYRM ile analiz edilmiştir. Bulgular incelendiğinde, akademik yazma becerisinin değerlendirilmesinde ÇYRM'in kullanılmasının geçerli sonuçlar vereceği anlaşılmıştır. Akademik yazma kapsamında akranlardan gelen puanlamalar modele oldukça iyi uyum göstermiş ve puanlayıcı, zaman, performans ve ölçüt yüzeylerinde örneklemden bağımsız ve güvenilir düzeyde ayırt ediciliği olan analizler sunulmuştur.

İkinci olarak, altı hafta süreyle yapılan puanlayıcı eğitiminin yanlılık davranışını azalttığını ve ölçüt ve performans değerlendirmelerinde katılık ve cömertlik açısından az oranda farklılıklar oluşturduğu tespit edilmiştir. Puanlayıcı davranışları bir bütün olarak düşünüldüğünde ön ve son uygulama arasında belirgin farklar oluşmamıştır. Bunun başlıca sebebi arasında puanlayıcıların alan uzmanlığı, puanlama sürecinin uzun sürmesi, eğitim modellerinin sınırlı tutulması ve eğitim sürenin görece olarak kısa olması düşünülmektedir. Kompozisyon yüzey analizleri incelendiğinde, kompozisyonlar seçilme kriterlerine ve düzeylerine uygun olarak puanlandırılmıştır. Toplam 16 kriterden bazı kriterler cömert ve bazı kriterler katı puanlanmıştır. Bu bulgular dikkate alınarak yazma becerileri derslerinde bu ölçütler üzerine ek eğitimler planlanabilir.

Görüldüğü üzere ÇYRM, yabancı dil ölçme ve değerlendirme alanında yazma becerilerinin ölçülmesinde ve sürece akranların dahil olmasında oldukça elverişli bir modeldir. Bu

modelin, farklı dil becerilerinin değerlendirilmesinde ve özellikle performans temelli alternatif ölçme ve değerlendirme uygulamalarında kullanılması tavsiye edilmektedir. Puanlayıcı eğitiminin yanlılık davranışlarını azalttığı gözlenmiştir. Bu tarz eğitimlerin ihtiyaca uygun olan bir model ile kavramsallaştırılarak ve puanlayıcıya oldukça zengin deneyim yaşatma fırsatı sunacak şekilde planlanması önerilmektedir.

Akranların uzmanlıkları, geçmiş yaşantıları ve puanlama sürecinin pratikliği gibi değişkenler eğitimin etkililiğini şekillendirmektedir. Bu nedenle puanlayıcı ve ölçme kaynaklı özelliklerinin, puanlayıcı eğitimi kapsamında araştırılmasına işaret edilmektedir. Puanlama eğitimi çalışmalarında katılımcıların deneyim kazanacağı süreçleri ön plana çıkaran karma modellerden yararlanılması tavsiye edilmektedir. Bu araştırma zayıf deneysel desen üzerine kurulduğu için kontrol grubunda yaşanılacak olası değişimleri gözlemlemek açısından sınırlı kalmıştır. Bu nedenle yapılacak çalışmaların deneysel modellerinde kontrol grubunun yer alması ve akademik yazma performansını etkileyecek farklı değişkenlerin kontrol altına alınması önerilmektedir.

KAYNAKLAR

Akiyama, T. (2001) The application of G-theory and IRT in the analysis of data from speaking tests administered in a classroom. *Melbourne Papers in Language Testing 10*(1), 1-21.

Aldabbus, S., & Almansouri, E. (2022). Academic writing difficulties encountered by university EFL learners. *British journal of English linguistics*, *10*(3), 1-11. https://doi.org/10.37745/bjel.2013/vol10n3111

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge University.

Anthony, C. J., Styck, K. M., Cooke, E., Martel, J. R., & E. Frye, K. (2022). Evaluating the impact of rater effects on behavior rating scale score validity and utility. *School Psychology Review*, *51*(1), 25-39. https://doi.org/10.1080/2372966X.2020.1827681

Atılgan, H. (2005). Genellenebilirlik kurami ve puanlayicilar arasi güvenirlik için örnek bir uygulama. *Eğitim Bilimleri ve Uygulama, 4*(7), 95-108.

Aryadoust, V. (2015). Self-and peer assessments of oral presentations by first-year university students. *Educational Assessment*, *20*(3), 199-225. https://doi.org/10.1080/10627197.2015.1061989

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Ayçiçek, B. (2025). Peer Assessment Analysis via the Many-Facet Rasch Model. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *21*(2), 631-646. https://doi.org/10.17860/mersinefd.1642676

Ayoobiyan, H., & Ahmadi, A. (2023). Detecting halo effects across rubric criteria in L2 writing assessment: a many-facet Rasch analysis. *Applied Research on English Language*, *12*(1), 159-176. https://doi.org/10.22108/are.2022.132503.1848

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University.

Bachman, L. F., & Purpura, J. E. (2008). Language Assessments: Gate-Keepers or Door-Openers?. *The handbook of educational linguistics*, 456-468.

Bailey, S. (2003). *Academic writing: A practical guide for students*. Psychology.

Baker, F. B. (2001). *The basics of item response theory.* ERIC.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, *2*(1), 49-58.

Birjandi, P., & Siyyari, M. (2010). Self-assessment and peer-assessment: A comparative study of their effect on writing performance and rating accuracy. *IJAL*, 13, 23-45.

Bortolotti, S. L. V., Tezza, R., de Andrade, D. F., Bornia, A. C., & de Sousa Júnior, A. F. (2013). Relevance and advantages of using the item response theory. *Quality & Quantity*, *47*(4), 2341-2360. https://doi.org/10.1007/s11135-012-9684-5

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading.* ASCD.

Brown, J. D., Hudson, T., Norris, J., & Bonk, W.J. (2002). *An investigation of second language task-based performance assessments*. Second Language Teaching & Curriculum Center.

Brown, J. D. (2004). *Language assessment: Principles and classroom practices.* Pearson/Longman.

Brown, J. D. (2005). *Testing in language programs.* McGraw-Hill.

Brown, J. D. (2013). *New ways of classroom assessment.* TESOL International.

Büyükkıdık, S., & Atar, H. (2018). Çok kategorili madde tepki kuramı modellerinin örneklem büyüklüğü açısından incelenmesi. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 38* (2), 665-694.

Büyüköztürk, Ş., Akgün, Ö. E., Demirel, F., Karadeniz, Ş., & Çakmak, E. K. (2014). *Bilimsel araştırma yöntemleri.* Pegem Yayınları.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* (1), 1-47

Canale, M. (1987). The measurement of communicative competence. *Annual Review Of Applied Linguistics*, *8*, 67-84.

Carr, N. (2011). *Designing and analyzing language tests.* Oxford University.

Carrió-Pastor, M. L. (2016). Should peer assessment be included in foreign language testing? The role of motivation in testing. *New Challenges For Language Testing: Towards Mutual Recognition of Qualifications*, 61-75.

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language testing*, *22*(1), 93-121.

Çelen, Ü. (2008). Comparison of validity and reliability of two tests developed by classical test theory and item response theory. *Elementary Education Online*, *7*(3), 758-768.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2014). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve Lisrel Uygulamalar.* Pegem Akademi.

Demir, S. (2023). Analysis of peer and self-assessments using the many-facet Rasch measurement model and student opinions. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(3), 266-286. https://doi.org/10.21031/epod.1344196

Douglas, D. (2010). *Understanding language testing.* Hodder Education.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197-221.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Eckes, T. (2017). Rater effects: Advances in item response modeling of human ratings–Part I. *Psychological Test and Assessment Modeling*, *59*(4), 443-452.

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment: Volume I – Fundamental techniques* (pp. 153-175). Routledge.

Engelhard J, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Esfandiari, R. (2015). Rater Errors among Peer-Assessors: Applying the Many-Facet Rasch Measurement Model. *Iranian Journal of Applied Linguistics*, *18*(2).

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *International Journal of Language Testing*, *1*(1), 1-16.

Farrokhi, F., & Esfandiari, R. (2011). A Many-facet Rasch Model to Detect Halo Effect in Three Types of Raters. *Theory & Practice in Language Studies (TPLS)*, *1*(11).

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, *34*(1), 79-101.

Fang, Z. (2021). *Demystifying academic writing: Genres, moves, skills, and strategies.* Routledge.

Fişne, F. N., Sata, M., & Karakaya, I. (2022). Standard setting in academic writing assessment through objective standard setting method. *International Journal of Assessment Tools in Education, 9*(1), 80-97. https://doi.org/10.21449/ijate.1059304

Folse, K. S., & Pugh, T. (2019). *Great writing 5: From great essays to research*. National Geographic Learning/Cengage Learning.

Gyagenda, I. S., & Engelhard Jr, G. (1998). Applying the Rasch model to explore rater influences on the assessed quality of students' writing ability. *ERIC,* 1-30.

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing, 30*, 21-31. https://doi.org/10.1016/j.asw.2016.07.004

Güler, N., & Gelbal, S. (2010). A study based on classic test theory and many facet Rasch model. Egitim Arastirmalari-Eurasian Journal of Educational Research, 38, 108-125.

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

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.

Hamid, M. O., Hoang, N. T., & Kirkpatrick, A. (2019). Language tests, linguistic gatekeeping and global mobility. *Current Issues in Language Planning*, *20*(3), 226-244.

Hamp-Lyons, L. (1989). Applying the partial credit method of Rasch analysis: language testing and accountability. *Language Testing, 6*(1), 109-118.

Hani, U., & Perlina, M. (2022). Investigating undergraduate students' outline and essay writing. *Wanastra: Jurnal Bahasa dan Sastra*, *14*(1), 38-43. https://doi.org/10.31294/wanastra.v14i1.11640

Harrington, K., Elander, J., & Lusher, J. (2006). Using core assessment criteria to improve essay writing. In *Innovative assessment in higher education* (pp. 130-140). Routledge.

He, T. H., Gou, W. J., Chien, Y. C., Chen, I. S. J., & Chang, S. M. (2013). Multi-faceted Rasch measurement and bias patterns in EFL writing performance assessment. *Psychological Reports, 112*(2), 469-485. https://doi.org/10.2466/03.11.PR0.112.2.469-485

Hughes, A. (2003). *Testing for language teachers.* Cambridge University.

Huisman, B., Saab, N., Van Driel, J., & Van Den Broek, P. (2018). Peer feedback on academic writing: Undergraduate students' peer feedback role, peer feedback perceptions and essay performance. *Assessment & Evaluation in Higher Education, 43*(6), 955-968. https://doi.org/10.1080/02602938.2018.1424318

İlhan, M., Güler, N., & Taşdelen Teker, G. (2024*). Puanlayıcı yargılarına dayalı ölçmelerde geçerlik ve güvenirlik.* Pegem.

Jin, K. Y., & Eckes, T. (2024). Measuring the impact of peer interaction in group oral assessments with an extended many-facet Rasch model. *Journal of Educational Measurement*, *61*(1), 47-68. https://doi.org/10.1111/jedm.12375

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505.

Kennedy, S., Foote, J. A., & Dos Santos Buss, L. K. (2015). Second language speakers at university: Longitudinal development and rater behaviour. *TESOL Quarterly*, *49*(1), 199-209.

Khamboonruang, A. (2023). EFL Thai students' rating performances in self-and peer-assessments of writing: a many-facets Rasch analysis. *LEARN Journal: Language Education and Acquisition Research Network*, *16*(1), 221-245.

Kim, H. (2020). Effects of rating criteria order on the halo effect in L2 writing assessment: a many-facet Rasch measurement analysis. *Language Testing in Asia*, *10*(1), 16.

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, *12*(3), 239-261.

Knoch, U., & McNamara, T. (2015). Rasch analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 275–304). Routledge.

Köse, İ. A., Usta, H. G., & Yandı, A. (2016). Sunum yapma becerilerinin çok yüzeyli Rasch analizi ile değerlendirilmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 16*(4), 1853-1864.

Kramer, L. (2022, August 17). *Composition writing: A step-by-step guide*. Grammarly. https://www.grammarly.com/blog/writing-process/composition-writing

Li, J., Huang, J., & Cheng, S. (2022). The reliability, effectiveness, and benefits of peer assessment in college EFL speaking classrooms: Student and teacher perspectives. *Studies in Educational Evaluation, 72*, 101-120. https://doi.org/10.1016/j.stueduc.2021.101120

Linacre, J. M. (1993, April). *Generalizability theory and many-facet Rasch measurement.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Linacre, J. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16, 878.

Linacre, J.M. (2017). *A user's guide to FACETS: Rasch-model computer programs.* MESA.

Linacre, J. M. (2023). *A user's guide to FACETS: Rasch-model computer programs* (Version 3.87.0) [Computer program manual].

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-180.

Madsen, H. S. (1983). *Techniques in language testing.* Oxford University.

Marsaulina, R. M. (2025). Can peer assessment be reliable as an evaluating instrument?. *Journal of English Language and Education*, *10*(4), 1441-1452. https://doi.org/10.31004/jele.v10i4.1255

McNamara, T., & Adams, R. J. (1991). Exploring rater behaviour with Rasch techniques. *ERIC Document Reproduction Service ED*, 345-498.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555-576.

Messick, S. (1996). Validity and washback in language testing. *Language testing, 13*(3), 241-256.

Mustafa, A., Arbab, A. N., & El Sayed, A. A. (2022). Difficulties in academic writing in English as a second/foreign language from the perspective of undergraduate students in higher education institutions in Oman. *Arab World English Journal, 13*(3), 41-53. https://dx.doi.org/10.24093/awej/vol13no3.3

Mumpuni, K. E., Priyayi, D. F., & Widoretno, S. (2022). How do students perform a peer assessment?. *International Journal of Instruction*, *15*(3), 751-766.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science, 37*, 375-401.

Norton, B. (1997). Accountability in language assessment. In C. Clapham and D. Corson (Eds.), *Encyclopedia of language and education,* Vol. 7, (pp. 313-22). Kluwer Academic.

Norton, L. S. (1990). Essay-writing: what really counts?. *Higher Education*, *20*(4), 411-442.

Ntumi, S., & Bulala, T. (2025). Multilevel psychometric analysis of clinician burnout using Bayesian G-theory and IRT fusion models. *Inquiry: The Journal of Health Care Organization, Provision, and Financing, 62*, 1–17. https://doi.org/10.1177/00469580251371389

O'Grady, S. (2023). Halo effects in rating data: Assessing speech fluency. *Research Methods in Applied Linguistics*, 2(2), 100048.

Oshima, A., & Hogue, A. (2007). *Introduction to academic writing*. Pearson/Longman.

Öç, U. (2024). *Açık uçlu soruların puanlanmasında puanlayıcı eğitiminin puanlayıcı davranışlarına etkisinin incelenmesi* [Doktora tezi, Gazi Üniversitesi]. Yükseköğretim Kurulu Ulusal Tez Merkezi.

Park, T. (2004). An investigation of an ESL placement test of writing using many-facet Rasch measurement. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics,* 4(1), 1–25.

Phung, D. V. (2013). Rater behaviors in assessing EFL compositions. *CNU Journal of Higher Education, 7*(1), 117-145.

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement: Issues and Practice, 10*(2), 15–16, 22, 25–26. https://doi.org/10.1111/j.1745-3992.1991.tb00188.x

Polat, M. (2020). A Rasch analysis of rater behaviour in speaking assessment. *International Online Journal of Education and Teaching, 7*(3), 1126-1141.

Ravand, S., Ravand, H., & Abbasi, A. (2016). Investigating the effect of self-, peer-, and teacher assessment in second language writing over time: A multifaceted Rasch approach. *Iranian Journal of Applied Language Studies, 8*(1), 91-112.

Razı, S. (2015). Development of a rubric to assess academic writing incorporating plagiarism detectors. *SAGE Open, 5*(2), 1–13. https://doi.org/10.1177/2158244015590162

Reinholz, D. (2016). The assessment cycle: A model for learning through peer assessment. *Assessment & Evaluation in Higher Education, 41*(2), 301-315.

Rentiana, L. H., Dhini, U. R., Yuliawati, L., & Wulandari, E. T. (2024). Item fit analysis for evaluating academic writing performance with Rasch measurement. *Attractive: Innovative Education Journal, 6*(1), 645-653.

Rotsaert, T., Panadero, E., Estrada, E., & Schellens, T. (2017). How do students perceive the educational value of peer assessment in relation to its social nature? A survey study in Flanders. *Studies in Educational Evaluation, 53*, 29-40.

Saddler, B., & Andrade, H. (2004). The writing rubric. *Educational Leadership, 62*(2), 48-52.

Sarıtaş-Akyol, S., & Karakaya, İ. (2021). Investigating the consistency between students' and teachers' ratings for the assessment of problem-solving skills with many-facet Rasch measurement model. *Eurasian Journal of Educational Research, 91*, 281-300.

Sayın, A., & Şata, M. (2022). Using Rasch analysis to examine raters' expertise Turkish teacher candidates' competency levels in writing different types of test items. *International Journal of Assessment Tools in Education, 9*(4), 998-1012. https://doi.org/10.21449/ijate.1058300

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33.

Spiller, D. (2012). Assessment matters: Self-assessment and peer assessment. *The University of Waikato, 13*, 2-18.

Stonewall, J., Dorneich, M., Dorius, C., & Rongerude, J. (2018, April). A review of bias in peer assessment. In *2018 CoNECD-The Collaborative Network for Engineering and Computing Diversity Conference*.

Styck, K. M., Anthony, C. J., Sandilos, L. E., & DiPerna, J. C. (2021). Examining rater effects on the classroom assessment scoring system. *Child Development, 92*(3), 976-993. https://doi.org/10.1111/cdev.13460

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239-261.

Sun, Q., Chen, F., & Yin, S. (2023). The role and features of peer assessment feedback in college English writing. *Frontiers in Psychology*, *13*, 1070618. https://doi.org/10.3389/fpsyg.2022.1070618

Şata, M. (2019). *Performans değerlendirme sürecinde puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisinin incelenmesi* [Doktora tezi, Gazi Üniversitesi]. Yükseköğretim Kurulu Ulusal Tez Merkezi.

Şahin, M. G., Teker, G. T., & Güler, N. (2016). An analysis of peer assessment through Many Facet Rasch Model. *Journal of Education and Practice, 7*(32), 172-181.

Şata, M., & Karakaya, İ. (2021). Investigating the effect of rater training on differential rater function in assessing academic writing skills of higher education students. *Journal of Measurement and Evaluation in Education and Psychology, 12(*2), 163-181. https://doi.org/10.21031/epod.842094

Thomas, E. (2012) Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270-292.

Topping, K. J. (2009). Peer assessment. *Theory into practice, 48*(1), 20-27.

Trace, J., Janssen, G., & Meier, V. (2017). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing, 34*(1), 3-22.

Tursynbayeva, N., Öç, U., & Karakaya, İ. (2024). The effect of rater training on rating behaviors in peer assessment among secondary school student**s**. *International Journal of Assessment Tools in Education, 11*(3), 507-523. https://doi.org/10.21449/ijate.1438798

Wang, B. (2010). On rater agreement and rater training. *English Language Teaching, 3*(1), 108-112.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language testing, 15*(2), 263-287.

Weir (2016). Academic Writing - Can we please all the people all the time? In Johnson, A. P. Ed. *Academic writing: Process and product*. Bloomsbury.

Weiss, D. J., & Yoes, M. E. (1991). Item response theory. In *Advances in educational and psychological testing: Theory and applications* (pp. 69-95). Springer.

West, H., Malcolm, G., Keywood, S., & Hill, J. (2019). Writing a successful essay. *Journal of Geography in Higher Education*, *43*(4), 609-617. https://doi.org/10.1080/03098265.2019.1655720

Wind, S. A. (2020). Exploring the impact of rater effects on person fit in rater-mediated assessments. *Educational Measurement: Issues and Practice*, *39*(4), 76-94. https://doi.org/10.1111/emip.12354

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*(3), 150-173.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of occupational and organizational psychology*, *67*(3), 189-205.

Wu, S. M., & Tan, S. (2016). Managing rater effects through the use of FACETS analysis: the case of a university placement test. *Higher Education Research & Development*, *35*(2), 380-394.

Yağız, O., & Yiğiter, K. (2012). Academic writing difficulties and challenges in advanced academic literacy. *The Journal of Academic Social Science Studies*, *5*(8), 1261-1272.

Yan, X., & Chuang, P. L. (2023). How do raters learn to rate? Many-facet Rasch modeling of rater performance over the course of a rater certification program. *Language Testing*, *40*(1), 153-179. https://doi.org/10.1177/02655322221074913

Yeşilçınar, S., & Şata, M. (2021). Examining rater biases of peer assessors in different assessment environments. *International Journal of Psychology and Educational Studies*, *8*(4), 136-151. https://doi.org/10.52380/ijpes.2021.8.4.622

Yu, C. H. (2020). Objective measurement: How rasch modeling can simplify and enhance your assessment. In M. S. Khine (Eds.). *Rasch measurement. Applications in Quantitative Educational Research*, 47-73. Springer.

Yu, S., & Lee, I. (2016). Peer feedback in second language writing (2005–2014). *Language Teaching*, *49*(4), 461-493.

Yüzüak, A., Yüzüak, B., & Kaptan, F. (2015). Performans görevinin akran gruplar ve öğretmen yaklaşımları doğrultusunda çok-yüzeyli rasch ölçme modeli ile analizi. *Journal of Measurement and Evaluation in Education and Psychology, 6*(1), 1-11.

Zhang, F., Schunn, C., Li, W., & Long, M. (2020). Changes in the reliability and validity of peer assessment across the college years. *Assessment & evaluation in higher education, 45*(8), 1073-1087. https://doi.org/10.1080/02602938.2020.1724260

EKLER

Ek 1. Kriter/Ölçüt Yüzeyi Ön Uygulama ve Son Uygulama Logit Değerleri

Kriter/Ölçüt Logit Değerlerinin Ön ve Son Uygulama Karşılaştırılması

Ön Uygulama	Logit	Son Uygulama	Logit
1. Başlık	-0,22	1. Başlık	-0,29
2. Organizasyon	-0,35	2. Organizasyon	-0,43
3. Tez Cümlesi	-0,83	3. Tez Cümlesi	-0,97
4. Konu Cümlesi	-0,56	4. Konu Cümlesi	-0,64
5. Destekleyici Cümleler	-0,22	5. Destekleyici Cümleler	-0,32
6. Uygun Uzunluk	0,32	6. Uygun Uzunluk	0,44
7. Konuya Uygunluk	0,49	7. Konuya Uygunluk	0,30
8. Fikir Gelişimi	-0,02	8. Fikir Gelişimi	-0,19
9. Tutarlık	0,33	9. Tutarlık	0,25
10. Bağlantı Unsurları	-0,33	10. Bağlantı Unsurları	-0,28
11. Dil Bilgisi	0,45	11. Dil Bilgisi	0,56
12. Sözdizimsel Karmaşıklık	-0,41	12. Sözdizimsel Karmaşıklık	-0,35
13. Kelime Seçimi	-0,55	13. Kelime Seçimi	-0,43
14. Kelime Çeşitliliği	-0,11	14. Kelime Çeşitliliği	-0,08
15. Yazım Kuralları/İmla	1,26	15. Yazım Kuralları/İmla	1,50
16. Noktalama	0,75	16. Noktalama	0,93

Ek 2. Yanlılık Test Sonuçları – Puanlayıcı Sırası

Yanlılık Test Sonuçları – Puanlayıcı Sırası – Ön Test

Puanlayıcı	Birey/	Gözlenen Puan	Beklenen Puan	Yanlılık (logit)	Standart hata	t
P5	Kompozisyon 2	97	107,54	-0,39	0,18	-2,14
P7	Kompozisyon 3	89	101,15	-0,41	0,18	-2,3
P7	Kompozisyon 2	100	110,69	-0,43	0,19	-2,26
P7	Kompozisyon 5	121	110,14	0,59	0,26	2,26
P8	Kompozisyon 1	108	119,77	-0,6	0,21	-2,91
P8	Kompozisyon 4	123	104,07	0,98	0,27	3,57
P9	Kompozisyon 3	112	100,7	0,46	0,22	2,13
P12	Kompozisyon 5	121	107,51	0,7	0,26	2,7
P14	Kompozisyon 3	93	105,75	-0,45	0,18	-2,53
P14	Kompozisyon 2	126	114,35	0,78	0,3	2,57
P15	Kompozisyon 2	96	108,84	-0,48	0,18	-2,64
P15	Kompozisyon 1	132	114,01	1,51	0,39	3,82
P17	Kompozisyon 3	72	96,83	-0,75	0,17	-4,47
P17	Kompozisyon 2	97	107,17	-0,38	0,18	-2,05
P17	Kompozisyon 1	127	112,58	0,96	0,31	3,05
P17	Kompozisyon 4	111	92,84	0,68	0,21	3,2
P18	Kompozisyon 2	117	100,03	0,74	0,24	3,13
P20	Kompozisyon 2	122	111,97	0,57	0,27	2,14
P22	Kompozisyon 4	70	84,88	-0,42	0,17	-2,55
P22	Kompozisyon 5	118	100,1	0,8	0,24	3,29
P23	Kompozisyon 5	121	107,14	0,71	0,26	2,76
P24	Kompozisyon 3	102	88,12	0,46	0,19	2,41
P25	Kompozisyon 1	111	119,31	-0,44	0,21	-2,05
P25	Kompozisyon 5	128	114,6	0,96	0,33	2,94
P26	Kompozisyon 4	75	90,24	-0,45	0,17	-2,67
P3	Kompozisyon 5	99	108,83	-0,38	0,19	-2,03
P3	Kompozisyon 2	120	109,4	0,55	0,25	2,19
P30	Kompozisyon 3	46	76,82	-0,87	0,17	-5,01
P30	Kompozisyon 5	101	89,04	0,4	0,19	2,09
P30	Kompozisyon 1	121	97,2	1,09	0,26	4,18
P31	Kompozisyon 1	89	107,85	-0,67	0,18	-3,79
P31	Kompozisyon 3	106	90,34	0,55	0,2	2,74
P32	Kompozisyon 1	112	123,66	-0,69	0,22	-3,2
P32	Kompozisyon 5	128	119,91	0,66	0,33	2,02
P33	Kompozisyon 1	107	117,46	-0,5	0,2	-2,47

P33	Kompozisyon 2	106	87,3	0,64	0,2	3,21
P34	Kompozisyon 1	106	118,39	-0,6	0,2	-2,97
P34	Kompozisyon 2	76	89,44	-0,4	0,17	-2,35
P34	Kompozisyon 3	121	105,29	0,79	0,26	3,04
P34	Kompozisyon 1	118	96,84	0,91	0,24	3,76
P35	Kompozisyon 4	94	105,07	-0,39	0,18	-2,18
P35	Kompozisyon 3	70	82,42	-0,35	0,17	-2,11
P35	Kompozisyon 2	106	94,87	0,4	0,2	2,01
P36	Kompozisyon 1	89	106,68	-0,62	0,18	-3,51
R36	Kompozisyon 3	108	88,78	0,68	0,21	3,31
P37	Kompozisyon 4	117	106,83	0,48	0,24	2,04
P38	Kompozisyon 2	106	114,89	-0,4	0,2	-2,01
P38	Kompozisyon 3	108	87,68	0,71	0,21	3,47
P39	Kompozisyon 2	82	98,88	-0,53	0,17	-3,12
P39	Kompozisyon 3	71	88,12	-0,5	0,17	-2,97
P39	Kompozisyon 4	66	82,59	-0,47	0,17	-2,81
P39	Kompozisyon 5	110	98,18	0,46	0,21	2,18
P39	Kompozisyon 1	121	105,34	0,79	0,26	3,04
P40	Kompozisyon 5	75	92,17	-0,51	0,17	-3,04
P40	Kompozisyon 3	66	83,51	-0,5	0,17	-2,98

Yanlılık Test Sonuçları – Puanlayıcı Sırası – Son Test

Puanlayıcı	**Birey/**	**Gözlenen Puan**	**Beklenen Puan**	**Yanlılık (logit)**	**Standart hata**	**t**
P7	Kompozisyon 2	60	67,84	-0,61	0,27	-2,27
P7	Kompozisyon 5	82	73,81	1,48	0,59	2,51
P8	Kompozisyon 2	61	68,56	-0,6	0,27	-2,22
P8	Kompozisyon 3	83	74,85	1,78	0,72	2,48
P11	Kompozisyon 1	73	65,54	0,68	0,33	2,08
P14	Kompozisyon 2	78	71,26	0,83	0,4	2,05
P15	Kompozisyon 1	84	73,89	2,6	1,01	2,59
P17	Kompozisyon 3	53	73,26	-1,61	0,26	-6,21
P17	Kompozisyon 1	83	75,31	1,73	0,72	2,41
P17	Kompozisyon 4	81	65,31	1,97	0,52	3,8
P18	Kompozisyon 3	66	72,37	-0,58	0,28	-2,03
P18	Kompozisyon 2	74	65,22	0,82	0,34	2,42
P20	Kompozisyon 1	72	78,87	-0,91	0,32	-2,84
P22	Kompozisyon 3	54	65,7	-0,85	0,26	-3,26
P24	Kompozisyon 1	66	72,6	-0,6	0,28	-2,11

P24	Kompozisyon 4	70	61,24	0,71	0,3	2,33
P26	Kompozisyon 4	39	60,08	-1,45	0,27	-5,33
P26	Kompozisyon 3	80	69,26	1,39	0,47	2,98
P27	Kompozisyon 3	59	67,02	-0,61	0,27	-2,29
P27	Kompozisyon 2	70	58,58	0,9	0,3	2,96
P3	Kompozisyon 5	67	73,63	-0,63	0,29	-2,18
P30	Kompozisyon 3	56	66,46	-0,77	0,26	-2,95
P30	Kompozisyon 5	78	65,73	1,32	0,4	3,26
P31	Kompozisyon 1	54	69,93	-1,21	0,26	-4,65
P31	Kompozisyon 4	57	67,54	-0,79	0,26	-3,02
P31	Kompozisyon 1	83	76,68	1,54	0,72	2,15
P31	Kompozisyon 3	82	74,85	1,36	0,59	2,3
P31	Kompozisyon 4	67	57,55	0,71	0,29	2,46
P31	Kompozisyon 3	80	67,21	1,57	0,47	3,36
P32	Kompozisyon 3	65	75,55	-1,02	0,28	-3,64
P32	Kompozisyon 1	72	79,29	-0,99	0,32	-3,09
P32	Kompozisyon 4	77	68,55	0,93	0,38	2,42
P33	Kompozisyon 1	66	75,46	-0,93	0,28	-3,27
P33	Kompozisyon 1	57	64,63	-0,56	0,26	-2,12
P33	Kompozisyon 2	63	52,26	0,75	0,27	2,74
P34	Kompozisyon 1	61	72,93	-1,02	0,27	-3,77
P34	Kompozisyon 3	78	70,55	0,9	0,4	2,22
P34	Kompozisyon 1	80	66,62	1,62	0,47	3,46
P35	Kompozisyon 4	60	68,81	-0,69	0,27	-2,58
P36	Kompozisyon 1	63	71,61	-0,73	0,27	-2,67
P38	Kompozisyon 2	55	64,98	-0,72	0,26	-2,77
P38	Kompozisyon 3	82	72,19	1,66	0,59	2,8
P39	Kompozisyon 3	47	65,51	-1,3	0,26	-5,02
P39	Kompozisyon 2	47	61,51	-0,99	0,26	-3,84
P39	Kompozisyon 4	47	60,31	-0,91	0,26	-3,5
P39	Kompozisyon 4	64	55,53	0,61	0,28	2,19
P39	Kompozisyon 1	83	71,94	2,11	0,72	2,93
P39	Kompozisyon 3	80	69,45	1,38	0,47	2,94
P40	Kompozisyon 3	48	63,6	-1,08	0,26	-4,19
P40	Kompozisyon 5	45	54,08	-0,61	0,26	-2,34

Ek 3. Veri Toplama Aracı (Kompozisyon Ödevi)

REACTION ESSAY GUIDELINES

1. You are supposed to write your reaction essay about the short story "Flowers for Algernon" written by Daniel Keyes.

2. Please submit the outline and the final draft of your essay by combining them into a single Word document.

3. Your essay should consist of at least 4 paragraphs and minimum 500 words. Use Times New Roman 12 font, 1.5 line spacing, and English language use on MS Word. (Please add word count.)

4. Avoid plagiarism at all costs. If plagiarism is detected, your paper will **NOT** be evaluated.

5. Submit your assignment to the Google Classroom.

6. Submissions sent after this date will **NOT** be accepted.

Submission files to be evaluated:

- ✓ Your outline
- ✓ Your final draft

Ek 4. Veri Toplama Aracı (Dereceli Puanlama Anahtarı)

ANALYTIC WRITING SCORING RUBRIC FOR ACADEMIC ESSAYS

	ORGANIZATION						CONTENT	
Point	**Title of Essay**	**Introduction-Body-Conclusion**	**Thesis Statement**	**Topic Sentence**	**Supporting Sentences**	**Appropriate Length**	**Topic Relevance**	**Idea Development**
4	Title of essay *comprehensively* represents the focus of the written text. It is *highly* relevant to the task.	The organization of introduction, body, and conclusion paragraphs is *highly* appropriate to written genre.	Thesis statement is *noticeably* given in introduction paragraph. It *comprehensively* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *comprehensively* addresses and supports the specific idea(s) given in thesis statement. It *extensively* demonstrates the main idea of the paragraph.	Supporting sentences *comprehensively* illustrate the main idea given in topic sentence.	There are *at least 250 words* in written text. It is constructed with *appropriate length.*	Written text is *highly* relevant to assigned topic in task. It *comprehensively* addresses all parts of the task.	*Extensive* details are provided to develop, support and illustrate information or ideas presented in written text.
3	Title of essay *adequately* represents the focus of the written text. It is relevant to the task.	The organization of introduction, body, and conclusion paragraphs is *largely* appropriate to written genre.	Thesis statement is *evidently* given in introduction paragraph. It *mostly* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *mostly* addresses and supports the specific idea(s) given in thesis statement. It *largely* demonstrates the main idea of the paragraph.	Supporting sentences *adequately* illustrate the main idea given in topic sentence.	Text length is between *200 and 249 words*. It is *slightly* shorter than required length.	Written text is *mostly* relevant to assigned topic in task. It *adequately* addresses the basic parts of the task.	*Adequate* details are provided to develop, support and illustrate information or ideas presented in written text.
2	Title of essay *moderately* represents the focus of the written text. It is relevant to the task in *some respects.*	The organization of introduction, body, and conclusion paragraphs is *moderately* appropriate to written genre.	Thesis statement is *less explicitly* given in introduction paragraph. It *moderately* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *moderately* addresses and supports the specific idea(s) given in thesis statement. It demonstrates the main idea of the paragraph in *some respects.*	Supporting sentences *moderately* illustrate the main idea given in topic sentence.	Text length is between 150 *and 199 words*. It is *seemingly* shorter than required length.	Written text is *moderately* relevant to assigned topic in task. It *partially* addresses the basic parts of task.	*Basic* details are provided to develop, support and illustrate information or ideas presented in written text.
1	Title of essay *slightly* represents the focus of the written text. It is *partially* relevant to the task.	There is *inadequate* organization of introduction, body, and conclusion paragraphs in the written text.	Thesis statement is *vaguely* given in introduction paragraph. It *slightly* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *partially* addresses and supports the specific idea(s) given in thesis statement. It *slightly* demonstrates the main idea of the paragraph.	Supporting sentences *partially* illustrate the main idea given in topic sentence.	Text length is between 100 *and 149 words*. It is *considerably* shorter than required length.	Written text is *slightly* relevant to assigned topic in task. It lacks addressing the basic parts of the task.	*Some details are* provided but they are not enough to develop, support and illustrate information or ideas presented in written text.
0	Written text does not include a title or title of essay is *completely* irrelevant.	Written text lacks organization of introduction, body and conclusion paragraphs.	Thesis statement is not given in introduction paragraph or it does not include any specific idea(s) to be elaborated in the written text.	Topic sentence is not included in written text, or it does not address the thesis statement or demonstrate the main idea of the paragraph.	Written text does not include supporting sentences or they do not illustrate the main idea given in topic sentence.	Text length is *below 99 words*. It does not meet the requirement of appropriate length.	Written text is irrelevant to assigned topic in task. It fails to address the task adequately.	Information or ideas are not *thoroughly* developed, supported or illustrated in written text.

	COHERENCE	COHESION	GRAMMAR		VOCABULARY		MECHANICS	
Point	**Coherence**	**Linking**	**Accuracy of Grammatical Forms**	**Syntactic Complexity**	**Word Choice**	**Lexical Range**	**Spelling**	**Punctuation**
4	Information or ideas sequenced in paragraphs are *highly* consistent. There is a *considerably* logical progression between sentences in written text.	A *wide* range of cohesive devices used to connect ideas in written text provides a smooth transition between sentences.	All grammatical forms are *accurately* used in written text. The communication is *successfully* established.	Complex and sophisticated sentences are *extensively* used in written text in which syntactic structures are *highly* diverse.	All the words and phrases are *appropriately* used. The intended meaning is *clearly* conveyed in written text.	There is a *wide range* of vocabulary used in written text which includes *highly* sophisticated words and phrases.	All the needed spelling rules are *accurately* used in written text.	All the needed punctuation rules are *accurately* used in written text.
3	Information or ideas sequenced in paragraphs are *mostly* consistent. There is an *adequately* logical progression between sentences in written text.	An *adequate* range of cohesive devices used to connect ideas in written text provides an easy transition between sentences.	The use of the grammatical forms is *mostly accurate* in the written text. There are *few grammatical errors* which do not impede communication.	Complex and sophisticated sentences are *widely* used in written text in which syntactic structures are *adequately* diverse.	The use of words and phrases is *mostly appropriate*. There are *few* misused words or phrases which cannot obscure the intended meaning.	There is an *adequate range* of vocabulary used in written text which includes *largely* sophisticated words and phrases.	All the needed spelling rules are *mostly accurate* in written text but there are *few errors* which violate these rules.	All the needed punctuation rules are *mostly accurate* in written text but there are few errors which violate these rules.
2	Information or ideas sequenced in paragraphs are *moderately* consistent but there are some inconsistencies which *partially* interrupt logical progression between sentences.	The use of cohesive devices *at basic level* to connect ideas in written text provides a complete transition between sentences.	It is attempted to use the grammatical forms accurately in written text but there are *occasional grammatical errors* which slightly impede communication.	Complex and sophisticated sentences are *moderately* used in written text in which syntactic structures are *partially* diverse.	It is attempted to use the words and phrases appropriately but there are *occasionally* misused words or phrases which *slightly* obscure the intended meaning.	The *basic* vocabulary is used in written text which includes *moderately* sophisticated words and phrases.	It is intended to use the needed spelling rules *accurately* in written text but there are *occasional errors* which violate these rules.	It is intended to use the needed punctuation rules *accurately* in written text but there are *occasional errors* which violate these rules.
1	Paragraphs are constructed with *slightly* consistent information or ideas which interrupt logical progression and sequence between sentences.	A *limited* range of cohesive devices used to connect ideas in written text makes transition between sentences fragmentary.	The use of the grammatical forms is *generally inaccurate* in written text. There are *frequent grammatical errors* which largely impede communication.	Complex and sophisticated sentences are *slightly* used in written text in which syntactic structures are diverse to some extent.	The use of words and phrases is *generally inappropriate*. There are *frequently* misused words or phrases which *largely* obscure the intended meaning.	There is a *limited range* of vocabulary used in written text which includes *slightly* sophisticated words and phrases.	The use of the needed spelling rules is *largely* inaccurate. There are *frequent errors* which violate these rules.	The use of the needed punctuation rules is *largely* inaccurate. There are *frequent errors* which violate these rules.
0	Written text lacks consistency and logical progression between sentences.	There is an *inadequate* use of cohesive devices in written text which lacks transition between sentences.	The use of grammatical forms is *completely inaccurate* in the written text. This causes a breakdown in communication.	Written text lacks sentential complexity, sophistication and syntactic variety.	The use of vocabulary is completely inappropriate in written text. The intended message is obscured.	A repetitive vocabulary is largely used in written text which lacks sophistication.	All the needed spelling rules are *inaccurately* used in written text.	All the needed punctuation rules are *inaccurately* used in written text.

Ek 5. Etik Kurul Onay Formu

Evrak Tarih ve Sayısı: 28.11.2024-E.1103478

T.C.
GAZİ ÜNİVERSİTESİ REKTÖRLÜĞÜ
Etik Komisyonu

Sayı : E-77082166-302.08.01-1103478 28.11.2024
Konu : Bilimsel ve Eğitim Amaçlı

DAĞITIM YERLERİNE

Üniversitemiz Eğitim Bilimleri Enstitüsü Eğitim Bilimleri Ana Bilim Dalı **Yüksek Lisans Öğrencisi Fatıma Nur FİŞNE'nin, Prof. Dr. İsmail KARAKAYA'nın** danışmanlığında yürüttüğü ***"İngilizce Akademik Yazma Becerilerinin Değerlendirilmesinde Puanlayıcı Eğitiminin Akran Puanlayıcı Davranışları Üzerine Etkisinin Çok Yüzeyli Rasch Modeli ile Belirlenmesi"*** adlı tez çalışması ile ilgili konu Komisyonumuzun **26.11.2024** tarih ve **19** sayılı toplantısında görüşülmüş olup,

İlgilinin çalışmasının, yapılması planlanan yerlerden izin alınması koşuluyla yapılmasında etik açıdan bir sakınca bulunmadığına oybirliği ile karar verilmiş ve karara ilişkin imza listesi ekte gönderilmiştir.

Bilgilerinizi rica ederim.

Araştırma Kod No: 2024 - 1773

Prof. Dr. İsmail KARAKAYA
Komisyon Başkanı

Ek:1 Liste
DAĞITIM
Gereği:
Sayın Prof. Dr. İsmail KARAKAYA

Bilgi:
Eğitim Bilimleri Enstitüsü Müdürlüğüne



GAZİLİ OLMAK AYRICALIKTIR...