



**PERFORMANSIN DEĞERLENDİRİLMESİNDE
FARKLILAŞAN PUANLAYICI DAVRANIŞLARININ
ÇOK YÜZEYLİ RASCH ÖLÇME MODELİ İLE İNCELENMESİ**

Cennet Tobaş

**YÜKSEK LİSANS TEZİ
EĞİTİMDE ÖLÇME VE DEĞERLENDİRME ANABİLİM DALI**

**GAZİ ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ**

HAZİRAN, 2020

TELİF HAKKI ve TEZ FOTOKOPİ İZİN FORMU

Bu tezin tüm hakları saklıdır. Kaynak göstermek koşuluyla tezin teslim tarihinden itibaren (2) yıldan sonra tezden fotokopi çekilebilir.

YAZARIN

Adı : Cennet

Soyadı : TOBAŞ

Bölümü : Eğitimde Ölçme ve Değerlendirme

İmza : 

Teslim tarihi: 30 / 06 /2020

TEZİN

Türkçe Adı: Performansın Değerlendirilmesinde Farklılaşan Puanlayıcı Davranışlarının Çok Yüzeyle Rasch Ölçme Modeli ile İncelenmesi

İngilizce Adı: Examination of The Differential Rater Behaviours in Performance Evaluation with Many Facet Rasch Measurement Model

ETİK İLKELERE UYGUNLUK BEYANI

Gazi Üniversitesi Eğitim Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Cennet TOBAŞ

30 / 06 / 2020

JÜRİ ONAY SAYFASI

Cennet TOBAŞ tarafından hazırlanan “Performansın Değerlendirilmesinde Farklılaşan Puanlayıcı Davranışlarının Çok Yüzeyle Rasch Ölçme Modeli ile İncelenmesi” adlı tez çalışması aşağıdaki jüri tarafından oy birliği / oy çokluğu ile Gazi Üniversitesi Eğitimde Ölçme ve Değerlendirme Anabilim Dalı’nda Yüksek Lisans tezi olarak kabul edilmiştir.

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

Başkan: Prof. Dr. Bayram ÇETİN
(Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Gazi Üniversitesi)

Üye: Dr. Öğr. Üyesi Mehmet ŞATA
(Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Ağrı İbrahim Çeçen Üniversitesi)

Tez Savunma Tarihi: 30 / 06 / 2020

Bu tezin Eğitimde Ölçme ve Değerlendirme Anabilim Dalı’nda Yüksek Lisans tezi olması için şartları yerine getirdiğini onaylıyorum.

Prof. Dr. Selma YEL
Eğitim Bilimleri Enstitüsü Müdürü

TEŞEKKÜR

Çalışmanın her aşamasında desteğini benden esirgemeyen, tezimin şekillendirmeme yardımcı olan, her konuda örnek aldığım, desteğine; varlığına her zaman ihtiyaç duyacağım danışmanım ve çok değerli hocam Sayın Prof. Dr. İsmail KARAKAYA'ya çok teşekkür ediyorum.

Yüksek lisans ders ve tez dönemlerinde edindiğim bilgi ve deneyimlere katkı sağlayan Gazi Üniversitesi Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda görev yapan kıymetli hocalarıma teşekkür ediyorum.

Görüşleriyle ve eleştirileriyle çalışmama değerli katkılarda bulunan jüri üyelerim Prof. Dr. Bayram ÇETİN ve Dr. Öğr. Üyesi Mehmet ŞATA hocalarıma teşekkür ediyorum.

Tezimin şekillenmesinde benden yardımlarını esirgemeyen adını sayamadığım değerli meslektaşlarıma teşekkürü borç bilirim.

Sevgisiyle ve desteğiyle, çalışmamın ortaya çıkmasındaki sabırlarıyla ve yardımlarıyla hiçbir zaman beni yalnız bırakmayan her daim yanımda olan başta babam olmak üzere anneme ve kardeşlerime çok teşekkür ederim.

PERFORMANSIN DEĞERLENDİRİLMESİNDE FARKLILAŞAN PUANLAYICI DAVRANIŞLARININ ÇOK YÜZEYLİ RASCH ÖLÇME MODELİ İLE İNCELENMESİ

(Yüksek Lisans Tezi)

Cennet TOBAŞ

GAZİ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Haziran 2020

ÖZET

Bu araştırmada, ilköğretim 7. sınıf öğrencilerinin performanslarının değerlendirilmesine yönelik yazma performansının belirlenmesi sürecinde öğretmen puanlamalarına karışan puanlayıcı davranışlarının yanı sıra puanlayıcı yanlılıklarının ve puanlayıcı güvenilirliğinin Çok Yüzeysel Rasch Ölçme modeli (ÇYRÖM) ile belirlenmesi amaçlanmaktadır. Performansa dayalı değerlendirmelerde puanlayıcı davranışları ve puanlayıcıların yanlı davranması birey performansının belirlenmesi aşamasında hem geçerliği hem de güvenilirliği olumsuz yönde etkilediği düşünüldüğünde bu davranışların belirlenmesi ve sonuçların bu bağlamda değerlendirilmesi önem taşımaktadır. Araştırmanın çalışma grubunu, 2019-2020 eğitim öğretim yılı ikinci döneminde Ankara ilinde bir devlet okulunda yedinci sınıfta öğrenim gören 57 ortaokul öğrencisi ile puanlayıcı eğitimi alan 11 öğretmen oluşturmaktadır. Mevcut çalışmada, öğrencilere yedi farklı ikna edici metin görevi verilerek bunlardan bir tanesini seçmelerini ve bu göreve uygun olarak ikna edici metinler oluşturmaları istenmiştir. Öğrencilerin yazmış olduğu metinlerden 9 tanesinin ikna edici metin yapısına ait olmadığı görülmüş ve 48 öğrenciye ait metinler 11 öğretmen tarafından altı ölçütten oluşan analitik dereceli puanlama anahtarı ile değerlendirilmiştir. Bu araştırmada öğretmen puanlamalarına bağlı olarak değerlendirmelerde en fazla ortaya çıkan puanlayıcıların davranışları (puanlayıcı katılığı/cömertliği davranışı, halo etkisi, merkeze yönelme davranışı), yanısıra farklılaşan puanlayıcı katılık/ cömertlik davranışının diğer bir deyişle puanlayıcı yanlılığının ve puanlayıcı güvenilirliğinin KTK, GK ve MTK kuramlarına dayalı sonuçları incelenmiştir. Çalışmanın bulguları incelendiğinde, öğrencilerin yazma becerilerinin

değerlendirilmesi sürecinde bireylerin performansında puanlayıcı katılığı ve cömertliği davranışına rastlanmış; merkeze eğilim ve halo etkisi gibi puanlayıcı davranışlarının bulunmadığı görülmüştür. Farklılaşan puanlayıcı katılığı ve cömertliği davranışı yani yanlılık etkileşimleri incelendiğinde; grup düzeyinde puanlayıcılarda yanlı puanlayıcı davranışlarının görülmediği fakat bireysel düzeyde puanlayıcıların yanlı puanlayıcı davranışı sergiledikleri görülmüştür. KTK, GK ve MTK kuramlarına dayalı puanlayıcı güvenilirlikleri incelendiğinde, farklı yöntemlerle hesaplanan güvenilirlik katsayılarının yüksek olduğu bulunmuş; puanlayıcıların birbirinden farklı puanlamalar yapmasına rağmen puanların birbiri ile tutarlı olduğu sonucuna varılmıştır. Sonuç olarak; performansın değerlendirilmesi sürecinde Çok Yüzeyle Rasch Ölçme modelinin farklılaşan puanlayıcı davranışlarının ve puanlayıcı güvenirliliğinin belirlenmesinde kullanılabilecek alternatif ölçme modeli olduğu açıklanmıştır.

Bilim kodu:6015

Anahtar Kelimeler: Farklılaşan Puanlayıcı Davranışları, Çok Yüzeyle Rasch Ölçme Modeli, Yazma Performansı, Analitik Dereceli Puanlama Anahtarı, Puanlayıcı Güvenirliliği.

Sayfa Adedi: xiv + 131

Danışman: Prof. Dr. İsmail KARAKAYA

EXAMINATION OF THE DIFFERENTIAL RATER BEHAVIOURS IN PERFORMANCE EVALUATION WITH MANY FACET RASCH MEASUREMENT

(M.S Thesis)

Cennet TOBAŞ

GAZI UNIVERSITY

GRADUATE SCHOOL OF EDUCATIONAL SCIENCES

June 2020

ABSTRACT

In this study, it is aimed to determine the rater bias and rater reliability with the Many-facet Rasch Measurement Model (MFRM) as well as the rater behaviors involved in the assessment of the writing performance for the evaluation of the performances of the 7th grade students of primary education. Considering that the rater behaviors and the biased behavior of the raters negatively affect both validity and reliability in the determination of individual performance, it is important to determine these behaviors and evaluate the results in this context. The study group of the research consists of 57 secondary school students studying at a seventh grade in a public school in Ankara in the second semester of the 2019-2020 academic year and 11 teachers who received rater education. In the current study, students were given seven different persuasive writing tasks and asked to choose one of them and create persuasive writings in accordance with this task. It was observed that 9 of the texts written by the students did not belong to the persuasive writing structure, and the texts of 48 students were evaluated by 11 teachers with an analytical grading key consisting of six criteria. In this research, the behaviors of the raters that appeared most in the evaluations based on teacher ratings (rater severity / leniency effect, central tendency effect, halo effect), as well as the effects of differential rater severity / leniency effect, namely scoring bias, and scoring reliability based on KTK, GK and MTK theories were examined. When the findings of the study were examined, the effect of the rater severity / leniency on the performance of the individuals was found during the evaluation of the students' writing skills; it was observed that there were no scoring behaviors such as central tendency effect and halo effect. When differential rater severity / leniency behavior, that is, bias interactions are examined; it was observed that biased scoring behaviors were not observed in group-level raters, but individual level raters showed biased scoring behavior.

When the rater reliability based on KTK, GK and MTK theories were examined, it was found that the reliability coefficients calculated using different methods were high; although the raters made different scores, it was concluded that the scores were consistent with each other. As a result; it has been explained that the Multi-Facet Rasch Measurement model is an alternative measurement model that can be used in determining different rater behaviors and rater reliability in the process of evaluating performance.

Science Code: 6015

Key Words: Differential Rater Behaviors, Many-Facet Rasch Measurement, Writing Performance, Analytic Rubric, Rater Reliability.

Page Number: xiv + 131

Supervisor: Prof. Dr. İsmail KARAKAYA

İÇİNDEKİLER

TELİF HAKKI ve TEZ FOTOKOPİ İZİN FORMU	i
ETİK İLKELERE UYGUNLUK BEYANI.....	ii
JÜRİ ONAY SAYFASI.....	iii
TEŞEKKÜR.....	iv
ÖZET	v
ABSTRACT	vii
İÇİNDEKİLER.....	ix
TABLolar LİSTESİ.....	xii
ŞEKİLLER LİSTESİ.....	xiii
SİMGELER VE KISALTMALAR LİSTESİ.....	xiv
BÖLÜM I	1
GİRİŞ.....	1
1.1. Problem Durumu.....	1
1.2. Araştırmanın Amacı.....	8
1.3. Araştırmanın Önemi	8
1.4. Problem İfadesi	10
1.5. Araştırma Soruları	10
1.6. Araştırmanın Sınırlılıkları.....	11
1.7. Araştırmanın Sayıltıları	11
BÖLÜM II.....	13
KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR.....	13
2.1. KURAMSAL ÇERÇEVE.....	13
2.1.1. Performans Değerlendirme	13
2.1.1.1. İkna Edici Metin Yazma Görevleri	14
2.1.1.2. Dereceli Puanlama Anahtarları (DPA-Rubrikler)	16
<i>Bütünsel Dereceli Puanlama Anahtarı (Holistik Rubrik)</i>	18

<i>Analitik Dereceli Puanlama Anahtarı (Analitik Rubrik)</i>	19
2.1.2. Puanlayıcı Davranışları.....	20
2.1.3. Puanlayıcı Güvenirliği (Inter-Rater Reliability)	24
2.1.3.1. Klasik Test Kuramına Dayalı Yöntemler	25
<i>Sınıf İçi Korelasyon (ICC) Katsayısı (Intraclass Correlation Coefficient)</i>	26
2.1.3.2. Genellenebilirlik Kuramına Dayalı Yöntemler	27
2.1.3.3. Madde Tepki Kuramına Dayalı Yöntemler	28
2.1.4. Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM)	29
<i>Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM)'nin Sayıtları</i>	31
<i>Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM)'nin Analizi ve Yorumlanması</i>	32
2.2. İLGİLİ ARAŞTIRMALAR.....	39
2.2.1 Puanlayıcı Davranışları ile İlgili Araştırmalar	39
2.2.2. Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM) ile İlgili Araştırmalar	46
BÖLÜM III	51
YÖNTEM.....	51
3.1. Araştırmanın Deseni.....	51
3.2. Çalışma Grubu.....	51
3.3. Veri Toplama Araçları.....	53
3.3.1. İkna Edici Metin Yazma Görevi.....	53
3.3.2. Analitik Dereceli Puanlama Anahtarları (Rubrik).....	54
<i>Geliştirilen Ölçme Araçlarının Geçerlik ve Güvenirliğine İlişkin Açıklamalar</i> ...	55
3.3.3. Puanlayıcı Formu (Öğretmen Değerlendirme Formu)	58
3.4. Verilerin Toplanması	58
3.5. Verilerin Analizi	59
BÖLÜM IV	61
BULGULAR VE YORUM	61
4.1. Birinci Araştırma Sorusuna İlişkin Bulgular ve Yorum	61
4.1.1. Alt Problem 1.a'ya İlişkin Bulgular ve Yorum	62
4.1.2. Alt Problem 1.b'ye İlişkin Bulgular ve Yorum	64
4.1.3. Alt Problem 1.c'ye İlişkin Bulgular ve Yorum.....	66
4.2. İkinci Araştırma Sorusuna İlişkin Bulgular ve Yorum	71
4.2.1. Alt Problem 2.a'ya İlişkin Bulgular ve Yorum	72
4.2.2. Alt Problem 2.b'ye İlişkin Bulgular ve Yorum	74

4.2.3. Alt Problem 2.c'ye İlişkin Bulgular ve Yorum.....	76
4.2.4. Alt Problem 2.d'ye İlişkin Bulgular ve Yorum	77
4.2.5. Alt Problem 2.e'ye İlişkin Bulgular ve Yorum.....	79
4.3. Üçüncü Araştırma Sorusuna İlişkin Bulgular ve Yorum	82
BÖLÜM V	87
SONUÇ VE ÖNERİLER.....	87
5.1. SONUÇ	87
5.1.1 Birinci Araştırma Sorusuna İlişkin Sonuçlar	87
5.1.2 İkinci Araştırma Sorusuna İlişkin Sonuçlar	89
5.1.3 Üçüncü Araştırma Sorusuna İlişkin Sonuçlar	90
5.2 ÖNERİLER.....	90
5.2.1. Uygulamaya Yönelik Öneriler.....	90
5.2.2. Yapılacak Çalışmalara Yönelik Öneriler	91
KAYNAKLAR.....	93
EKLER.....	109
EK 1. Evcil Hayvan Besleme	110
EK 2. Eve İnternet Bağlatmak	111
EK 3. Çanakkale'ye Gezi	112
EK 4. Okul Takımı Seçmeleri.....	113
EK 5. Geri Dönüşüm (Katı Atıklar).....	114
EK 6. Kişisel Temizlik	115
EK 7. Benim Kararım	116
EK 8. Yönerge	117
EK 9. Öğretmen Değerlendirme Formu.....	118
EK 10. Analitik Puanlama Anahtarı.....	120
EK 11. Uzman Değerlendirme Formu	121
EK 12. Meb İzin Belgesi	124
EK 13. Öğrencilerin Yazma Çalışmalarından Örnekler	125
EK 14. Öğrenci Yüzeyine İlişkin Ölçüm Raporu	128
EK 15. “Puanlayıcı x Ölçüt” Etkileşimine Yönelik Ölçüm Raporu.....	129
EK 16. “Puanlayıcı x Görev” Etkileşimine Yönelik Ölçüm Raporu	130

TABLolar LİSTESİ

Tablo 2.1 <i>Klasik Test Kuramına Dayalı Puanlayıcı Güvenirliği Hesaplama Yöntemleri...</i>	25
Tablo 3.1 <i>Öğrencilerin Cinsiyet ve Sınıf Değişkenlerine Göre Dağılımı</i>	52
Tablo 3.2 <i>Puanlayıcıların Cinsiyet ve Sınıf Değişkenlerine Göre Dağılımı</i>	52
Tablo 3.3 <i>Yazma Görevleri ve Puanlama Anahtarı Kapsam Geçerlik İndeksleri (KGI)</i>	56
Tablo 4.1 <i>Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu</i>	62
Tablo 4.2 <i>Ölçüt Güçlükleri Eşitlendiğinde Ölçüt Yüzeyine İlişkin Ölçüm Raporu</i>	65
Tablo 4.3 <i>Analitik Puanlama Anahtarının Puanlama Kategorilerine İlişkin Ölçüm Raporu</i>	67
Tablo 4.4 <i>Öğrenci Yüzeyine İlişkin Ölçüm Raporu</i>	68
Tablo 4.5 <i>Ölçüt Yüzeyine İlişkin Ölçüm Raporu</i>	69
Tablo 4.6 <i>“Puanlayıcı x Öğrenci” Yüzeylerinin Etkileşimine Yönelik Ölçüm Sonuçları....</i>	72
Tablo 4.7 <i>“Puanlayıcı x Ölçüt” Yüzeylerinin Etkileşimine Yönelik Ölçüm Sonuçları.....</i>	74
Tablo 4.8 <i>“Puanlayıcı x Cinsiyet” Yüzeylerinin Etkileşimine Yönelik Ölçüm Sonuçları....</i>	76
Tablo 4.9 <i>“Puanlayıcı x Görev” Yüzeylerinin Etkileşimine Yönelik Ölçüm Sonuçları.....</i>	78
Tablo 4.10 <i>“Puanlayıcı x Öğrenci x Ölçüt” Yüzeylerinin Etkileşimine Yönelik Ölçüm Sonuçları</i>	80
Tablo 4.11 <i>Farklı Ölçme Kuramlarına Dayalı Olarak Hesaplanan Puanlayıcılar-Arası Güvenirlik Katsayıları</i>	83

ŞEKİLLER LİSTESİ

<i>Şekil 2.1.</i> Dereceli puanlama anahtarı (Kutlu, Doğan ve Karakaya, 2014, s.54).....	17
<i>Şekil 4.1.</i> Dereceli puanlama anahtarının ölçütlerine ilişkin kategori olasılık eğrileri	70

SİMGELER VE KISALTMALAR LİSTESİ

AFA	Açımlayıcı Faktör Analizi
ÇYRÖM	Çok Yüzeyli Rasch Ölçme Modeli
DIF	Differential Item Function
DRF	Differential Rater Function
DPA	Dereceli Puanlama Anahtarı
GK	Genellenebilirlik Kuramı
KTK	Klasik Test Kuramı
MTK	Madde Tepki Kuramı
MKE	Madde Karakteristik Eğrisi
KGİ	Kapsam Geçerlik İndeksi
MEB	Milli Eğitim Bakanlığı
RMSE	Hata kareleri ortalamasının karekökü
ω	Omega
χ^2	Ki-kare
Φ	Phi (Fi)
N	Birey sayısı
ss	Standart sapma
sd	Serbestlik derecesi
p, α	Anlamlılık düzeyi

BÖLÜM I

GİRİŞ

Bu bölümde araştırmanın problem durumuna, amacına, önemine, problem ifadesine, araştırma sorularına, sınırlılıklarına ve sayıtlarına ilişkin bilgilere yer verilmiştir.

1.1. Problem Durumu

Öğrencilerin üst düzey zihinsel becerilerine dayalı olarak performansının belirlenmesinde en temel öge eğitimde ölçme ve değerlendirmedir. Bilim ve teknolojiye gelişmeler eğitim ve öğretimi doğrudan etkilemiş, geleneksel ölçme ve değerlendirme yöntemlerinden (yazılı yoklama, çoktan seçmeli, D-Y tipi sorular...vs.) daha çok, üst düzey zihinsel becerilerin ölçülmesine ve değerlendirilmesine yönelik performansa dayalı durum belirleme çalışmalarının (proje, performans görevi, e-portfolyo gibi) kullanımı yaygınlaşmaya başlamıştır. 21. yüzyılda bilim ve teknolojiye yaşanan hızlı değişim ile birlikte bireyin ve toplumun değişen ihtiyaçları, öğrenme-öğretme yaklaşımlarındaki gelişmeler bireylerden beklenen rolleri doğrudan etkilemiştir. Bu değişim bilgiyi üreten, öğrendiği bilgileri gerçek yaşam durumlarına aktarabilen, problem çözebilen, eleştirel ve yaratıcı düşünebilen, girişimci, kararlı, iletişim becerilerine sahip, empati yapabilen, topluma ve kültüre katkı sağlayan bireyleri tanımlamaktadır (MEB, 2019). Bireylerin bu özelliklerine bağlı olarak öğretim programlarında üst düzey zihinsel becerilerin geliştirilmesi gerekliliği ortaya çıkmıştır. Üst düzey zihinsel beceriler, bireyin birden fazla beceriyi kendi bireysel özellikleri ile ilişkilendirmesidir (Marzano, 2001). Üst düzey zihinsel becerileri gelişmiş bireyler, öğrenmiş olduğu bilgileri gerçek yaşama aktarabilmekte ve karşılaştığı sorunları çözmede bu bilgileri etkili biçimde kullanabilmektedir (Kutlu, Doğan ve Karakaya, 2014,

s.15). Bu bağlamda yenilenen öğretim programlarında daha çok üst düzey zihinsel becerilerin ölçülmesine ve değerlendirilmesine odaklanıldığı görülmektedir.

Üst düzey zihinsel becerilerin ölçülmesinde ve değerlendirilmesinde eğitim ortamında performansa dayalı durum belirleme çalışmalarından (proje, performans görevi, e-portfolio gibi) yararlanılmaktadır. Öğrenci başarısının gelişimini sağlamak amacıyla yapılan üst düzey zihinsel çaba gerektiren çalışmalar *performansa dayalı durum belirleme* olarak adlandırılmaktadır (Kutlu, Doğan ve Karakaya, 2014, s.30). Performansa dayalı durum belirleme, öğrencinin gerçek yaşam koşullarında, karmaşık görevleri yaparken öğrendiği bilgileri ne kadar iyi kullandığını ölçmeye çalışmaktadır (Erman Aslanoğlu ve Kutlu, 2003). Performansa dayalı durum belirleme çalışmalarında öğrencilerin basit, yalın, alt düzey düşünme gerektiren görevleri değil, üst düzey düşünme gerektiren karmaşık yapıdaki bazı görevleri yerine getirmeleri istenmektedir (Kutlu, Doğan ve Karakaya, 2014; s.30). Karmaşık yapıdaki bu üst düzey zihinsel becerilerin değerlendirilmesinde performans değerlendirme araçları kullanılmaktadır. Üst düzey zihinsel becerilerin değerlendirilmesinde performans değerlendirme araçları; karmaşık durumların ölçülmesine yardımcı olmakla birlikte bireyin öğrendiklerini gerçek yaşam durumlarına aktarabilmesine olanak tanımaktadır. Ayrıca performansa dayalı değerlendirmeler sadece öğretim sonunda değil öğretim süreci devam ederken bile kullanılabilir. Bu anlamda öğrenci tamamen öğrenme sürecinin içerisinde aktif olarak yer aldığından öğrenmeleri daha kalıcı hale gelmektedir. Bu haliyle performansa durum belirleme çalışmaları ile geleneksel ölçme ve değerlendirme yöntemlerine nazaran daha geçerli ve güvenilir sonuçlar elde edilebilmektedir.

Her ne kadar performansa dayalı durum belirleme araçlarının kullanımının, geleneksel ölçme ve değerlendirme araçlarına göre çok sayıda avantajı bulunmasına rağmen; bu tip becerilerin ölçülmesi, değerlendirilmesi oldukça karmaşık ve zaman alıcıdır. Bu yüzden performansın değerlendirilmesinde; puanlayıcılardan, ölçme aracından, bireylerden veya ölçmenin yapıldığı ortamın koşullarından kaynaklanan pek çok hata karışabilmektedir. Bu durum yapılan ölçümlerin geçerliği ve güvenilirliği açısından tehdit oluşturmaktadır. Yani; üst düzey zihinsel performansın ölçülmesinde ve değerlendirilmesinde daha alt düzey zihinsel becerilerin ölçümüne olanak sağlayan geleneksel ölçme ve değerlendirme uygulamalarına göre daha fazla ölçme hatası karıştığı görülmektedir. Ölçme sonuçlarına karışan hatalar genel olarak; ölçmeyi yapan kişilerden, ölçmenin yapıldığı araçtan, ölçme ortamından ve ölçmenin yapıldığı gruptan kaynaklanabilir. (Atılğan, Kan ve Doğan, 2011;

s.27). Bu hata kaynaklarından özellikle puanlayıcıdan kaynaklı hatalar yapılan ölçümlerin geçerliğini ve güvenilirliği olumsuz yönde etkilemektedir. Ölçmeyi yapan kişinin yani puanlayıcının; ölçme işlemi esnasında dikkatsiz olması, yorgun olması veya kağıdını okuduğu öğrencinin ismini görünce yanlış not vermesi gibi hatalı davranışlarda bulunması söz konusu olabilir. Ölçümlerin güvenilir olmaması, ölçme hatalarının miktarının göstergesidir ve elde edilen puanların yorumlanmasında ölçme hatalarının miktarının bilinmesi, o sonuçlara dayanarak doğru karar vermek açısından çok büyük öneme sahiptir (Atılğan, Kan ve Doğan, 2011; s. 33-34).

Performansa dayalı durum belirlemede yapılan ölçümlerin geçerliği ve güvenilirliğini sağlama açısından, kontrol listeleri ve dereceli puanlama anahtarları kullanılmaktadır. Puanlama ölçeklerinden dereceli puanlama anahtarları performansa dayalı durum belirlemede sıklıkla kullanılan araçlardandır (Russell ve Airasian, 2011, s.223). Performansa dayalı durum belirleme çalışmalarında yapılan ölçümlerin geçerli ve güvenilir olması için sıklıkla dereceli puanlama anahtarlarından yararlanılmaktadır. Performansa dayalı değerlendirmelerde bireylerin hangi ölçütlere göre nasıl puanlanacağını belirlemek için dereceli puanlama anahtarları kullanılmaktadır. Dereceli puanlama anahtarları (DPA) bir göreve ilişkin ölçüt listesini ve bu ölçütlere ilişkin niteliklerin derecesini içeren puanlama araçlarıdır (Goodrich, 1997). Dereceli puanlama anahtarları diğer bir adıyla rubrikler bireylerin performanslarının değerlendirilmesi açısından birçok avantaja sahiptir. Dereceli puanlama anahtarları ölçülecek performansın yapısına ve değerlendirmelerin amacına bağlı olarak, holistik (bütüncül) dereceli puanlama anahtarı ve analitik dereceli puanlama anahtarı olmak üzere iki çeşittir (Atılğan, Kan ve Doğan, 2011, s.282; Mertler, 2001; Korkmaz, 2004, s.232; Parlak ve Doğan, 2014; Popham, 1997). Holistik (bütüncül) DPA; ölçülecek performans farklı boyutlara ayırmadan, özelliklerin tümünün farklı düzeyler için tanımlaması yapıldığında kullanılan bir ölçme aracıdır (Nitko ve Brookhart, 2014). Buna karşılık analitik DPA, gözlemlere ait puanları, tanımlanmış kategorilerden (ölçüt ya da ölçütler) uygun düşen boyuta kaydetmemizi sağlayan bir ölçme aracıdır (Haladyna, 1997). Analitik DPA bireylerin performansını süreç temelli olarak holistik DPA'ya göre daha ayrıntılı değerlendirmesinden dolayı bireylerin performansındaki eksik ve zayıf noktalarının belirlenmesinde daha etkilidir. Bu bağlamda performansa dayalı durum belirleme çalışmalarında değerlendirme sürecinde puanlamaların geçerliği ve güvenilirliği açısından kontrol listeleri ve dereceli puanlama anahtarlarının kullanımı önem taşımaktadır.

Performansa dayalı durum belirlemede en önemli hata kaynağı, *puanlayıcılardan kaynaklı hatalar*'dır (Goodwin, 2001). Puanlayıcılardan kaynaklı hatalara literatürde “*puanlayıcı davranışları*” veya “*puanlayıcı etkileri*” de denilmektedir. Bu hatalar; bir puanlayıcının aynı bireylerin puanını farklı zamanlardaki ölçümü arasındaki tutarsızlıklar, farklı puanlayıcıların aynı bireylerin performanslarını puanlarken ölçümler arasındaki tutarsızlıklar, halo etkisi, merkeze eğilim, ranj daralması, puanlayıcı katılık/cömertlik, farklılaşan puanlayıcı katılık/cömertlik davranışı (puanlayıcı yanlılığı) (Myford ve Wolfe, 2003; 2004)'dir. Bu tip davranışların ortaya çıkmasında sebep olarak; puanlayıcıların ölçme aracını tam olarak anlayamaması, puanlamanın objektif olmaması ve puanlayıcı eğitimlerinin yapılmaması gösterilebilir. Bu puanlayıcı davranışları puanlayıcılar arası tutarlılığı dolayısıyla “*puanlayıcılar arası güvenilirliği*” doğrudan etkilemektedir. Özellikle üst düzey zihinsel özelliklerin ölçülmesinde, ölçmenin geçerliğinin ve güvenilirliğinin artırılması için puanlayıcı güvenilirliğinin sağlanması gereklidir. Bunun için puanlamalarda dereceli puanlama anahtarlarının kullanılması (Aslanoğlu, 2003; Dunbar, Brooks ve Miller, 2006; Kutlu, Doğan ve Karakaya, 2014; Wolf ve Stevens, 2007), birden fazla puanlayıcı kullanılması (Ebel, 1951; Kubiszyn ve Borich, 2013), puanlama anahtarlarının değerlendirme ölçütlerinin açık ve anlaşılır olması (Moskal, 2000) puanlayıcı güvenilirliğinin sağlanması için önem taşımaktadır.

Puanlayıcı güvenilirliği genel olarak puanlayıcılar tarafından verilen puanların tutarlığına dayanır ve en az iki bağımsız puanlayıcı ya da farklı zamanlarda aynı puanlayıcıların puanlandırılmasıyla ilişkilidir (Ada, 2015). Puanlayıcı güvenilirliği, *puanlayıcı içi güvenilirlik* (intrarater reliability) ve *puanlayıcılar arası güvenilirlik* (interrater reliability) olmak üzere iki temel başlıkta incelenmektedir (Jonsson ve Svingby, 2007). Birbirinden bağımsız en az iki puanlayıcının puanları arasındaki tutarlılığın göstergesi puanlayıcılar-arası güvenilirlik (interrater reliability) olarak tanımlanırken; aynı puanlayıcının farklı zamanlarda puanları arasındaki tutarlılığın göstergesi ise puanlayıcı-içi güvenilirlik (inrarater reliability) olarak tanımlanmaktadır (Crocker ve Algina, 2008, s.104). Performans değerlendirmede bir öğrencinin performansına yönelik ölçümler birden fazla puanlayıcı tarafından puanlanmakta olup, puanlayıcılar arasındaki tutarlılık *puanlayıcılar arası güvenilirlik katsayısı* ile yorumlanmaktadır. Yapılan ölçümlerin geçerliği ve güvenilirliği açısından hayati bir öneme sahip olan puanlayıcılar arası güvenilirlik katsayısı, ölçmenin üç temel kuramı olan klasik test kuramı (KTK), madde tepki kuramı (MTK) ve genellenebilirlik kuramı (GK)'na bağlı olarak farklı yöntemlerle belirlenmeye çalışılmaktadır. Klasik test

kuramına dayalı güvenilirlik katsayısı hesaplamada aynı bireyin tekrarlı ölçmeler sonucu tek bir hata kaynağına yönelik hesaplamalar yapılırken genellenebilirlik kuramında güvenilirlik hesaplamada birden fazla hata kaynağı ve bunlar arasındaki etkileşim birlikte göz önünde bulundurulmaktadır. Bu iki kuramın sınırlılıklarını ortadan kaldırmaya yarayan madde tepki kuramı ise bireyin gözlenemeyen yetenekleri ile ölçme aracındaki maddeler arasındaki ilişkiyi matematiksel bağıntılarla açıklayarak güvenilirlik katsayısı hesaplayabilmektedir. Güvenirlik hesaplamada madde tepki kuramına dayalı olarak bir parametrelili Rasch ölçme modeli kullanılmaktadır. Bu model ile hesaplanan güvenilirlik katsayısı klasik test kuramı ve genellenebilirlik kuramına dayalı yöntemlerle hesaplanan güvenilirlik katsayısından daha yüksektir.

Madde tepki kuramı (MTK) ve özellikle Rasch ölçme teorisindeki son gelişmeler, derecelendirme kalitesini değerlendirmek ve puanlayıcı hatalarını tespit etmek için teorik olarak sağlam ve model tabanlı yöntemler geliştirme fırsatı sunmaktadır (Tindal ve Haladyna, 2002; s.262). Bu yöntemlerden bir tanesi de Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM)'dir. Güvenirlik hesaplamada değişkenlik kaynaklarının sayısına göre Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM) de kullanılmaktadır. Bu model ikiden fazla değişkenlik kaynağını ele alarak Rasch modelinin daha da genişletilmiş bir halidir (Nakamura, 2000). Ayrıca bu model her bir değişkenlik kaynağının her bir yüzeyine ilişkin hesaplamalar yapar ve böylece her bir değişkenlik kaynağındaki yüzeylerin (bireyler, maddeler ve puanlayıcılar) beklenmedik değerleri ya da istenmedik durumlardaki performansların belirlenmesini sağlar (Alharby, 2006). Diğer ölçme kuramlarında (klasik test kuramı ve genellenebilirlik kuramında) tek bir ölçme durumu için bir güvenilirlik katsayısı kestirilirken, madde tepki kuramına dayalı olarak Çok Yüzeyli Rasch Ölçme modelinde her bir değişkenlik kaynağı için aynı anda ayrı ayrı güvenilirlik katsayıları kestirilebilmektedir (Güler, 2008). ÇYRÖM; puanlayıcıların performans değerlendirmede çeşitli değişkenleri modellemelerini, puanlar üzerindeki etkilerini tahmin etmelerini ve karşılaştırma için aynı logit ölçeğine yerleştirmelerini sağlar. Her bir değişkenlik kaynağı, bir dereceli puanlama anahtarı kullanıldığında ve tüm yüzeyler (teste giren, görev, puanlayıcı, vb.) değişken veya kalibrasyon haritası adı verilen tek bir doğrusal skalaya yerleştirildiğinde ham puanlar, potansiyel olarak sıralamalı ölçek düzeyinde kalibre edilir (Barkaoui, 2013). Örneğin, çok sayıda yazma görevine cevap vermeyi ve daha sonra çok sayıda puanlayıcı tarafından puanlanan makaleleri dereceli puanlama anahtarı ile puanlanan bir yazma çalışmasında, dört tane yüzey vardır: teste girenler, görevler,

puanlayıcılar ve değerlendirme ölçütleri. ÇYRÖM, her bir skoru, teste girenlerin yeteneği, görev zorluğu, ölçüt güçlüğü ve puanlayıcının katılığı etkileşiminin bir fonksiyonu olarak görür (McNamara, 1996). Modele bağlı olarak kalibrasyon haritası üzerinde puanlayıcı katılığını/cömertliğini, teste girenlerin yeteneğini, görev zorluğunu, ölçüt zorluğunu ve ölçek adım zorluğunu tahmin etmektedir. Bu bağlamda ÇYRÖM, araştırmaya dahil edilen bütün değişkenlik kaynakları (yüzeyler) hakkında ayrıntılı bilgi sağlaması açısından önemli bir yere sahiptir.

Madde tepki kuramına dayalı olarak geliştirilen ÇYRÖM, diğer ölçme kuramlarına bağlı yöntemlere göre bir takım üstün özelliklere sahiptir. ÇYRÖM, öğrencilerin performansının değerlendirilmesinde, bireylerin puanlarına etki edebilecek faktörler bireylerin yetenek düzeyleri ve madde güçlük düzeyleri ile sınırlı kalmamakla birlikte puanlayıcılar ile ilgili diğer değişkenlik kaynakları da bireylerin performansına etki edebilmektedir (Baird, Hayes, Johnson, Johnson ve Lamprianou, 2013). Buna ek olarak ÇYRÖM; ölçümlere ait her bir yüzey için hem bireysel hem de grup bazında bilgi sağlamasının yanı sıra, puanlayıcı ve değerlendirmeye alınan her bir birey için grup düzeyinde bilgi vermek yerine bireysel düzeyde yorumlamalar yapılmasına olanak tanımaktadır. Bu özelliği ile ÇYRÖM puanlayıcı davranışından etkilenen performans değerlendirmeleri için uygun bir model olmaktadır (Mulqueen, Baker ve Dismukes, 2000). Bu model test geliştiricilerin ve kullanıcıların, çeşitli tipte puanlayıcı davranışlarını; halo etkisi, puanlayıcı katılık/cömertlik davranışı, ranj daralması, merkeze eğilim davranışı, tutarsızlık, farklılaşan puanlayıcı katılık/cömertlik davranışlarının belirlenebilmesine olanak tanımaktadır (Myford ve Wolfe, 2003, 2004). Buna ek olarak bu model ile farklı hata kaynakları arasındaki etkileşimler de belirlenebilmektedir (Haiyang, 2010). “Bireyler, görevler, maddeler, puanlayıcılar, zaman” gibi değişkenlik kaynakları ve bunlar arasındaki etkileşim performansın ölçülmesiyle elde edilen sonuçların güvenilirliğini etkileyebilmektedir (Güler, 2008). Bu sayede aynı anda birden fazla hata kaynağı dikkate alınarak geçerlik için daha yüksek yetenek kestirimleri üretilebilmektedir (İlhan, 2016). Bu özellikleri ile ÇYRÖM, klasik test kuramı ve genellenebilirlik kuramına göre psikometrik açıdan daha güçlü bir model olarak kabul edilmektedir.

En önemli hata kaynağı olarak görülen puanlayıcıdan kaynaklı yanlı puanlayıcı etkileşimleri (rater bias effects) diğer bir adıyla farklılaşan puanlayıcı katılık/cömertlik davranışı (differential rater severity/leniency function- DRF) yapılan ölçümlerin geçerliği ve güvenilirliğini olumsuz yönde etkilemektedir. Puanlayıcı yanlılığı (rater bias);

puanlayıcının değerlendirme işlemini yaparken bireylerin cinsiyet, yaş, kültürel faktörler gibi çeşitli özelliklerine bağlı olarak bazı bireylere diğerlerine göre daha yüksek ya da daha düşük puanlar verme eğiliminde olması şeklinde tanımlanmaktadır (Kumar, 2005). Bu hata kaynağı da genel olarak yapılan ölçümlerde sistematik hatalara yol açmaktadır. Yani; puanlayıcıların bir takım çevresel faktörlere dayalı olarak bazı öğrencilere performansının altında veya üstünde puanlar vermesi, yapılan ölçümlerin tutarlılığını etkilemektedir. Puanlamadaki tutarsızlıklar öğrenci performansının geçerliğini ve güvenilirliğini olumsuz yönde etkilemektedir. Bu olumsuz durumun ortadan kaldırılmasına yönelik; değerlendirme sürecinde birden fazla puanlayıcının kullanılması, puanlama anahtarının ölçütlerinin açık ve anlaşılır olması, puanlayıcıların değerlendirme öncesi puanlayıcı eğitimine alınması gibi öneriler sunulmaktadır. Yapılan ölçümlerde yanlış puanlayıcı etkileşimlerinin tespitinde madde tepki kuramı (MTK)'na dayalı olarak Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM) kullanılmaktadır. Bu modele ait istatistiksel veriler FACETS 3.81.0 paket programı ile analiz edilmektedir. Puanlayıcıdan kaynaklı hataların tespitinde puanlayıcı yüzeyi ile diğer değişkenlik kaynakları (Birey, görevler, maddeler, ölçütler, zaman, vs.) arasındaki ikili ve daha fazla değişkenlik kaynağı arasındaki etkileşime bakılmaktadır. Puanlayıcı yüzeyi ile diğer yüzeylerin etkileşimlerinin sonucunda güvenirliliğin düşük ve χ^2 (Ki-kare) istatistiğinin anlamlı olmaması beklenir. Bu durum puanlayıcıların değerlendirmelerinde tutarlı puanlamalar yaptığını, yanlış puanlamalar yapmadığını yani puanlamaların geçerliği ve güvenirliliğinin sağlandığını göstermektedir.

Alanyazın incelendiğinde, birey performansının birden fazla puanlayıcı tarafından değerlendirilmesi sonucu birey performansına puanlayıcı kaynaklı ilişkisiz varyansın ölçümlere karıştığı görülmektedir. Birey performansının belirlenmesinde en önemli hata kaynağı olan puanlayıcılardan kaynaklı hatalardır. Buna bağlı olarak ölçümlerin geçerliği ve güvenirliliğinin olumsuz yönde etkilenmektedir. Bu bağlamda mevcut çalışmada ilköğretim 7. sınıf öğrencilerinin üst düzey zihinsel becerilerin gelişimine olanak tanıyan ikna edici metin yazma performansına ait ölçümlerin geçerliği ve güvenirliliğini olumsuz yönde etkileyen puanlayıcı davranışlarından; puanlayıcı katılık/cömertlik davranışı, halo etkisi, ranj daralması ve merkeze eğilim davranışı madde tepki kuramına dayalı olarak hesaplanan ÇYRÖM ile analiz edilecektir. Ayrıca puanlayıcıların değerlendirme sürecinde ölçüm sonuçlarını etkileyebilecek diğer değişkenlik kaynakları ile olan etkileşim etkileri, farklılaşan puanlayıcı katılık/cömertlik davranışı yani puanlayıcı yanlılığı açısından ayrıntılı olarak yorumlanacaktır. Ölçümlerin güvenirliliğine yönelik olarak klasik test kuramına

dayalı yöntemlerden sınıf içi korelasyon katsayısı (ICC), genellenebilirlik kuramına bağlı olarak hesaplanan genellenebilirlik katsayısı ve madde tepki kuramında Rasch analizine dayalı olarak hesaplanan güvenilirlik katsayısı hesaplanarak sonuçlar karşılaştırılması planlanmaktadır.

1.2. Araştırmanın Amacı

Bu çalışmada ilköğretim öğrencilerinin yazma performanslarının değerlendirilmesi sürecinde öğretmen puanlamalarına karışan puanlayıcı davranışlarının yanı sıra farklılaşan puanlayıcı katılık/cömertlik davranışının (puanlayıcı yanlılıklarının) ve puanlayıcı güvenilirliğinin belirlenmesi amaçlanmaktadır. Bu amaç doğrultusunda öğrencilerin ikna edici yazma performanslarının değerlendirilmesine yönelik araştırmacı tarafından geliştirilen ilköğretim 7. sınıf öğrencilerine ikna edici metin yazma görevi verilip öğrencilerin yazma performansları analitik dereceli puanlama anahtarı kullanılarak puanlayıcı eğitimi alan öğretmenler tarafından puanlanmıştır. Bu çalışmada öğretmen puanlarından elde edilen veriler FACETS 3.81.0 (Linacre, 2018) paket programı kullanılarak ÇYRÖM ile analiz edilip sonuçlar puanlayıcı davranışları, farklılaşan puanlayıcı katılık/cömertlik davranışı (puanlayıcı yanlılık etkileşimleri) ve puanlayıcı güvenilirliği bağlamında yorumlanmıştır. Ayrıca, puanlayıcı güvenilirliğine yönelik KTK, GK ve MTK bağlamında ayrı ayrı güvenilirlik katsayıları hesaplanıp sonuçlar karşılaştırılmıştır.

1.3. Araştırmanın Önemi

Son yıllarda eğitim alanında meydana gelen gelişmeler, üst düzey zihinsel süreçlerinin belirlenmesini amaçlayan, öğretim ve durum belirleme süreçlerini birbirine yakınlaştıran ve okul öğrenmelerinin yaşam durumlarda gözlenmesine olanak tanıyan performansa dayalı durum belirleme (Doğan ve Kutlu, 2011); proje, performans görevi, e-portfolyo gibi yeni durum belirleme yöntemlerinin önemini artırmıştır. Üst düzey zihinsel özelliklerin ölçülmesinde etkili olan performansa dayalı durum belirleme araçlarından yazma performansı bireylerin yazılı anlatım, sözlü anlatım, sunu yapma, kompozisyon oluşturma, makale yazma (Arı, 2010) gibi etkinlikleri içermektedir. Bu etkinlikler bireylerin eleştirel düşünme, yaratıcı düşünme ve yansıtıcı düşünme becerilerinin gelişmesini sağlamaktadır. Bu becerilerin puanlanmasında eğitimde sıklıkla performansa dayalı ölçme araçlarından

dereceli puanlama anahtarları kullanılmaktadır. Üst düzey zihinsel özelliklerin ölçülmesinde bireylerin yazma performanslarının ölçülmesine yönelik dereceli puanlama anahtarlarının kullanılması ölçümlerin geçerliği ve güvenirliği açısından önem taşımaktadır. Mevcut çalışmada performansın değerlendirilmesine yönelik ilköğretim öğrencilerinin yazma performanslarının belirlenmesinde ölçümlerin geçerlik ve güvenirliğinin sağlanması açısından analitik dereceli puanlama anahtarları kullanılmıştır.

Üst düzey zihinsel becerilerin belirlenmesine yönelik bireylerin yazma performansının değerlendirilmesinde birden fazla puanlayıcının kullanıldığı durumlarda puanlamalar arasında tutarsızlıklar olabilmektedir. Bu durum puanlayıcılar-arası güvenirliği olumsuz yönde etkilemektedir. Puanlayıcılar-arası güvenirliğin KTK, GK ve MTK'ya dayalı yöntemlerle belirlenmesi önem taşımaktadır. Bu bağlamda mevcut çalışmada; puanlayıcılar-arası güvenirliğin belirlenmesinde KTK'ya dayalı güvenirlik belirleme yöntemlerinden ölçme aracının ve veri setinin yapısına uygun olarak sınıf-içi korelasyon (ICC) katsayısı, GK'nda ANOVA'ya dayalı genellenebilirlik katsayısı ve MTK'ya dayalı yöntemlerden Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM) kullanılmıştır. Elde edilen sonuçlar puanlayıcı güvenirliği açısından yorumlanmış ve karşılaştırmalar yapılmıştır. Bu bağlamda mevcut çalışma farklı ölçme kuramlarına dayalı olarak hesaplanan güvenirlik katsayılarının puanlayıcı güvenirliği açısından ayrıntılı bilgiler sağlayacağı düşünülmektedir.

Performansın değerlendirilmesine yönelik ölçüm sonuçları incelendiğinde, birden fazla puanlayıcının tüm performansları veya bir puanlayıcının aynı performansı farklı zamanlarda puanladığı durumlarda puanlamalar arasında oluşan tutarsızlıklarda puanlayıcılar önemli bir hata kaynağı olarak görülmektedir. Puanlayıcılardan kaynaklı hatalara literatürde *puanlayıcı hataları*, *puanlayıcı davranışları*, *puanlayıcı etkileri* veya *puanlayıcı yanlılığı* denilmektedir. Ölçümlerin geçerliği ve güvenirliği açısından puanlayıcı davranışlarının belirlenmesi ve hata kaynaklarının düzeltilmesi önem taşımaktadır. Bu bağlamda mevcut çalışmada puanlayıcı davranışlarından; puanlayıcı katılık/cömertlik davranışı, halo etkisi, merkeze eğilim davranışı ve farklılaşan puanlayıcı katılık/cömertlik davranışı (yanlılık etkileşimi) ÇYRÖM ile analiz edilecektir. Ayrıca puanlayıcıların değerlendirme sürecinde ölçüm sonuçlarını etkileyebilecek diğer değişkenlik kaynakları ile olan etkileşim etkileri, puanlayıcı yanlılığı açısından yorumlanacaktır. Mevcut çalışmada farklı puanlayıcıların öğrencilerin yazma becerilerini değerlendirmesi sürecinde çeşitli faktörlerin (cinsiyet, ırk, yazı güzelliği vs.) etkisinde

kalarak kimi öğrencilere beklenen performansından daha yüksek kimi öğrencilere ise beklenen performansından daha düşük notlar vermesi yani puanlayıcı yanlılık etkileşimlerinin belirlenmesi önem taşımaktadır. Buna ek olarak, gerek yurtiçinde gerekse yurtdışında üst düzey zihinsel becerilerin geliştirilmesine yönelik performansa dayalı durum belirleme çalışmalarında yapılan analizlerde madde tepki kuramına dayalı olarak ÇYRÖM'nin son yıllarda sıklıkla kullanıldığı görülmektedir. Bu çalışmada da performansın belirlenmesinde yazma becerilerinin değerlendirilmesinde madde tepki kuramı bağlamında ÇYRÖM kullanılarak puanlayıcı davranışları, puanlayıcı yanlılık etkileşimleri ve puanlayıcı güvenilirliği ayrıntılı bir şekilde incelenecektir. Mevcut çalışmanın ilköğretim öğrencilerinin üst düzey zihinsel becerilerin gelişimine olanak tanıyan ikna edici metin yazma performansının değerlendirilmesinde farklı puanlayıcıların puanlamalarına bağlı olarak puanlayıcı hatalarının belirlenmesinde ÇYRÖM kullanılarak puanlayıcı davranışları, farklılaşan puanlayıcı katılık/cömertlik davranışı (puanlayıcı yanlılık etkileşimleri) ve puanlayıcı güvenilirliği hakkında ayrıntılı bilgi sağlaması açısından alanyazına katkı sağlayacağı düşünülmektedir.

1.4. Problem İfadesi

İlköğretim öğrencilerinin yazma performanslarının değerlendirilmesinde puanlayıcı davranışlarının, farklılaşan puanlayıcı katılık/cömertlik davranışının (puanlayıcı yanlılık etkileşimlerinin) ve puanlayıcı güvenilirliğinin Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM) ile incelenmesi.

1.5. Araştırma Soruları

Bu çalışmada araştırmanın amacı ve problem ifadesi doğrultusunda aşağıdaki sorulara yanıt aranmıştır:

- 1) Öğretmen puanlamaları bağlamında puanlayıcı davranışlarına yönelik;
 - a) Puanlayıcı katılık ve cömertlik davranışı durumları nedir?
 - b) Halo etkisi durumları nedir?
 - c) Merkeze eğilim davranışı durumları nedir?

2) Öğretmen puanlamalarına ilişkin olarak “farklılaşan puanlayıcı katılığı ve cömertliği davranışı” yani “puanlayıcı yanlılıklarının” belirlenmesinde;

- a) “*Puanlayıcı x Öğrenci*” yüzeylelerinin etkileşimine yönelik ölçüm sonuçları nedir?
- b) “*Puanlayıcı x Ölçüt*” yüzeylelerinin etkileşimine yönelik ölçüm sonuçları nedir?
- c) “*Puanlayıcı x Cinsiyet*” yüzeylelerinin etkileşimine yönelik ölçüm sonuçları nedir?
- d) “*Puanlayıcı x Görev*” yüzeylelerinin etkileşimine yönelik ölçüm sonuçları nedir?
- e) “*Puanlayıcı x Öğrenci x Ölçüt*” yüzeylelerinin etkileşimine yönelik ölçüm sonuçları nedir?

3) Öğretmen puanlamalarına ilişkin olarak KTK, GK ve MTK’ya göre kestirilen puanlayıcılar arası güvenilirlik katsayıları ne düzeydedir?

1.6. Araştırmanın Sınırlılıkları

Bu çalışma;

- 1) İlköğretim 7. sınıf öğrencilerinin yazma performansları ile sınırlıdır.
- 2) 2019-2020 eğitim öğretim yılı ikinci yarıyılı ile sınırlıdır.
- 3) Performansın değerlendirilmesi sürecinde yapılan puanlamalarda öz ve akran puanlama yapılmamış olup, yapılan çalışma sadece öğretmen puanları ile sınırlıdır.

1.7. Araştırmanın Sayıtları

- 1) Çalışmada araştırmacı tarafından puanlayıcı eğitimi alan öğretmenlerin ciddiyetle ve birbirlerinden bağımsız olarak puanlama yaptıkları varsayılmıştır.
- 2) Çalışmaya katılan öğrencilerin yazma becerilerine yönelik gerçek bilgi ve becerilerini yansıtacak şekilde performans gösterdiği varsayılmıştır. Bunun için uygulama sürecinde öğrencilere seçimli bir görev listesi verilmiş ve bu görevlerden bir tanesini seçmesi istenmiştir. Bu süreçte öğrencilerin gerçek performanslarını göstermesi sağlanmaya çalışılmıştır.

BÖLÜM II

KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

Bu bölüm araştırmanın problem ifadesine uygun olarak kuramsal çerçeve ve ilgili araştırmalar olmak üzere iki kısımdan oluşmaktadır.

2.1. KURAMSAL ÇERÇEVE

Kuramsal çerçeve kısımda performans değerlendirme, puanlayıcı davranışları, puanlayıcı güvenilirliği, Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM)'ne ilişkin bilgilere yer verilmiştir.

2.1.1. Performans Değerlendirme

Performans görevi (performance task), öğrencinin listelenmiş yanıtlardan birini seçmesi yerine, kendi yanıtını oluşturmasını sağlayan bir durum belirleme çalışması olarak tanımlanmaktadır (Stark, 1998; Zhu, 2009). Bu çalışmalarda amaç, öğrencilerin sahip oldukları üst düzey zihinsel becerilerin değerlendirilmesine katkı sağlamaktır. Performans görevleri öğrencilere gerçek yaşamda karşılaşılabilecekleri problem durumları sunan ve öğrencilerin üst düzey zihinsel becerilerinin geliştirilmesini ve ölçülmesini amaçlayan etkinliklerdir (Kutlu, Doğan ve Karakaya, 2014, s. 32). Bu etkinlikler öğretmenlerin, öğrencilerinin başarıları ile ilgili bilgi toplarken geleneksel ölçme yöntemlerinin kullanılmadığı (Kaplan, 2006); problem çözme, yazılı anlatım, sözlü anlatım, sunu yapma, kompozisyon yazma, deney düzeneği oluşturma, deney yapma, resim yapma, grafik çizme, bir araç yapma, model oluşturma gibi işlemlere sahip becerileri içerir (Arı,

2010). Geleneksel ev ödevlerinin tersine performans görevlerinde, öğrenciden çözüme ulaşmasının yanı sıra çözüme nasıl ulaştığını açıklaması ve cevabını gerekçelendirmesi de istenmektedir. Yani; performansa dayalı görevler ile öğrenciden var olan bilgiyi farklı açılardan ele alarak yeniden yapılandırması beklenmektedir. Bu beklentiler dahilinde öğrencinin üst düzey zihinsel becerilerinin ölçülmesi ve değerlendirilmesi geleneksel ölçme ve değerlendirme yaklaşımlarına nazaran daha karmaşık ve zaman alıcı olduğundan, üst düzey zihinsel performansın belirlenmesinde “*performans değerlendirme*” yapılmaktadır.

Öğrencilerin aktif öğrenme yoluyla süreç içerisinde gerçekleştirdikleri çalışmaların, etkinliklerin ve süreç sonunda ortaya koydukları ürünlerin değerlendirilmesi “*performans değerlendirme*” olarak adlandırılmaktadır (Tekindal vd. 2010, s. 130). Performans değerlendirmeleri, iletişim kurma, problem çözme ve eleştirel düşünme becerilerini kullanma gibi çoktan seçmeli bir testle ölçülmesi zor veya imkânsız olan yeteneklerin etkili bir şekilde ölçülmesini sağlayabilmektedir (Perlman, 2003). Bu tip becerilerin değerlendirmesinde öğrencinin güçlü ve zayıf yönlerini belirlenebilmekte ve öğrenciye kendi öğrenmelerine ilişkin daha ayrıntılı bilgi sahibi olma imkânı sunulmaktadır. Bu anlamda öğrencinin akademik başarısının belirlenmesinde; portfolyo değerlendirme, performans değerlendirme, öz değerlendirme, akranını değerlendirme gibi yöntemler ile DPA, proje, kavram haritaları, vb. tamamlayıcı ölçme ve değerlendirme araçlarının kullanılması bir gereklilik haline gelmiştir (Korkmaz, 2009). Buna bağlı olarak mevcut çalışmada performansın belirlenmesinde yönelik üst düzey zihinsel becerilerden ikna edici metin yazma görevleri kullanılmış ve yazma performansının değerlendirilmesinde analitik dereceli puanlama anahtarı (ADPA) kullanılmıştır.

2.1.1.1. İkna Edici Metin Yazma Görevleri

İkna; bir bireyin özgür düşünce ortamında istenen mesajın iletilmesi yoluyla birtakım gerekçeler sunarak karşısındaki kişinin inancını, davranışını veya tutumunu değiştirmeye çalıştığı sosyal ve iletişimsel bir süreç olarak tanımlanmaktadır (Lonigro vd., 2017; Perloff, 2010). İkna etmede temel amaç bireylerin, hedef kitlenin veya karşısındaki kişinin düşüncesinde değişikliğe gidilmesini sağlamaktır (Lonigro vd., 2017; Okur, Göçen ve Süğümlü, 2013). Bireylerin karşısındaki kişinin veya hedef kitlenin düşüncesini değiştirmeye yönelik sosyal ve iletişimsel bir süreç olarak görülen ikna, sadece konuşma dilinde değil; aynı zamanda yazılı dilde de karşımıza çıkmaktadır. Günlük hayatta ikna

etmeye yönelik ikna odaklanmış metinler; mektuplar, dilekçeler, reklam, film, gösteri vb. afişleri, kitap kapakları, seri ilanlar, makaleler ve hayatımızda yer alan birçok yazılı materyalde ikna ile karşı karşıya kalınmaktadır (Kaptan, 2015). Bu metinlere genel olarak literatürde *ikna edici metin* denilmektedir.

21. yüzyılda en çok kullanılan ikna türlerinden biri olan ve eleştirel düşünmeye olanak tanıyan *ikna edici metinler*, yazarın belli bir bakış açısına dayalı olarak bir düşünceyi benimseyerek karşısındaki kişiyi bu yönde ikna etmeye çalıştığı bir metin türü olarak karşımıza çıkmaktadır (Kurudayıoğlu ve Yılmaz, 2014). İkna edici metin; bireyin herhangi bir konu hakkında hedef kitlenin düşünce yapısında değişiklik yapmayı amaçlayan, herhangi bir şeyin olması gereken veya olmaması gereken şeklinin belirtildiği yazı türüdür (Kurt, 2019). İkna edici yazıda birey, bir fikri çeşitli argümanlarla desteklemeye çalışır. Bu sayede ikna edici yazma, öğrencilerin ikna yeteneğinin gelişmesini ve ikna tekniklerinin çözümlemesini sağlar (Kurudayıoğlu ve Yılmaz, 2014). İkna edici yazılar bu anlamda üst düzey zihinsel becerilerin gelişmesi için, bilgi üretmek veya bilgiyi daha anlamlı hale getirmek için bir araç olarak da düşünülebilir. İkna edici bir yazının özelliklerini ele alırken öncelikle uygun tartışılabilir bir konuyu ele alarak, karşı düşüncede olan okurların olduğunu farz etmeli, okur ve yazar kabul edilebilir, anlaşılabilir bir ortak nokta üzerinde başlayıp okurun ilgilerini, düşüncelerini ve kendi görüşlerini yansıtmalı daha sonra karşıt olan düşünceleri çürütme yoluna gitmeli, detaylarla sebep sonuç ilişkisi belirten gerekçeler vs. sunmalı, başladığı veya bitirdiği cümleler iddiası ile ilgili olmalı, okurun reddedebileceğini tahmin ederek bir plan hazırlamalı, mantıksal nedenlerle güven kazanılmalı, duygusal, psikolojik başvurular yapılmalıdır (Kaptan, 2015).

İkna edici metinlerin ve yapısının öğretimine ilköğretim 3. sınıf seviyesinden itibaren başlanabilmektedir (Caine, 2008). Sınıf seviyesi arttıkça öğrenciden istenen metinlerin niteliği ve niceliği doğru orantılı olarak değişmeli yani; sınıf seviyesi arttıkça öğrencilerden oluşturdukları ikna edici metinlerdeki iddiaların daha kuvvetli ve sayıca daha fazla olması beklenmektedir. Öğrencilere ikna edici metin oluşturmaları için de basitten karmaşığa (sınıf seviyesine göre) birçok konu verilebilmektedir (Kurudayıoğlu ve Yılmaz, 2014). Örneğin bir ilkokul 5. sınıf öğrencisinden arkadaşları ile oynadığı bir oyun ya da sevdiği bir kitap hakkında bir metin oluşturmaları istenirken daha üst sınıftaki bir öğrenciden daha kültürel ya da bilimsel bir konuda ikna edici metin oluşturmaları istenebilir. Bu konuda bazı araştırmacılar farklı okul türlerinde ve farklı sınıf seviyelerinde ikna edici yazma

uygulamalarının geliştirilmesi gerektiğini vurgulamışlardır (Newell, Beach, Smith ve Van Der Heide, 2011; Newell, Bloome, Kim ve Goff, 2019).

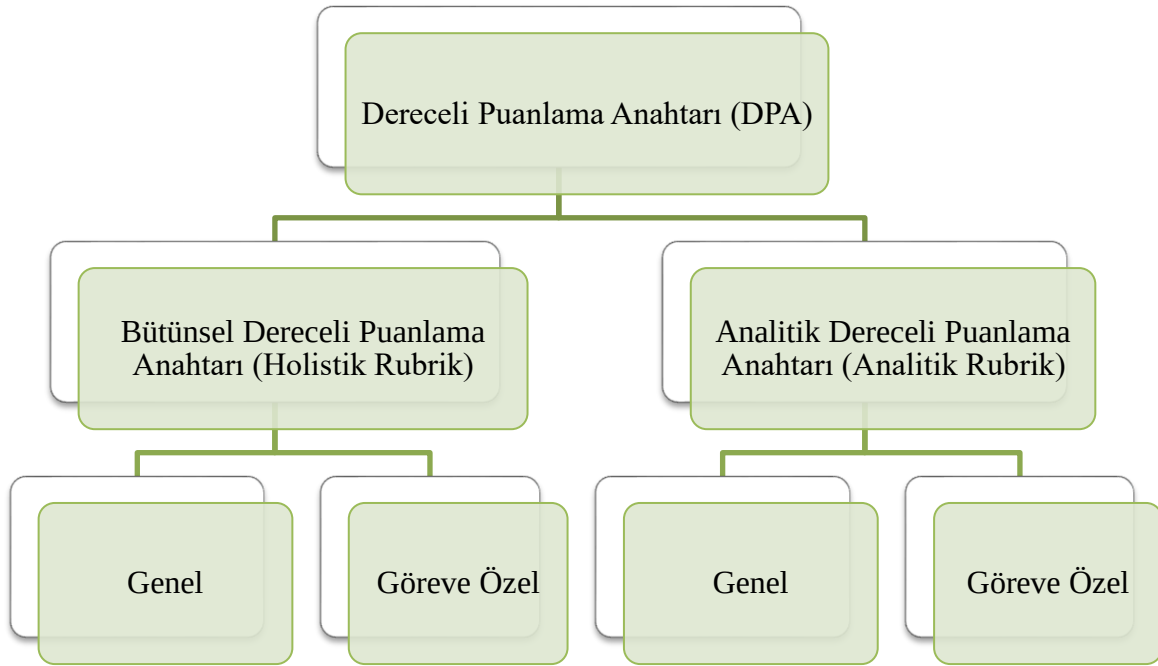
Öğrencilerin ikna edici metin yazma uygulamalarının geliştirilmesi önemli bir yere sahip olmakla beraber, öğrencilerin yazma performanslarının geçerli ve güvenilir bir şekilde puanlanıp değerlendirilmeleri de bu anlamda önemlidir. Eleştirel düşünmeye olanak tanıyan ikna edici metin yazma becerilerinin üst düzey zihinsel bir beceri olduğu göz önüne alındığında yazma performansının geleneksel ölçme ve değerlendirme yöntemleri ile değerlendirilmeyeceğinden bu tip becerilerin performansa dayalı ölçme ve değerlendirme yöntemlerinden dereceli puanlama anahtarı ile puanlanarak değerlendirilmesi gerekmektedir. Bu bağlamda nitelikli ve özgün bir yazılı anlatım ürünü ortaya çıkarmak önemli olduğu kadar bu ürünlerin objektif kriterlere göre değerlendirilmesi de önemlidir. Buna dayanarak mevcut çalışmada öğrencilerin üst düzey zihinsel gelişimine olanak tanıyan ikna edici metin türündeki yazma becerilerinin değerlendirilmesinde performansa dayalı ölçme ve değerlendirme yöntemleri kullanılmış, puanlamaların geçerliği ve güvenilirliği için birden fazla puanlayıcı puanlama yapmış, yapılan puanlamaların geçerliği ve güvenilirliği hakkında yorumlamalar yapılmıştır.

2.1.1.2. Dereceli Puanlama Anahtarları (DPA-Rubrikler)

Dereceli puanlama anahtarı (DPA) diğer bir adıyla rubrik; öğretmenlerin, öğrencilerin belli bir alandaki bilgisini göstermesi, bir görevi yerine getirmesindeki performansını ya da yetkinlik düzeyini belirlemek için kullandığı, hangi durumda hangi puanın verileceğinin önceden belirlenmesini gerektiren bir puanlama aracıdır (Brookhart, 1999). Dereceli puanlama anahtarları (DPA), önceden belirlenmiş ölçütlerle öğrenci performansını ölçmek için kullanılan ayrıntılı puanlama rehberleridir (Korkmaz, 2009). Dereceli puanlama anahtarları (DPA); grup etkinliklerinin, portfolyoların, performans görevlerinin, projelerin ya da sözel sunumların puanlanmasının yanında matematik, İngilizce, fen, Türkçe derslerindeki birçok etkinlik için kullanılabilir (Bıkmaz, 2011). Popham(1997) performans dayalı durum belirlemenin değerlendirme kısmını oluşturan dereceli puanlama anahtarını(DPA); öğrencilerin yazılı kompozisyonlarının, sözel sunumlarının veya bilimsel projelerinin değerlendirilmesinde yani üst düzey zihinsel becerilerin değerlendirilmesinde bir çeşit puanlama ölçüsü olduğunu belirtmektedir.

Öğrencilere geri bildirim verilmesinde, objektif olarak istenen sonuçların tanımlanmasında,

öğrencilerin bireysel değerlendirme yeteneklerini sürekli geliştirmelerinde, planlı olarak öğretime odaklanmada ve öğrencilerin düşünsel olarak çalışmalarda daha etkin roller almalarını sağlamada kullanılan dereceli puanlama anahtarları (DPA) (Relearning, 2000'den aktaran Korkmaz, 2009), alan yazında bütünsel (holistik) ve analitik olarak iki türde karşımıza çıkmaktadır (Atılğan, Kan ve Doğan, 2011, s.282; Mertler, 2001; Korkmaz, 2004, s.232; Parlak ve Doğan, 2014; Popham, 1997). Performansa dayalı durum belirleme çalışmasında kullanılacak dereceli puanlama anahtarının türüne karar vermek hangi ders veya hangi sınıf düzeyine uygulandığından çok, değerlendirmenin amacı ile ilgilidir (Moskal, 2000). Yani dereceli puanlama anahtarlarından hangisinin kullanılacağı değerlendirmenin amacına bağlı olarak belirlenmektedir (Popham 1997).



Şekil. 2.1. Dereceli puanlama anahtarı

“Kutlu, Ö., Doğan, D. C. & Karakaya, İ. (2014). *Öğrenci başarısının belirlenmesi: performansa ve portfolyoya dayalı durum belirleme*. Ankara: Pegem Akademi, s.54” tarafından yapılan çalışmadan alınmıştır.

Analitik ve bütünsel dereceli puanlama anahtarları genel amaçlı olabileceği gibi belirlenmiş özel bir görev için de geliştirilmiş olabilir (Kutlu, Doğan ve Karakaya, 2014, s.54). Yukarıdaki şekilde bu durum, genel ve göreve özel olarak belirtilmiştir. Üst düzey zihinsel becerilerin ölçülmesinde performans görevlerinden yararlanılmaktadır. Bu görevler dereceli puanlama anahtarı kullanılarak değerlendirilmektedir. Eğer

değerlendirme sürecinde tüm görevler için aynı dereceli puanlama anahtarı kullanılacak ise, genel dereceli puanlama anahtarı; tüm görevler için değil de öğretmen özel bir durum ile ilgili görevler veriyor ise, göreve özel dereceli puanlama anahtarı kullanılması gerekmektedir.

Dereceli puanlama anahtarı (DPA); ölçütlerden ve performans düzeylerinden oluşan yapısı ile puanlama sırasında oluşabilecek yanlışlıkları en aza indirgerken, öğrencilere performanslarıyla ilgili daha gerçekçi ve ayrıntılı geri bildirim verme imkânı sağlar (Parlak ve Doğan, 2014). Performansa dayalı durum belirleme sürecinde DPA'ları kullanan öğretmenlerin, DPA'ların bahsedilen katkılarından yararlanabilmeleri için bu araçların hazırlama, uygulama, puanlama ve geribildirim verme özellikleri konusunda yeterli bilgiye sahip olmaları gerekmektedir (Kutlu, Yıldırım ve Bilican, 2009). Ayrıca değerlendirmelerde objektifliğin ve tutarlılığın sağlanması için rubriklerin puanlanmasında iyi tanımlanmış olan bir performans veya öğrencinin yaptığı performansı açık ve net bir biçimde ortaya koyabilen görevler belirlenmelidir (Yazıcı, 2013). Bu bağlamda Kutlu, Doğan ve Karakaya (2014)' göre dereceli puanlama anahtarı geliştirmek için;

1. Beklenen performansa göre amacı belirlemek,
2. Beklenen performansa göre ölçütleri tanımlamak,
3. Tanımlanan ölçütlerin üst-orta-alt nitelik düzeylerine karar vermek,
4. Nasıl bir puanlama anahtarı (analitik, bütünsel) kullanılacağına karar vermek,
5. Öğrencilere, hangi ölçütlere göre değerlendirilecekleri hakkında bilgi vermek ve ölçütleri öğrencilere açıklamak,
6. Belirlenen ilkelere göre uygulamayı yapmak ve çalışmalarını puanlama anahtarına göre puanlamak,
7. Sonuçları öğrencilerle paylaşmak, öncelikle yapabildiklerini sonra varsa eksiklikleri görmelerini ve kendilerini geliştirmelerini sağlamak (s.66-67).

şeklinde önerilerde bulunmuşlardır.

Bütünsel Dereceli Puanlama Anahtarı (Holistik Rubrik)

Bütünsel dereceli puanlama anahtarları, öğrencilerin her bir görev ya da maddede göstermiş oldukları performansın bir bütün olarak alt bileşenlerine ayrılmaksızın

puanlanmasında kullanılır (Güler, 2008). Bu tip puanlama anahtarları öğrenci performansındaki bazı küçük hataların göz ardı edilebileceği ve performansın bütününe odaklanıldığı durumlarda kullanılmaktadır (Kutlu, Doğan ve Karakaya, 2014, s.56). Bu bağlamda bütünsel dereceli puanlama anahtarını kullanan bir öğretmen; bireyin performansını genel hatlarıyla bir bütün olarak ele alıp, o performansın hangi başarı düzeyine girdiğine karar vermelidir.

Bütünsel puanlama anahtarları, bireyin performansını bir bütün olarak ele aldığından kullanımı kolay ve kullanışlı olmasına rağmen; bireylerin performansları hakkında ayrıntılı bilgi sağlamamaktadır. Bu yüzden bireylerin performansındaki eksik veya yanlış öğrenmelerinin belirlenmesi mümkün olamamaktadır. Bütünsel dereceli puanlama anahtarının bu sınırlılığı yapılan ölçümlerin geçerliğini ve güvenilirliğini olumsuz yönde etkilemektedir. Bu sınırlılığın önüne geçebilmek için; puanlama anahtarının ölçütlerinin düzeylerinin açık ve net bir şekilde ifade edilmesi, puanlamanın güvenilirliğini arttırabilmek için farklı puanlayıcılara değerlendirme yaptırılması ve öğrenci performansının bir bütün olarak ele alındığı durumlarda bu tip puanlama anahtarının kullanılması daha uygun olacaktır.

Analitik Dereceli Puanlama Anahtarı (Analitik Rubrik)

Analitik dereceli puanlama anahtarı, gözlemlere ait puanların, tanımlanmış ölçütlerden (kriterlerden) uygun olan boyuta kaydedilmesini sağlayan bir ölçme aracıdır (Haladyna, 1997). Bu tür dereceli puanlama anahtarları, ölçülmek istenen performansı ayrıntılı olarak her bir temel özellik için alt boyutlara ayırarak bireylerin başarı düzeyleri hakkında ayrıntılı bilgi sağlamanın yanı sıra bireylerin eksik öğrenmelerine yönelik geri dönüt verebilmesine olanak tanımaktadır. Bu tür puanlama anahtarları alt boyutlara sahip olduğu için öğrencilerin belirli bir alandaki güçlü ve zayıf yönlerini göstermekte, dolayısıyla öğrenci performansı hakkında daha ayrıntılı bilgi sunmaktadır (Kutlu ve diğerleri, 2014, s.60). Bireylerin performansı hakkında ayrıntılı bilgi sunan bu puanlama anahtarları; bütünsel dereceli puanlama anahtarlarına nazaran daha geçerli ve güvenilir sonuçlar vermesine rağmen, hazırlanması ve puanlanması bütünsel dereceli puanlama anahtarına nazaran daha uzun sürmektedir.

Bu iki tip puanlama anahtarı incelendiğinde bütünsel dereceli puanlama anahtarı; performansın oluşumunda etkili olan süreçten çok ortaya çıkan ürüne, ürünün niteliğine ve

sonucuna odaklanırken (Atılğan, Kan ve Doğan, 2011, s.282); analitik dereceli puanlama anahtarı ise, ölçülen bir yetenek boyutunun farklı ölçütlere ayırabildiğinde ve daha ayrıntılı puanlama yapmak istendiğinde kullanılmaktadır (Haladyna, 1997). Yani genel anlamda analitik dereceli puanlama anahtarları daha çok performansın alt parçalarının puanlanmasını içerirken, bütünsel dereceli puanlama anahtarları davranışı bir bütün olarak ele alıp ona göre değerlendirip, ürünü veya sonucu değerlendirmede kullanılmaktadır (Yazıcı, 2013). Puanlama sürecinde bütünsel dereceli puanlama anahtarının (holistik rubriklerin) analitik dereceli puanlama anahtarına (analitik rubriklerin) göre, daha kullanışlı olması ve kısa sürede sonuçlar elde edilebilmesi açısından çok tercih edilmektedir (Nitko ve Brookhart, 2014, s.222). Bütünsel DPA 'larının çok tercih edilmesinde temel faktör; öğretmenin, öğrencinin başarabildiği şey hakkında "genel" bir fikir edinmek için, öğrenci ürününü veya performansını bir kez okuması veya başka şekilde incelemesi gerektiğinden kaynaklanmaktadır (Mertler, 2001).

Mevcut çalışmada, ilköğretim 7. sınıf öğrencilerin üst düzey zihinsel becerilerinin gelişimine olanak tanıyan yazma (ikna edici metin yazma) performansının değerlendirilmesinde farklı puanlayıcıların öğrencilerin performansına yönelik puanlayıcı davranışlarının belirlenmesinde yani puanlayıcıların ayrıntılı olarak hangi ölçüt veya ölçütlerde puanlama hatası yaptıklarının tespit edilmesi amaçlanmaktadır. Bu bağlamda çalışmanın amacına uygun olarak, öğrenci performansına karışan puanlayıcı davranışları ve puanlayıcı yanlılık etkileşimlerinin belirlenmesine yönelik istenen performansın ayrıntılı değerlendirilmesine olanak tanıyan yazma performansını ortaya çıkaracak şekilde 6 ölçütten oluşan 4'lü derecelemeyle sahip analitik dereceli puanlama anahtarı kullanılmıştır.

2.1.2. Puanlayıcı Davranışları

Performansa dayalı ölçme ve değerlendirme çalışmalarında araştırma yapma, sunum yapma, problem çözme, etkili iletişim kurabilme gibi üst düzey zihinsel becerilerin ölçülmesine olanak sağlayan görevler; puanlayıcılar tarafından değerlendirilirken puanlayıcının kendi ön yargısı ölçme sonuçlarını etkileyerek ölçümlerin geçerliğini ve güvenilirliğini düşürebilmektedir. Geçerliği ve güvenilirliği olumsuz yönde etkileyen bu durum "*puanlayıcı davranışları*" tarafından açıklanmaktadır. Alan yazınına bakıldığında yabancı literatürde "*Rater Effects*", "*Rater Bias*" "*Rater Behavior*" veya "*Rater Errors*" olarak bilinen bu kavramın Türkiye'de yapılan araştırmalarda "*puanlayıcı etkileri*",

“*puanlayıcı davranışları*” veya “*puanlayıcı hataları*” olarak farklı isimlerde kullanıldığı görülmektedir. İlhan ve Çetin (2014) ve Köse, Usta ve Yandı (2016) araştırmalarında performans değerlendirmeye karışan hataları “*puanlayıcı etkisi*” olarak adlandırmışlardır. Anadol (2018) ve Şata (2019)’ün araştırmalarında bu kavram “*puanlayıcı davranışları*” olarak ele alınırken, Börkan (2017a) çalışmasında bu kavramı “*puanlayıcı hataları*” olarak tanımlamaktadır.

Öğrencilerin performans puanlarını olumsuz yönde etkileyen puanlayıcı kaynaklı faktörlere “*puanlayıcı davranışı*” denilmektedir (Farrokhi, Esfandiari ve Vaez Dalili, 2011; Schaefer, 2008). Puanlayıcı değerlendirmelerinin niteliği, performans değerlendirmelerinden alınan puanların yorumlanmasında ve kullanılmasında merkezi bir rol oynadığından, puanlayıcı davranışlarının öğrenci başarısı tahminlerinin hesaplanmasındaki ve sınıflandırma kararları üzerindeki etkilerini dikkate alma açısından önemli bir yere sahiptir (Wind ve Guo, 2019). Araştırmacılar belirli puanlayıcı etkilerini analiz etmek ve yorumlamak için farklı metodolojiler geliştirilmişlerdir (DeCarlo, Kim ve Johnson, 2011; Engelhard, 2002; Linacre, 1989; Patz, Junker ve Johnson, 2002; Raymond ve Viswesvaran, 1993; Wolfe ve McVay, 2012). Performansa dayalı değerlendirmelerde puanlamaya karışan çok sayıda puanlayıcı hatası bulunmasına rağmen alanyazında bu etkilerin sadece birkaç tanesi analiz edilebilmekte ve yorumlanabilmektedir. Alanyazınında çok sık karşılaştığımız puanlayıcı davranışları; *puanlayıcı katılığı ve cömertliği davranışı*, *halo etkisi*, *merkeze eğilim davranışı*, *ranj daralması* (Eckes, 2017; Saal, Downey ve Lahey, 1980), *farklılaşan puanlayıcı katılığı/cömertliği davranışı (yanlılık)* ve *tutarsızlık* (Myford ve Wolfe, 2003, 2004) gibi farklı başlıklar altında ele alınmaktadır.

Puanlayıcı katılığı ve cömertliği davranışı (Rater Leniency and Severity Effect), puanlayıcılardan birinin sürekli olarak diğer puanlayıcılardan veya puanlama ölçütlerinden daha düşük ya da daha yüksek puan verme eğiliminde olmasıdır (Jackson, Schuler ve Werner, 2009; Myford ve Wolfe, 2004; Wu, 2017). Performansı, katı bir puanlayıcı tarafından değerlendirilen bir öğrencinin, ölçülen özellik açısından kendisinden daha az yetenekli olan ancak daha cömert bir puanlayıcı tarafından değerlendirilen bir öğrenciye göre daha düşük bir puan almasına neden olabilmektedir (Wiseman, 2012).

Halo etkisi (Halo Effect), puanlayıcının öğrencinin performansının kavramsal olarak farklı ve bağımsız yönlerini ayırt etmede başarısız olması ve bu özelliklere benzer puanlar vermesidir (Engelhard, 2002; Knoch, Read ve Randow, 2007). Myford ve Wolfe’a (2004) göre, halo etkisi, “puanlayıcının, öğrencinin performansını puanlarken, performansın

kavramsal olarak farklı yönlerine aynı puanı verme eğiliminde olması” şeklinde tanımlanmaktadır.

Merkeze eğilim davranışı (Central Tendency Effect), puanlama ölçeğinin orta kategorisinin (veya kategorilerinin) daha fazla kullanılması olarak tanımlanmaktadır (Landy ve Farr, 1983; Myford ve Wolfe, 2003, 2004; Wolfe ve McVay, 2010). Bazı puanlayıcılar, bir puanlama ölçeğinin orta kategorisini aşırı kullanma eğiliminde olabilirken, ölçeğin hem yüksek hem de düşük uçlarında daha az puan atar ve böylece merkeze yığılma gösterir (Eckes, 2019, s.169). Yani; merkeze eğilim davranışında belirli bir puanlayıcının, derecelendirme ölçeğinin orta kategorisinde (veya kategorilerinde), öğrencilerin yeterlilik seviyelerinin “0 logit” etrafında dizilme eğiliminde olduğunu göstermektedir (Wolfe ve Song, 2015; Wu, 2017). Örneğin; ölçülen performansın puanlanması konusunda yeterli bilgisi olmayan bir puanlayıcı düşük ve yüksek performans düzeyindeki öğrencileri birbirinden ayırt edebilir. Ancak ortalama performans gösteren öğrencileri birbirinden ayırt etmede başarısız olabilir ve bu öğrenciler için ölçeğin orta kategorilerini daha fazla kullanabilir.

Ranj daralması veya ranj sınırlaması (Restriction-of-Range Effect) öğrencinin performansından bağımsız olarak, puanlayıcının puanlama ölçeğinin sadece belli kategorilerini kullanma eğiliminde olmasıdır (Moore, 2009). Puanlayıcılar, bir derecelendirme ölçeğinin en düşük, orta veya en yüksek kategorilerini aşırı kullandıklarında ranj daralması görülmektedir (Wind, 2018b). Bu anlamda ranj daralması, merkeze eğilim davranışına benzemektedir (Engelhard 1994; Engelhard ve Stone, 1998). Ancak merkeze eğilim davranışında, yığılma puanlama ölçeğinin orta noktasında olurken; ranj daralması puanlama ölçeğinin herhangi bir noktasında olabilmektedir (Lottridge, Schulz ve Mitzel, 2013; Wu, 2017). Örneğin, bazı puanlayıcılar puanlama ölçeğinin alt ucunu aşırı kullanırken; diğer puanlayıcılar ölçeğin üst ucunu daha yoğun bir biçimde kullanabilmektedir, bu durumda başarılı ve başarısız öğrenciler birbirlerinden ayırt edilememektedir. Gerek merkeze eğilim davranışı gerekse de ranj daralması öğrencilerin performanslarının doğru bir şekilde ayrılması konusunda başarısızlığa neden olduğundan, değerlendirmelerin geçerliğini tehdit eden unsurlardır (Saal, Downey ve Lahey, 1980).

Farklılaşan puanlayıcı katılığı ve cömertliği davranışı (Differential Rater Leniency/Severity Effect) yani diğer bir deyişle *puanlayıcı yanlılığı (Rater Biases)* ise, puanlayıcının değerlendirme işlemini yaparken, cinsiyet, yaş, kültürel faktörler gibi çeşitli özelliklerine bağlı olarak bazı öğrencilere diğerlerine göre daha yüksek ya da daha düşük

puanlar verme eğiliminde olmasıdır (Kumar, 2005). Yani, puanlayıcının beklenmedik bir biçimde cömert ya da katı puanlamalar yapmasıdır (Knock, Read ve Randow, 2007). *Yanlılık analizi*, belirli bir puanlayıcının derecelendirme durumunun belirli bir yüzeyi (örneğin derecelendirme ölçütleri, görev, sınav katılımcısı) ile etkileşiminden kaynaklanan herhangi bir sistematik alt davranış kalıplarını tanımlamayı ve bu etkileşimlerin test puanları üzerindeki etkilerini tahmin etmeyi amaçladığı için *Değişen Madde Fonksiyonu* (*Differential Item Functioning -DIF*) analizine benzemektedir (Kondo-Brown, 2002; Lumley ve McNamara, 1995). Bond ve Fox (2015), yanlılık analizini *Değişen Puanlayıcı Fonksiyonu* (*Differential Rater Functioning-DRF*) olarak adlandırmaktadır.

Bir diğer puanlayıcı davranışı olan *tutarsızlık* (*Inaccuracy or Randomness effect*) ise puanlayıcılardan herhangi birinin puanlama ölçeğini diğer puanlayıcılara göre daha farklı şekilde yorumlayıp kullanmasıdır (Myford ve Wolfe, 2004). Tutarsız puanlayıcı, bazen düşük puanları ve düşük puanları hak eden öğrencilere yüksek puanlar vererek yüksek puanları hak edenlere ise düşük puanlar vererek puanları başarılı ve başarısız olan öğrenciler arasında güvenilir bir şekilde ayırt etme yeteneğini azaltır.

Performans değerlendirmede kabul edilebilir bir güvenilirlik düzeyine ulaşılması ancak sıralanan puanlayıcı etkilerinin minimum düzeyde tutulması ile mümkün olabilir. Performans değerlendirmeye karışan puanlayıcı etkilerini azaltabilmek için; *puanlama sürecinde birden fazla puanlayıcının kullanılması* veya *bir puanlayıcıya farklı zamanlarda birden fazla puanlama yaptırılması* (Atılğan, 2004; Ebel, 1951; Güler, 2008, Güler ve Gelbal, 2010; Wind, 2018a; Wind ve Guo, 2019; Yazıcı, 2013), *öğrenci çalışmalarının daha güvenilir ve yansız puanlanabilmesi için dereceli puanlama anahtarlarının yani rubriklerin kullanılması* (Kutlu, Doğan ve Karakaya, 2014, s.52) ve *puanlayıcı eğitimi verilmesi* (Çetin ve İlhan, 2014; Han, 2013; Loignon, Woehr, Thomas, Loughry, Ohland ve Ferguson, 2017; Şata, 2019; Wang ve Engelhard, 2019; Weigle, 1998; Woehr ve Huffcutt, 1994; Wind, 2018b) gibi farklı öneriler getirilmiştir.

Bu bağlamda mevcut çalışmada; ilköğretim öğrencilerinin ikna edici metin yazma performanslarının puanlayıcı eğitimi alan öğretmenler tarafından analitik dereceli puanlama anahtarı ile puanlanarak değerlendirilmesi sonucu öğrencilerin performanslarına karışan puanlayıcı davranışları ve puanlayıcı yanlılık etkileşimlerinin belirlenmesi amaçlanmaktadır. Burada ölçümlerin daha güvenilir sonuçlar verebilmesi ve değerlendirme sürecinde öğretmen puanlarına karışan puanlayıcı davranışlarından sırasıyla; *puanlayıcı katılık/cömertlik davranışı*, *halo etkisi*, *merkeze eğilim*, *puanlayıcı yanlılık etkileşimleri*

(farklılaşan puanlayıcı katılık/cömertlik davranışı) incelenmiştir. Farklılaşan puanlayıcı katılık/cömertlik davranışı yani yanlılık etkileşimlerine yönelik analizlerde; puanlayıcı yüzeyi ile diğer tüm yüzeylerin ikili etkileşimleri ile “Puanlayıcı $\times$ Öğrenci $\times$ Ölçüt” yüzeylerinin üçlü etkileşimleri analiz edilmiştir.

2.1.3. Puanlayıcı Güvenirliği (Inter-Rater Reliability)

Performansa dayalı durum belirleme çalışmalarında bireylerin puanları doğrudan ölçme araçlarıyla değil, performans sergilendiği sırada veya sonrasında önceden hazırlanmış derecelendirme ölçeklerinin puanlayıcılar tarafından değerlendirilmesi ile ölçülebilmektedir. Puanlayıcı, performansın ölçülmesinde güvenilirliğini düşüren önemli bir hata kaynağı olmakla birlikte, görev ya da zaman gibi diğer faktörlerle olan etkileşimi de en az o kadar önemli bir hata kaynağıdır. Puanlamayı yapan kişinin ölçme yaparken; öznel puanlama eğilimi, dikkatsizliği, yorgunluğu vb. ölçme sonuçlarına hata karışmasına neden olabilir (Atılğan, Kan ve Doğan, 2011, s.27). Bu tür durumlarda, değerlendirme ölçeklerinin güvenilirliği ve geçerliğinin yanı sıra puanlayıcılardan da ölçmeye hata karıştığı için puanlayıcı güvenirliliğinin de incelenmesi gerekir (Erkuş, 2017, s.38). Ölçmeye karışan puanlayıcı hataları; bir puanlayıcının bireyin performansına farklı zamanlarda vermiş olduğu puanlar veya farklı puanlayıcıların aynı bireyin performansına verdiği puanlar arasındaki ilişkiye dayalı olarak belirlenebilmektedir.

Puanlayıcı güvenirliliği, iki ya da daha fazla puanlayıcı farklı bireylere ve farklı maddelere ilişkin yaptıkları puanlamalar arasındaki tutarlılığın derecesidir (Atılğan, Kan ve Doğan, 2011, s.49). Puanlayıcı güvenirliliğini genel anlamda, puanlayıcı-içi (intrarater) güvenilirlik ve puanlayıcılar-arası (interrater) güvenilirlik hesaplama yöntemleri olarak ikiye ayırmak mümkündür (Alharby, 2006). *Puanlayıcı-içi (intrarater) güvenilirlik* bir puanlayıcının farklı zamanlarda aynı bireylerin performanslarını birden fazla kez puanlaması ile elde edilen puanları arası tutarlılıklarının derecesi iken *puanlayıcılar-arası (inter-rater) güvenilirlik* ise birden çok puanlayıcının birbirinden bağımsız olarak her bir öğrencinin performansına aynı puanı verme tutarlılıklarının derecesidir (Johnson ve Swingby'den aktaran Bıkmaz, 2011; Crocker ve Algina, 2008, s.104). Sadece puanlayıcılar arası (inter-rater) ya da puanlayıcının kendi içindeki puanlayıcı içi (intrarater) tutarlılığı sağlamak puanlamanın güvenirliliği için yeterli olmamakta, hata kaynakları arasındaki etkileşimin de güvenirliliğin hesaplanmasında göz önünde bulundurulması gerekmektedir. Performansın değerlendirilmesinde ölçümlerin geçerliği ve güvenirliliği açısından hayati bir öneme sahip

olan puanlayıcı güvenilirliği, ölçmenin üç temel kuramı olan Klasik Test Kuramı (KTK), Madde Tepki Kuramı (MTK) ve Genellenebilirlik Kuramı (GK)'na dayalı olarak farklı yöntemlerle belirlenmeye çalışılmaktadır.

2.1.3.1. Klasik Test Kuramına Dayalı Yöntemler

Performansa dayalı değerlendirmelerde puanlayıcılar-arası güvenilirliği hesaplamada, ölçek türü (sınıflama, sıralama, eşit oranlı ve eşit aralıklı) ve puanlayıcı sayısı (2 ve daha fazla)'na bağlı olarak farklı yöntemlerle belirlenebilmektedir. Puanlayıcılar-arası güvenilirliğin belirlenmesinde araştırmanın amacına ve yöntemine bağlı olarak; *Pearson Moment Çarpım korelasyon katsayısı*, *Cohen' in Kappa istatistiği*, *ağırlıklandırılmış Kappa (Weighted Kappa) istatistiği*, *Fleiss Kappa istatistiği*, *Kendall' in uyum katsayısı*, *Kendall τ* , *Krippendorff'un Alpha katsayısı*, *Goodman ve Kruskal's Y*, *Log-linear analiz tekniği*, *puanlayıcılar arası kesin uyumun yüzdesi*, *Sınıf içi korelasyon (ICC)*, *Spearman sıra korelasyon katsayısı* gibi farklı yöntemler kullanılmaktadır. Tablo 2.1'de Klasik Test Kuramı (KTK) bağlamında puanlayıcı güvenilirliğinin ölçek türü ve puanlayıcı sayısına göre farklı hesaplama yöntemleri verilmiştir (Kılıç, 2009).

Tablo 2.1

Klasik Test Kuramına Dayalı Puanlayıcı Güvenirliği Hesaplama Yöntemleri

Yöntemler	Ölçek Türü				Puanlayıcı Sayısı	
	Sınıflama	Sıralama	Eşit Oranlı	Eşit Aralıklı	2	3+
Pearson korelasyon			+	+	+	
Spearman sıra korelasyon			+	+	+	
Weighted Kappa		+			+	
Cohen' in Kappa istatistiği	+				+	
Fleiss Kappa istatistiği	+					+
Kendall'ın uyum katsayısı(W)		+				+
Kendall τ		+			+	
Krippendorff'un Alpha	+	+	+	+	+	+
Sınıf İçi Korelasyon (ICC)		+	+	+	+	+
Goodman ve Kruskal's Y		+			+	
Log-linear analiz tekniği	+				+	+
Puanlayıcılar arası kesin uyumun yüzdesi	+				+	+

Not: "Kılıç, S. (2009). *Ölçümlerin uyumluluğu ve tıptaki uygulamaları*. Yüksek Lisans Tezi, Sağlık Bilimleri Enstitüsü, Biyoistatistik ABD, Çukurova Üniversitesi" tarafından yapılan çalışmadan uyarlanmıştır.

Tablo 2.1’de görüldüğü üzere, puanlayıcılar arası uyum veya daha doğru ifade ile puanlayıcılar arası güvenilirliğe ilişkin çok sayıda yöntem bulunmaktadır (Shoukri, 2004). Performans değerlendirme sürecinde puanlayıcılardan kaynaklanan değişkenliğinin/hataların belirlenmesi için kullanılacak olan puanlayıcı güvenirlik yöntemi; yapılacak çalışmanın amacına, desenine, kullanılan ölçme aracının ölçek yapısına ve türüne, puanlayıcı sayısına bağlı olarak değişebilmektedir. Yukarıda adı geçen bu yöntemlerin her birinin kendi içinde avantaj ve dezavantajları, birbirine göre üstün ve zayıf yönleri bulunmaktadır. Ancak güvenirliğin hesaplanmasında KTK ‘ya dayalı tüm yöntemler hem puanlayıcı hem diğer hata kaynaklarıyla olan etkileşimi birlikte ele almamaktadır. KTK’nın bu sınırlılığına yönelik farklı varyans kaynaklarının büyüklüğünün kestirilmesine, çoklu analizlerin yapılmasına, farklı varyans kaynakları arasındaki etkileşimin kestirilmesine olanak tanıyan genellenebilirlik kuramı (GK) geliştirilmiştir. Yukarıda belirtilen KTK’nın farklı yöntemleriyle farklı anlamlarla yorumlanan güvenirlik katsayılarının birbirinden farklı olmasından hareketle, Cronbach ve arkadaşları, Genellenebilirlik Kuramını (GK) ortaya atmışlardır (Atılğan, 2005). Mevcut çalışmada KTK bağlamında birden fazla puanlayıcının puanlayıcılar arası güvenirliğinin hesaplanmasında veri setinin yapısına uygun olarak *Sınıf İçi Korelasyon (ICC) katsayısı* ile güvenirlik hesaplanmıştır.

Sınıf İçi Korelasyon (ICC) Katsayısı (Intraclass Correlation Coefficient)

Puanlayıcılar arası güvenirlik katsayısının hesaplanmasında ve öğrencilerin herhangi bir alandaki performansını değerlendirmek amacıyla kullanılan bir dereceli puanlama anahtarının güvenirliğinin belirlenmesinde (Kutlu, Doğan ve Karakaya, 2014, s.86) sınıf içi korelasyon katsayısı hesaplanabilmektedir. *Sınıf içi korelasyon katsayısı (ICC)*, iki ya da daha fazla puanlayıcının en az sıralama ölçek düzeyinde toplam puanları üzerinden puanlayıcılar arası uyum düzeyine yani puanlayıcılar arası güvenirliğe bakar. Sınıf içi korelasyon katsayısı (ICC), birden fazla puanlayıcının güvenirliğini tahmin etmek için varyans analizine ve varyans bileşenlerinin tahminine dayanmaktadır. Sınıf içi korelasyon (ICC) katsayısı, puanlayıcılar arası güvenirliğin belirlenmesinde kullanılan bir çeşit korelasyon katsayısıdır ve 0 ile 1 arasında değer alır. Bulunan korelasyon katsayısının bu değeri;

$r < 0,40$ ise zayıf,

$r = 0,41 - 0,59$ arası orta,

$r = 0,60 - 0,74$ arası iyi,

$r > 0,75$ ise mükemmel düzeyde uyum olduğu belirtilir (Howell, 2010, s.492).

Bu korelasyon katsayısının kabul edilebilir seviyeleri verilerin özelliğine ve kullanıldığı alana göre farklılık göstermektedir (Erdoğan, 2004). Performansa dayalı değerlendirmelerde birden fazla puanlayıcı arasında bir uyum olduğunu yani puanlayıcılar arası güvenilirliğin sağlandığını söyleyebilmek için sınıf içi korelasyon katsayısının *en az* 0,70 olması gerekmektedir. *Sınıf içi korelasyon katsayısı* (ICC); 0,95 ile 1,00 arasında bir değer alıyorsa puanlayıcılar arası güvenilirlik “mükemmel”, 0,85 ile 0,94 arasında bir değer alıyorsa puanlayıcılar arası güvenilirlik “yüksek”, korelasyon değerinin 0,70’in altında olması durumunda ise puanlayıcıların birbirleriyle hiçbir şekilde uyumlu olmadığı yani ölçümlerin güvenilirliğinin düşük olduğu yorumu yapılabilmektedir (Alper, 2001).

2.1.3.2. Genellenebilirlik Kuramına Dayalı Yöntemler

Genellenebilirlik kuramı (GK), özellikle farklı hata kaynaklarını konu edinen ölçmelerde, bu hata kaynakları ile bunların etkileşiminden kaynaklı hataların kestirimini sağlayan ve temelinde varyans analizine (ANOVA) dayalı olarak geliştirilen istatistiksel bir kuramdır (Atılğan, 2019a; Shavelson ve Webb, 1991; Brennan, 2001). Genellenebilirlik kuramı (GK), Klasik Test kuramının sınırlılıklarına ek olarak ortaya çıkmıştır. Klasik Test kuramı güvenilirliğin sadece bir hata kaynağına bağlı kestirimine izin veren gerçek puan modeline dayalı bir kuram iken, G kuramı Klasik Test kuramının en açık sınırlılıklarından biri olan tek hata kaynağı içerme durumuna tepki olarak geliştirilmiştir (Güler, Uyanık ve Taşdelen Teker, 2012). G kuramı sadece yapılan ölçümlerin güvenilirliği hakkında tahmin yapmakla kalmaz, aynı zamanda gelecekteki uygulamalarda ölçme işlemleri geliştirmek için kullanılacak hata kaynakları hakkında bilgi sağlamaktadır. Genellenebilirlik kuramının en önemli yararı, araştırmacıların aynı anda birden fazla ölçümü istatistiksel olarak modellemesine izin vermesidir (Grabowski ve Lin, 2019, s.55).

Shavelson ve Webb (1991), Genellenebilirlik kuramının Klasik Test kuramının genişletilmiş bir uzantısı olduğunu dört maddeyle açıklamaktadır:

1. Tek bir analizle birçok hata/varyans kaynağı aynı anda kestirilebilmektedir.

2. Varyans/hata kaynaklarının her birinin büyüklüğünü belirlenebilmektedir.
3. Bireylerin performanslarına yönelik hem bağıl hem de mutlak kararlar alınabilmekte ve buna bağlı olarak iki farklı güvenirlik katsayısı hesaplanabilmektedir. Bağıl değerlendirmelerde “*Genellenebilirlik (G) katsayısı*” kullanılırken, mutlak değerlendirmelerde “*Phi(ϕ) Katsayısı*” kullanılmaktadır (Atılgan, 2019a; Brennan 2001; Shavelson ve Weeb, 2003).
4. İstenilen ölçme durumlarında, en uygun güvenirlik katsayısının elde edilebileceği Karar (D) çalışmaları yapılabilmektedir. Burada bulunan güvenirlik katsayısı Sperman-Brown formülüne benzer şekilde hesaplanabilmektedir (Mushquash ve O’Connor, 2006).

Puanlayıcılar arası güvenirliği belirlemek için değişkenlik kaynağı olarak puanlayıcıları ve puanlayıcılardan gelen potansiyel hata kaynakları dikkate alınarak puanlayıcılar arası uyuma dayalı puanlayıcılar arası güvenirliği belirlemek için G kuramından yararlanılabilir (Cronbach, Gleser, Nanda ve Rajaratham, 1963’ten aktaran Atılgan, Kan ve Doğan, 2011, s.50). Bu kurama dayalı olarak hesaplanan *Genellenebilirlik (G) katsayısı* aynı zamanda güvenirlik katsayısı bağlamında yorumlanabilmektedir. Genellenebilirlik Kuramı (GK)’nda farklı varyans bileşimlerine dayalı olarak bulunan genellenebilirlik/güvenirlik katsayısı mGENOVA, GENOVA ve EduG yazılımları ile belirlenebilmektedir (Atılgan, 2019b, s.4).

2.1.3.3. Madde Tepki Kuramına Dayalı Yöntemler

Performans değerlendirmede puanlamaya karışan *puanlayıcı davranışları/farklı değişkenlik kaynakları* genellikle klasik test kuramı (KTK) ve genellenebilirlik kuramı (GK)’na dayalı yöntemlerle belirlenebilmektedir. Ancak bu kuramlara dayalı yöntemlerin ölçme araçlarında ve değerlendirmelerde kullanımı bazı sınırlılıkları da beraberinde getirdiği görülmektedir. Bu sınırlılıklar; (a) madde ve test istatistiklerinin testin uygulandığı gruba ve teste bağımlı olması, (b) tüm hata kaynakları için tek bir güvenirlik katsayısının kestirilmesi ve (c) değerlendirmelerin toplam test puanları üzerinden yapılarak bireylerin yeteneklerin tam olarak kestirilememesidir (Crocker ve Algina, 2008, s.143; Embretson ve Reise, 2000, s.53; Hambleton ve Jones, 1993; Lord, Novick ve Birnbaum, 2008; Hambleton, Swaminathan ve Rogers, 1991). Bu sınırlılıkları ortadan kaldıran Madde tepki kuramı (MTK), temelde bireyin gözlenemeyen yetenekleri ile ölçme aracındaki maddeler arasındaki ilişkiyi matematiksel bağıntılarla açıklamaya çalışan bir kuramdır

(Güler, 2008). Hattie, Jaeger ve Bond (1999)’a göre “Örtük Özellik Teorisi olarak bilinen bu kuram, KTK’ya dayalı yöntemlerin sınırlılıklarını ortadan kaldıran güvenilir bir model” olarak görülmektedir (s. 399). Son zamanlarda Madde Tepki Kuramı(MTK) ve özellikle Rasch ölçüm teorisindeki gelişmeler, değerlendirme kalitesini arttırmak ve puanlayıcı hatalarını/davranışlarını tespit etmek için teorik olarak farklı yöntemler geliştirme fırsatı sunmaktadır (Engelhard, 2002, s.262).

Madde tepki kuramı (MTK)’nda, puanlayıcılar arası güvenilirliğin ya da değişkenlik kaynağının birden fazla olduğu durumda güvenilirlik hesaplanmasında kullanılabilen bir parametrelili model, “*Rasch Ölçme Modeli*” olarak bilinmektedir (Güler, 2008). Bir parametrelili bu model, 1980 yılında Danimarkalı bir matematikçi olan George Rasch tarafından geliştirilmiştir. Rasch ölçme modeli; Madde Tepki Kuramı (MTK)’nın *tekboyutluluk* ve *yerel bağımsızlık* sayılına ek olarak, Madde Karakteristik Eğrisi (MKE) için sadece madde güçlük indeksi olan “ β ” parametresinin var olmasının yeterli görüldüğü, madde ayırıcılık indeksi olarak bilinen “ α ” parametrelerinin tüm maddeler için eşit ve bir kabul edildiği ve maddelerin şansla doğru yanıtlandığının göstergesi olan “ γ ” parametrelerinin tüm maddeler için sıfır olarak alındığı bir modeldir.

Linacre, Engelhard, Tatum ve Myford (1994), Rasch modelini çok yüzeyli değişkenlerin birleşimi olarak sunmaktadır. Madde Tepki Kuramı (MTK)’nın bir parametrelili ölçüm modeli olan Rasch Ölçme modelinin geliştirilmesi ile Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM) oluşturulmuştur (Linacre, 1993b). Bu model bir parametrelili Rasch ölçme modelinin sınırlılıklarına ek olarak geliştirilmiştir.

2.1.4. Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM)

Bir parametrelili Rasch ölçme modelinin genişletilmiş hali olan ve bu modelin sınırlılıklarına ek olarak geliştirilen model “Çok Yüzeyli Rasch Ölçme modeli (ÇYRÖM)”dir. Bu model yabancı literatürde “multi-facet(ed) or many-faceted Rasch measurement” (Engelhard, 1994; McNamara, 1996; Eckes, 2015) veya “many-faceted conjoint measurement” (Linacre, Engelhard, Tatum, ve Myford, 1994) veya multi-facet Rasch modeling (Bahrouni, 2016; Lianzhen ve Jie; 2008; Lunz ve Linacre, 1998) yer alırken Türkçe kaynaklarda genellikle “Çok yüzeyli Rasch ölçme/ölçüm modeli” veya “Çok yüzeyli Rasch modeli” (Çetin ve İlhan, 2017; Güler, 2014; İlhan, 2016; Karakaya, 2015; Şata ve Karakaya, 2020) olarak kullanılmaktadır.

Çok yüzeyli Rasch ölçme modeli, madde tepki kuramına dayalı olarak madde güçlüğüne ve bireylerin yetenek düzeyine ek olarak, puanlayıcılar, durumlar veya koşullar gibi faktörlerin getirdiği değişkenliği hesaba katmayı amaçlayan bir parametrelili Rasch ölçme modelinin bir uzantısıdır (Bond ve Fox, 2015). Bir parametrelili Rasch ölçme modelinde bireyler ve maddeler değerlendirilir, bireylerin yetenek düzeyleri ve madde güçlükleri eşit aralıklı bir ölçek üzerine yerleştirilir (Linacre, 1993a). Örneklemden bağımsız madde güçlük düzeyleri ve testten bağımsız olarak bireylerin yetenek düzeyleri kestirilir (Elhan ve Atakurt, 2005). Çok yüzeyli Rasch Ölçme modelinde ise çok sayıda değişkenlik/varyans kaynağı (birey, madde, puanlayıcı, görev, ölçüt, zaman, uygulama sayısı gibi) tek bir eşit aralıklı ölçek üzerine yerleştirilir. Bir parametrelili modelden farklı olarak bu model bireylerin test puanlarını etkilediği düşünülen tüm değişkenlik/varyans kaynaklarını dikkate alınır ve bu değişkenlik kaynaklarının birbiri ile olan etkileşimi analiz edilir (Abu Kassim, 2007). Örneğin; test puanlarında değişkenliğe neden olabilecek faktörler arasında yer alan *birey* ve *madde* değişkenleri arasında anlamlı bir etkileşim olup olmadığı belirlenebilir. Bu sayede, bir grup bireyin belli bir maddede, sistematik olarak diğer maddelerden farklı bir performans gösterip göstermediği ortaya konulabilir. Benzer olarak, diğer değişkenlik kaynaklarının birbiriyle olan etkileşimleri *çok yüzeyli Rasch ölçme modeli* ile belirlenebilir. Daha genel bir ifadeyle, “*birey*×*madde*”, “*puanlayıcı*×*birey*” ve “*puanlayıcı*×*madde*” gibi test puanlarını etkileme olasılığı bulunan değişkenlik kaynaklarının ikili kombinasyonları arasındaki etkileşimler belirlenebilmektedir (İlhan, 2015). Bu etkileşimlerin analizindeki temel amaç, değerlendirmelerde bireylerin performansını etkileyen sistematik hata terimlerinin ve puanlayıcı etkilerinin belirlenmesini sağlamaktır. Downing (2005), ÇYRÖM’ni sadece puanlayıcı davranışını izleyememekle kalmayıp aynı zamanda sistematik hata ayarını yapabileceğini vurgulamaktadır.

Çok yüzeyli Rasch Ölçme modelinde bireylerin performansını etkileyen değişkenlik kaynaklarının herbirine *yüzey (facet)* denilmektedir (Koons, 2008; Sudweeks, Reeve ve Bradshaw, 2004). Bu model adından da anlaşılacağı üzere analizinde en az iki yüzey yada değişkenlik kaynağı içermektedir. Bilgisayar programı olan FACETS (Linacre, 2018) ÇYRÖM’ni operasyonel hale getirmiştir. FACETS paket programı ile Çok yüzeyli Rasch Ölçme modeline dayalı olarak; tüm yüzeylere (birey, puanlayıcı, ölçütler, zaman, uygulama sayısı gibi) ait ölçümler tek seferde ve örneklemden bağımsız olarak analiz edilebilmekte, yüzeylere ait bireysel ve grup düzeyinde değerlendirmeler yapılabilmekte,

ölçüm sonuçlarının güvenilirliği tüm değişkenlik kaynakları ve bunların etkileşimleri için ayrı ayrı hesaplanabilmekte ve ölçüm sonuçlarını etkileyebilecek sistematik hatalara yol açan puanlayıcı etkileri kolaylıkla yorumlanabilmektedir. Bireylerin performansına etki edebilecek puanlayıcı davranışlarından; puanlayıcı katılığı/cömertliği davranışı, halo etkisi, ranj daralması, merkeze eğilim ve farklılaşan puanlayıcı katılığı/cömertliği davranışı (yanlılık) gibi etkiler *çok yüzeyli Rasch ölçme modeli (ÇYRÖM)* ile analiz edilebilmekte ve yorumlanabilmektedir.

Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM)'nin Sayıltıları

Çok yüzeyli Rasch Ölçme modeli temelinde Madde Tepki Kuramına dayalı olarak geliştirilen bir istatistiksel bir modeldir. Çok yüzeyli Rasch Ölçme modelinin karşılanması gerekli temel sayıltılar;

1. *Puanlar ve yetenek düzeyleri arasında pozitif bir ilişki (Model-veri uyumu);* Yüksek yetenekli öğrencilerin daha az yetenekli öğrenciden daha fazla maddeden daha yüksek puan alma olasılığı daha yüksek olması,
2. *Yerel bağımsızlık;* Bireylerin performansını etkileyen yetenekler sabit tutulduğunda, bireylerin herhangi bir madde çiftine verdikleri tepkilerin bağımsız olması,
3. *Tek boyutluluk;* Tek seferde sadece bir gizil yeteneğin ölçülmesidir. Yani tek boyutluluk, bir ölçme aracındaki tüm maddelerin tek bir değişkeni/yapıyı ölçmesidir (Baker ve Kim, 2004, s.115; Embretson ve Reise, 2000, s.65; Erkuş vd.,2017, s.161-162; Farrokhi, Esfandiari ve Dalili, 2011).

Burada model-veri uyumu sayıltısının sağlanması için; ± 2 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı toplam gözlem sayısının %5'inden fazla olmayacak ve ± 3 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı toplam gözlem sayısının %1'inden fazla olmayacaktır (Linacre, 2018, s.171). Yerel bağımsızlık sayıltısının sağlanması için tek boyutluluk sayıltısının mutlaka sağlanmış olması gerekir (Erkuş vd., 2017, s.126). Bu durumda tek boyutluluk sayıltısı sağlanıyorsa yerel bağımsızlık sayıltısı da sağlanacaktır. Verilerin tek boyutluluk sayıltısını sağlayıp sağlamadığı faktör analizi ile test edilmektedir. Dolayısıyla araştırma verilerinin tek boyutluluk sayıltısını sağlayıp sağlamadığı *Açımlayıcı Faktör Analizi (AFA)* ile

belirlenebilmektedir. Mevcut çalışmada Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM)'nin bütün sayıltıları sağlandıktan sonra veri analizine geçilmiştir.

Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM)'nin Analizi ve Yorumlanması

Temel Rasch analizi için bir parametrelili Rasch ölçme modeli bireylerin yetenek düzeyi ve madde güçlük düzeyleri arasındaki etkileşimi belirlemeye çalışmaktadır. Genel olarak George Rasch'ın geliştirdiği 1-0 şeklinde ikili puanlanan test maddeleri için temel Rasch modeli aşağıdaki gibidir.

$$\log\left(\frac{P_{ni}}{1-P_{ni}}\right) = B_n - D_i \quad (1)$$

P_{ni} : n bireyinin i maddesinde başarılı olma olasılığı

B_n : n bireyinin yeteneği

D_i : i maddesinin güçlüğü

Yukarıdaki formülde birey ve madde yüzeylerine ait *iki yüzeyle Rasch ölçme analizi* yapılabilmektedir. Burada B parametresi birey yüzeyine ait yetenek düzeyini ifade ederken, D parametresi ise madde yüzeyine ait güçlük analizini göstermektedir. Bu model aynı matematiksel değerler korunarak derecelendirilmiş ölçekler için geliştirilmiştir. Buna bağlı olarak dereceli puanlamalar için iki parametrelili Rasch modeli (Linacre, 2018, s. 248):

$$\log\left(\frac{P_{nik}}{P_{nik-1}}\right) = B_n - D_i - F_k \quad (2)$$

P_{nik} : n bireyinin i kategorisi veya görevinde k puanını alma olasılığı

P_{nik-1} : n bireyinin i kategorisi veya görevinde k-1 puanını alma olasılığı

B_n : n bireyinin yeteneği

D_i : i kategorisi ya da görevi için güçlük düzeyi

F_k : k-1 puanından k puanına geçişin zorluğu

Yukarıdaki formüle göre, birey ve madde yüzeylerine ait derecelendirilmiş ölçekler için iki yüzeyli Rasch ölçüm analizi yapılabilmektedir. Bu modelin 1-0 şeklinde puanlanan iki yüzeyli Rasch modelinden tek farkı, bu modelde derecelendirme ölçütlerinin kategorilerinin güçlük düzeylerinin bireylerin test puanlarını etkilemesidir.

Ölçme sürecine puanlayıcıların da katılması durumunda (1) ve (2) nolu formüller yeterli olmayabilir. Bu durumda çok değişken kaynaklı modellere ihtiyaç vardır. Bu modellerden ÇYRÖM genel olarak üç yüzeyli (birey, madde ve puanlayıcı) olarak karşımıza çıkmaktadır. Linacre(1993a) tarafından geliştirilen üç yüzeye ait geliştirilen model:

$$\log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = B_n - D_i - C_j - F_k \quad (3)$$

P_{nijk} : n bireyinin j puanlayıcısından i kategorisi veya görevinde k puanını alma olasılığı

P_{nijk-1} : n bireyinin j puanlayıcısından i kategorisi veya görevinde k-1 puanını alma olasılığı

B_n : n bireyinin yeteneği

D_i : i kategorisi yada görevi için güçlük düzeyi

C_j : j puanlayıcısının katılık düzeyi

F_k : j puanlayıcısı için k-1 puanından k puanına geçişin zorluğu

şeklinde tanımlanmaktadır (Eckes, 2015, s.36; Linacre, 2018, s.272). Bu formüle dayanarak bireysel ve grup düzeyinde istatistiklere dayalı olarak değerlendirmeler yapılabilmektedir. İlk olarak grup düzeyindeki istatistiksel göstergeler sonrasında ise bireysel düzeydeki istatistiksel göstergeler incelenmiştir. Grup düzeyindeki değerlendirmeler için her bir yüzeye ilişkin oluşturulan ölçüm raporlarına ait; standart hata (SE), ayırma oranı, ayırma indeksi, ayırma indeksinin güvenilirliği, sabit etkili χ^2 değeri yorumlanırken; bireysel düzeydeki değerlendirmeler için; gözlenen ortalama, düzeltilmiş ortalama, logit değerleri, standart hata, uygunluk ölçütlerine ait istatistiksel değerler (uyum-içi ve uyum-dışı değerler ve yüzdeleri) yorumlanmaktadır (Bahrouni, 2016; Barkaoui, 2013; Linacre, 2018, s.199-203; Myford ve Wolfe, 2003,2004).

Grup düzeyindeki istatistiklerden birisi de “Ki-kare(χ^2)” istatistiğidir. Çok yüzeyli Rasch ölçme modelinde her bir yüzey için, “sabit etkili χ^2 ” ve “normal χ^2 ” olmak üzere iki farklı değer hesaplanmaktadır. Herhangi bir yüzeyin bileşenlerinin normal dağılıma sahip bir evrenden tesadüfi olarak seçilen bir örnekleme temsil edip etmediğine karar vermek için “normal χ^2 ” değeri; yüzeyin bileşenleri arasında anlamlı fark olup olmadığını belirlemek için ise “sabit etkili χ^2 ” değeri incelenmektedir (Linacre, 2018, s.204). Sabit etkili ki-kare testine göre; puanlayıcının katılık değerinin anlamlı bir şekilde farklı olmadığı kabul edilmektedir. Yani; ölçmenin standart hatası hesaplandıktan sonra tüm puanlayıcıların aynı katılık düzeyinde olduğu varsayılır. Sabit etkili ki-kare testi(χ^2) sonucunun anlamlı olması puanlayıcıların birey performansını değerlendirme sürecinde farklı puanlamalar yaptığı şeklinde yorumlanmaktadır. Ki-kare testi(χ^2) örneklem büyüklüğüne karşı çok duyarlı olduğundan dolayı bir yüzeyin grup düzeyindeki istatistikleri yorumlandığında diğer istatistiklerinde verilmesi gerekmektedir (Myford ve Wolfe, 2004; Weigle,1998).

Grup düzeyindeki istatistiklerden bir diğeri *ayırma oranı (separation ratio)*’dir. Bu oran, yüzeylerin farklılıklarının ölçüm hatasından gerçekten daha büyük olup olmadıklarını belirlemeye yardımcı olmaktadır. Ayırma oranı; ölçüm hatası ayarlandıktan sonra yüzey ölçüsünün “gerçek” standart sapma oranının ($SDt(j)$), ortalama yüzey standart hatası ($RMSEj$) üzerinden ayarlanmasıyla tahmin edilebilir. Ayırma oranı hesaplanmasında kullanılan formül (Eckes, 2015, s.62; Linacre, 2018, s.338):

$$G_j = SDt(j) / RMSE_j \quad (6)$$

şeklindedir. G_j , yüzey ölçümleri için hata birimlerindeki yayılımını göstermektedir. Daha yüksek bir puanlayıcı ayırma oranı (G_j), yüzey ölçümlerinin daha geniş bir alana yayıldığını gösterir. Ayırma oranı değeri kullanılarak hesaplanan bir diğer istatistik *ayırma indeksi (separation index)*’tir. Ayırma indeksi, yüzeyleri daha iyi yorumlamak için kullanılabilir. Bu indeks ayırma oranı ile benzer şekilde yorumlanabilmektedir. Ayırma indeksi hesaplanmasında kullanılan formül (Eckes, 2015, s. 63; Linacre, 2018, s.338):

$$H_j = (4.G_j + 1) / 3 \quad (7)$$

Grup düzeyindeki istatistiklerden bir diğeri ise *Ayırma indeksinin güvenilirliği (Separation index reliability)*’dir. Hesaplanan ayırma oranı (G_j) güvenilirliği hesaplamak için kullanılır. Ayırma indeksinin güvenilirliğinin hesaplanmasında kullanılan formül (Eckes, 2015, s. 63; Linacre, 2018, s.338):

$$R_j = G_j^2 / (1 + G_j^2) \quad (8)$$

Grup düzeyindeki istatistiklerden bir diğeri de *Rasch-Cohen's Kappa indeksi*'dir. Çok yüzeyli Rasch ölçme modelinde puanlayıcı yüzeyine ait puanlayıcı güvenilirliği her bir puanlayıcı için hem bireysel hemde grup düzeyinde beklenen ve gözlenen uyum yüzdeleri hesaplanmaktadır. Bir puanlayıcının beklenen ve gözlenen uyum yüzdeleri eşit olduğunda puanlayıcının bireyin performansını değerlendirirken hiçbir puanlayıcı davranışı göstermediği yani mükemmel puanlayıcı olduğu şeklinde yorumlanır (Eckes, 2015, s.92). Rasch ölçme modeline göre puanlayıcı güvenilirliğinin belirlenmesinde kullanılan formül [-1, 1] arasında değişen değerler almaktadır ve bu değer 0.00'a yakın değerler alması beklenir (Eckes, 2015, s.92; Linacre, 2018, s.306). Bu formül;

$$Rasch - Cohen's Kappa = \left(\frac{gözlenen uyum yüzdesi - beklenen uyum yüzdesi}{100 - beklenen uyum yüzdesi} \right) \quad (9)$$

şeklindedir. Performansın değerlendirilmesinde gözlenen ve beklenen uyum değerinin birbirine yakın olması önemlidir. Bu istatistiksel değer, hem bireysel düzeyde hem de grup düzeyinde hesaplanabilmektedir. Mevcut çalışmada bu değer, sadece puanlayıcı yüzeyine ait ölçüm raporunda grup düzeyindeki istatistiklere dayalı olarak hesaplanmıştır.

Bireysel düzeydeki değerlendirmeler için ilk olarak modeldeki her bir yüzeyin bileşenleri için *logit değerleri* yorumlanmaktadır. Genel anlamda bu değerler puanlayıcı yüzeyi için en katı veya en cömert puanlayıcının belirlenmesinde, madde düzeyi için en kolay veya en zor maddenin ne olduğunun tespitinde, birey yüzeyi için ise istenen yeterliliğe sahip en başarılı ve en başarısız bireylerin belirlenmesine yardımcı olmaktadır. Bu logit değerleri *logit ölçeği* veya *kalibrasyon haritası (değişken haritası)* üzerinde gösterilmektedir. Üç yüzeyli Rasch ölçme modelinde bütün yüzeyler aynı anda bağımsız olarak tek bir doğrusal ölçek üzerinde yani logit ölçeği üzerinde kalibre edilir. Kalibrasyon haritası üzerindeki logit değerleri, analizdeki tüm yüzeyler ve çeşitli yüzeyler arasında karşılaştırma yapmanıza yardımcı olur. Herbir yüzeyin bileşenleri için oluşturulan logit cetvelleri, verilerin eşit aralıklı ölçek düzeyinde değerlendirilmesine olanak tanır. Logit ölçeğinde değerler $+\infty$ ile $-\infty$ arasında değişebilmektedir (İlhan, 2015). Logit ölçeğinde yüzeyler için pozitif olarak tanımlanan değerler; bireylerin yüksek yetenek düzeyinde, maddelerin kolay güçlükte ve puanlayıcıların değerlendirmelerinde daha katı olduğunu ifade ederken; negatif olarak tanımlanan ölçümler ise bireylerin düşük yetenek düzeyinde, maddelerin zor güçlükte ve puanlayıcıların değerlendirmelerinde daha cömert olduğunu ifade etmektedir

(Eckes, 2015, s.58; Güler, 2008; İlhan, 2015). Çok yüzeyli Rasch ölçme modelinde logit ölçeğindeki her bir veri için *standartlaştırılmış hata değeri* hesaplanmaktadır. Bu değer logit ölçeğindeki veriler için t_değerine dönüştürülür. Puanlayıcılara ait logit değerleri, bu logit değerlerinin ortalaması ve standartlaştırılmış hata değeri kullanılarak hesaplanan t_değeri puanlayıcı katılığı ve cömertliği davranışının belirlenmesindeki bireysel düzeydeki göstergelerden biridir (Çetin ve İlhan, 2017). Örneğin; puanlayıcı yüzeyi için, hesaplanan t_değeri t dağılım tablosundaki t_değerinden büyük yani t_değeri anlamlı ise, bulunan t_değerinin + veya – oluşuna göre bireylerin başarılı veya başarısız olduğu söylenebilir.

Logit ölçeğinin t_değeri aşağıdaki formül yardımıyla hesaplanabilmektedir:

$$t_değeri = \frac{\text{puanlayıcının logit değeri} - \text{logit değerlerinin ortalaması}}{\text{logit ölçümünün standart hatası}} \quad (4)$$

Bireysel düzeydeki istatistiksel göstergelerden *uygunluk ölçütleri*; puanlayıcı uyum istatistikleri, belirli bir puanlayıcı tarafından sağlanan değerlendirmelerin, ÇYRÖM tarafından beklenen değerlendirmelerle ne derece uyumunu göstermektedir (Eckes, 2019, s.165). Her bir yüzeye ait ölçüm raporlarında Infit MNSQ (uyum-içi) ve Outfit MNSQ (uyum-dışı) olarak adlandırılan sütunlar uygunluk ölçütlerini göstermektedir. Burada her bir yüzeye ait tüm veriler için ayrı ayrı uygunluk ölçüt değeri hesaplanmaktadır. Infit ve Outfit MNSQ istatistiklerinin beklenen değeri “1” olup, $[0, +\infty]$ arasında değişen değerler alabilmektedir (Linacre, 2002; Myford ve Wolfe, 2003; Eckes, 2019). Linacre (2002, 2018), uyum-içi ve uyum-dışı istatistik değeri için 0,50’i alt sınır ve 1,50’i üst sınır olarak önermiştir. 1,5’den büyük ve 0,5’den küçük uyum değeri değerlendirmelerde beklenen değer ile gözlenen değer arasında daha fazla değişim olduğunu gösterir; bu durum *uyumsuzluk* olarak adlandırılırken; 1,5’den küçük ve 0,5’den büyük olan uyum değeri beklenen değer ile gözlenen değer arasında daha az değişim olduğunu gösterir; bu durum değerlendirmelerin çok fazla tahmin edilebilir olduğu veya fazla bilgi sağladığı anlamına gelir; bu durum *uyumluluk* olarak ifade edilmektedir (Barkaoui, 2013; Bond ve Fox, 2015; Güler, 2008). Buna dayanarak uyum istatistikleri 0,5 ile 1,5 aralığının içinde ise bireyin, maddenin ve puanlayıcının uygun olduğu kabul edilir, bu aralığın dışında değerlerin analizden çıkarılması önerilmektedir (Myford ve Wolfe, 2003; Wright ve Linacre, 1994).

Bireysel düzeydeki değerlendirmelerden *ölçüm (Meausure) istatistiği*; logit cetvelinde olduğu gibi modeldeki her bir yüzeyin bileşenleri için yorumlanmaktadır. Burada her bir veri için ayrı ayrı ölçüm değeri hesaplanmaktadır. Bu ölçüm değeri logit ölçütü'dür. Bu istatistik yardımıyla; puanlayıcı etkilerinden puanlayıcı katılığı/cömertliği davranışı ve farklılaşan puanlayıcı katılığı/cömertliği davranışı diğer bir adıyla yanlılık etkileşimleri belirlenebilmektedir. Logit değeri; puanlayıcı yüzeyi için puanlayıcı katılığının ve cömertliğinin, madde yüzeyi için kolay ya da zorluğunun, birey yüzeyi için başarılı veya başarısız öğrencilerin yorumlanmasında yardımcı olmaktadır. Örneğin; puanlayıcı yüzeyi için negatif değerler cömert puanlayıcıları gösterirken, pozitif değerler alan puanlayıcılar katı puanlayıcıları göstermektedir.

Denklem (3)'den yola çıkarak; grup düzeyinde hesaplanan istatistiklerden biri olan “puanlayıcı x birey” etkileşimi için yanlılık analizi (*Bias analysis*):

$$\log\left(\frac{P_{nik}}{P_{nik-1}}\right) = B_n - D_i - C_j - F_k - I_{nj} \quad (5)$$

P_{nik} : n bireyinin j puanlayıcısından i kategorisi veya görevinde k puanını alma olasılığı

P_{nik-1} : n bireyinin j puanlayıcısından i kategorisi veya görevinde k-1 puanını alma olasılığı

B_n : n bireyinin yeteneği

D_i : i kategorisi ya da görevi için güçlük düzeyi

C_j : j puanlayıcısının katılık düzeyi

F_k : j puanlayıcısı için k-1 puanından k puanına geçişin zorluğu

I_{nj} : “puanlayıcı x birey” yüzeyi için yanlılık etkileşimi terimi

olarak tanımlanmaktadır (Linacre, 2018, s.272). Çok yüzeyli Rasch ölçme analizinde puanlayıcı etkilerinin belirlenmesinde *yanlılık indeksi* önemli bir yere sahiptir (Linacre, 2018). Denklem (4)'de tanımlanan I_{nj} , “*puanlayıcı x birey*” etkileşimi için yanlılık terimi (Bias indexs) olarak ifade edilmektedir. ÇYRÖM'ne dayalı analizlerde ikili ve üçlü yanlılık etkileşimlerinde; grup düzeyindeki istatistiksel göstergelerden sabit etkili χ^2 değerinin anlamlı olması, bireysel düzeydeki istatistiksel göstergelerden t_değerinin ± 2

aralığı dışında olması farklılaşan puanlayıcı katılık/cömertliği davranışının yani puanlayıcı yanlılığının göstergesi olarak yorumlanmaktadır (Barkaoui, 2013; Bond ve Fox, 2015; Linacre, 2018; s.180).

Performansın değerlendirilmesinde kullanılan dereceli puanlama ölçeğine ilişkin olarak hesaplanan *kategori istatistikleri*; puanlama ölçeğinin ölçütlerinin veya derecelendirmelerinin uygunluğuna karar vermek için hesaplanmaktadır (Myford ve Wolfe, 2004). Bu istatistikler hem bireysel düzeyde hem de grup düzeyinde hesaplanabilmektedir. Puanlayıcı davranışlarından merkeze eğilim ve ranj daralması davranışlarının belirlenmesinde kategori istatistikleri kullanılmaktadır. Kategori istatistiklerine ilişkin ölçüm raporunda; ortalama logit ölçüsü, beklenen logit ölçüsü ve uyum-dışı istatistikleri incelenir. Ortalama değerlerinin uyum-dışı istatistiksel değerlerinin 0,5 ile 1,5 değerleri arasında olması puanlama ölçeğinin uygun olduğunu göstermektedir (Bond ve Fox, 2015; Linacre, 2018, s.278).

2.2. İLGİLİ ARAŞTIRMALAR

Bu bölümde araştırma konusu hakkında genel çerçeveyi oluşturmak amacıyla ilgili kaynaklardan elde edilen bilgilere yer verilip, *puanlayıcı davranışları* ve *Çok Yüzeyli Rasch ölçme modeli (ÇYRÖM)* ile ilgili bazı araştırma ve uygulamalardan bahsedilmektedir.

2.2.1 Puanlayıcı Davranışları ile İlgili Araştırmalar

Yabancı literatürde “*rater effect*” veya “*rater bias*” olarak geçen “*puanlayıcı davranışları*” ile ilgili geçmişten günümüze yapılan çalışmalar incelenmiştir. İncelenen çalışmalarda çoğunlukla mevcut çalışma bağlamında temel puanlayıcı davranışlarının (puanlayıcı katılık/cömertlik davranışı, halo etkisi, merkeze eğilim, ranj daralması vs.) yanı sıra özellikle puanlayıcı yanlılıkları yani farklılaşan puanlayıcı katılık/cömertlik davranışı ile ilgili çalışmalara yer verilmiştir.

Hoyt (2000)’a göre puanlayıcı yanlılığı (rater bias), psikolojik araştırmalarda önemli bir hata kaynağı olarak görülmektedir. Çalışmasında çok değişkenli genellenebilirlik kuramına dayanan bir model kullanarak, değerlendirme sürecinde puanlayıcı yanlılığının ve puanlayıcılar-arası güvenilirliğin araştırma sonuçlarını nasıl etkilediğini göstermektedir. Model, geleneksel olarak katılık/cömertlik davranışı ve halo hataları olarak adlandırılan yanlı davranışlar olmak üzere puanlayıcı derecelendirmelerini kullanan araştırmalardaki bulguları etkileyebilecek iki yanlılık çeşidini tanımlamaktadır. Yanlılık etkileşimi analizi sonucunda, iki yanlılık çeşidinden hangisinin kullanıldığına ve bu yanlılıklarının etkisini azaltmaya yönelik formüller üretilmiştir. Puanlayıcı yanlılığının belirlenmesine yönelik oluşturulan bu model, yanlılık çalışma bulguları üzerindeki olumsuz etkilerini en aza indirmeyi amaçlayan araştırmacılar için prosedürler önermektedir.

Kondo-Brown (2002) tarafından yapılan çalışmada; FACETS paket programını kullanarak, öğretmen puanlayıcılarının değerlendirmelerine yönelik, Japonca ikinci dil (L2) yazısının değerlendirilmesinde belirli öğrencilere ve belirli ölçütlere karşı nasıl yanlı olduğu araştırılmaktadır. Çalışmasında “puanlayıcı x aday” etkileşiminin yanlılığı incelemiştir. Araştırmanın çalışma grubunu 234 üniversite öğrencisi ve üç eğitimli öğretmen oluşturmaktadır. Puanlayıcılar yüksek korelasyonlu değerler elde edilmiş ve puanlamaların tutarlı olduğu görülmüştür. Ancak genel olarak yanlılık analizinde anlamlı farklılıklar ortaya çıkmıştır. Puanlayıcılar, belirli adayları ve değerlendirme ölçütlerini daha cömert

veya katı bir şekilde puanlamış ve her puanlayıcının yanlılık şekli farklı olarak bulunmuştur. Önemli derecede yanlı “puanlayıcı x aday” etkileşimlerinin en yüksek yüzdesi, yetenek düzeyi son derece yüksek veya düşük olan adaylar arasında bulunmuştur.

Youn (2007) çalışmasında KFL öğrencilerinin farklı puanlayıcılar tarafından değerlendirmesinde; test türleri, konuşma eylemleri, aday grupları ve test maddeleri gibi çeşitli faktörlerin KFL öğrencilerinin yazmaya yönelik pragmatik test yeterliliği konusundaki değerlendirmelerini etkileyip etkilemediğini araştırmaktadır. Bu amaçla çalışmada FACETS (Linacre, 1996) bilgisayar programını kullanarak test sonuçları ile puanlayıcılar arasındaki yanlılık etkileşimlerini analiz edilmiştir. “Puanlayıcı x test” türleri, “puanlayıcı x konuşma eylemleri”, “puanlayıcı x madde zorluğu” ve “puanlayıcı x aday” yüzeyleri arasındaki etkileşimlere yönelik hesaplamalarda puanlayıcı yanlılıkları tespit edilmeye çalışılmıştır. Bu çalışmada Hudson, Detmer ve Brown’ın (1995) yazmaya yönelik bazı testlerden uyarlanmış üç farklı yazma testi kullanılmıştır: “Açık Yazılı Anlatım Tamamlama Görevi, Dil Laboratuvarı ve Rol Oynama”. Bu araştırmanın sonuçları, üç puanlayıcının da test türüne ve konuşma eylemine bağlı olarak derecelendirmelerinde puanlayıcıların farklı derecelerde puanlama katılığı davranışı gösterdiğini göstermektedir. Ek olarak, her puanlayıcı etkileşimleri içinde benzersiz yanlılık etkileşimleri bulunmuştur. Belirli konuşma eylemlerinin ve test türlerinin puanlayıcı değerlendirmelerini nasıl etkilediğini, puanlayıcıların çeşitli faktörlerde gösterdiği sistematik hata kaynaklarının ne olduğunu yani puanlayıcıların farklı yüzeyler altındaki etkileşimlerinde yanlı puanlayıcı davranışlarına sahip oldukları görülmüştür.

Schaefer (2008) tarafından yapılan çalışmada, EFL makalelerini değerlendirilmesinde NES puanlayıcılarının puanlama eğilimi modellerini keşfetmek için çok yüzeyli Rasch ölçme modeli (MFRM) kullanılmaktadır. 40 NES puanlayıcısı Japon üniversite öğrencileri tarafından TOEFL Yazılı İngilizce Testinden (TWE) uyarlanan tek bir konuda yazılan 40 kompozisyonu değerlendirmiştir. Yazılan makaleler altı kategorili derecelendirme ölçeği kullanılarak değerlendirildi (İçerik, Organizasyon, Anlatım Tarzı ve Kalitesi, Dil Kullanımı, Yapı ve Akıcılık). MFRM, puanlayıcıların alt gruplar arasında birkaç tekrarlayan yanlılık etkileşimini ortaya çıkarmıştır. Derecelendirme kategorisi yanlılık etkileşimlerinde, eğer İçerik ve Organizasyon katı şekilde derecelendirilmişse, Dil Kullanımı ve Mekaniği cömert yani tam tersi olarak değerlendirilmiştir. “Puanlayıcı x yazar” yanlılık etkileşimlerinde, daha az yetenekli yazarlara göre daha yüksek yetenekli yazarlara karşı daha katı ya da zayıf bir yanlılık eğilimi vardır. Bazı puanlayıcılar ayrıca,

daha fazla yetenekli yazarları daha katı şekilde değerlendirmiş ve daha fazla yetenekli yazarları beklenenden daha cömertçe değerlendirmiştir. Bu çalışmada sonuç olarak, L2 yazma değerlendirmesinde puanlayıcı yanlılığının etkileri bulunduğu görülmüştür.

Tamanini (2008) çalışmasında performans değerlendirme sürecinde puanlamadaki hataların sıklıkla puanlama yanlılığından kaynaklandığını dile getirmektedir. Performans değerlendirme araştırmalarının birçoğu; puanlayıcı hatalarını ortadan kaldırmanın, tespit etmenin veya kontrol etmenin yollarına odaklanmıştır. Bu çalışmalardan elde edilen sonuçlar puanlamadaki yanlılıkların açıklanmasında yetersiz kalmaktadır. Bu bağlamda çalışma, puanlayıcı hatalarıyla nasıl başa çıkılacağı hakkında anlayış sağlamaktadır. Bu çalışma, bireylerin performansının değerlendirmesi sırasında puanlayıcı yanlılığının tespitine olanak sağlayan “farklılaşan puanlayıcı katılık ve cömertlik (DRF)” davranışının belirlenmesini amaçlamaktadır. DRF tekniği, belirli bir bireyin farklı madde grupları için farklı olan cevaplarını tanımlamak için kullanılmaktadır. Çalışmada DRF tekniğinin duyarlılığı ve tutarlılığını tespit etmek için, katılık/cömertlik davranışı sonuçları ile karşılaştırılmıştır. Sonuç olarak DRF yönteminin araştırmacılara performans değerlendirmelerinde puanlayıcı yanlılıkları tespit etmek ve ortadan kaldırmak için kullanılabilecek bir araç olduğu görülmüştür.

Choi ve Young-Min (2011) çalışmalarında, Koreli dil öğretmenlerinin puanlayıcı özelliklerini ve yanlılıklarını Korece dil öğretmenlerinin Çok Yönlü Rasch Ölçme Modeli (MFRM) kullanarak değerlendirmelerinin sonuçlarını analiz ederek açıklığa kavuşturmayı amaçlamışlardır. Bu çalışmada, 35 öğrencinin yazılı metinleri Koreli 68 öğretmen tarafından puanlanmıştır ve burada elde edilen sonuçlar analiz edilmiştir. Korece dil öğretmenleri, ayrı ayrı sunulan değerlendirme ve ayrı ayrı değerlendirmelere tabi tutulmadan yapılan bireysel analiz ve yazma değerlendirmelerine dayalı olarak, öğrenci ölçeğini beş ölçütü (1-5) değerlendirmektedir. Değerlendirme ölçütleri; içerik, organizasyon, tablo, resmi deyimler ve kelime seçimidir. Öğrenci makalelerinin değerlendirilmesine katılan Koreli öğretmenlerin cinsiyet ve kariyer açısından değerlendirmede katılık/cömertlik davranışları ve etkileşim etkilerini analiz edilmiştir. Kadın öğretmenlerin katılığı, erkek öğretmenlerinkinden daha yüksekti ve en üst düzeyde katılık, 1 ila 5 yıllık deneyime sahip gruplarda bulundu. Değerlendirme ölçütlerinde Koreli öğretmenler kelime seçim ölçütünde en yüksek sayıyı vermişlerdir ve aşağıdakiler tablo, organizasyon, biçimsel kullanım ve içerik açısından ortaya çıkmıştır. 68 Koreli öğretmen içerisinde 48 öğretmenin değerlendirmesinde, öğrencilere yönelik puanlayıcı yanlılığı

anlamli olarak analiz edildi, minimum yanlilik 1, minimum yanlilik 17 idi. Korece dil retmenleri ile deęerlendirme ltleri arasında 42 deęerde yanlilik vardı, ierik lt ve resmi dilbilgisi lt ilgili puanlayıcı yanlilıęı tespit edilmiřtir. Sonu olarak, Koreli retmenlerin ęrenci deęerlendirmesi iin Koreli retmenlerin zellikleri ve puanlayıcı yanlilıęı hakkında bilgi saęlanmıřtır.

Farrokhi, Esfandiari ve Schaefer (2012a) alıřmalarında  deęerlendirme trnde farklılařan puanlayıcı katılık/cmertlik davranıřını ok Yzeyli Rasch Modeli kullanarak incelemiřlerdir. Puanlayıcı etkisinin bir bileřeni olarak puanlayıcı yanlilıęı ya da etkileřim etkisi alıřmaları, deneyimli retmenleri puanlayıcı olarak istihdam ederken, bu alıřma,  puanlayıcı tipi arasında puanlayıcı yanlilik etkileřimini ya da farklılařan puanlayıcı katılık/cmertlięini arařtırmak iin ok yzeyli Rasch modeli (MFRM) kullanmaktadır: z puanlayıcı, akran puanlayıcı ve retmen puanlayıcısı. İran'daki iki devlet niversitesinde 188 ęrencinin İngilizce yazdıęı metinler, hem ęrencilerin hem z hem de akran puanlayıcılar ve retmenleri tarafından 6 puan analitik derecelendirme lęi kullanılarak deęerlendirilmiřtir. MFRM,  deęerlendirme tr arasında farklılařan puanlayıcı katılık/cmertlięi etkisi arařtırılmıřtır. rneęin, z ve retmen puanlayıcılar, en yksek ve en dřk yetenekli ęrencileri deęerlendirirken akran deęerlendiricilere kıyasla katı puanlama davranıřı gstermiřtir. Bu alıřmada sonu olarak; L2 yazma deęerlendirmesinde akran ve z deęerlendirmede farklılařan puanlayıcı katılık/cmertlięi etkisi olduęu grlmřtir.

Farrokhi, Esfandiari ve Schaefer (2012b) alıřmalarında, puanlayıcı etkilerini veya puanlayıcı hatalarını beř farklı řekilde gruplandırılmıřtır: “puanlayıcı katılıęı/cmertlięi etkisi”, “halo etkisi”, “merkezi eęilim etkisi”, “tutarsızlık”ve “yanlilik veya etkileřim etkisi”. Bu alıřmada yanlilik etkisi, L2 testi alanında kapsamlı bir řekilde alıřılmıřtır. Ayrıca, ok yzeyli Rasch modelini (MFRM) kullanarak; z deęerlendirme, akran deęerlendirme ve retmen deęerlendirmedeki puanlayıcı katılıęı/cmertlięi etkisini incelemek amalanmaktadır. alıřma iki devlet niversitesinde ileri yazma derslerine katılan İngilizce ęrenimini grmekte olan 188 ęrenci zerinde yapıldı. alıřmada, 188 makale toplandı ve 15 ltten oluřan 6 puanlık bir analitik deęerlendirme lęi kullanılarak veriler z, akran ve retmen puanlayıcıları tarafından deęerlendirildi. Toplanan veriler, arařtırma sorularını cevaplamak iin FACETS 3.68.0'ın bir srm kullanarak analiz edildi. Analizin sonuları,  deęerlendirme trnde bazı tekrar eden kalıplar gstermiřtir. Bulgular ayrıca z deęerlendiricilerin maddelere ynelik en fazla

yanlı davranışa sahip olma eğiliminde oldukları öğretmen puanlayıcıların ise, öğrencilere karşı en hafif yanlılığa sahip bulunmuştur. Ayrıca, “kelime seçiminin” puanlayıcı türüne göre en zor olan ölçüt olduğu ve “yazım”ın puanlayıcı türüne göre en kolay ölçüt olduğu tespit edilmiştir.

İlhan (2015) tarafından yapılan araştırmada, standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi amaçlanmıştır. Araştırmanın katılımcılarını, sekizinci sınıfa devam eden 104 ortaokul öğrencisi ile çalışmada puanlayıcı olarak görev alan yedi matematik öğretmeni oluşturmaktadır. Çalışmada veri toplama aracı olarak; araştırmacı tarafından geliştirilen ve açık uçlu sekiz sorudan oluşan matematik başarı testi, standart rubrikler, SOLO taksonomisine dayalı rubrikler, standart rubrikler ile ilgili düşünceler anketi (SRDA) ve SOLO taksonomisine dayalı rubrikler ile ilgili düşünceler anketi (STDRDA) kullanılmıştır. Araştırmada, açık uçlu matematik sorularına verilen yanıtların standart ve SOLO taksonomisine dayalı rubrikler kullanılarak puanlanmasıyla elde edilen veriler çok yüzeyli Rasch modeline göre analiz edilmiştir. Puanlayıcıların SRDA ile STDRDA’da yer alan kapalı uçlu maddelere verdikleri yanıtların analizinde aritmetik ortalama değerlerinden faydalanılmış; açık uçlu maddelere verilen yanıtlar ise betimsel analiz yaklaşımıyla çözümlenmiştir. Araştırmada; hem standart hem de SOLO taksonomisine dayalı rubriklere göre yapılan puanlamalarda yetenek düzeyi farklı olan öğrencilerin yüksek güvenirlikte birbirinden ayırt edilebildiği ve maddelerin güçlük düzeyleri arasında anlamlı fark olduğu belirlenmiştir. Standart rubrikler kullanılarak yapılan puanlamalarda; *halo etkisi*, *tutarsızlık*, “*puanlayıcı×birey*” ve “*puanlayıcı×birey×madde*” yanlılıklarının bulunmadığı saptanmıştır. Diğer taraftan; puanlamalarda *merkeze eğilim etkisi* ile “*puanlayıcı×madde*” yanlılığının söz konusu olduğu, *katılık* ve *cömertlikleri* yönüyle puanlayıcılar arasında anlamlı fark bulunduğu ve puanlayıcı güvenirliğinin düşük olduğu tespit edilmiştir. SOLO taksonomisine dayalı rubrikler kullanılarak yapılan puanlamalarda ise, puanlayıcı katılığı ve cömertliği, merkeze eğilim etkisi, halo etkisi, tutarsızlık ve yanlılık şeklinde sıralanan puanlayıcı etkilerinden hiçbirine rastlanmamış; puanlayıcı güvenirliğinin yüksek olduğu ve puanlayıcıların benzer katılık/cömertlikte puanlamalar yaptığı sonucuna ulaşılmıştır.

Wesolowski, Wind ve Engelhard (2015) çalışmasında, çok yüzeyli Rasch Modeli kullanılarak büyük gruplarda müzik performansının değerlendirmesi bağlamında model-veri uyumu ve farklılaşan puanlayıcı etkisini yani puanlayıcı yanlılığını incelemiştir.

Özellikle, uzman puanlayıcıların (N = 24) şiddetinin dört okul düzeyinde değişmez olup olmadığını belirlemeye çalışmışlardır. Yanlılık etkileşimi analizleri, hem puanlayıcı grubu hem de bazı bireysel puanlayıcılar için puanlayıcı yanlılığının logit ölçeğinde beklenen konumlarına dayanarak var olduğunu göstermiştir. Bulgularda uzman puanlayıcıların, dört okul düzeyinde toplulukların alt gruplarını değerlendirirken puanlayıcı yanlılığı göstermediği görülmüştür. İncelenen 92 potansiyel ikili etkileşimin 14'ünde (% 15.2) etkileşim istatistiksel olarak anlamlı bulunmuştur, burada 10 bireysel puanlayıcı en az bir okul düzeyinde farklılaşan katılık etkisi göstermiştir. Çift yönlü etkileşimler araştırıldıktan sonra bazı puanlayıcılar için bireysel olarak farklılaşan puanlayıcı etkisi yani puanlayıcı yanlılığı davranışı ortaya çıkmıştır. Çalışmada genel olarak büyük gruplarda müzik performans değerlendirmelerinde ölçümlerin geçerliğini ve güvenilirliğini arttırmanın etkileri tartışılmaktadır.

Jin ve Wang (2017) çalışmalarında çok yüzeyli verilerde farklı yüzeyler arasındaki etkileşimi inceleyerek DRF adı verilen yanlı puanlayıcı davranışlarının belirlenmesi amaçlanmıştır. Çalışmalarında diferansiyel puanlayıcı fonksiyonu (DRF) olarak adlandırılan farklılaşan puanlayıcı katılık/ cömertlik etkisini yani puanlayıcı yanlılığını; puanlayıcılar arasındaki etkileşim ile tespit etmişlerdir. Mevcut DRF çalışmalarında çoğunlukla, cinsiyet veya etnik köken gibi gruplarda etkileşim etkileri tahmin edilmektedir. Bununla birlikte, grup üyeliği bilinmediğinde ve dolayısıyla verilerden tahmin edilmesi gerektiğinde DRF oluşabilir. Bu sorunu çözmek için bu çalışmada, grup üyeliğine bağlı olarak DRF'yi değerlendirmek için yeni bir yüzey modeli geliştirilmiştir ve uygulamalarını göstermek için iki ampirik örnek sunulmuştur. Yeni modelin DRF değerlendirmesindeki performansını Bayesian çerçevesindeki değerlendirmek için bir dizi inceleme yapıldı. Sonuç olarak, farklı yüzeyler arasındaki etkileşimlerle ölçüm sonuçlarına yönelik DRF olup olmadığı tespit edilmiştir.

Soo Park ve Xing (2019)'a göre; puanlayıcı yanlılığını diferansiyel puanlayıcı fonksiyonu (DRF) olarak ele almaktadır. Puanlayıcı yanlılığı; puanlayıcıların farklı alt gruplar içindeki puanlamaları incelendiğinde farklı bireylere yönelik katılık veya cömertlik etkisi olduğunda ortaya çıkmaktadır. Bu çalışmada, DRF çalışmaları alt grupları belirlemek için değişkenler arasındaki etkileşimi kullanarak puanlayıcı yanlılığı incelenmiştir. Bu değişkenler; cinsiyet ve sınav katılımcının etnik kökenini içermektedir. Örneğin, bir puanlayıcı erkek öğrencileri daha katı şekilde puanlayabilir. İdeal olarak, her bir puanlayıcının katılığı alt gruplar arasında değişmez olmalıdır. Bu çalışma, DRF'yi,

puanlayıcıların farklılaşan puanlama davranışına dayanarak olası DRF kaynaklarını sınıflandıran gizil alt gruplar bağlamında incelemektedir. DRF'yi tanımlamak için (LC-SDT) modelinin bir uzantısı, gerçek dünya verileri ve simülasyonları kullanılarak önerilmiş ve incelenmiştir. DRF'yi tanımlamak için farklı örneklem büyüklüklerine ait veri setlerinde puanlayıcı yüzeyine bağlı olarak diğer değişkenler arasındaki etkileşimler incelenmiştir. Sonuçlar, LC-SDT'nin DRF uzantısının, puanlayıcıların özelliklerini incelemek ve puanlayıcı eğitime yardımcı olabilecek bilgileri eklemek için kullanışlı bir model olabileceğini düşündürmektedir.

Wind ve Guo (2019) çalışmasında puanlayıcı etkileri veya puanlayıcıların garanti edilen performansların puanlarından farklı performanslara puan verme eğilimlerini yani puanlayıcı yanlılıklarını incelemişlerdir. Puanlayıcı etkileriyle ilgili birçok gerçek veri çalışmasında araştırmacılar; puanlayıcıların, öğrenci alt gruplarıyla ilgili uyumsuzluk ve sistematik hataların bir kombinasyonu gibi birden fazla etki gösterdiğini bildirmiştir (diğer bir deyişle, [DRF] çalışan farklılaşan puanlayıcı katılık/cömertlik etkisi yada puanlayıcı yanlılığı). Bu çalışmanın amacı, puanlayıcılar birden fazla etki gösterdiğinde, puanlayıcı etkisi göstergelerinin puanlayıcı etkilerine ne derece duyarlı olduklarını araştırmak ve bu duyarlılığın farklı veri toplama araçları altında değişme derecesini araştırmaktır. DRF ve puanlayıcı uyumsuzluğu kombinasyonlarını keşfetmek için bir simülasyon çalışması kullanılmıştır. Genel olarak, bu çalışmanın bulguları DRF'nin ortak sayısal ve grafiksel göstergelerini kullanmanın ve puanlayıcıların her iki etki gösterdiğinde puanlayıcı uyumsuzluğunun mümkün olduğunu göstermekte, ancak araştırmacılara göre bu etkilerin yalnızca sayısal göstergeler kullanarak ayırt edilmesinin zor olabilmektedir. Puanlama sürecinde analitik veya holistik puanlama anahtarları kullanıldığında, puanlayıcı etkilerinin yani puanlayıcı yanlılıklarının tanımlanmasının daha kolay olduğunu görülmektedir.

Yurt içinde ve yurt dışında yapılmış olan çalışmaları incelediğimizde; mevcut çalışmanın amacına uygun olarak temel puanlayıcı davranışlarının (puanlayıcı katılık/cömertlik davranışı, halo etkisi, ranj sınırlaması, merkeze eğilim davranışı) yanısıra farklılaşan puanlayıcı davranışlarından puanlayıcı yanlılığına yönelik çalışmalar incelenmiştir. Bu çalışmalarda puanlayıcı yanlılığına ilişkin olarak, puanlayıcı yüzeyi ile çalışmadaki diğer tüm değişkenlik kaynaklarının ikili etkileşimleri sonucunda farklılaşan puanlayıcı katılık/cömertlik davranışı diğer bir adıyla puanlayıcı yanlılığı belirlenmeye çalışılmıştır. Mevcut çalışmada da ilköğretim öğrencilerinin ikna edici metinleri analitik puanlama anahtarı kullanılarak çok sayıda puanlayıcı tarafından değerlendirilmiş ve puanlayıcı

davranışları (puanlayıcı katılık/cömertlik davranışı, halo etkisi, ranj daralması, merkeze eğilim davranışı, farklılaşan puanlayıcı katılık/cömertlik davranışı) ile açıklanmaya çalışılmıştır.

2.2.2. Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM) ile İlgili Araştırmalar

Bu bölümde mevcut çalışmanın istatistiksel yapısını oluşturan madde tepki kuramına (MTK)'na dayalı Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM) ile ilgili geçmişten günümüze yer alan çalışmalar incelenmiştir. Aşağıda üst düzey zihinsel becerilerin ölçümünde performansa dayalı değerlendirmelerde Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM)'nin kullanımı ilgili bazı çalışmalara yer verilmiştir.

Wolfe (2004) tarafından yapılan çalışmada; çok yüzeyli Rasch modelinin performans değerlendirmede puanlayıcı etkilerini tanımlamak için nasıl uygulanabileceği açıklanmaktadır. Çalışma grubu, Amerika Birleşik Devletleri'ndeki İleri Düzey Yerleştirme (AP) kurslarına katılan 28 lise öğrencisi ve 101 puanlayıcıdan oluşmaktadır. Çalışmanın verileri, 28 öğrencinin yazmış olduğu kompozisyonların dereceleme ölçeği kullanılarak 101 puanlayıcı tarafından puanlanması ile elde edilmiştir. Çalışmada genel olarak üç temel puanlayıcı etkisi üzerinde durulmaktadır: Katılık/Cömertlik etkisi, Merkeze eğilim etkisi ve Tutarsızlık. Puanlayıcıların %5'inin katı puanlamalar yaptığı, puanlamalarda orta kategorilerin daha fazla kullanıldığı yani merkeze eğilim etkisinin görüldüğü, ayrıca puanlayıcıların puanlamaları arasında bariz tutarsızlıklar olduğu görülmüştür. Sonuç olarak; çok yüzeyli Rasch modelinin performans değerlendirmede puanlayıcı etkilerinin tanımlamasında kullanılabilecek bir model olduğu sonucuna varılmıştır.

Baştürk (2010) tarafından yapılan çalışmada, öğrencilerin bilimsel araştırma ile ilgili teorik kazanımlarını göstermek için performansa dayalı olarak oluşturdukları bilimsel araştırma ödevlerini çok yüzeyli Rasch ölçme yöntemi ile değerlendirilmesi amaçlanmıştır. Rasch ölçme modeli bir soruya doğru cevap verme olasılığının, o kişinin yeteneğine bağlı olduğunu ifade eder. Çok yüzeyli Rasch ölçme modeli bu modele, puanlayıcı katılığı/cömertliği, soru güçlüğü, gibi faktörleri de ekleyerek tek parametrelili model çok parametrelili bir model haline getirmiştir. Araştırmada öğrencilerin Bilimsel araştırma becerileri, puanlayıcıların katılık/cömertliklerine ve davranışı sergileme güçlüğüne göre değerlendirilmiştir. Araştırmada Pamukkale Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Rehberlik ve Psikolojik Danışmanlık anabilim dalında “Bilimsel

Araştırma Yöntemleri” dersine devam eden 20 öğrencinin katılımcı olarak kullanılmıştır. Verilerin analizi sonucunda, hazırlanan bilimsel araştırma ödevleri içerisinde “*uygunluk içi*” ve “*uygunluk dışı*” istatistiklere sahip ödevlerin olmadığı, ödevlerin istatistiksel olarak anlamlı yaklaşık 3 ödev kalite tabakasına ayrılabilceği gözlenmiştir. Puanlayıcıların yapmış oldukları puanlamalarda ise istatistiksel olarak katılık/cömertlik bakımından birbirlerinden farklılıklar göstermediği, yalnızca bir puanlayıcı için “*uygunluk dışı*” kareler toplamının beklenen değerden farklı olduğu belirlenmiştir. Araştırmada, ödevlerinin kalitelerini belirlemede kullanılan “Bilimsel Araştırma Ödevlerini Değerlendirme Formu” içerisinde yer alan soruların da amaca hizmet eder nitelikte oldukları belirlenmiştir. Sonuç olarak yüksek öğrenimde özellikle performansa dayalı öğrenmelerin değerlendirilmesinde Çok-Yüzeyli Rasch ölçme modelinin etkin olarak kullanılabileceği sonucuna varılmıştır.

Semerci (2011) tarafından yapılan araştırmada, doktora yeterlikler çerçevesinde öğretim üyesi, akran ve öz değerlendirmelerin Rasch ölçme modeliyle analizinin yapılması amaçlanmaktadır. Araştırmada tarama yöntemi kullanılmış olup öğretim üyesi ve araştırmacılar tamamen gönüllülük esasına göre seçilmişlerdir. Rasch ölçme modelinde üç yüzey bulunmaktadır. Bu yüzeyler, 11 jüri (5 Öğretim üyesi ve 6 doktora öğrencisi) olarak belirtilen değerlendirmeci, 15 yeterlik (Bu yeterlikler YÖK tarafından kabul edilen ve halen geliştirilmeye çalışılan ölçütlerdir) ve değerlendirilen 6 doktora öğrencisidir. Çalışma grubu, Fırat Üniversitesinde Eğitim programları ve Öğretim (EPÖ) Anabilim dalında 2011 Mayıs itibarıyla görev yapan 5 öğretim üyesi ve bu dalın 6 doktora öğrencisidir. Araştırma bulgularına göre, belirlenen ölçütler çerçevesinde en yüksek performansı gösteren bir başka deyişle diğerlerine göre P1 ve P2 kodlu öğrenciler daha yeterli iken P5 ve P6 kodlu öğrenciler daha az yeterli olarak görülmüştür. J5 No’lu jürinin en cömert ve J4 No’lu jürinin de en katı değerlendirmeleri yaptıkları tespit edilmiştir. En zor gerçekleşen görev B3 (Özgün ve disiplinler arası çalışmalarda liderlik yapar) ve en kolay gerçekleşen görev ise İ3 (Bir yabancı dili en az Avrupa Dil Portföyü C1 Genel Düzeyinde kullanarak ileri düzeyde yazılı, sözlü ve görsel iletişim kurar ve tartışır)’dür. Bu çalışmada, doktora eğitimi alan öğrencilerin yeterlikleri arasında farklılıklar görülmektedir. Bu farklılıkların giderilmesi için öğrenci ve öğretim elemanlarının hizmet içi eğitim programlarından geçirilmesi gerektiği üzerinde durulmaktadır.

Akın ve Baştürk (2012) tarafından yapılan çalışmada, Güzel Sanatlar ve Spor Lisesi öğrencilerinin keman eğitiminde kazanmış oldukları temel becerileri çok yüzeyli Rasch ölçme yöntemi ile değerlendirilmesi amaçlanmaktadır. Çok yüzeyli Rasch ölçme modeli;

öğrencilerin bir davranışı gerçekleştirebilme durumuna ilave olarak, puanlayıcı katılığı/cömertliği, davranışın güçlüğü, gibi faktörleri de ekleyerek tek parametrelili modeli çok parametrelili bir model haline getirmiştir. Bu araştırmada müzik öğretmenliği alanında eğitim gören öğretmen adaylarının keman eğitimi ile ilgili temel becerileri, puanlayıcıların katılık/cömertliklerine ve davranışı sergileme gücüne göre değerlendirilmiştir. Araştırmada 3 farklı Anadolu Güzel Sanatlar Lisesinde “Keman eğitimi” dersine devam eden 27 öğrenci örneklem olarak belirlenmiştir. Araştırmada verileri toplamak amacı ile “Keman Çalma Becerilerini Değerlendirme Ölçeği” kullanılmıştır. Sonuç olarak; müzik eğitiminde özellikle performansla dayalı öğrenmelerin değerlendirilmesinde Çok-Yüzeyli Rasch ölçme modelinin etkin olarak kullanılabileceği sonucuna varılmıştır.

Yüzüak, Yüzüak ve Kaptan (2015) tarafından yapılan araştırmada, Genetik ve Biyoteknoloji konulu performans görevinin (poster) akran grupları ve öğretmen tarafından çok-yüzeyli Rasch ölçme modeliyle değerlendirilmesi amaçlanmıştır. Araştırmada tarama modeli kullanılmıştır. Araştırma 2013-2014 eğitim yılı güz döneminde Bartın ilinde çalışan bir öğretmen ve 7. sınıfta öğrenim gören 50 öğrenci ile yürütülmüştür. Değerlendirme için beşer kişilik 10 öğrenci grubu oluşturulmuştur. Her grup kendi içerisinde birer materyal hazırlamıştır. Araştırmada veri toplama aracı olarak öğretmen tarafından revize edilen form kullanılmıştır. Verilerin analizi için FACETS programı kullanılmıştır. Ölçütler çerçevesinde elde edilen analiz sonuçlarına göre Grup 74’ün en cömert puanlayıcı grup olduğu, öğretmenin ise en katı puanlayıcı olduğu; P4’ün ölçütleri en iyi sağlayan görev olduğu, P6’nın ise ölçütleri daha iyi sağlamayan görev olduğu; “Bilgi-doğruluk” kodlu ölçütün en kolay karşılandığı, “Kaynakça” kodlu ölçütün ise en zor karşılandığı ifade edilebilir. Çok-yüzeyli Rasch ölçme modelinin fen eğitiminde performans ve akran grup değerlendirmeleri için etkili bir şekilde kullanılabileceği sonucuna ulaşılmıştır.

İlhan (2016) çalışmasında açık uçlu sorularla yapılan ölçmelerde klasik test kuramı (KTK) ve çok yüzeyli Rasch modeline (ÇYRM) göre hesaplanan yetenek kestirimlerinin karşılaştırılması amaçlamıştır. Araştırma sekizinci sınıfa devam eden 100 öğrenci ile çalışmada puanlayıcı olarak görev alan dört matematik öğretmenin katılımıyla gerçekleştirilmiştir. Çalışmanın verileri, açık uçlu altı matematik sorusundan oluşan bir başarı testi ve bu soruların puanlanmasında kullanılan bütüncül bir rubrik yardımıyla toplanmıştır. Açık uçlu matematik sorularına verilen yanıtların puanlanmasıyla elde edilen veriler, hem KTK’ya hem de ÇYRM’ye göre analiz edilmiştir. Dört puanlayıcının verdiği puanların ortalaması alınarak, KTK’ya ilişkin yetenek kestirimleri hesaplanmıştır. Ardından

puanlayıcı, birey ve madde şeklinde üç yüzeyli bir desen ile çok yüzeyli Rasch analizi uygulanmıştır. Rasch analizinde ulaşılan ve logit cetvelinde rapor edilen yetenek kestirimlerinin puanlamada kullanılan rubriğin birimlerine dönüştürülmesiyle, iki kurama göre hesaplanan yetenek kestirimleri karşılaştırmaya hazır hale gelmiştir. Araştırma sonucunda; KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimleri arasındaki göreceli uyumun son derece yüksek olduğu belirlenmiştir. İki kurama göre hesaplanan yetenek kestirimlerine ait ortalamalar arasında anlamlı fark bulunduğu ve dolayısıyla mutlak bir uyumun söz konusu olmadığı saptanmıştır. ÇYRM'de rapor edilen yetenek kestirimlerinin ölçüt geçerliğinin KTK'dan elde edilen yetenek kestirimlerine kıyasla daha yüksek olduğu tespit edilmiştir.

İlhan ve Güler (2018) tarafından yapılan araştırmada, Likert tipi ölçeklerin psikometrik özelliklerinin incelenmesinde ve bu ölçeklerden alınan puanlar üzerinden gerçekleştirilen farka dayalı istatistiklerde Rasch modelinin kullanımına örnek olabilecek bir çalışmanın alanyazına kazandırılması amaçlanmıştır. Araştırma, 367 ortaöğretim öğrencisinden oluşan bir çalışma grubu üzerinde yürütülmüştür. Araştırmanın verileri Olumsuz Değerlendirilme Korkusu Ölçeği-Öğrenci Formu (ODKÖ-ÖF) ile Akademik Beklentilere İlişkin Stres Envanteri (ABSE) kullanılarak toplanmıştır. Çalışma kapsamında toplanan veriler FACETS paket programından yararlanılarak Rasch modeline göre analiz edilmiştir. Araştırmadan elde edilen bulgular; ODKÖ-ÖF'ye ilişkin 16 maddeden oluşan tek boyutlu modelde öğrenciler ve maddelerin yüksek güvenirlikte birbirinden ayırt edildiğini, ölçekte kullanılan beşli derecelendirmenin etkin bir biçimde çalıştığını, uyum istatistiklerinin kabul edilebilir sınırlar içerisinde kaldığını ve gözlenen ile beklenen test karakteristik eğrilerinin büyük ölçüde örtüştüğünü göstermiştir. Çok yüzeyli Rasch analizi sonucunda, öğrencilerin sınıf ortamında öğretmenleri ya da arkadaşları tarafından olumsuz değerlendirileceklerine dair yaşadıkları korkunun ABSE puanlarına göre farklılaştığı saptanmış ve bu bulgu ODKÖ-ÖF'nin ölçüt geçerliğine yönelik bir kanıt olarak yorumlanmıştır.

Karakaya Özyer (2018) "Açıköğretim sistemindeki açık uçlu soruların Çok Yüzeyli Rasch Modeli analizi yöntemiyle puanlanması"na yönelik çalışmada açık uçlu soruların oluşturduğu sınavların puanlanmasında, bir puanlayıcıdan beklenen soruların zorluk derecesinden ve değerlendirme yöntemlerinden bağımsız olarak değerlendirme aşamasını nesnel ve adil olarak gerçekleştirmeyi amaçlamıştır. 2016-2017 eğitim öğretim yılı itibarıyla Açıköğretim Fakültesindeki bazı sınavlarda açık uçlu sorular kullanılmaya başlanmış ve bu denli nesnel yargılamaya açık sınavların değerlendirilmesinde yeni

yöntemler arayışı içine girilmiştir. Mevcut yöntemlerden en etkililerinden biri de farklı yüzeyleri analize dâhil edebilen Çok Yüzeyli Rasch Modelidir. Çok yüzeyli Rasch modeli yaklaşımı açık uçlu soruların sorulduğu sınavların değerlendirmesinde önemli katkılar sağlamıştır. Ek olarak, farklı değişkenlerin analize katılması sonucunda ortaya çıkabilecek ölçme hatalarıyla başa çıkılmasına olanak tanımaktadır. Bu model, birden fazla değişkenin (öğrenci, puanlayıcı, etkinlikler gibi) aynı anda kalibre edilmesini ve tek bir ölçek üzerinde temsil edilmesini sağlamaktadır. Açık öğretim sınavlarındaki açık uçlu soruların yanıtlarının çok yüzeyli Rasch analizi yöntemiyle analiz edilmesi olası puanlayıcı tutarsızlıklarının belirlenmesine ve öğrencilerin performanslarının daha güvenilir bir biçimde ölçülmesine olanak sağlamıştır.

Govindasamy, Salazar, Lerner ve Green (2019) çalışmasında, öğretmen yetiştirme öğretim programının etkililiğini değerlendirmek için tasarlanmış yeni geliştirilen bir yöntemin ampirik bir değerlendirmesi yapılmıştır. Bu çalışmada, Çok Yüzeyli Rasch Ölçme modeli ile, puanlama anahtarına dayalı değerlendirme aracı olan “Eşitlikli ve Etkili Öğretim Çerçevesinin (FEET)” psikometrik kalitesini değerlendirmek için kullanılmıştır. Analizlerde, (1) puanlayıcıların puanlamadaki katılık/cömertlik derecesini, (2) madde güçlük seviyesini, (3) değerlendirme zamanını ve (4) öğretmen adaylarının verimliliğini incelemeye odaklanılmıştır. Ek olarak analizlerde puanlamaların geçerlilik kanıtı olarak, öğretmen adaylarının öğretim öz değerlendirme raporları ile orta düzeyde korelasyon göstermiştir. Çalışmada, Çok Yüzeyli Ölçme Rasch modelinin öğretmen yetiştirme öğretim programının etkililiğini değerlendirmek için etkili bir model olduğu sonucuna varılmıştır.

Yurt içinde ve yurt dışında yapılmış olan çalışmaları incelediğimizde; Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM)’nin puanlayıcı yanlılıklarının veya hatalarının belirlenmesinde, açık uçlu soruların değerlendirilmesinde, ölçeklerin psikometrik özelliklerinin incelenmesinde, performansa yönelik üst düzey zihinsel becerilerin değerlendirilmesinde etkili olduğu görülmektedir. Ayrıca modelde puanlayıcılar arasında öz, akran ve öğretmen değerlendirmelerin beraber bulunduğu çalışmaların da olduğu görülmüştür. Literatürde ilköğretim öğrencilerinin ikna edici metin yazma becerilerinin belirlenmesinde yani performanslarının değerlendirilmesinde birden fazla puanlayıcının analitik puanlama anahtarı ile puanlanması sürecinde ölçümlerin geçerliği ve güvenilirliğini olumsuz yönde etkileyen puanlayıcı davranışlarının yanı sıra farklı puanlayıcılar arasındaki yanlı puanlayıcı etkileşimlerinin çok yüzeyli Rasch ölçme modeli (ÇYRÖM) ile birlikte analiz edildiği bir çalışmaya rastlanmamıştır.

BÖLÜM III

YÖNTEM

Bu bölümde araştırmanın deseni, çalışma grubu, veri toplama araçları hakkında bilgiler verilmiş olup, verilerin toplanması ve verilerin analizi ile ilgili açıklamalar yapılmıştır.

3.1. Araştırmanın Deseni

Araştırma ilköğretim öğrencilerinin yazma performanslarının değerlendirilmesinde öğretmen puanlayıcılarının puanlama sürecinde göstermiş oldukları puanlayıcı davranışlarının özellikle puanlayıcı yanlılıklarının (diğer bir deyişle farklılaşan puanlayıcı katılık/cömertlik davranışının) ortaya çıkarılmasını hedeflediğinden var olan bir durumun betimlenmesinden dolayı tarama modelinden betimsel bir araştırma özelliği göstermektedir. Tarama modeli, bir grubun belli özelliklerini belirlemek için verilerin toplanmasını amaçlayan çalışmalardır (Büyüköztürk, Kılıç Çakmak, Akgün, Karadeniz ve Demirel, 2017, s.15). Araştırmada tarama modeli kullanılmış olup, var olan yapı olduğu şekliyle betimlenmeye çalışılmıştır. Bu nedenle araştırma, tarama modelinde yürütülen betimsel bir çalışmadır (Balcı, 2015, s.15; Karasar, 2017, s.109).

3.2. Çalışma Grubu

Araştırmanın çalışma grubu, öğrenciler ve puanlayıcılar olmak üzere iki kısımdan oluşmaktadır. Öğrencilerden 7 farklı ikna edici metin yazma görevi içinden bir tanesini seçerek yönergeye uygun olarak yazılar yazmaları, puanlayıcılardan ise öğrencilerin yazmış oldukları çalışmalarını analitik dereceli puanlama anahtarı kullanarak

değerlendirmeleri istenmektedir.

Çalışma grubunun öğrenci kısmını, 2019-2020 eğitim-öğretim yılı ikinci döneminde, Ankara ili, Keçiören ilçesinde yer alan bir ilköğretim okulunda öğrenim gören 7. sınıf düzeyindeki iki şubedeki 32'si kız, 31'i erkek olmak üzere toplam 63 öğrenci oluşturmaktadır. Fakat, uygulama sırasında BEP'li öğrencilerin (yani kaynaştırma öğrencileri) bulunması, sürekli devamsız öğrencilerin olması ve araştırmaya katılımın gönüllülük esasına göre yapılmasından dolayı çalışmaya 57 öğrenci katılmıştır. 57 öğrenciye 7 farklı ikna edici metin yazma görevi içerisinde bir tanesini seçmelerini ve seçmiş oldukları görev için ikna edici yazılar yazmaları istenmiştir. Çalışmaya 57 öğrenci arasından 26'sı kız, 22'si erkek olmak üzere toplam 48 öğrencinin yazmış oldukları metinler dahil edilmiştir. Çalışma grubunun öğrenci kısmına ait cinsiyet ve sınıf değişkenlerine göre dağılımları frekans ve yüzdeleri Tablo 3.1'de verilmiştir.

Tablo 3.1

Öğrencilerin Cinsiyet ve Sınıf Değişkenlerine Göre Dağılımı

Sınıf	<u>7/E</u>		<u>7/İ</u>		<u>TOPLAM</u>	
	f	%	f	%	f	%
Cinsiyet						
Kız	13	%50	14	%63,6	27	%56,25
Erkek	13	%50	8	%36,4	21	%43,75
TOPLAM	26	%100	22	%100	48	100

Çalışma grubunun puanlayıcı kısmını ise, MEB bünyesinde çalışan 8 Türkçe öğretmeni ve 3 alan uzmanı (*Ölçme değerlendirme yüksek lisans eğitimine devam eden 2 kişi ve Türk Dili alanında yüksek lisans yapmış 1 kişi*) toplam 11 puanlayıcıdan oluşturmaktadır. Çalışma grubunun puanlayıcı kısmına ait cinsiyet ve sınıf değişkenlerine göre dağılımları frekans ve yüzdeleri Tablo 3.2'de verilmiştir.

Tablo 3.2

Puanlayıcıların Cinsiyet ve Sınıf Değişkenlerine Göre Dağılımı

Puanlayıcı	<u>Uzman</u>		<u>Öğretmen</u>		<u>TOPLAM</u>	
	f	%	f	%	f	%
Cinsiyet						
Kadın	1	%33,3	5	%62,5	6	%54,5
Erkek	2	%66,7	3	%37,5	5	%45,5
Toplam	3	%27,3	8	%72,7	11	100

Öğretmenlere ve uzman puanlayıcılara öğrencilerin yazmış oldukları ikna edici metinlerin

puanlanması için analitik puanlama anahtarının nasıl kullanılacağına yönelik ve olası puanlayıcı hatalarının ortaya çıkmaması açısından araştırmacı tarafından puanlayıcı eğitimi verilmiştir. Seçilen öğretmenlerin; MEB bünyesinde çalışıyor olması, en az 5 yıllık tecrübeye sahip olması ve Türkçe alan yeterliliğine sahip olması şartı aranmıştır. Öğretmenler ve uzman puanlayıcılar araştırma konusunda bilgilendirilmiş olup, katılımlarda gönüllülük esas alınmıştır.

3.3. Veri Toplama Araçları

Araştırmanın amaçları doğrultusunda veri toplama aracı olarak öğrencilerin yazma becerilerini (ikna edici metin yazma) ölçmek üzere araştırmacı tarafından geliştirilen yazma performansına yönelik yedi tane ikna edici metin yazma görevi oluşturulmuştur. Söz konusu görevlerin öğretmenler tarafından değerlendirilmeleri için yine araştırmacı tarafından geliştirilen 6 ölçütten oluşan analitik dereceli puanlama anahtarları (rubrik) kullanılmıştır.

3.3.1. İkna Edici Metin Yazma Görevi

İlköğretim öğrencilerinin ikna edici metin yazma becerilerinin ölçülmesinde araştırmacı tarafından hazırlanan “*ikna edici metin yazma görevleri*” oluşturulmuştur. Yazma performansına yönelik araştırmacı tarafından 11 tane görev yazılmış, her bir görev için öğrencilerin ikna edici metin yazma becerisini ortaya çıkaracak şekilde görsellerden yararlanılmıştır. Sonrasında her bir yazma görevinin altında yazma becerisini ölçmeye yardımcı olacak şekilde açıklamalar hazırlanmıştır. Hazırlanan 11 görev içerisinde yazma performansının belirlenmesine yönelik uzman görüşleri doğrultusunda araştırma için 7 ikna edici metin yazma görevi seçilmiştir. Mevcut çalışmada yer alan ikna edici metin yazma görevleri;

Görev-1: Evcil hayvan besleme (EK 1)

Görev-2: Eve internet bağlatma (EK 2)

Görev-3: Çanakkale’ye gezi (EK 3)

Görev-4: Okul takımı seçmeleri (EK 4)

Görev-5: Geri dönüşüm (Katı atık toplama) (EK 5)

Görev-6: Kişisel temizlik (EK 6)

Görev-7: Benim kararım (EK 7) şeklinde belirlenmiştir.

Yazma becerilerinin ölçülmesine yönelik bu çalışmada öğrencilere 7 farklı ikna türünde yazma görevi verilmiş, öğrenciden bu görevlerden sadece bir tanesi ile ilgili kompozisyon yazması istenmiştir. Yazma görevi için çalışmanın öncesinde araştırmacı tarafından öğrencilere ikna türündeki metinlerle ilgili 10 dk. süre ile açıklama yapılmış, kompozisyon yazmaları için 25 dk. süre verilmiş ve bu süre içerisinde öğrencilerden yönergeye uygun olarak ikna edici en az 10 cümle yazmaları istenmiştir. Yazma görevi araştırmacı ve alan uzmanı öğretmen tarafından 1 ders saati (40 dk.) içerisinde uygulanmış olup, çalışmaya 7.sınıfta öğrenim gören iki şubeden 63 öğrenciden 57'si katılmıştır. Fakat, 9 öğrencinin yazmış oldukları metinler yönergeye uygun yazılmadığından dolayı araştırmacı tarafından 48 öğrencilerin metinleri çalışmaya dahil edilmiştir.

3.3.2. Analitik Dereceli Puanlama Anahtarları (Rubrik)

İlköğretim öğrencilerinin yazdıkları ikna edici metin yazma çalışmalarının değerlendirilmesi amacıyla Ölçme ve Değerlendirme uzmanlarından ve Türk dili konusunda çalışan uzmanlardan görüşler alınarak araştırmacı tarafından analitik dereceli puanlama anahtarı geliştirilmiştir. Dereceli puanlama anahtarları, öğrencilere yaptıkları çalışmaların hangi ölçütlere göre değerlendirileceğini ve performanslarının hangi düzeydeki puana denk geleceğini gösteren puanlama araçlarıdır (Kutlu, Doğan ve Karakaya, 2014, s.51). Mevcut çalışmada dereceli puanlama anahtarı geliştirmek için Haladyna (1997) tarafından önerilen adımlar dikkate alınmıştır;

1. Dereceli puanlama anahtarının kullanım amacının belirlenmesi,
2. Ölçütlerin açıkça belirtilmesi,
3. Bütünsel ya da analitik dereceli puanlama anahtarının kullanılacağına karar verilmesi,
4. Dereceli puanlama anahtarının genel görev için mi özel görev için mi kullanılacağına karar verilmesi,
5. Dereceli puanlama anahtarının taslağının oluşturulması,
6. Düzeltmelerin yapılması,

7. Deneme uygulamasının yapılması,
8. Sonuçların değerlendirilmesi,
9. Geçerlik ve güvenirlik çalışmalarının yapılması,
10. Gelecek kullanımlar için düzeltmelerin yapılması.

Buna dayalı olarak puanlama anahtarı için ölçme aracının genel yapısı dikkate alınarak araştırmacı tarafından analitik dereceli puanlama anahtarının kullanılması uygun görülmüştür. Analitik puanlama anahtarının ölçütleri belirlenirken Kutlu, Doğan ve Karakaya (2014) tarafından;

- Ölçülmek istenen davranışlarla DPA arasında uyum olması
- Çok sayıda göreceli anlatımların kullanılması
- Puanlama anahtarında bulunan ölçütlerin binişik olmasından kaçınılması

gibi durumlar dikkate alınarak; üst düzey zihinsel becerilerin gelişimine olanak tanıyan yazma becerilerinin değerlendirilmesinde analitik dereceli puanlama anahtarında ikna edici metin yapısına uygun olarak 6 ölçüt belirlenmiştir. Araştırmacı tarafından ikna edici metin yazma becerilerinin değerlendirilmesine yönelik “*görüş sunma, içerik, yapı, dil ve anlatım, yazım ve noktalama, sayfa düzeni*” gibi 6 ölçüt oluşturulmuştur. Ölçütler analitik dereceli puanlama anahtarı üzerinde istenen davranışın doğasına uygun olarak *Oldukça Başarısız (1), Orta Düzey (2), İyi (3), Oldukça İyi (4)* şeklinde puanlanmaktadır.

Geliştirilen Ölçme Araçlarının Geçerlik ve Güvenirliğine İlişkin Açıklamalar

Araştırmacı tarafından hazırlanan ikna edici metin yazma görevleri ile analitik dereceli puanlama anahtarının geçerlik ve güvenirliği uzman görüşleri alınarak sağlanmaya çalışılmıştır. İlk olarak düzenlenen Yazma Görevleri ve Analitik Dereceli Puanlama Anahtarı; Uzman Değerlendirme Formu (EK-11) ile birlikte üçü Türkçe eğitimi alanında uzman öğretim üyesi, biri Türkçe eğitimi doktora öğrencisi, biri ise Ölçme ve Değerlendirme alanında uzman öğretim üyesi toplam beş kişi tarafından incelenmiştir.

Elde edilen verilerle ikna edici metin yazma görevlerinin ve analitik dereceli puanlama anahtarının kapsam geçerlik indeksleri (KGİ) “*Davis tekniği*” kullanılarak hesaplanmıştır. Bu teknikte en az 3, en fazla 20 uzman önerilmektedir. Bu doğrultuda mevcut çalışmada 5 uzmandan görüş alınmıştır. Ölçme ve değerlendirme araçlarının son hali uzman görüşleri

doğrultusunda son şeklini almıştır.

Uzman Görüşü: Geliştirilen yedi yazma görevi ve analitik dereceli puanlama anahtarı Türkçe dil kurallarına uygunluk, bilimsel doğruluk, ölçme ve değerlendirme ilkelerine uygunluk, yaş düzeyine uygunluk ve görevlerin eş güçlük düzeyinde olup olmadıklarını konularında 4 Türkçe alan uzmanı ve bir Ölçme ve Değerlendirme uzmanı olmak üzere 5 uzmanın yazılı görüşleri alınarak düzenlenmiştir. Alınan görüşler doğrultusunda 11 tane ikna edici metin yazma görevi içerisinde 7 tane yazma görevi seçilmiştir.

Kapsam Geçerlik İndeksi (KGİ): Kapsam geçerliği için kapsam geçerlik indeksi (KGİ), *Davis tekniği* ile belirlenmiştir. Davis (1992) tekniği uzman görüşlerini (4) “Uygun”, (3) “Madde hafifçe gözden geçirilmeli”, (2) “Madde ciddi olarak gözden geçirilmeli” ve (1) “Madde uygun değil” şeklinde dördü derecelendirmektedir. Bu teknikte (4) ve (3) seçeneğini işaretleyen uzmanların sayısı toplam uzman sayısına bölünerek maddeye ilişkin “*kapsam geçerlik indeksi (KGİ)*” elde edilmektedir ve bu değer 0,80 değeri ölçüt olarak kabul edilmektedir (Yurdugül 2005).

Tablo 3.3

Yazma Görevleri ve Puanlama Anahtarı Kapsam Geçerlik İndeksleri (KGİ)

KGİ	
Yazma Görevi	Ölçütler
	1
	2
	3
	4
	5
Puanlama Anahtarı	Ölçütler
	1
	2
	3
	4
	5

* KGİ $\geq$ 0,80 olanlar

Uzman Değerlendirme formunda, yazma görevleri ve puanlama anahtarında olması gereken uygunluğuna bakılacak 10 ölçüt ve her ölçütün altında öneriler kısımları bulunmaktadır. Ayrıca formun en altında genel önerilerin yazılabileceği bir kısım bulunmaktadır. Uzman değerlendirme formunda, 4’lü Likert tipi ölçek türü kullanılmakta olup derecelendirmeler “(4) tamamen uygun, (3) uygun, (2) bazı değişiklik gerekiyor ve son olarak (1) tamamen değişiklik gerekiyor” anlamına gelmektedir. Yazma Görevleri ve

puanlama anahtarı değerlendirme formunda 4'lü Likert tipi dereceleme kullanıldığından, çalışmada Davis tekniğinden faydalanılmıştır. Yazma görevleri ve puanlama anahtarı için hesaplanan KGİ'ler Tablo 3.3'te gösterilmiştir. Tablo 3.3'te görüldüğü gibi, yazma görevlerinin ve puanlama anahtarlarının kapsam geçerlik indeksleri(KGİ) 0.80 ile 1 arasında değiştiği görülmektedir. Bu değerlere bakıldığında KGİ değerlerinin 0.80 ve üzerinde değerler almasından dolayı; ölçme ve değerlendirme araçlarının geçerliğine ilişkin yeterli kanıtlar toplandığına karar verilmiş olup, veri toplama sürecine geçilmiştir.

Yapı Geçerliği: Araştırmacı tarafından hazırlanan analitik dereceli puanlama anahtarının uzman görüşleri alınarak kapsam geçerliği sağlandıktan sonra ölçme aracının yapı geçerliğine bakılmıştır. Bunun içi dereceli puanlama anahtarının yapı geçerliğine yönelik *açımlayıcı faktör analizi (AFA)* yöntemine başvurulmuştur. Açımlayıcı faktör analizi yapmadan önce gerekli sayıtların sağlanıp sağlanmadığı kontrol edilmiştir. İlk olarak veri setinin faktörleşmeye uygun olup olmadığının belirlenmesi amacıyla Kaiser-Meyer-Olkin (KMO) ve Bartlett's küresellik test sonuçları incelenmiştir. Analiz sonucunda Bartlett's küresellik testinin istatistiksel olarak anlamlı olduğu (χ^2 (sd) =1030,137(15), $p<.001$) ve KMO değerinin 0,804 olduğu bulunmuştur. Bu değer 0,70'den yüksek olması iyi, 0,50-0,70 arasında olması yeterli, 0,50'nin altı ise yeterli ilişkiyi sağlayacak örneklem ihtiyacı olduğunu göstermektedir (Çokluk, Şekercioğlu, Büyüköztürk, 2012, s. 206; Can, 2017, s.325). AFA analizi sonucunda ölçütlere ait faktör yük değerleri (*Görüş Sunma, İçerik, Yapı, Dil ve Anlatım, Yazım ve Noktalama, Sayfa Düzeni*) 0,729; 0,689; 0,721; 0,838; 0,685; 0,658 olarak bulunmuştur. Ayrıca ölçme aracına ait faktör özdeğerleri incelendiğinde; birinci faktör yük değerine ait açıklanan varyans oranının %30'un üzerinde olduğu bulunmuştur. Büyüköztürk (2010) faktör analizinde, ilk faktörün açıkladığı varyans oranı %30 ve üzeri olduğu durumlarda ölçeğin tek boyutlu bir yapı sergileyebileceğini belirtmektedir. Yapılan AFA analizleri sonucunda derecelendirme ölçeğinin ölçütlerinin tek bir faktör altında toplandığı ve açıklanan varyans oranının %52,173 olduğu bulunmuştur. İlgili ölçme aracının tek faktörlü bir yapıya sahip olması, ölçme aracının ikna edici metin yazma becerilerini ölçmeye yarar bir araç olduğu göstermektedir.

Ölçümlerin güvenilirliği için McDonald (1999) tarafından geliştirilen güvenilirlik katsayısı (ω) kullanılmıştır. Mevcut çalışmada analitik puanlama anahtarının ölçütlerinin faktör yük değerleri birbirinden farklı olduğundan dolayı bu tür ölçümlerde daha tutarlı kestirimlerin elde edilebilmesi (Osburn, 2000) için McDonald ω katsayısının kullanılması gerekmektedir. Yapılan analizler doğrultusunda McDonald ω katsayısının 0,817 (%95

Güven Aralığı: 0,754 – 0,823) olduğu bulunmuştur. Bu sonuca göre; ilgili ölçme aracından elde edilen sonuçların geçerli ve güvenilir sonuçlar verebileceği görülmektedir.

3.3.3. Puanlayıcı Formu (Öğretmen Değerlendirme Formu)

Puanlayıcıların (Türkçe öğretmenlerinin) ilköğretim 7. sınıf öğrencilerinin yazma performansını ölçmeye yönelik analitik dereceli puanlama anahtarına ek olarak araştırmacı tarafından geliştirilen puanlama anahtarının tüm ölçütlerini, görev numaralarını ve öğrenci numaralarını içeren bir form oluşturulmuştur (EK-9). Bu formda ek olarak; puanlayıcıların eğitim düzeyi, kıdem yılı, çalıştığı okul türü gibi bilgilerin yer aldığı kişisel bilgilerine ait doldurulması gereken alanlar da bulunmaktadır.

3.4. Verilerin Toplanması

Mevcut çalışma, 2019-2020 eğitim ve öğretim yılı ikinci yarısında Ankara ili Keçiören ilçesindeki bir devlet okulunda yürütülmüştür. Çalışmada 7. sınıflar içerisinde iki farklı şubede yazma çalışması yürütülmüştür. Her sınıf için ayrılan süre ortalama birer ders saati (40dk + 40dk)'dir. Uygulama için ilk olarak sınıflarda ikna edici metin yapısı ile ilgili bilgilendirmeler yapılmış ve ikna edici metin örnekleri gösterilmiştir. Sonrasında öğrencilere 7 farklı ikna edici yazma görevi gösterilmiş, öğrencilerden bu görevlerden kendilerine uygun olan bir görev seçmeleri istenmiş ve yönergeler doğrultusunda bir kompozisyon yazmaları istenmiştir. Çalışmanın verileri toplam 2 ders saati içerisinde toplanmıştır. Toplanan verilerde her bir öğrenciye ait kompozisyonlar değerlendirilmeden önce puanlamanın geçerliği ve güvenilirliğini sağlamaya yönelik olarak yazma görevleri G1, G2, G3, ... şeklinde; öğrencilerin yazmış olduğu metinler ise Ö1, Ö2, ... şeklinde; puanlayıcılar ise uzman puanlayıcılar için PU1, PU2,... ve öğretmen puanlayıcılar için PÖ1, PÖ2,... şeklinde numaralandırılmıştır.

Çalışmada toplanan veriler devlet okullarında çalışmakta olan 8 Türkçe öğretmeni ve 3 uzman (puanlayıcılar) tarafından değerlendirilmiştir. Değerlendirme öncesi araştırmacı tarafından puanlayıcılara puanlamaların geçerliği ve güvenilirliğini sağlamaya yönelik analitik dereceli puanlama anahtarının kullanımı hakkında puanlayıcı eğitimi verilmiştir. Puanlayıcı eğitimi olarak; puanlamaya karışan ölçme hataları (halo etkisi, merkeze eğilim, puanlayıcı katılık/cömertlik, tutarsızlık, yanlılık, vs.) hakkında bilgi verilmiş, yazma performansına yönelik öğrenci çalışmaları incelenmiş, puanlama anahtarının ölçütleri ve

derecelendirme sistemi hakkında ayrıntılı açıklamalar yapılmış ve analitik puanlama anahtarının kullanımına yönelik araştırmacı tarafından örnek puanlamalar gösterilmiştir. Puanlayıcı eğitimi sonrasında öğretmen puanlayıcılarına; öğrenci kompozisyonları, analitik puanlama anahtarı ve ikna edici metin yazma performansına ilişkin öğretmen değerlendirme formunu içeren bir dosya verilmiş, öğretmenlerden bir hafta içerisinde kompozisyonları değerlendirmeleri istenmiştir. Öğretmen puanlamaları sonrasında elde edilen veriler düzenlenerek FACETS 3.81.0 (Linacre, 2018) paket programında Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM)'ne göre analiz edilmiştir.

3.5. Verilerin Analizi

Verilerin analizinde performansının değerlendirilmesinde; puanlayıcı davranışları ve farklılaşan puanlayıcı katılık/cömertlik davranışının (puanlayıcı yanlılıklarının) belirlenmesinde madde tepki kuramı (MTK)'na bağlamında Çok yüzeyle Rasch Ölçme Modeli (ÇYRÖM)'ne dayalı olarak hesaplanan çıktılar yorumlanmış, puanlayıcı güvenilirliğinin belirlenmesinde ise KTK (klasik test kuramı) bağlamında sınıf-içi korelasyon (ICC) katsayısı, GK (genellenebilirlik kuramı) bağlamında G katsayısı ile MTK (madde tepki kuramı) bağlamında Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM)'ne bağlı olarak hesaplanan güvenilirlik katsayısı ve puanlayıcılar arası uyuma bağlı olarak hesaplanan Rasch Cohen's-Kappa değeri sonuçları karşılaştırılmıştır.

Çalışmada; puanlayıcı (Türkçe öğretmenleri ile uzmanlar), puanlayıcı türü, öğrenci, cinsiyet, görevler (7 görev), ölçütler (6 ölçüt) olmak üzere altı yüzey bulunmaktadır. 7. sınıf öğrencilerinin ikna edici metin yazma çalışmaları sonucunda öğretmen puanlamaları sonrasında elde edilen veriler FACETS 3.81.0 (Linacre, 2018) paket programı kullanılarak *Çok Yüzeyle Rasch Ölçme Modeli* (ÇYRÖM)'ne göre analiz edilmiştir. Araştırmada puanlayıcılar tüm öğrenci metinlerini bütün ölçütler üzerinden değerlendirdiğinden tamamen çaprazlanmış desen kullanılmıştır. Toplanan verilerin Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM)'ne göre analiz edilebilmesi için gerekli üç temel sayıltının sağlanıp sağlanmadığı kontrol edilmiştir. Bu temel sayıltılar; *tek boyutluluk*, *yerel bağımsızlık* ve *model veri uyumu*'dur.

Ölçme aracının tek boyutluluk sayıltısı için araştırmacı tarafından geliştirilen analitik dereceli puanlama anahtarının tüm ölçütlerinin tek bir yapıyı ölçmesi gerekmektedir. Bu bağlamda ölçme aracının geliştirilmesi kısmında puanlama anahtarının tüm ölçütlerine

yönelik *açımlayıcı faktör analizi (AFA)* sonucunda ölçme aracının tek bir yapıyı ölçtüğü yani; tek boyutluluk sayılıtısının sağlandığı görülmektedir.

Toplanan verilerde yerel bağımsızlık sayılıtısının sağlanması için tek boyutluluk sayılıtısının mutlaka sağlanmış olması gerekir (Erkuş vd., 2017, s.126). Yerel bağımsızlık; bireyin performansını etkileyen değişkenler sabit tutulduğunda, bireylerin testteki maddelerin herbirine verdiği tepkilerin birbirinden bağımsız olması, tek boyutluluk varsayımının karşılanması yerel bağımsızlık varsayımının da sağlandığı şekilde yorumlanabilmektedir (Hambleton ve Swaminathan, 2013; s. 23). Bu açıdan bakıldığında tek boyutluluk sayılıtısı sağlandığından dolayı analitik puanlama anahtarının tüm ölçütlerinin yerel bağımsızlık sayılıtısını sağladığı söylenebilir.

Son olarak model veri uyumu için ± 2 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısının toplam gözlem sayısının % 5'inden az ve ± 3 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısının ise toplam gözlem sayısının % 1'inden az olması gerekmektedir (Linacre, 2018; s.171). Ölçme aracının uygulaması için toplam gözlem sayısı $11 \times 48 \times 6 = 3168$ iken ± 2 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısının 126 (%3,97'si) ve ± 3 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısının 14 (%0,44'ü) olduğu görülmüştür. Belirlenen aralıklardaki standartlaştırılmış artık değerlerine bakıldığında verilerin Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM) analizinin yapılabilmesi için model-veri uyumu sayılıtısını sağladığı görülmüştür. Sonuç olarak; gerekli tüm sayılıtların sağlandığı görülmüş ve analizlere geçilmiştir.

BÖLÜM IV

BULGULAR VE YORUM

Bu bölümde her bir araştırma sorusu ve alt problemlerine ilişkin bulgulara ve alanyazına dayalı yorumlamalara yer verilmiştir.

4.1. Birinci Araştırma Sorusuna İlişkin Bulgular ve Yorum

İlköğretim öğrencilerinin performanslarının değerlendirilmesi sürecinde öğretmen puanlamalarına karışan puanlayıcı davranışlarından “*Puanlayıcı katılık ve cömertlik davranışı*”, “*Halo etkisi*”, “*Merkeze eğilim davranışı*”na yönelik bulgular incelenmiştir. Puanlayıcı davranışlarının belirlenmesinde her bir değişkenlik kaynağına yönelik yüzeyler için ilk olarak grup düzeyindeki istatistiksel göstergeler, sonrasında ise bireysel düzeyde istatistiksel göstergeler incelenmiştir. Puanlayıcı davranışlarının belirlenmesinde yorumlanan grup düzeyindeki istatistiksel göstergeler; Ki-kare istatistikleri (Sabit etkili ve Normal Ki-Kare), ayırma oranı, ayırma indeksi ve ayırma indeksinin güvenilirliği’dir. Bireysel düzeydeki istatistiksel göstergeler ise; Uygunluk ölçütleri (uyum-İçi ve uyum-dışı değerler), logit ölçütleri’dir. Mevcut çalışmada 11 farklı puanlayıcı, 48 öğrencinin yazma performansını, analitik dereceli puanlama anahtarının 6 değerlendirme ölçütünü dikkate alarak puanlamıştır. Çalışmada öğrencilerden 7 farklı yazma görevi içerisinde bir tanesini seçip ikna edici metinler oluşturmaları istenmiştir. Çalışmada verilere dayalı olarak öğretmen puanlamalarına ilişkin puanlayıcı davranışlarının belirlenmesine yönelik bulgular ve yorumlamalara yer verilmiştir.

4.1.1. Alt Problem 1.a'ya İlişkin Bulgular ve Yorum

Puanlayıcı katılık ve cömertlik davranışı durumları nedir?

Performansın değerlendirilmesi sürecinde puanlamadaki farklılıkların belirlenmesine yönelik her bir puanlayıcının katılık ve cömertlik davranışı, grup düzeyinde ve bireysel düzeydeki istatistiksel göstergelere dayalı olarak incelenmiştir. Bunun için puanlayıcı yüzeyine ilişkin ÇYRÖM analizi çıktısı Tablo 4.1'de verilmiştir.

Tablo 4.1

Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu

Puanlayıcı No	Model		Uygunluk Ölçütü		Uyum %	
	Logit Ölçütü	Std. Hata	Uyum içi (INFIT)	Uyum dışı (OUTFIT)	Gözlenen Uyum %	Beklenen Uyum %
PU2	0,66	0,09	1,19	1,02	63,9	43,2
PÖ4	0,46	0,08	0,95	0,90	52,2	39,1
PU3	0,35	0,09	0,97	0,91	62,0	43,7
PÖ5	0,17	0,08	0,91	0,87	50,9	39,7
PÖ2	0,07	0,08	0,96	0,91	53,0	39,7
PÖ1	0,06	0,08	0,82	0,86	53,4	39,7
PÖ8	-0,15	0,08	0,92	0,94	49,6	39,7
PU1	-0,31	0,08	1,21	1,17	44,3	40,1
PÖ3	-0,39	0,08	1,00	1,01	45,4	38,5
PÖ7	-0,45	0,08	1,10	1,10	47,6	38,2
PÖ6	-0,49	0,08	1,12	1,10	43,5	38,0
Ortalama	0,73	0,08	1,01	0,98		
Ss(Evren)	0,37	0,00	0,12	0,10		
Ss(Örnekleme)	0,39	0,00	0,12	0,11		
Model, Evren: RMSE=0,08 Ss= 0,36 Ayırma Oranı=4,44 Ayırma İndeksi= 6,25 Güvenirlik=0,95						
Model, Örneklem: RMSE=0,08 Ss=0,38 Ayırma Oranı=4,66 Ayırma İndeksi= 6,55 Güvenirlik=0,96						
Model, Sabit etkili Ki-kare=219,1 sd= 10 p= 0,00						
Model, Normal Ki-kare= 9,6 sd= 9 p= 0,39						
Puanlayıcılar arası gözlenen uyum: %50,1						
Puanlayıcılar arası beklenen uyum: %39,4						
Puanlayıcılar arası güvenilirliğe ilişkin Kappa İstatistik değeri:0,18						
Not: sd: serbestlik derecesi, ss: standart sapma, RMSE:hata kareleri ortalamasının karekökü						

İlk olarak grup düzeyindeki istatistiksel göstergeler incelenmiştir. Grup düzeyindeki istatistiksel göstergeler sonucunda, puanlayıcıların puanlamada farklılık gösterdikleri belirlendikten sonra bu farklılığın hangi puanlayıcı veya puanlayıcılardan kaynaklandığı belirlenmesi gerekmektedir (Çetin ve İlhan, 2017; İlhan, 2015 ve Myford ve Wolfe, 2004).

Tablo 4.1'de grup düzeyindeki istatistiksel göstergeler incelendiğinde; ayırma oranı 4,66, ayırma indeksi 6,55 ve ayırma indeksinin güvenirliliği 0,96 olarak bulunmuştur. Grup düzeyindeki istatistiksel göstergelerden ayırma indeksine ait güvenirlilik katsayısı yüksek olduğundan puanlayıcıların öğrenci performanslarında puanlamada farklılık gösterdiği

söylenbilir. Buna ek olarak, grup düzeyindeki istatistiklerden “Sabit etkili χ^2 ” değeri yorumlanmaktadır. Mevcut çalışmada veri setinin doğasına uygun olarak Tablo 4.1’de Sabit etkili χ^2 değerine bakılmış ve χ^2 istatistiğinin anlamlı olduğu ($\chi^2(sd)= 219,1 (10)$, $p=.00 < .01$) bulunmuştur. Buradan χ^2 istatistiğinin anlamlı olması, puanlayıcıların (Türkçe öğretmenlerinin ve uzman puanlayıcıların) puanlama yaparken farklı davranışlar sergilediğini göstermektedir. Ayrıca, puanlayıcıların farklı davranışlar gösterdiğinin diğer bir kanıtı da puanlayıcılar arası uyum değerlerine ait yüzdeler kullanılarak elde edilen Rasch-Cohen’s Kappa istatistiği’dir. Puanlayıcılar arası gözlenen uyum ile beklenen uyum yüzdesi arasındaki farka dayalı olarak hesaplanan Rasch-Cohen’s Kappa istatistiği -1 ile +1 arasında değişen değerler almaktadır ve Rasch modeli koşullarında bulunan bu değer 0’a yakın olması beklenmektedir (Eckes, 2015: s.92; Linacre, 2018: s.306). Rasch-Cohen’s Kappa istatistiği sonucu bulunan bu değer için 0,00’ın üzerindeki değerlerde farklı puanlayıcıların puanlamaları arasında tutarlı sonuçlar elde edildiği yani puanlayıcılar arasındaki güvenilirliğin (uyumun) yüksek olduğu; 0,00’ın altındaki değerlerde ise farklı puanlayıcıların puanlamaları arasında tutarsız sonuçlar elde edildiği yani puanlayıcılar arasındaki güvenilirliğin (uyumun) düşük olduğu söylenebilir. Mevcut verilere dayanarak puanlayıcılar arası gözlenen uyum %50,1 ve puanlayıcılar arası beklenen uyum %39,4 olup, puanlayıcılar arası güvenilirliğe ilişkin Rasch-Cohen’s Kappa istatistiği değeri 0,18 olarak bulunmuştur. Bulunan bu değer 0,00’dan büyük olması puanlayıcıların performansın değerlendirilmesi sürecinde kısmen de olsa tutarlı değerlendirmeler yaptığı yani puanlayıcılar arası uyumun dolayısıyla puanlayıcılar arası güvenilirliğin sağlandığı yorumu yapılabilir.

Tablo 4.1’de bireysel düzeydeki istatistiksel göstergeler incelendiğinde; en cömert puanlayıcının PÖ6 (logit = -0,49) ve en katı puanlayıcının ise PU2 (logit = 0,66) olduğu görülmektedir. Bu durumda logit değeri negatif olan PÖ6 puanlayıcısı tüm öğrencilerin performanslarını değerlendirirken diğer puanlayıcılara nazaran daha çok puan verme eğilimindeyken; logit değeri pozitif olan PU2 puanlayıcısı tüm performanslara daha düşük puanlar verme eğilimindedir. Performansın değerlendirilmesi sürecinde 48 öğrencinin değerlendirildiği bu çalışmada genel olarak 6 puanlayıcının katı (logit > 0), 5 puanlayıcının ise cömert (logit < 0) puanlama davranışı sergilediği görülmektedir. Puanlayıcıların uygunluk istatistiklerine dayalı olarak hesaplanan uyum içi ve uyum dışı değerlerinin kabul edilebilir bir aralıkta (0,5 ile 1,5 aralığında) olduğu görülmektedir.

Alanyazın incelendiğinde araştırmaların sonuçlarını destekleyen çok sayıda çalışmanın olduğu görülmektedir. Köse, Usta ve Yandı (2016) tarafından yapılan çalışmada, bütün puanlayıcıların puanlama davranışlarının güvenilir olduğu, katılık ve cömertlik yönünden güvenilir bir şekilde sıralandıkları ve birbirlerinden farklılıklar gösterdiği görülmüştür. Mevcut çalışmada Tablo 4.1'deki uyum istatistiklerine bakıldığında puanlayıcıların birbirlerine göre farklı puanlamalar yapmasına rağmen, ÇYRÖM'ne göre puanlayıcılar öğrencilerin performanslarını tüm değerlendirme ölçütlerini dikkate alarak tutarlı bir şekilde puanlama yaptıkları yorumu yapılabilir. Benzer bir çalışma da Polat (2018) tarafından yapılmış, ilgili araştırmada puanlayıcıların ne kadar deneyimli veya iyi eğitilmiş olsalar bile puanlayıcı katılık/cömertlik davranışının görülebileceğini sonucuna varmıştır. Mevcut çalışmada öğretmen puanlayıcılarının puanlama öncesi araştırmacı tarafından puanlayıcı eğitimi almasına rağmen, puanlamalarda farklılıkların olması yani puanlayıcı katılık ve cömertlik davranışının görülmesi bu düşüncüyü desteklemektedir.

4.1.2. Alt Problem 1.b'ye İlişkin Bulgular ve Yorum

Halo etkisi durumları nedir?

Performansın değerlendirilmesi sürecinde ortaya çıkma olasılığı yüksek olan diğer bir puanlayıcı davranışı da *Halo etkisi*'dir. Puanlayıcıların halo davranışı sergileyip sergilemediğinin belirlenmesi için ilk olarak ölçme aracının değerlendirme ölçütlerinin güçlükleri eşitlenerek ÇYRÖM analizi tekrarlanmalıdır. Eğer bu analiz çıktıları sonucunda yani değerlendirme ölçütlerinin güçlüklerinin eşitlenmesinden sonra oluşan çıktılarda uyum içi ve uyum dışı istatistik değerleri mükemmel uyum (1,0'a eşit olma) gösteren puanlayıcılar varsa, ilgili puanlayıcıların halo davranışı göstermediği kabul edilmektedir (Myford ve Wolfe, 2004). Mevcut çalışmada da yazma performansının değerlendirilmesinde analitik puanlama anahtarının değerlendirme ölçütlerinin güçlükleri eşitlenip, ÇYRÖM ile analiz edilmiştir. Genel olarak halo etkisini belirleyebilmek için değerlendirme ölçütlerinin yüzeyine ilişkin ölçüm raporu incelenmektedir. Halo etkisini belirleyebilmek için önce grup düzeyindeki istatistiksel göstergeler sonrasında ise halo davranışı göstergesi var ise bireysel düzeyde istatistiksel göstergelere bakılmaktadır. Mevcut çalışmada analitik DPA'nın 6 ölçütünün güçlükleri eşitlenerek ölçüt yüzeyine ilişkin ÇYRÖM analizi çıktısı Tablo 4.2'deki gibidir.

Tablo 4.2

Ölçüt Güçlükleri Eşitlendiğinde Ölçüt Yüzeyine İlişkin Ölçüm Raporu

Ölçütler	Ortalamalar		Model		Uygunluk Ölçütü	
	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Ölçütü	Std. Hata	Uyum içi (INFIT)	Uyum dışı (OUTFIT)
<i>Görüş Sunma</i>	3,11	3,26	0,49	0,06	0,91	0,84
<i>İçerik</i>	3,02	3,17	0,32	0,06	1,14	1,11
<i>Dil ve Anlatım</i>	2,98	3,12	0,23	0,06	0,70	0,68
<i>Sayfa Düzeni</i>	2,89	3,02	0,06	0,06	1,06	1,05
<i>Yazım ve Noktalama</i>	2,70	2,78	-0,29	0,06	1,22	1,17
<i>Yapı</i>	2,40	2,42	-0,82	0,06	1,02	1,04
Ortalama	2,85	2,96	0,00	0,06	1,01	0,98
Ss(Evren)	0,24	0,28	0,44	0,00	0,17	0,17
Ss(Örnekleme)	0,26	0,31	0,48	0,00	0,18	0,18
Model, Evren: RMSE= 0,06 Ss= 43 Ayırma Oranı= 7,23 Ayırma İndeksi= 9,98 Güvenirlilik=0,98						
Model, Örnekleme: RMSE= 0,06 Ss= 48 Ayırma Oranı= 7,94 Ayırma İndeksi= 10,91 Güvenirlilik=0,98						
Model, Sabit etkili Ki-kare= 325,0 sd= 5 p= 0,00						
Model, Normal Ki-kare= 4,9 sd= 4 p= 0,29						
Not: sd: serbestlik derecesi, ss: standart sapma, RMSE:hata kareleri ortalamasının karekökü						

Tablo 4.2 incelendiğinde, grup düzeyindeki istatistiksel göstergelerden ayırma oranının 7,94; ayırma indeksinin 10,91; ayırma indeksinin güvenilirliğinin 0,98 ve sabit etkili Ki-kare değerinin istatistiksel olarak anlamlı ($\chi^2(sd)=325,0$ (5), $p= .00 < .01$) olduğu görülmektedir. Ki-Kare değerinin anlamlı bulunması puanlayıcıların halo davranışı göstermediğinin bir kanıtıdır. Bu değerlere göre analitik puanlama anahtarının değerlendirme ölçütlerinin güçlük düzeylerinin birbirinden farklı olduğu dolayısıyla yapılan puanlamalarda puanlayıcıların halo davranışı sergilemedikleri söylenebilir. Puanlayıcıların halo davranışı göstermediklerinin kanıtı olarak bireysel düzeydeki istatistiksel göstergelerden puanlayıcılara ait uyum içi ve uyum dışı değerleri incelenmiştir. Myford ve Wolfe (2004) değerlendirme ölçütlerin güçlükleri eşitlendikten sonra yapılan analizlerde uygunluk değerleri 1'den önemli derecede farklılık gösteren puanlayıcı varsa, o puanlayıcıda halo etkisi görüldüğünü ifade eder. Tablo 4.2 incelendiğinde ölçütlere ilişkin ölçüm raporunda; ölçütlerin güçlük düzeyleri arasındaki logit farkının $(0,49-(-0,82) = 1,31)$ büyük olduğu yani 1'den önemli derecede farklılık gösteren puanlayıcı bulunmadığı bundan dolayı puanlamada halo etkisinin bulunmadığı görülmektedir. Uyum istatistiklerine bakıldığında genel olarak uyum-İç ve uyum-dış değerlerin kabul edilebilir bir aralıkta (yani 0,5 ile 1,5 aralığında) olduğu görülmektedir. Yapılan analizler sonucunda hiçbir puanlayıcının uyum istatistiklerinin mükemmel uyuma sahip olmadığı puanlayıcıların halo davranışı göstermediği belirlenmiştir.

Elde edilen bulgulara grup ve bireysel düzeydeki istatistikler incelendiğinde, puanlayıcılarda halo etkisi görülmemektedir. Alanyazına bakıldığında araştırmanın bulgularına paralellik gösteren çalışmalara rastlanmaktadır. İlhan (2015) tarafından yapılan çalışmada, açık uçlu matematik sorularının değerlendirilmesi sürecinde standart rubrik kullanılarak yapılan puanlamalarda mevcut çalışmanın bulgularına paralel olarak hem grup düzeyinde hem de bireysel düzeydeki istatistiksel göstergelerde halo etkisi görülmemiştir. Diğer çalışmalardan farklı olarak mevcut çalışmanın bulgularında grup düzeyinde istatistiklerde halo etkisi görülmez iken, bireysel düzeydeki istatistiklerde bazı puanlayıcılarda halo etkisi görülmüştür. Nalbantoğlu Yılmaz (2017) çalışmasında değerlendirme ölçütlerinin bireylerin performanslarının değerlendirilmesinde güvenilir sonuçlar verdiğini ve her ne kadar bazı puanlayıcılarda halo etkisi görülse de puanlayıcıların genel olarak puanlama ölçütlerini tutarlı bir şekilde değerlendirdiğini yani halo etkisinin görülmediği sonucuna ulaşmıştır. Mevcut çalışmada puanlayıcıların öğrencilerin yazma performanslarını tüm analitik DPA'nın değerlendirme ölçütleri için uyum içerisinde puanladıkları yani puanlayıcı davranışlarından halo etkisi göstermedikleri yorumu yapılabilir.

4.1.3. Alt Problem 1.c'ye İlişkin Bulgular ve Yorum

Merkeze eğilim davranışı durumları nedir?

Performansın değerlendirilmesi sürecinde sıklıkla ortaya çıkan diğer bir puanlayıcı davranışı da *merkeze eğilim davranışı*'dır. Analitik DPA'nın puanlama kategorilerinden orta kategorilerin aşırı kullanılması olarak tanımlanan merkeze eğilim davranışının belirlenmesinde ÇYRÖM'ne dayalı olarak yapılan hesaplamalarda “uyum-dışı (OUTFIT)” değerlerinin 1,0'dan önemli derecede büyük değerler alması beklenir (Eckes, 2015; s.85). Puanlayıcıların merkeze eğilim davranışının belirlenebilmesi için, ilk olarak grup düzeyindeki istatistiksel göstergeler, sonrasında ise bireysel düzeyde istatistiksel göstergeler incelenmiştir.

Grup düzeyindeki istatistiksel göstergelerden biri kategori istatistikleridir. Mevcut çalışmada kullanılan analitik puanlama anahtarının puanlama kategorileri 1-4 arası olup, bu puanlama kategorilerine ilişkin ölçüm raporu Tablo 4.3'deki gibidir.

Tablo 4.3

Analitik Puanlama Anahtarının Puanlama Kategorilerine İlişkin Ölçüm Raporu

Puanlama Kategorileri	Frekans	Yüzde (%)	Yığılmış Yüzde (%)	Ortalama Logit değeri	Beklenen Logit değeri	Uyum Dışı (OUTFIT)
1	336	%11	%11	-0,66	-0,77	1,1
2	811	%26	%36	-0,13	-0,08	0,8
3	1017	%32	%68	0,72	0,74	1,0
4	1004	%32	%100	1,74	1,71	1,0

Tablo 4.3’de puanlama kategori istatistikleri incelendiğinde; frekans ve yüzde değerleri baz alındığında puanlayıcıların genel olarak 3 ve 4 numaralı kategorileri daha çok tercih ettikleri 1 ve 2 numaralı kategorileri ise daha az kullandıkları görülmektedir. Bu durumun olası iki sebebi; merkeze eğilim davranışının görülmesi veya birey performanslarının orta düzeyde olmasının bir sonucu olabilmektedir. Bu durumun gerçek sebebinin belirlenmesinde yani merkeze eğilim davranışının belirlenmesinde; grup düzeyindeki istatistiksel göstergelerinden biri olan kategori istatistiklerine ek olarak grup düzeyindeki diğer istatistiksel göstergelerin de incelenmesi gerekmektedir. Bunun için öğrenci ve ölçüt yüzeylerine ilişkin ölçüm raporlarındaki grup ve bireysel düzeydeki istatistiksel göstergeler incelenmiştir.

İlk olarak Tablo 4.4’deki öğrenci yüzeyine ilişkin ölçüm raporu incelenmiştir. Öğrenci yüzeyine ilişkin bütün değerlerin verilmesi mümkün olmadığından logit değerleri en yüksek 7 öğrenci ile en düşük 7 öğrenci örnek olarak verilmiştir (EK-14).

Tablo 4.4’de öğrenci yüzeyine ilişkin grup düzeyindeki istatistiksel göstergeler incelendiğinde; ayırma oranı 5,73, ayırma indeksi 7,90 ve ayırma indeksinin güvenilirliği 0,97 olduğundan, yetenek düzeyi farklı olan öğrencilerin geçerli ve güvenilir bir şekilde ayırt edilebildiği yorumu yapılabilir. Öğrenci yüzeyi için grup düzeyindeki istatistikler incelendiğinde Sabit etkili Ki-kare değerinin anlamlı olması ($\chi^2(sd)= 1371,0 (47)$, $p=.00<.01$) ve ayırma indeksinin güvenilirliğinin yüksek olması; öğrencilerin yetenek düzeyine göre geçerli ve güvenilir bir şekilde ayırt edilebildiğini dolayısıyla puanlayıcılarda performansın belirlenmesi sürecinde yaptığı puanlamalarda; puanlayıcı davranışlarından merkeze eğilim davranışının bulunmadığını göstermektedir.

Tablo 4.4

Öğrenci Yüzeyine İlişkin Ölçüm Raporu

Öğrenci No	Ortalamalar		Model		Uygunluk Ölçütü	
	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Ölçütü	Std. Hata	Uyum-içi (INFIT)	Uyum-dışı (OUTFIT)
Ö35	3,79	3,84	2,29	0,29	0,74	0,68
Ö27	3,82	3,81	2,13	0,31	0,85	0,83
Ö26	3,79	3,78	1,95	0,29	1,11	0,76
Ö12	3,65	3,72	1,69	0,23	1,00	0,86
Ö21	3,70	3,66	1,48	0,25	0,85	0,86
Ö43	3,71	3,65	1,45	0,25	0,80	0,58
Ö33	3,33	3,54	1,12	0,18	0,71	0,70
...	...	...	...	...	...	...
Ö25	2,23	2,22	-1,09	0,15	0,70	0,76
Ö47	2,24	2,20	-1,12	0,15	0,88	0,85
Ö16	2,15	2,19	-1,13	0,16	0,78	0,89
Ö20	1,95	2,04	-1,37	0,16	0,70	0,81
Ö9	2,02	2,02	-1,40	0,16	0,65	0,70
Ö17	1,97	1,87	-1,64	0,16	0,80	0,93
Ö1	1,76	1,59	-2,19	0,17	0,51	0,50
Ortalama	2,85	2,89	0,00	0,18	1,00	0,98
Ss(Evren)	0,58	0,59	1,05	0,04	0,33	0,33
Ss(Örnekleme)	0,58	0,59	1,06	0,04	0,34	0,33
Model, Evren: RMSE= 0,18 Ss= 1,03 Ayırma Oranı= 5,67 Ayırma İndeksi= 7,90 Güvenirlilik= 0,97						
Model, Örnekleme: RMSE= 0,18 Ss= 1,04 Ayırma Oranı= 5,73 Ayırma İndeksi= 7,98 Güvenirlilik= 0,97						
Model, Sabit etkili Ki-kare= 1371,0 sd= 47 p= 0,00						
Model, Normal Ki-kare= 45,2 sd= 46 p= 0,51						
Not: sd: serbestlik derecesi, ss: standart sapma, RMSE:hata kareleri ortalamasının karekökü						

Merkeze eğilim davranışının incelenmesinde son olarak Tablo 4.5'deki ölçüt yüzeyine ilişkin ölçüm raporu sonuçları da incelenmiştir. Tablo 4.5'deki ölçüt yüzeyine ilişkin ölçüm raporu sonuçları incelendiğinde, grup düzeyindeki istatistiksel göstergelerden ayırma oranının, ayırma indeksinin ve ayırma indeksinin güvenilirliğinin yüksek olması yetenek düzeyi farklı olan öğrencilerin geçerli ve güvenilir bir şekilde ayırt edilebildiğini göstermektedir. Yazma performansının değerlendirilmesi sürecinde analitik puanlama anahtarının değerlendirme ölçütlerinin bireylerin performansını istatistiksel olarak birbirinden ayırt edilebilmesi için yapılan sabit etkili Ki-kare testinin manidar olduğu görülmektedir ($\chi^2(sd)= 325,0 (5), p=.00<.01$). Buna ek olarak bireysel düzeydeki istatistiksel göstergeler incelendiğinde, analitik puanlama anahtarının değerlendirme ölçütlerine ilişkin uyum değerlerinin kabul edilebilir aralıkta olduğu ve standart hata değerlerinin düşük olduğu görülmektedir.

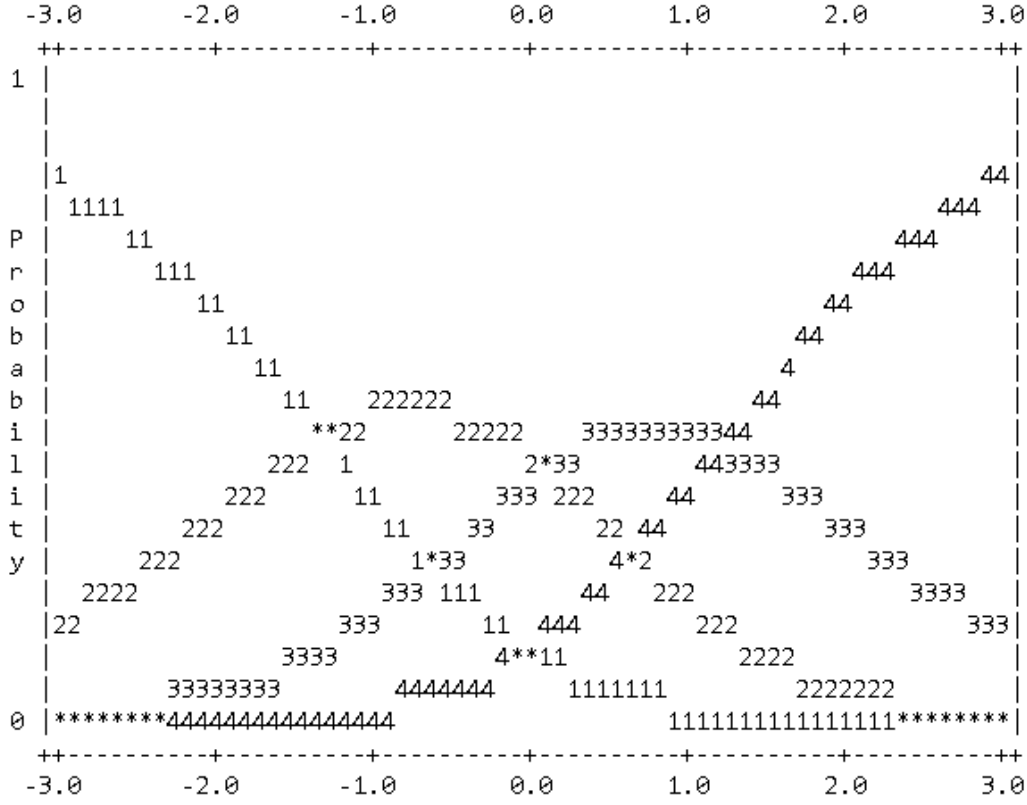
Tablo 4.5

Ölçüt Yüzeyine İlişkin Ölçüm Raporu

Ölçütler	Ortalamalar		Model		Uygunluk Ölçütü	
	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Ölçütü	Std. Hata	Uyum içi (INFIT)	Uyum dışı (OUTFIT)
Görüş Sunma	3,11	3,26	0,49	0,06	0,91	0,84
İçerik	3,02	3,17	0,32	0,06	1,14	1,11
Dil ve Anlatım	2,98	3,12	0,23	0,06	0,70	0,68
Sayfa Düzeni	2,89	3,02	0,06	0,06	1,06	1,05
Yazım ve Noktalama	2,70	2,78	-0,29	0,06	1,22	1,17
Yapı	2,40	2,42	-0,82	0,06	1,02	1,04
Ortalama	2,85	2,96	0,00	0,06	1,01	0,98
Ss(Evren)	0,24	0,28	0,44	0,00	0,17	0,17
Ss(Örnekleme)	0,26	0,31	0,48	0,00	0,18	0,18
Model, Evren: RMSE= 0,06 Ss= 43 Ayırma Oranı= 7,23 Ayırma İndeksi= 9,98 Güvenirlilik=0,98						
Model, Örnekleme: RMSE= 0,06 Ss= 48 Ayırma Oranı= 7,94 Ayırma İndeksi= 10,91 Güvenirlilik=0,98						
Model, Sabit etkili Ki-kare= 325,0 sd= 5 p= 0,00						
Model, Normal Ki-kare= 4,9 sd= 4 p=0,29						
Not: sd: serbestlik derecesi, ss: standart sapma, RMSE:hata kareleri ortalamasının karekökü						

Tüm bu bulgular incelendiğinde, puanlayıcıların ilköğretim 7. sınıf öğrencilerinin yazma performansının değerlendirilmesi sürecinde merkeze eğilim davranışı göstermediği söylenebilir. Değerlendirme ölçütlerine ilişkin kategori istatistikleri incelendiğinde ölçütlerin derecelendirilmesinde merkeze eğilim davranışının görülmediği, dereceli puanlama anahtarının ölçütlerinin öğrenci performansını ayırt edebilmede geçerli ve güvenilir ölçümler sağladığı söylenebilir. Elde edilen bulgular; grup ve bireysel düzeydeki istatistiksel göstergelerin, puanlayıcıların analitik dereceli puanlama anahtarının kategorilerini kullanırken bireyleri performanslarına göre daha etkili bir şekilde ayırt edebildiklerini göstermektedir. Sonuç olarak; ilköğretim öğrencilerinin yazma performansının değerlendirilmesi sürecinde puanlayıcılarda puanlayıcı davranışlarından merkeze eğilim davranışının görülmediği yorumu yapılabilir.

Merkeze eğilim davranışının görülmediğine ek olarak dereceli puanlama anahtarının ölçütlerine ilişkin olasılık eğrileri Şekil 4.1'deki gibidir.



Şekil 4.1. Dereceli puanlama anahtarının ölçütlerine ilişkin kategori olasılık eğrileri

Şekil 4.1 incelendiğinde; düşük yetenek düzeyindeki öğrencilerin alt kategoride bulunma olasılığı yüksek iken, üst kategorilerde daha düşük olduğu görülmektedir. Benzer şekilde; yüksek yetenek düzeyindeki öğrencilerin alt kategoride bulunma olasılığı düşük iken, üst kategorilerde daha yüksektir.

Alanyazın incelendiğinde çoğu çalışmada genel olarak performansın değerlendirilmesi sürecinde puanlayıcılarda merkeze eğilim davranışı görülmemiştir. Bu durum mevcut çalışmanın bulguları ile paralellik göstermektedir. Linlin (2020) çalışmasında grup düzeyindeki istatistiklerden kategori istatistiklerini inceleyerek puanlayıcıların genel olarak üst kategorileri kullandığını sonuç olarak puanlayıcıların merkezi eğilim davranışı göstermediğini bulmuştur. Börkan (2017b) çalışmasında, ölçek düşük veya yüksek performanslı öğrenciler için değil, sadece ortalama performans gösteren öğrenciler için işlevsel olduğunda merkezi eğilim davranışının görülebileceğini ifade etmiştir. Üçpınar ve Ünaldı (2018) tarafından yapılan çalışmada merkeze eğilim davranışı görülmediği durumlarda, puanlayıcıların değerlendirmelerinde bütün ölçek kategorilerini dengeli bir şekilde kullandıklarını ve kategorilerin farklı yeterlilik düzeylerindeki öğrencilerin performanslarını ayırt edebildiklerini göstermişlerdir. Bu bağlamda sonuçlar genel olarak

değerlendirildiğinde mevcut çalışma, puanlayıcıların analitik dereceli puanlama anahtarının kategorilerini kullanarak öğrencilerin yazma performansını ayırt etmede başarılı olduğunu göstermektedir.

4.2. İkinci Araştırma Sorusuna İlişkin Bulgular ve Yorum

Performansın değerlendirilmesi sürecinde sıklıkla ortaya çıkan diğer bir puanlayıcı davranışı da *farklılaşan puanlayıcı katılığı/cömertliği davranışı (puanlayıcı yanlılıkları)*'dır. Bu puanlayıcı davranışına göre, genel olarak puanlayıcılar öğrenci performansını değerlendirirken bazı öğrencilere daha düşük puan verme eğiliminde iken (örneğin; siyahi öğrencilere daha düşük not verme); bazı öğrencilere performansından daha yüksek puan verme (örneğin; erkek öğrencilere daha yüksek not verme) eğiliminde olabilmektedir (Myford ve Wolfe, 2004). Puanlayıcıların, öğrenci performansının değerlendirilmesinde yanlı puanlar verme eğiliminde olması geçerliği ve güvenilirliği olumsuz yönde etkilemektedir. Bu bağlamda, öğrenci performansının değerlendirilmesinde farklılaşan puanlayıcı katılığı ve cömertliği davranışı (başka bir deyişle puanlayıcı yanlılığının) belirlenmesi önemlidir. Buna dayalı olarak mevcut çalışmada ikna edici metin yazma performanslarının değerlendirilmesine yönelik öğretmen puanlayıcılarının tüm öğrencileri ve analitik dereceli puanlama anahtarının tüm değerlendirme ölçütlerini değerlendirirken yanlı puanlama davranışı gösterip göstermediğini belirlemek için; FACETS paket programı kullanılarak ÇYRÖM analizi sonucunda ilk olarak grup düzeyindeki istatistiksel göstergeler daha sonra ise bireysel düzeyde istatistiksel göstergeler incelenmiştir. Grup düzeyindeki istatistiksel göstergelerden Sabit etkili Ki-Kare değerinin anlamlı olması; bireysel düzeydeki istatistiksel göstergelerden t değerinin ± 2 aralığı dışında olması farklılaşan puanlayıcı katılık/cömertliği davranışının yani puanlayıcı yanlılığının göstergesidir (Barkaoui, 2013; Bond ve Fox, 2015; Linacre, 2018; s.180). t değerlerinin işaretlerine bakıldığında 0'dan büyük değerlerde gözlenen puan beklenen puandan daha yüksek olup puanlayıcı cömertlik davranışı görülürken, 0'dan küçük t değerlerinde gözlenen puan beklenen puandan daha düşük olup puanlayıcı katılık davranışı görülmektedir (Eckes, 2015, s.134).

ÇYRÖM'nin en önemli avantajlarından biri de çalışmaya dahil edilen bütün yüzeyler arasındaki etkileşim etkilerini analiz ederek puanlayıcı yanlılıklarının belirlenebilmesidir. Mevcut çalışmada puanlayıcı, öğrenci, ölçüt, cinsiyet, görev olmak üzere 5 farklı yüzey bulunmaktadır. Bu çalışma dahilinde puanlayıcı yüzeyine ilişkin olarak; "*Puanlayıcı x*

Öğrenci”, “Puanlayıcı x Ölçüt”, “Puanlayıcı x Cinsiyet”, “Puanlayıcı x Görev” ve “Puanlayıcı x Öğrenci x Ölçüt”, etkileşim etkileri dikkate alınarak puanlayıcı yanlılığı davranışı olup olmadığı incelenmiştir. Öğretmen puanlamalarına ilişkin olarak farklılaşan puanlayıcı katılığı/cömertliği davranışı yani puanlayıcı yanlılıklarının belirlenmesine yönelik bulgular;

4.2.1. Alt Problem 2.a’ya İlişkin Bulgular ve Yorum

“Puanlayıcı x Öğrenci” yüzeylerinin etkileşimine yönelik ölçüm sonuçları nedir?

“Puanlayıcı x Öğrenci” etkileşimine yönelik farklılaşan puanlayıcı katılık/cömertlik davranışı yani puanlayıcı yanlılığının tespiti için, ilk olarak grup düzeyindeki istatistiksel göstergeler yorumlanmış, sonrasında bireysel düzeydeki istatistiksel göstergeler incelenmiştir. Buna ek olarak; gözlenen puan ile beklenen puan arasındaki fark ve yanlılık ölçümüne ait değerlerin işaretleri, puanlayıcının puanlama yaparken bazı öğrencilere ne kadar cömert veya ne kadar katı olduğunun belirlenmesinde etkilidir.

Tablo 4.6

“Puanlayıcı x Öğrenci” Yüzeylerinin Etkileşimine Yönelik Ölçüm Sonuçları

Puanlayıcı	Öğrenci	Gözlenen Puan	Beklenen Puan	Yanlılık	Std. Hata	t-değeri
PÖ6	45	20	14,76	1,44	0,58	2,49
PÖ6	16	16	11,01	1,41	0,51	2,78
PÖ3	18	19	14,53	1,38	0,58	2,37
PÖ7	28	18	13,55	1,23	0,55	2,23
PU1	17	16	11,34	1,22	0,48	2,53
PÖ6	17	14	10,19	1,19	0,52	2,29
PÖ3	13	19	15,17	1,18	0,58	2,04
...	...	...	...	...	...	...
PÖ6	37	13	17,05	-1,07	0,53	-2,01
PÖ7	46	13	17,65	-1,27	0,52	-2,43
PÖ6	23	13	17,86	-1,28	0,53	-2,41
PÖ8	14	11	15,63	-1,29	0,56	-2,32
PÖ1	26	20	22,51	-1,35	0,63	-2,14
PÖ3	26	18	21,85	-1,52	0,56	-2,72
PÖ3	38	11	15,85	-1,60	0,62	-2,58
PU1	4	11	17,37	-1,65	0,58	-2,85
PÖ7	15	10	17,13	-2,03	0,59	-3,41
PÖ6	4	8	15,35	-2,58	0,80	-3,24

Sabit etkili Ki-kare= 437,5 sd= 528 p=1,00

Mevcut çalışmada 11 puanlayıcı, 48 öğrencinin ikna edici metin yazma performansına ait metinleri değerlendirmiş ve bunun sonucunda “Puanlayıcı x Öğrenci” etkileşimine yönelik

toplam 528 (11 puanlayıcı x 48 öğrenci) olası etkileşim meydana gelmiştir. “Puanlayıcı x Öğrenci” etkileşimine yönelik ölçüm sonuçlarına ait istatistiksel göstergeler Tablo 4.6’deki gibi olup, bu etkileşime yönelik tüm değerlerin verilmesi mümkün olmadığından sadece yanlı davranışlar gösteren etkileşimlere yer verilmiştir.

Tablo 4.6 incelendiğinde grup düzeyindeki istatistiksel göstergelerden Sabit etkili Ki-kare değerinin ($\chi^2(sd) = 437,5(528)$, $p = 1,00 > 0,01$) manidar olmadığı görülmektedir. Bulunan bu değere göre, puanlayıcıların genel olarak puanlama sürecinde farklılaşan puanlayıcı katılık/cömertlik davranışı göstermediği yani puanlamada yanlılık davranışı göstermedikleri yorumu yapılabilir. Puanlayıcıların puanlama sürecinde yanlı puanlama yapıp yapmadıkları konusunda grup düzeyindeki istatistiksel göstergeler bazen yeterli olmayabilir. Bunun için bireysel düzeydeki istatistiksel göstergelerin incelenmesi gereklidir. Yanlılık davranışının belirlenmesinde bireysel düzeydeki istatistiksel göstergelerden t_istatistiklerine bakıldığında bazı puanlayıcıların yanlı etkileşimlerinin anlamlı olduğu yani; t_istatistiklerinin ± 2 aralığı dışında bulunduğu görülmüştür. Tablo 4.6 incelendiğinde, “Puanlayıcı x Öğrenci” etkileşimine yönelik toplam 528 (11 x 48) olası etkileşim arasından 17 tanesinin (%3,22) t_değerlerinin belirlenen aralık dışında olduğu bulunmuştur. Mevcut çalışmada performansın değerlendirilmesinde 17 anlamlı etkileşimin yanlılık değerlerinin işaretleri incelendiğinde yanlılık değerleri pozitif işaretli olan 7 etkileşimde farklılaşan puanlayıcı cömertliği, yanlılık değerleri negatif işaretli olan 10 etkileşimde ise farklılaşan puanlayıcı katılığı olduğu görülmektedir. Yanlılık etkileşim değerleri ve t_değerleri pozitif ise Gözlenen puan > Beklenen puan olup puanlayıcılar bazı öğrencilerin performanslarını değerlendirirken daha cömert davranmakta iken, negatif olduğunda durum tam tersidir. Yani; Gözlenen puan < Beklenen puan olup puanlayıcılar bazı öğrencilerin performanslarını değerlendirirken daha katı davranışlar sergilemektedir. Bunu Tablo 4.6 üzerinden örnekleyecek olursak; PÖ6 puanlayıcısının öğrencilerinden bazılarının olması gerekenden daha fazla puan verdiği (cömert davrandığı) bazılarının ise olması gerekenden daha az puanlar verdiği (katı davrandığı) görülmektedir. PÖ6 numaralı puanlayıcı ile 16 numaralı öğrencinin etkileşimi incelendiğinde, “Gözlenen puan 16”, “Beklenen puan 11,01” ve “Yanlılık 1,41” olarak bulunmuştur. Gözlenen puan > Beklenen puan ve yanlılık terimi pozitif olduğundan dolayı, PÖ6 puanlayıcısı 16 numaralı öğrencinin performansına olması gerekenden daha fazla puan verdiği yani cömert davrandığı yorumu yapılabilir. Tablo 4.6’da PÖ6 puanlayıcısı için 7 yanlı etkileşim gözlemlenmiş olup, bu etkileşimler içerisinde 3 öğrenciye (16, 17 ve 45 numaralı öğrenciler) puanlama yaparken

olması gerekenden daha fazla puanlar vermiş, 4 öğrencisine (4, 18, 23 ve 37 numaralı öğrencilere) ise puanlama yaparken olması gerekenden daha az puanlar vermiştir. Ayrıca; 11 puanlayıcı arasından PÖ6 numaralı puanlayıcının, öğrenci performanslarının değerlendirilmesinde yapılan ölçümlerde diğer puanlayıcılara göre daha fazla yanlı etkileşime sahip olduğundan dolayı ölçümlerin geçerliği ve güvenilirliğini olumsuz yönde etkilemektedir. Sonuç olarak, performansın değerlendirilmesi sürecinde PÖ3 ve PÖ6 puanlayıcılarında bazı öğrencilerin performanslarını puanlarken diğer puanlayıcılara nazaran daha çok farklılaşan puanlayıcı katılımı ve cömertliği davranışı görüldüğü yani bu puanlayıcıların puanlama sürecinde diğer puanlayıcılara göre daha yanlı davrandıkları yorumu yapılabilmektedir.

4.2.2. Alt Problem 2.b'ye İlişkin Bulgular ve Yorum

“Puanlayıcı x Ölçüt” yüzeylelerinin etkileşimine yönelik ölçüm sonuçları nedir?

“Puanlayıcı x Ölçüt” etkileşimine yönelik farklılaşan puanlayıcı katılım/cömertlik davranışı yani puanlayıcı yanlılığının tespiti için, ilk olarak grup düzeyindeki istatistiksel göstergeler sonrasında ise bireysel düzeydeki istatistiksel göstergeler incelenmiştir. Mevcut etkileşimde; gözlenen puan ile beklenen puan arasındaki fark ve yanlılık ölçümüne ait değerlerin işaretleri, puanlayıcının puanlama yaparken analitik puanlama anahtarının bazı ölçütlerini değerlendirirken ne kadar cömert veya ne kadar katı olduğunun belirlenmesinde etkilidir.

Tablo 4.7

“Puanlayıcı x Ölçüt” Yüzeylelerinin Etkileşimine Yönelik Ölçüm Sonuçları

Puanlayıcı	Ölçüt	Gözlenen Puan	Beklenen Puan	Yanlılık	Std. Hata	t-değeri
PÖ1	Yazım ve Noktalama	138	129,42	0,35	0,20	1,72
PÖ3	Yazım ve Noktalama	126	117,68	0,35	0,21	1,71
PÖ6	Görüş Sunma	143	134,81	0,32	0,20	1,60
PU2	Yapı	144	135,71	0,27	0,18	1,48
...	...	...	...			...
PÖ1	Yapı	110	116,25	-0,25	0,20	-1,24
PU2	Yazım ve Noktalama	144	151,28	-0,25	0,18	-1,37
PU3	Yazım ve Noktalama	137	144,62	-0,28	0,19	-1,49
PÖ3	Yapı	94	105,60	-0,54	0,22	-2,44

Sabit etkili Ki-kare=42,9 sd= 66 p=0,99

Mevcut çalışmada 11 puanlayıcı, analitik dereceli puanlama anahtarının 6 değerlendirme ölçütünü kullanarak ikna edici metin yazma performansına ait metinleri değerlendirmiş ve bunun sonucunda “Puanlayıcı x Ölçüt” etkileşimine yönelik toplam 66 (11 puanlayıcı x 6 ölçüt) olası etkileşim meydana gelmiştir. “Puanlayıcı x Ölçüt” etkileşimine yönelik ölçüm sonuçlarına ait istatistiksel göstergeler Tablo 4.7’deki gibi olup, bu etkileşime yönelik tüm değerlerin verilmesi mümkün olmadığından (EK-15); t_değeri en yüksek ve en düşük 4 puanlayıcıya ait etkileşimlere yer verilmiştir.

Tablo 4.7 incelendiğinde grup düzeyindeki istatistiksel göstergelerden Sabit etkili Ki-kare değerinin ($\chi^2(sd) = 42,9(66)$, $p = 0,99 > 0,01$) manidar olmadığı görülmektedir. Bulunan bu değere göre, puanlayıcıların genel olarak puanlama sürecinde değerlendirme ölçütlerini kullanırken farklılaşan puanlayıcı katılık/cömertlik davranışı göstermediği yani puanlamada yanlılık davranışı göstermedikleri yorumu yapılabilir. Puanlayıcıların puanlama sürecinde yanlı puanlama yapıp yapmadıkları konusunda grup düzeyindeki istatistiksel göstergeler bazen yeterli olmayabilir. Bunun için bireysel düzeydeki istatistiksel göstergelerin incelenmesi gereklidir. Yanlılık davranışının belirlenmesinde bireysel düzeydeki istatistiksel göstergelerden t_istatistiklerine bakıldığında bazı puanlayıcıların yanlı etkileşimlerinin anlamlı olduğu yani; t_istatistiklerinin ± 2 aralığı dışında bulunduğu görülmüştür. Tablo 4.7 incelendiğinde, “Puanlayıcı x Ölçüt” etkileşimine yönelik toplam 66 (11 x 6) olası etkileşim arasından sadece 1 tanesinin (%1,52) t_değerinin belirlenen aralık dışında (yani $t = -2,44$) olduğu bulunmuştur. t_değerinin işaretinin negatif olması puanlayıcının puanlama yaparken o ölçüte olması gerekenden ne kadar az puanlar verdiğini yani katı puanlamalar yaptığının göstergesidir. Mevcut çalışmada da yanlı etkileşime sahip “PÖ3” numaralı puanlayıcısı ile “Yapı” ölçütü arasındaki etkileşim incelendiğinde, “Gözlenen puan 94”, “Beklenen puan 105,60” ve “Yanlılık -0,54” olarak bulunmuştur. Gözlenen puan < Beklenen puan ve yanlılık terimi negatif olduğundan dolayı, PÖ3 puanlayıcısı Yapı ölçütünü puanlarken olması gerekenden daha az puan verdiği yani katı davrandığı yorumu yapılabilir. Grup ve bireysel düzeydeki istatistiksel göstergeler incelendiğinde, puanlayıcıların performansın değerlendirilmesinde analitik dereceli puanlama anahtarının değerlendirme ölçütlerini puanlarken farklılaşan puanlayıcı katılığı/cömertliği davranışı yani puanlayıcı yanlılığı davranışı göstermedikleri sonucuna ulaşılmıştır.

4.2.3. Alt Problem 2.c'ye İlişkin Bulgular ve Yorum

“Puanlayıcı x Cinsiyet” yüzeylelerinin etkileşimine yönelik ölçüm sonuçları nedir?

“Puanlayıcı x Cinsiyet” etkileşimine yönelik farklılaşan puanlayıcı katılık/cömertlik davranışı yani puanlayıcı yanlılığının tespiti için, ilk olarak grup düzeyindeki istatistiksel göstergeler yorumlanmış, sonrasında bireysel düzeydeki istatistiksel göstergeler incelenmiştir. Gözlenen puan ile beklenen puan arasındaki fark ve yanlılık ölçümüne ait değerlerin işaretleri, puanlayıcının puanlama yaparken kız veya erkek öğrencilere ne kadar cömert veya ne kadar katı olduğunun belirlenmesinde etkilidir. Mevcut çalışmada “Puanlayıcı x Cinsiyet” etkileşimine yönelik toplam 22 (11 puanlayıcı x 2 Kız-Erkek) olası etkileşim meydana gelmiştir. “Puanlayıcı x Cinsiyet” etkileşimine yönelik ölçüm sonuçlarına ait istatistiksel göstergeler Tablo 4.8’de verilmiştir.

Tablo 4.8

“Puanlayıcı x Cinsiyet” Yüzeylelerinin Etkileşimine Yönelik Ölçüm Sonuçları

Puanlayıcı	Cinsiyet	Gözlenen Puan	Beklenen Puan	Yanlılık	Std. Hata	t-değeri
PÖ3	Erkek	330	314,97	0,20	0,12	1,73
PU1	Kız	478	466,28	0,14	0,11	1,29
PÖ6	Kız	434	424,54	0,11	0,11	1,02
PÖ1	Erkek	356	348,04	0,11	0,12	0,93
PÖ5	Kız	484	479,42	0,06	0,11	0,51
PU2	Kız	542	538,72	0,05	0,12	0,41
PÖ7	Kız	432	427,73	0,05	0,11	0,46
PU3	Erkek	394	390,80	0,05	0,12	0,40
PÖ4	Erkek	380	377,01	0,04	0,12	0,36
PÖ8	Erkek	336	332,88	0,04	0,12	0,36
PÖ2	Erkek	350	348,98	0,01	0,12	0,12
PÖ2	Kız	470	471,01	-0,01	0,11	-0,11
PÖ8	Kız	450	453,11	-0,04	0,11	-0,34
PÖ4	Kız	499	501,98	-0,04	0,11	-0,4
PU3	Kız	514	517,19	-0,04	0,12	-0,38
PU2	Erkek	407	410,28	-0,06	0,13	-0,43
PÖ7	Erkek	306	310,27	-0,06	0,12	-0,49
PÖ5	Erkek	352	356,58	-0,06	0,12	-0,54
PÖ1	Kız	462	469,96	-0,10	0,11	-0,88
PÖ6	Erkek	298	307,46	-0,13	0,12	-1,09
PU1	Erkek	333	344,72	-0,16	0,12	-1,36
PÖ3	Kız	418	433,03	-0,18	0,11	-1,63

Sabit etkili Ki-kare= 15,1 sd=22 p=0,86

Tablo 4.8 incelendiğinde grup düzeyindeki istatistiksel göstergelerden Sabit etkili Ki-kare değerinin ($\chi^2(sd) = 15,1(22)$, $p = 0,86 > 0,01$) manidar olmadığı görülmektedir. Bulunan bu değere göre, puanlayıcıların genel olarak puanlama sürecinde kız ve erkek öğrencilerin

performanslarını puanlarken farklılaşan puanlayıcı katılık/cömertlik davranışı göstermediği yani puanlamada yanlılık davranışı göstermedikleri yorumu yapılabilir. Puanlayıcıların puanlama sürecinde yanlı puanlama yapıp yapmadıkları konusunda grup düzeyindeki istatistiksel göstergeler bazen yeterli olmayabilir. Bunun için bireysel düzeydeki istatistiksel göstergelerin incelenmesi gereklidir. Yanlılık davranışının belirlenmesinde bireysel düzeydeki istatistiksel göstergelerden t -istatistiklerine bakıldığında tüm puanlayıcıların kız ve erkek öğrencilerin performanslarını puanlarken yanlı etkileşimlerinin anlamlı olmadığı yani; t -istatistiklerinin ± 2 aralığı içinde olduğu görülmüştür. Mevcut çalışmada, “Puanlayıcı x Cinsiyet” etkileşimine yönelik Gözlenen puanlar ile Beklenen puanlar arasındaki farka dayalı olarak hesaplanan yanlılık değerlerinin 0’a yakın olduğu, yani puanlayıcıların hem kız hem de erkek öğrencilerin performanslarını değerlendirirken olması gerekenden daha az veya daha fazla puanlamalar yapmayıp yanlı davranışlar sergilemediklerini yani farklılaşan puanlayıcı katılık/cömertlik davranışı göstermedikleri yorumu yapılabilir. Sonuç olarak performansın değerlendirilmesi sürecinde “Puanlayıcı x Cinsiyet” etkileşimine yönelik değerlendirmelerde grup ve bireysel düzeydeki istatistiklerde yanlılık davranışı gözlenmediği, ölçümlerin geçerliği ve güvenilirliğinin olumsuz yönde etkilenmediği görülmektedir.

4.2.4. Alt Problem 2.d’ye İlişkin Bulgular ve Yorum

“Puanlayıcı x Görev” yüzeylerinin etkileşimine yönelik ölçüm sonuçları nedir?

“Puanlayıcı x Görev” etkileşimine yönelik farklılaşan puanlayıcı katılık/cömertlik davranışı yani puanlayıcı yanlılığının tespiti için, ilk olarak grup düzeyindeki istatistiksel göstergeler yorumlanmış, sonrasında bireysel düzeydeki istatistiksel göstergeler incelenmiştir. Gözlenen puan ile beklenen puan arasındaki fark ve yanlılık ölçümüne ait değerlerin işaretleri, puanlayıcının puanlama sürecinde bazı görevleri değerlendirirken ne kadar cömert veya ne kadar katı olduğunu göstermektedir.

Mevcut çalışmada 11 puanlayıcı, 7 farklı ikna edici metin yazma performansına ait metinleri değerlendirmiş ve bunun sonucunda “Puanlayıcı x Görev” etkileşimine yönelik toplam 77 (11 puanlayıcı x 7 görev) olası etkileşim meydana gelmiştir. “Puanlayıcı x Görev” etkileşimine yönelik ölçüm sonuçlarına ait istatistiksel göstergeler Tablo 4.9’daki gibi olup, bu etkileşime yönelik tüm değerlerin verilmesi mümkün olmadığından (EK-16) t -değeri en düşük ve en yüksek 4 puanlayıcıya ait yanlılık etkileşimlerine yer verilmiştir.

Tablo 4.9

“Puanlayıcı x Görev” Yüzeylerinin Etkileşimine Yönelik Ölçüm Sonuçları

Puanlayıcı	Görev	Gözlenen Puan	Beklenen Puan	Yanlılık	Std. Hata	t-değeri
PÖ6	G6	50	41,93	0,78	0,31	2,49
PU1	G6	55	47,11	0,75	0,32	2,34
PU3	G2	137	126,56	0,44	0,22	2,05
PÖ3	G7	87	82,04	0,35	0,27	1,30
...	...	...	...	...	...	...
PÖ6	G7	75	80,90	-0,37	0,25	-1,48
PÖ6	G2	86	95,45	-0,39	0,21	-1,88
PÖ6	G1	118	127,22	-0,40	0,21	-1,91
PÖ5	G6	44	48,84	-0,44	0,30	-1,47
Sabit etkili Ki-kare= 71,4 sd=77 p=0,66						
NOT: Görevler G1: Evcil Hayvan Besleme, G2: Eve İnternet Bağlatma, G3: Çanakkale'ye Gezi, G4: Okul Takımı Seçmeleri, G5: Geri Dönüşüm, G6: Kişisel Temizlik, G7: Benim Kararım						

Tablo 4.9 incelendiğinde grup düzeyindeki istatistiksel göstergelerden Sabit etkili Ki-kare değerinin ($\chi^2(sd) = 71,4(77)$, $p = 0,66 > 0,01$) manidar olmadığı görülmektedir. Bulunan bu değere göre, puanlayıcıların genel olarak puanlama sürecinde farklılaşan puanlayıcı katılık/cömertlik davranışı göstermediği yani puanlamada yanlılık davranışı göstermedikleri yorumu yapılabilir. Puanlayıcıların puanlama sürecinde yanlı puanlama yapıp yapmadıkları konusunda grup düzeyindeki istatistiksel göstergeler bazen yeterli olmayabilir. Bunun için bireysel düzeydeki istatistiksel göstergelerin incelenmesi gereklidir. Yanlılık davranışının belirlenmesinde bireysel düzeydeki istatistiksel göstergelerden t_istatistiklerine bakıldığında bazı puanlayıcıların yanlı etkileşimlerinin anlamlı olduğu yani; t_istatistiklerinin ± 2 aralığı dışında bulunduğu görülmüştür. Tablo 4.9 incelendiğinde, “Puanlayıcı x Görev” etkileşimine yönelik toplam 77 (11 x 7) olası etkileşim arasından 3 tanesinin (%3,89) t_değerlerinin belirlenen aralık dışında olduğu bulunmuştur. Mevcut çalışmada performansın değerlendirilmesinde 3 anlamlı etkileşimin yanlılık değerlerinin işaretleri incelendiğinde yanlılık değerlerinin pozitif işaretli olduğu yani farklılaşan puanlayıcı cömertliği olduğu görülmektedir. Yanlılık etkileşim değerleri pozitif ve Gözlenen puan > Beklenen puan olduğundan puanlayıcılar öğrencilerin performanslarını değerlendirirken daha cömert davranışlar sergilemektedir. Bunu Tablo 4.9 üzerinden örnekleyecek olursak; PÖ6 numaralı puanlayıcı ile G6 (Kişisel Temizlik) görevi arasındaki yanlı etkileşimi inceleyecek olursak, “Gözlenen puan 50”, “Beklenen puan 41,93” ve “Yanlılık 0,78” ($t=2,49$) olarak bulunmuştur. Gözlenen puan > Beklenen puan ve yanlılık terimi pozitif olduğundan dolayı, PÖ6 puanlayıcısı G6 (Kişisel Temizlik) görevini puanlarken olması gerekenden daha fazla puan verdiği yani bu görevi puanlarken daha

cömert davrandığı yorumu yapılabilir. Tablo 4.9’da puanlayıcıların görevlere yönelik değerlendirmelerinde 3 yanlı etkileşim (“PÖ6 x G6”, “PU1 x G6” ve “PU3 x G2”) gözlemlenmiş olup, bu etkileşimlerde puanlayıcıların “Kişisel Temizlik (G6)” ve “Eve İnternet Bağlatma (G2)” görevlerini puanlama yaparken olması gerekenden daha fazla puanlar vermiştir.

4.2.5. Alt Problem 2.e’ye İlişkin Bulgular ve Yorum

“Puanlayıcı x Öğrenci x Ölçüt” yüzeylerinin etkileşimine yönelik ölçüm sonuçları nedir?

Puanlayıcıların, öğrencilerin bazılarına hangi değerlendirme ölçütlerine dayalı olarak performansına beklenenden daha düşük ya da yüksek puanlar verip vermediğini saptamak için puanlayıcı, öğrenci ve ölçüt yüzeyleri arasındaki üçlü etkileşim (*Puanlayıcı × Öğrenci × Ölçüt*) incelenmiştir. “*Puanlayıcı x Öğrenci x Ölçüt*” etkileşimine yönelik farklılaşan puanlayıcı katılığı/cömertliği davranışı yani puanlayıcı yanlılığının tespiti için, ilk olarak grup düzeyindeki istatistiksel göstergeler yorumlanmış, sonrasında bireysel düzeydeki istatistiksel göstergeler incelenmiştir. Gözlenen puan ile beklenen puan arasındaki fark ve yanlılık ölçümüne ait değerlerin işaretleri, puanlayıcının puanlama yaparken bazı öğrencilere ne kadar cömert veya ne kadar katı olduğunu göstermektedir. Puanlayıcı, öğrenci ve ölçüt yüzeyleri arasındaki 3168 (11 puanlayıcı x 48 öğrenci x 6 ölçüt) etkileşimin tümüne ilişkin analiz çıktılarının Tablo 4.10’da gösterilmesi mümkün olmadığından, tabloda sadece üç yüzey arasındaki anlamlı etkileşimlere yer verilmiştir.

Tablo 4.10 incelendiğinde grup düzeyindeki istatistiksel göstergelerden Sabit etkili Ki-kare değerinin ($\chi^2(sd) = 1603,8(3168)$, $p = 1,00 > 0,01$) manidar olmadığı görülmektedir. Bulunan bu değere göre, puanlayıcıların genel olarak puanlama sürecinde farklılaşan puanlayıcı katılığı/cömertliği davranışı göstermediği yani puanlamada yanlılık davranışı göstermedikleri yorumu yapılabilmektedir. Puanlayıcıların performansın değerlendirilmesine yönelik yanlılık analizleri için; grup düzeyindeki istatistiksel göstergelerin yanı sıra bireysel düzeydeki istatistiksel göstergeler de incelenmelidir.

Tablo 4.10

“Puanlayıcı x Öğrenci x Ölçüt” Yüzeylerinin Etkileşimine Yönelik Ölçüm Sonuçları

Puanlayıcı	Öğrenci	Ölçüt	Gözlenen Puan	Beklenen Puan	Yanlılık	Std. Hata	t-değeri
PÖ6	17	Yapı	3	1,33	3,00	1,26	2,39
PÖ7	20	Yapı	3	1,34	2,99	1,26	2,38
PÖ6	16	Yapı	3	1,43	2,69	1,26	2,14
PÖ7	32	Yapı	3	1,45	2,63	1,26	2,09
PÖ7	8	Yapı	3	1,50	2,51	1,26	2,00
...	...	...	...	...	...	...	...
PÖ1	35	Görüş Sunma	3	3,88	-2,55	1,26	-,202
PÖ1	19	İçerik	2	3,55	-2,57	1,29	-2,00
PU2	10	Görüş Sunma	2	3,56	-2,61	1,29	-2,03
PÖ5	41	Sayfa Düzeni	2	3,57	-2,64	1,29	-2,06
PU1	12	Yazım	2	3,58	-2,65	1,29	-2,06
PÖ1	26	Yapı	2	3,61	-2,74	1,29	-2,13
PU3	21	Görüş Sunma	3	3,90	-2,75	1,26	-2,19
PÖ4	41	Sayfa Düzeni	2	3,67	-2,93	1,29	-2,28
PÖ4	27	İçerik	3	3,92	-2,95	1,26	-2,35
PÖ3	26	Sayfa Düzeni	2	3,73	-3,18	1,29	-2,47
PU2	41	Sayfa Düzeni	2	3,79	-3,45	1,29	-2,68
PU2	42	Yazım	1	3,44	-3,63	1,77	-2,05
PU1	19	İçerik	1	3,53	-3,74	1,71	-2,19

Sabit etkili Ki-kare= 1603,8 sd=3168 p=1,00

Yanlılık davranışının belirlenmesinde bireysel düzeydeki istatistiksel göstergelerden t_istatistiklerine bakıldığında bazı puanlayıcıların yanlı etkileşimlerinin anlamlı olduğu yani; t_istatistiklerinin ± 2 aralığı dışında bulunduğu görülmüştür. Tablo 4.10 incelendiğinde, “Puanlayıcı x Öğrenci x Ölçüt” etkileşimine yönelik toplam 3168 (11 x 48 x 6) olası etkileşim arasından 18 tanesinin (%0,56) t_değerlerinin belirlenen aralık dışında olduğu bulunmuştur. Performansın değerlendirilmesinde 18 anlamlı etkileşimin yanlılık değerlerinin işaretleri incelendiğinde yanlılık değerleri pozitif işaretli olan 5 etkileşimde farklılaşan puanlayıcı cömertliği, yanlılık değerleri negatif işaretli olan 13 etkileşimde ise farklılaşan puanlayıcı katılığı olduğu görülmektedir.

Mevcut çalışmada da yanlı etkileşime sahip “PÖ6” numaralı puanlayıcının 17 numaralı öğrencinin performansını değerlendirdiğinde “Yapı” ölçütünü puanlarken, “Gözlenen puan 3,00” ; “Beklenen puan 1,33” ; “Yanlılık 3,00” ve “t_değeri 1,26” olarak bulunmuştur. Gözlenen puan > Beklenen puan ve yanlılık terimi pozitif olduğundan dolayı, PÖ6 puanlayıcısı, 17 numaralı öğrencinin performansını değerlendirirken “Yapı” ölçütüne olması gerekenden daha fazla puan verdiği yani cömert davrandığı söylenebilir. Benzer şekilde; “PU1” numaralı puanlayıcının 19 numaralı öğrencinin performansını değerlendirdiğinde “İçerik” ölçütünü puanlarken, “Gözlenen puan 1,00” ; “Beklenen puan 3,53” ; “Yanlılık -3,74” ve “t_değeri -2,19” olarak bulunmuştur. Yanlılık terimi negatif ve

Gözlenen puan < Beklenen puan olduğundan dolayı, “PU1” puanlayıcısı, 19 numaralı öğrencinin performansını değerlendirirken “İçerik” ölçütüne olması gerekenden daha az puan verdiği yani katı davrandığı söylenebilir. Sonuç olarak; grup ve bireysel düzeydeki istatistiksel göstergeler incelendiğinde, bireysel düzeyde istatistiklere bağlı olarak bazı puanlayıcıların bazı öğrencilerin performanslarını değerlendirirken ölçütleri puanlamada farklılaşan puanlayıcı katılığı/cömertliği davranışı yani puanlayıcı yanlılığı davranışı göstermesine rağmen, grup düzeyindeki istatistiksel değerler incelendiğinde puanlayıcı yanlılığı davranışı göstermedikleri belirlenmiştir.

Araştırmanın ikinci alt problemine ilişkin bulgulara tüm ikili ve üçlü yanlılık etkileşimlerinde grup düzeyindeki istatistiksel göstergelerin yanısıra bireysel düzeydeki istatistikler de incelenmiştir. Alanyazın incelendiğinde ÇYRÖM’ne dayalı olarak yapılan araştırmalarda genel olarak hem grup düzeyinde hem de bireysel düzeyde yanlılık tespit edilirken, mevcut çalışmada sadece bireysel düzeyde yanlılıklarının olduğu görülmüştür. Bu sebepten farklılaşan puanlayıcı katılık/cömertlik davranışı yani yanlılık etkileşimlerinin tespit edilmesinde, puanlayıcıların yüzeyler ile arasındaki etkileşimlerinin araştırılması, puanlayıcı yanlılığının tam olarak nerede oluştuğunun belirlenmesi gerekmektedir (He, 2019). Yanlılık etkileşimlerinin belirlenebilmesi için araştırmacılar farklı değerleri yorumlamaktadır. Eckes (2019) ve Yüzüak Erten ve Kara (2019) yanlılık analizin tespitinde bireysel düzeydeki istatistiklerden beklenen ve gözlenen puanlar arasındaki farkı yanlılık olarak yorumlarken; Weselowski, Wind ve Engelhard (2015) bireysel düzeydeki istatistiklerden t değerinin ± 2 aralığı dışındaki değerleri yanlılık olarak ele almıştır. Mevcut çalışmada bireysel düzeydeki istatistikler anlamlı bulunmuş, yanlılıkların belirlenmesinde beklenen ile gözlenen puanlar, t değerleri buna ek olarak yanlılık terimleri yorumlanmıştır. Elde edilen bulgular, Şata ve Karakaya (2020) tarafından yapılan araştırma sonuçları ile tutarlılık göstermektedir. Araştırmacılar tarafından, performans değerlendirilmesinde puanlayıcıların farklılaşan puanlayıcı katılık ve cömertlik davranışı sergiledikleri diğer bir deyişle puanlayıcıların puanlama sürecinde yanlı davrandıkları belirlenmiştir. Goodwin (2016) çalışmasında grup düzeyinde ve bireysel düzeyde yanlılık tespit etmiş, bu durumun görülmesinde puanlayıcıların puanlama konusunda yeteri kadar bilgiye sahip olmaması şeklinde yorumlamıştır. Wan (2017) tarafından yapılan çalışmada diğer araştırmacılardan farklı olarak cinsiyete dayalı (“performans x cinsiyet” ile “performans x cinsiyet x madde”) etkileşimler incelenmiş, grup düzeyinde etkileşimin anlamlı ve bireysel düzeyde 1 veride yanlılık olduğu bulunmuştur. Çalışmanın bulgularına

bakıldığında ikili etkileşimlerden sadece cinsiyet yüzeyine ait etkileşimine yönelik yanlılık bulunmamıştır. Mevcut çalışmanın tersine Shirazi (2019) çalışmasında puanlayıcı yüzeyine bağlı üçlü etkileşimde yanlılık etkileşimlerinin hem grup hem de bireysel düzeyde istatistiksel olarak anlamsız olduğunu bulmuş, sonuç olarak da puanlayıcıların performansa dayalı değerlendirmelerinde deneyime bağlı olmaksızın tutarlı puanlamalar yaptıklarını ifade etmiştir.

4.3. Üçüncü Araştırma Sorusuna İlişkin Bulgular ve Yorum

Öğretmen puanlamalarına ilişkin olarak KTK, GK ve MTK'ya göre kestirilen puanlayıcılar arası güvenilirlik katsayıları ne düzeydedir?

Performansın değerlendirilmesi sürecinde en önemli hata kaynağı puanlayıcılardır. Özellikle üst düzey zihinsel becerilerin değerlendirilmesinde birden fazla puanlayıcıya dayalı olarak yapılan değerlendirmelerde puanlamalardaki farklılıkların belirlenmesine yönelik puanlayıcılar arası uyum veya puanlayıcılar arası güvenilirlik katsayısı hesaplanmalıdır. Puanlayıcılar arası güvenilirlik; performansın değerlendirilmesinde birden fazla puanlayıcının farklı öğrencilere ve farklı maddelere ilişkin yaptıkları puanlamalar arasındaki tutarlılığın derecesi olarak tanımlanmaktadır (Atılğan, Kan ve Doğan, 2011, s.49). Farklı puanlayıcıların puanlamaları arasındaki tutarlılığın derecesini temsilen hesaplanan puanlayıcılar arası güvenilirlik katsayısı; Klasik Test Kuramı (KTK), Genellenebilirlik Kuramı (GK) ve Madde Tepki Kuramı (MTK)'na dayalı olarak farklı yöntemlerle analiz edilmektedir. Mevcut çalışmada performansın değerlendirilmesi sürecinde 11 farklı puanlayıcının puanlamadaki farklılıkların belirlenmesine yönelik puanlayıcılar uyumun yani puanlayıcılar arasındaki güvenirliliğin belirlenmesinde; Klasik Test Kuramı (KTK), Genellenebilirlik Kuramı (GK) ve Madde Tepki Kuramı (MTK)'na dayalı ÇYRÖM'ne bağlı olarak hesaplanan güvenilirlik katsayıları Tablo 4.11'de verilmiştir.

Tablo 4.11 incelendiğinde puanlayıcı güvenirliliğine ilişkin, KTK'ya bağlı olarak SPSS 25 paket programı kullanılarak Sınıf içi Korelasyon Katsayısı (ICC) ve Cronbach's Alpha iç tutarlılık katsayıları; GK'na bağlı olarak EduG 6.1 paket programı kullanılarak G ve ϕ katsayıları; MTK'ya bağlı olarak FACETS 3.81 paket programı kullanılarak puanlayıcı yüzeyine ait ayırma indeksinin güvenirliliği ve Rasch-Cohen's Kappa istatistiği hesaplanmıştır.

Tablo 4.11

Farklı Ölçme Kuramlarına Dayalı Olarak Hesaplanan Puanlayıcılar-Arası Güvenirlik Katsayıları

<u>Ölçme Kuramı</u>	<u>Paket Program</u>	<u>Güvenirlik</u>
KTk (Klasik Test Kuramı)	SPSS 25	Sınıf içi Korelasyon Katsayısı (ICC) = 0,974 Cronbach's Alpha = 0,974
GK (Genellenebilirlik Kuramı)	EduG 6.1	G_bağıl = G katsayısı = 0,96 G_mutlak = ϕ katsayısı = 0,74
MTK (Madde Tepki Kuramı)	FACETS 3.81	Rasch-Cohen's Kappa = 0,18 Puanlayıcı yüzeyine ait ayırma indeksinin güvenilirliği = 0,96

KTk'nda, en az iki puanlayıcının belirli bir davranışı veya performansı puanlaması durumunda, puanlayıcılar arası güvenilirliğin belirlenmesinde istatistiksel tekniklerden Sınıf içi Korelasyon Katsayısı (ICC) ve Cronbach's Alpha iç tutarlılık katsayısı hesaplanmaktadır. KTK'ya bağlı olarak 11 farklı puanlayıcıya ait ölçümlerde, Sınıf içi Korelasyon Katsayısı (ICC) ve Cronbach's Alpha güvenirlilik katsayısı 0,974 olarak bulunmuştur. Bulunan bu değerlerin 0,70 ve üzerinde değerler almasından dolayı performansın değerlendirilmesinde farklı puanlayıcılara ait ölçümlerde puanlayıcılar arası tutarlılığın sağlandığı yani puanlayıcılar arası güvenilirliğin oldukça yüksek olduğu yorumu yapılabilir. Sonuç olarak KTK'ya dayalı ölçmelerde; performansının değerlendirilmesi sürecinde yer alan 11 puanlayıcının öğrencilerin yazma performanslarına verdikleri puanlar paralellik gösteriyorken, puanlayıcıların katılık-cömertlik davranışları birbirlerinden farklılık göstermektedir.

GK'nda, en az iki puanlayıcının öğrencilerin performansını puanlaması durumunda farklı değişkenlik kaynaklarını dikkate alarak puanlayıcılar arası güvenilirliğin belirlenmesinde istatistiksel tekniklerden G_katsayısı hesaplanmaktadır. GK'ya bağlı olarak 11 farklı puanlayıcıya ait ölçümlerde, iki farklı G_katsayısı hesaplanmıştır. Bunlardan biri bağıl değerlendirmeler için hesaplanan G_katsayısı, diğeri ise mutlak değerlendirmeler için hesaplanan ϕ katsayısıdır. Mevcut çalışmada 11 farklı puanlayıcıya ait ölçümler sonucu hesaplanan G_katsayısı (Genellenebilirlik katsayısı) puanlayıcılar arası güvenirlilik katsayısı olarak yorumlanmış olup, bu değer 0,96 olarak bulunmuştur. Bulunan bu değer 0,90 ve üzerinde değerler almasından dolayı performansın değerlendirilmesinde farklı puanlayıcılara ait ölçümlerde puanlayıcılar arası tutarlılığın sağlandığı yani puanlayıcılar arası güvenilirliğin oldukça yüksek olduğu yorumu yapılabilmektedir. Sonuç olarak GK'na dayalı ölçmelerde; performansının değerlendirilmesi sürecinde yer alan 11 puanlayıcının

öğrencilerin yazma performanslarına verdikleri puanların benzer olduğu yani puanlayıcılar arası değişkenliğin çok az olduğu, farklı puanlayıcıların verdikleri puanlar arasında tutarlılığın olması ise puanlayıcılar arası güvenilirliğin yüksek olduğunun göstergesidir.

Bu bulgular; Güler ve Taşdelen Teker (2015)'in araştırma sonuçlarıyla benzerlik göstermektedir. Güler ve Taşdelen Teker (2015)'in tarafından yapılan araştırmanın bulgularına göre, KTK ve GK'na dayalı olarak farklı yöntemlerle puanlayıcılar arası güvenilirlik katsayısı hesaplanmış, sonuçların birbiri ile paralellik gösterdiğini ifade etmişlerdir. KTK ve GK'ya dayalı güvenilirlik hesaplamalarında güvenilirlik katsayılarının oldukça yüksek olması puanlayıcılara bağlı güvenilirliğin yani puanlamalar arası tutarlılığın oldukça yüksek olduğunu ifade ederken; MTK kuramına bağlı olarak ÇYRÖM'ne dayalı analizlerde hesaplanan güvenilirlik katsayısı farklı şekilde yorumlanmaktadır. İlhan (2016); oldukça yüksek bulunan güvenilirlik katsayısının puanlayıcılarda puanlamaların tutarlılığın göstergesinin değil de puanlamalardaki farklılıkların güvenilirliği olduğunu ileri sürmektedir. Bu durumda; birbirlerine göre puanlamada katılık veya cömertlik açısından farklılık bulunan puanlayıcıların verdikleri puanlar birlikte değişiyorsa, verilen ortalamalar arası fark büyük olmasına rağmen korelasyon katsayısı yüksek çıkacaktır (Güler ve Taşdelen Teker, 2015). Dolayısıyla mevcut çalışmanın bulgularında puanlayıcı yüzeyine ait ayırma indeksinin güvenilirliğinin oldukça yüksek olması, puanlamalar arasında farklılıkların olduğunu yani; puanlayıcı davranışlarından puanlayıcı katılık ve cömertlik davranışının görüldüğü yorumu yapılabilir.

MTK'nda, en az iki puanlayıcının belirli bir davranışı veya performansı puanlaması durumunda, puanlayıcılar arası tutarlılığının göstergesi olarak hesaplanan puanlayıcılar arası güvenilirliğin belirlenmesinde istatistiksel tekniklerden ÇYRÖM'ne dayalı olarak "Rasch-Cohen's Kappa istatistiği" ve "Puanlayıcı yüzeyine ait ayırma indeksinin güvenilirliği" hesaplanmaktadır. ÇYRÖM'ne bağlı olarak 11 farklı puanlayıcıya ait ölçümlerde, puanlayıcılar arası beklenen ve gözlenen uyumun yüzdesine bağlı olarak hesaplanan Rasch-Cohen's Kappa istatistiği 0,18 olarak bulunmuş; puanlayıcı yüzeyine ait ayırma indeksinin güvenilirliği ise 0,96 olarak bulunmuştur. Rasch-Cohen's Kappa istatistiğinin 0,00'a yakın değerler alması beklenir (Eckes, 2015, s.92; Linacre, 2018, s.306). Bulunan bu değer 0,00'dan büyük olması; puanlayıcıların performansın değerlendirilmesi sürecinde tutarlı değerlendirmeler yaptığını yani puanlayıcılar arası uyumun dolayısıyla puanlayıcılar arası güvenilirliğin sağlandığını göstermektedir. Puanlayıcı yüzeyine ait ayırma indeksinin güvenilirliğinin 0,90'dan büyük değerler alması

ise, 11 farklı puanlayıcının öğrencilerin yazma performanslarını değerlendirmesi sürecinde her ne kadar farklı puanlamalar yapmış olsalar bile, puanlayıcıların puanlamaları arasında tutarlılıkların olduğu söylenebilir. Bu sonuçlara göre, ÇYRÖM analizine dayalı olarak 11 puanlayıcının, öğrencilerin performanslarını tutarlı bir şekilde ölçtüğü yani puanlayıcılar arası güvenilirliğin yüksek olduğu söylenebilir. Elde edilen bu bulgulara göre, yazma performansının değerlendirilmesi sürecinde puanlama yapan 11 farklı puanlayıcının puanlayıcı katılık-cömertlik açısından farklılık gösterdiği yani puanlayıcıların öğrencilere verdikleri puanların birbiriyle tutarlı olduğu yorumu yapılabilir. Şata (2019) tarafından yapılan çalışmada Rasch-Cohen's Kappa güvenilirlik katsayısının mevcut çalışmanın sonucu ile paralellik gösterdiği, puanlayıcılar-arası güvenilirliğin orta düzeyde ve istatistiksel olarak anlamlı bir etkiye sahip olduğu bulunmuştur.

BÖLÜM V

SONUÇ VE ÖNERİLER

Bu bölümde araştırmanın bulgularına dayalı olarak her bir araştırma sorusuna ilişkin alt problemlere ait sonuçlara ve önerilere yer verilmiştir.

5.1. SONUÇ

Bu araştırmada, performansın değerlendirilmesinde puanlayıcı davranışları, farklılaşan puanlayıcı katılık/cömertlik davranışları ÇYRÖM ile incelenmiş; puanlayıcılar arası kestirilen güvenilirlik katsayı değerleri ölçmenin üç temel kuramı olan KTK, GK ve MTK'ya bağlı olarak değerlendirilmiştir.

Performansın değerlendirilmesinde (*ilköğretim öğrencilerinin yazma performansının belirlenmesinde*) birden fazla puanlayıcının puanlamaları arasındaki farklılıkların belirlenmesinde puanlayıcı davranışlarına yönelik ÇYRÖM'ne göre analizinin yapılabilmesi için gerekli sayıltılar kontrol edilmiş; *tek boyutluluk, yerel bağımsızlık ve model-veri uyumu* incelenmiş, verinin model için uygun olduğu sonucuna varılmıştır.

5.1.1 Birinci Araştırma Sorusuna İlişkin Sonuçlar

ÇYRÖM'ne dayalı olarak yapılan analizlerde literatürde en çok karşılaşılan puanlayıcı davranışlarından; *puanlayıcı katılık/cömertlik davranışı, halo etkisi, merkeze eğilim davranışı* durumları incelenmiştir. Performansın değerlendirilmesi sürecinde tüm yüzeyler

ve bileşenleri için, grup düzeyindeki ve bireysel düzeydeki istatistikler yorumlanmıştır. Puanlayıcı davranışlarının belirlenmesine yönelik grup düzeyindeki istatistiklerden ayırma oranı, ayırma indeksi, ayırma indeksine bağlı güvenilirlik katsayısı ve Sabit etkili Ki-Kare değerleri; bireysel düzeydeki istatistiklerden uygunluk istatistiklerinden uyum-içi ile uyum-dışı değerler ve logit değerleri yorumlanmıştır. Bulgularda yüzeylere ait ölçüm raporları incelendiğinde Sabit etkili Ki-Kare değerlerinin anlamlı sonuçlar verdiği, uygunluk istatistiklerine dayalı olarak hesaplanan uyum içi ve uyum dışı değerlerinin tüm yüzeyler için kabul edilebilir bir aralıkta (0,5 ile 1,5 aralığında) olduğu görülmüştür.

Puanlayıcıların katılık/cömertlik davranışı için; puanlayıcı yüzeyine ilişkin ölçüm raporu incelenmiştir. Performansın değerlendirilmesi sürecinde puanlayıcıların öğrencilerin yazma çalışmalarını puanlarken grup düzeyindeki ve bireysel düzeydeki istatistiklerde puanlayıcı katılık/cömertlik davranışı görülmüştür. Elde edilen bulgularda puanlayıcı yüzeyine ait ölçüm raporunda sabit etkili Ki-kare değerinin anlamlı ve güvenilirliğin yüksek olması; puanlayıcıların performansın değerlendirilmesi sürecinde her ne kadar birbirinden bağımsız olarak puanlama yapmış olsalar da tutarlı puanlamalar yaptıkları sonucuna ulaşılmıştır.

Halo etkisi için; analitik dereceli puanlama anahtarının değerlendirme ölçütlerinin güçlüklerinin eşitlenmesinden sonra yapılan analizlerde grup düzeyindeki istatistiklerden ayırma indeksine bağlı güvenilirlik katsayısının yüksek ve Sabit etkili Ki-Kare değeri anlamlı bulunmuştur. Mevcut çalışmanın bulgularına dayalı olarak; yapılan analizlerde hem grup hem de bireysel düzeyde halo etkisi görülmemiştir. Sonuç olarak performansın değerlendirilmesi sürecinde puanlayıcıların analitik DPA'nın ölçütlerini tutarlı bir şekilde puanladığı sonucuna ulaşılmıştır.

Merkeze eğilim davranışı için; “Öğrenci” yüzeyi; “Ölçüt” yüzeyi; “Puanlama Kategorileri”nin istatistiksel değerlerine bakılmıştır. Puanlama kategori istatistikleri incelendiğinde; frekans ve yüzde değerleri baz alındığında puanlayıcıların genel olarak üst kategorileri daha çok tercih ettiği, alt kategorileri ise daha az kullandıkları görülmüştür. “Öğrenci” ve “Ölçüt” yüzeylerine ilişkin ölçüm raporları incelendiğinde; hem grup düzeyinde hem de bireysel düzeyde merkeze eğilim davranışı görülmemiştir. Sonuç olarak mevcut çalışmanın bulguları incelendiğinde analitik dereceli puanlama anahtarının ölçütlerinin başarılı ve başarısız öğrencileri etkili bir şekilde ayırt edebildiği yani; ölçütlerin öğrenci performansını ayırt edebilmede geçerli ve güvenilir sonuçlar sağlamıştır.

5.1.2 İkinci Araştırma Sorusuna İlişkin Sonuçlar

Puanlayıcıların bazı öğrencilerin performanslarını puanlarken ne kadar katı veya ne kadar cömert olduklarını belirleyebilmek için “*Puanlayıcı x Öğrenci*” yüzeylelerinin etkileşimine yönelik ölçüm sonuçları; puanlayıcıların analitik puanlama anahtarının bazı ölçütlerini puanlarken ne kadar katı veya ne kadar cömert olduklarını belirleyebilmek için “*Puanlayıcı x Ölçüt*” yüzeylelerinin etkileşimine yönelik ölçüm sonuçları; puanlayıcıların bazı görevleri puanlarken ne kadar katı veya ne kadar cömert olduklarını belirleyebilmek için “*Puanlayıcı x Görev*” yüzeylelerinin etkileşimine yönelik ölçüm sonuçları incelenmiştir. Mevcut çalışmanın bulgularına dayalı olarak bu etkileşimlerde puanlayıcılarda grup düzeyinde anlamlı farklılık bulunmamış; fakat bireysel düzeydeki istatistiklerde bazı puanlayıcıların *farklılaşan puanlayıcı katılık/cömertlik davranışı* sergiledikleri görülmüştür.

Puanlayıcıların kız veya erkek öğrencileri puanlarken ne kadar katı veya ne kadar cömert olduklarını belirleyebilmek için “*Puanlayıcı x Cinsiyet*” yüzeylelerinin etkileşimine yönelik ölçüm sonuçları incelendiğinde; diğer etkileşimlerden farklı olarak hem grup hem de bireysel düzeyde anlamlı bir farklılık görülmemiştir. Mevcut çalışmanın bulgularına dayalı olarak; puanlayıcıların kız ve erkek öğrencileri değerlendirirken yanlı davranmadıkları, tutarlı puanlamalar yaptığı sonucuna ulaşılmıştır.

Puanlayıcıların bazı öğrencilerin performanslarını bazı analitik puanlama anahtarının değerlendirme ölçütlerine göre puanlarken yanlı davranışlar sergilediklerini belirleyebilmek için “*Puanlayıcı x Öğrenci x Ölçüt*” yüzeylelerinin etkileşimine yönelik ölçüm sonuçları incelendiğinde; puanlayıcılarda grup düzeyinde anlamlı farklılık bulunmamış; fakat bireysel düzeydeki istatistiklerde bazı puanlayıcıların farklılaşan puanlayıcı katılık/cömertlik davranışı sergiledikleri görülmüştür. Elde edilen bulgulara dayalı olarak; performansın değerlendirilmesi sürecinde birden fazla puanlayıcının aynı öğrenciye yönelik aynı değerlendirme ölçütünden farklı davranışlar sergilemediği yani “*Puanlayıcı x Öğrenci x Ölçüt*” yüzeylelerinin etkileşimlerine yönelik ortak bir yanlı davranımın söz konusu olmadığı sonucuna varılabilir.

5.1.3 Üçüncü Araştırma Sorusuna İlişkin Sonuçlar

Performansın değerlendirilmesi sürecinde öğretmen puanlamalarına ilişkin olarak KTK, GK ve MTK'ya göre kestirilen puanlayıcılar arası güvenilirlik katsayıları incelendiğinde genel olarak; kestirilen güvenilirlik katsayılarının 0,90 ve üzerinde olduğu, bulunan sonuçların birbiri ile paralellik gösterdiği sonucuna varılmıştır. Alanyazın incelendiğinde mevcut çalışmanın sonuçlarının benzer çalışmanın sonuçları ile paralellik gösterdiği bulunmuştur.

Buna ek olarak; beklenen ve gözlenen uyum yüzdeleri arasındaki farka dayalı olarak hesaplanan bir diğer güvenilirlik ölçütü de *Rasch-Cohen's Kappa güvenilirlik katsayısı* 'dır. Mevcut çalışmanın bulgularına dayanarak elde edilen puanlayıcılar arası güvenirliliğe ilişkin Rasch-Cohen's Kappa istatistiği değeri 0,18 olarak bulunmuştur. Bulunan bu değerin 0,00'dan büyük olması performansın değerlendirilmesi sürecinde puanlayıcıların tutarlı değerlendirmeler yapıldığını yani puanlayıcılar arası uyumun dolayısıyla puanlayıcılar arası güvenirliliğin sağlandığını göstermektedir.

5.2 ÖNERİLER

Mevcut çalışma sonucunda öneriler; uygulamaya yönelik öneriler ve yapılacak çalışmalara yönelik öneriler olmak üzere iki başlık altında yorumlanmıştır.

5.2.1. Uygulamaya Yönelik Öneriler

1. Mevcut çalışmada performansın değerlendirilmesi sürecinde puanlayıcı yüzeyine ait tüm ikili ve üçlü etkileşimlerde grup düzeyinde farklılaşan puanlayıcı katılık/cömertlik etkisi davranışı görülmemiştir. Bu çalışmaya ek olarak farklılaşan puanlayıcı davranışlarına yönelik farklı yüzeylere ait ikili ve üçlü yanlılık etkileşimlerine bakılabilir.
2. Çalışmada analitik dereceli puanlama anahtarının değerlendirme ölçütlerine yönelik ölçüm raporlarına ait bulgularda halo etkisi ve merkeze eğilim davranışlarına rastlanmamıştır. Öğrencilerin performansları öğretmen puanlayıcıları tarafından analitik DPA'nın ölçütleri değiştirilmeden dereceleme kategorilerinin sayısı arttırılarak halo etkisi ve merkeze eğilim davranışları incelenebilir.

3. Puanlayıcı sayısı arttırılarak veya öğrencilerin de puanlama sürecine dahil edilerek (yani öz ve akran puanlama yaptırılıp) çalışmanın kapsamı genişletilebilir.

4. Çalışmanın sonuçlarını desteklemek amacıyla ek olarak nitel bir çalışma yapılabilir.

5.2.2. Yapılacak Çalışmalara Yönelik Öneriler

1. Mevcut çalışmada öğrencilerin Türkçe dersindeki başarı durumları dikkate alınmamıştır. Benzer bir çalışmada öğrencilerin Türkçe dersindeki başarıları ile olan ilişkilerine bakılarak farklılaşan puanlayıcı davranışları incelenebilir.
2. Mevcut çalışmada farklılaşan puanlayıcı davranışlarının belirlenmesinde sadece öğretmen puanlamaları dikkate alınmış olup, benzer bir çalışmada öz, akran ve öğretmen puanlamaları arasındaki ilişki incelenebilir.
3. Mevcut çalışmada performansın değerlendirilmesi sürecinde analitik DPA kullanılmıştır. Benzer bir çalışmada performansın değerlendirilmesi sürecinde analitik DPA yerine, bütünsel DPA kullanılarak farklılaşan puanlayıcı davranışları incelenebilir.
4. Mevcut çalışmada performansın değerlendirilmesinde farklılaşan puanlayıcı davranışlarının belirlenmesinde sadece puanlayıcı yüzeyine ait yanlılık etkileşim analizleri incelenmiştir. Benzer bir çalışmada farklı yüzeyler arasındaki etkileşimler incelenebilir.

KAYNAKLAR

- Abu Kassim, N. L. (2007). *Exploring rater judging behaviour using the many-facet rasch model*. The Second Biennial International Conference on Teaching and Learning of English in Asia: Exploring New Frontiers (TELiA2), Faculty of Communication and Modern Languages, University Utara Malaysia, pp. 1-13. ISBN 978-983-42061-2-3.
- Ada, S. (2015). *Klasik test kuramına göre sayısal ve sözel alanlar için puanlama güvenilirliğinin kayıp veri kapsamında incelenmesi*. Yüksek Lisans Tezi. Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı. Gazi Üniversitesi, Ankara.
- Akın, Ö. & Baştürk, R. (2012). *Keman eğitiminde temel becerilerin Rasch ölçme modeli ile değerlendirilmesi*. Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, Sayı 31, ss. 175-187.
- Alharby, E.R. (2006). *A comparison between two scoring methods, holistic vs. analytic using two measurement models, generalizability theory and the many facet Rasch measurement within the context of performance assessment*. Unpublished Doctoral Dissertation. Pennsylvania: Pennsylvania State University.
- Alper, R. (2001). *Spor bilimlerinde uygulamalı istatistik*. Ankara: Nobel Yayın Dağıtım.
- Arı, A. (2010). *Öğretmenlere göre proje ve performans görevlerinin uygulanmasında karşılaşılan sorunlar*. Elektronik Sosyal Bilimler Dergisi, 9(34), 32-55.
- Anadol, H.Ö. & Doğan.C.D. (2018). *Dereceli puanlama anahtarlarının güvenilirliğinin farklı deneyim yıllarına sahip puanlayıcıların kullanıldığı durumlarda incelenmesi*. İlköğretim Online, 2018; 17(2): s.1066-1076. Web: <http://ilkogretim-online.org.tr> doi 10.17051/ilkonline.2018.419355
- Aslanoğlu, A. E. (2003). *Öğretimde sunu becerilerinin değerlendirilmesinde dereceli puanlama anahtarı (rubric) kullanılmasına ilişkin bir araştırma*. Yüksek Lisans

- Tezi. Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı. Ankara Üniversitesi.
- Atılğan, H. (2004). *Genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch modelinin karşılaştırılmasına ilişkin bir araştırma*. Doktora Tezi. Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü. Ankara, Türkiye.
- Atılğan, H. (2005). *Genellenebilirlik kuramı ve puanlayıcılar arası güvenirlik için örnek bir uygulama*. Eğitim Bilimleri ve Uygulama, 4(7), 95-108.
- Atılğan, H., Kan, A., & Doğan, N. (2011). *Eğitimde ölçme ve değerlendirme*. Anı Yayıncılık.
- Atılğan, H. (2019a). *Generalizability theory*. Ö.T. Ahmet Doğanay in, Educational Sciences, (s.111-123). Ankara: Academician Publishing House Inc.
- Atılğan, H. (2019b). *Genellenebilirlik kuramı ve uygulaması*. Anı Yayıncılık. Ankara.
- Bahrouni, F. (2016). *Using multi-facet Rasch model (MFRM) in rater-mediated assessment*. The Journal of Teaching English For Specific And Academic Purposes. Volume 4(1). s.195-212.
- Baird, J.A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability a comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling*. Oxford University Centre For Educational Assessment.
- Baker, F. B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.
- Balcı, A. (2015). *Sosyal bilimlerde araştırma: Yöntem, teknik ve ilkeler*. Ankara: PegemA Yayınevi.
- Barkaoui, K. (2013). *Multifaceted Rasch analysis for test evaluation*. The companion to language assessment, 3, 1301-1322.
- Baştürk, R. (2010). *Bilimsel araştırma ödevlerinin çok yüzeyli Rasch ölçme modeli ile değerlendirilmesi*. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1(1), 51-57.
- Bıkmaz, Ö. (2011). *Üst düzey zihinsel özelliklerin ölçülmesinde puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması*, Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara.

- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.)*. Routledge, New York.
- Börkan, B. (2017a). *Akran değerlendirilmesinde puanlayıcı katılığı kayması*. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi. 8(4); 469-489. ISSN: 1309 – 6575.
- Börkan, B. (2017b). *Exploring variability sources in student evaluation of teaching via many-facet Rasch model*. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8(1), 15-33.
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.
- Brookhart, S. M. (1999). *The art and science of classroom assessment: The missing part of pedagogy*. ASHE-ERIC Higher Education Report, 27(1).
- Büyüköztürk, Ş. (2010). *Sosyal bilimler için veri analizi el kitabı* (12. baskı). Ankara: Pegem Akademi Yayıncılık.
- Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2017). *Bilimsel araştırma yöntemleri*. Pegem A Yayıncılık, 1-360.
- Can, A. (2017). *SPSS ile bilimsel araştırma sürecinde nicel veri analizi*. Pegem A Yayıncılık.
- Caine, K. (2008). *Writing to persuade minilessons to help students plan, draft, and revise, Grades 3–8*, Heinemann Portsmouth, NH.
- Choi, S. K., & Young-Min, P. (2011). *The rater characteristics and raters bias in korean language teacher's persuasive writing assessment*. Journal of Curriculum Evaluation, 14(1), 201-228.
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.
- Çelik, M. E. (2012). *İlköğretim sekizinci sınıf öğrencilerinin yazılı anlatım becerilerinin farklı değişkenler açısından değerlendirilmesi*. Turkish Studies- International Periodical for the Languages, Literature and History of Turkish or Turkic. 7 (1), 727-743.
- Çetin, B., & İlhan, M. (2017). *Standart ve SOLO Taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı katılığı ve cömertliğinin incelenmesi*. Eğitim ve Bilim, 42(189).

- Çokluk, Ö., Şekercioğlu, G. & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve Lisrel uygulamaları*, Pegem Akademi Yayıncılık, Ankara.
- DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). *A hierarchical rater model for constructed responses, with a signal detection rater model*. Journal of Educational Measurement, 48 (3), 333–356.
- Doğan, C. D., & Kutlu, Ö. (2011). *Öğretmen adaylarının yeni durum belirleme yöntemlerini tercihlerinde etkili olan öğrenmeye ilişkin özellikler*. Kastamonu Eğitim Dergisi, 19(2), 459-474.
- Downing, S.M. (2005). *Threats to the validity of clinical teaching assessments: What about rater error?*. Medical Education, 39(4), 353-355. DOI: 10.1111/j.1365-2929.2005.02138.x
- Dunbar, N.E., Brooks, C.F., & Miller, T.K. (2006). *Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills*. Innovative Higher Education, 31(2), 2006, 115-128.
- Ebel, R.L. (1951). *Estimation of the reliability of ratings*. Psychometrika, 16(4), 407- 424.
- Eckes, T. (2015). *Introduction to Many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang GmbH, Internationaler Verlag der Wissenschaften.
- Eckes, T. (2017). *Rater effects: Advances in item response modeling of human ratings – Part I*. Psychological Test and Assessment Modeling, Volume 59 (4), 443-452.
- Eckes, T. (2019). *Many-Facet Rasch measurement: Implications for rater-mediated language assesment*. Quantitative Data Analysis for Language Assessment Volume I, Chapter 7. Fundamental Techniques, Edited by Vahid Aryadoust and Michelle Raquel, 153-175.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Multivariate Applications Books Series. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Engelhard, G., Jr. (1994). *Examining rater errors in the assessment of written composition with a many-faceted Rasch model*. Journal of Educational Measurement , 31 (2), 93- 112.

- Engelhard, G. (2002). *Monitoring raters in performance assessments*. Large-scale assessment programs for all students: validity, Technical Adequacy, and Implementation. Volume 11, s. 621-288. Editor: Tindal, G. & Haladyna, T.M. Mahwah, NJ: Lawrence Erlbaum.
- Engelhard, G., & Stone, G. E. (1998). *Evaluating the quality of ratings obtained from standard-setting judges*. Educational and Psychological Measurement, 58(2), 179–196. <https://doi.org/10.1177/0013164498058002003>
- Elhan, A. H., & Atakurt, Y. (2005). *Ölçeklerin değerlendirilmesinde niçin Rasch analizi kullanılmalıdır?*. Ankara Üniversitesi Tıp Fakültesi Mecmuası, 58(01), 47-50.
- Erdoğan, S. (2004). *Sürekli değişkenler için güvenilirlik analizinde kullanılan ilişki katsayılarının değişken varyansından ve etki büyüklüğünden etkilenme durumları*. Yüksek Lisans Tezi. Biyoistatistik anabilim dalı. Mersin Üniversitesi.
- Erkuş, A., Sünbül, Ö., Ömür Sünbül S., Yormaz S. & Aşiret S. (2017). *Psikolojide ölçme ve ölçek geliştirme – 2: Ölçme araçlarının psikometrik nitelikleri ve ölçme*. Pegem Akademi, Ankara.
- Erman Aslanoğlu, A., & Kutlu, Ö. (2003). *Öğretimde sunu becerilerinin değerlendirilmesinde dereceli puanlama anahtarı (rubric) kullanılmasına ilişkin bir araştırma*. Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, Vol. 36, issue:1-2, 25-36.
- Farrokhi, F., Esfandiari, R., & Vaez Dalili, M. (2011). *Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment*. World Applied Sciences Journal, 15(11), 70-77.
- Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012a). *A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment*. JALT Journal, 34(1), 79-101.
- Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012b). *A Many-Facet Rasch measurement of differential rater severity/leniency in self-assessment, peer-assessment, and teacher assessment*. Journal of Basic and Applied Scientific Research, 2(9)8786-8798.
- Goodrich, H. (1997). *Understanding rubric*. Educational Leadership, 54(4), 14-17. 18.01.2020 tarihinde <http://www.ascd.org/publications/educational-leadership/dec96/vol54/num04/understanding-rubrics.aspx> adresinden erişilmiştir.

- Goodwin, L.D. (2001). *Interrater agreement and reliability*. Measurement in Physical Education and Exercise Science, 5(1), 13-34.
- 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.
- Govindasamy, P., del Carmen Salazar, M., Lerner, J., & Green, K. E. (2019). *Assessing the reliability of the framework for equitable and effective teaching with the many-facet Rasch model*. Frontiers in Psychology, 10:1363. doi: 10.3389/fpsyg.2019.01363.
- Grabowski, K.C. & Lin, R. (2019). *Multivariate generalizability theory*. Quantitative Data Analysis for Language Assessment Volume I, Chapter 3. Fundamental Techniques, Edited by Vahid Aryadoust and Michelle Raquel, 54-80.
- Güler, N. (2008). *Klasik test kuramı, genellenebilirlik kuramı ve Rasch modeli üzerine bir araştırma*. Doktora Tezi, Hacettepe Üniversitesi, Ankara, Türkiye.
- Güler, N. ve Gelbal, S. (2010). *Açık uçlu matematik sorularının güvenilirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi*. Kuram ve Uygulamada Eğitim Bilimleri, 10 (2), 989-1019. 23.12.2019 tarihinde <http://www.idealonline.com.tr/IdealOnline/pdfViewer/index.xhtml?uId=2120&ioM=Paper&preview=true&isViewer=true#pagemode=bookmarks> adresinden erişilmiştir.
- Güler, N., Uyanık, G. K. & Taşdelen Teker, G. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi.
- Güler, N., & Taşdelen Teker, G. (2015). *Açık uçlu maddelerde farklı yaklaşımlarla elde edilen puanlayıcılar arası güvenilirliğin değerlendirilmesi*. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6(1), 12-24.
- Güler, N. (2014). *Analysis of open-ended statistics questions with many facet Rasch model*. Eurasian Journal of Educational Research, 55, 73-90.
- Haiyang, S. (2010). *An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates*. Chinese Journal of Applied Linguistics, 33(2), 87-102.
- Haladyna, M. T. (1997). *Writing test items to evaluate higher order thinking*. Needham Heights: Allyn and Bacon.

- Hambleton, R. K. & Jones, R. W. (1993). *Comparison between classical test theory and item response theory and their applications to the test development*. Educational Measurement, 12, 38-47.
- Hambleton, R. K., Swaminathan, H. ve Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Sage Publications. The International Professional Publishers.
- Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.
- Han, T. (2013). *The impact of rating methods and rater training on the variability and reliability of EFL students' classroom-based writing assessments in turkish universities: An investigation of problems and solutions*. Doktora Tezi, Atatürk Üniversitesi, Sosyal Bilimler Enstitüsü, Erzurum.
- Hattie, J., Jaeger, R.M. & Bond, L.(1999). *Chapter 11: Persistent methodological questions in educational testing*. Review of Research in Education, AERA. Volume: 24 issue: 1, page(s): 393-446.
- He, T. H. (2019). *The impact of computers on marking behaviors and assessment: A Many-Facet Rasch measurement analysis of essays by EFL college students*. SAGE Open, 9(2), 2158244019846692.
- Hoyt, W. T. (2000). *Rater bias in psychological research: When is it a problem and what can we do about it?*. Psychological methods, 5(1), 64.
- Howell, C. D. (2010). *Statistical methods for psychology*. Belmont, CA: Thomson Wadsworth.
- İlhan, M., & Çetin, B. (2014). *Performans değerlendirmeye karışan puanlayıcı etkilerini azaltmanın yollarından biri olarak puanlayıcı eğitimleri: Kuramsal bir analiz*. Journal of European Education, volume. 4, issue. 2, p. 29-38, ISSN 2146-2674.
- İlhan, M. (2015). *“Standart ve Solo taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi”*. Doktora Tezi, Eğitim Bilimleri Enstitüsü, Eğitim Bilimleri ABD, Gaziantep Üniversitesi.
- İlhan, M. (2016). *Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması*.

- İlhan, M. & Güler, N. (2018). *Likert tipi ölçeklerde Rasch modelinin kullanımı: Olumsuz değerlendirilme korkusu ölçeği-öğrenci formu (ODKÖ-ÖF) üzerinde bir uygulama*. Trakya Eğitim Dergisi, Cilt 8, Sayı 4, 756-775.
- Jackson, S.E., Schuler, R.S., & Werner, S. (2009). *Managing human resources*. Mason, OH: Cengage/Southwestern Publisher.
- Jin, K. Y., & Wang, W. C. (2017). *Assessment of differential rater functioning in latent classes with new mixture facets models*. Multivariate behavioral research, 52(3), 391-402.
- Jonsson, A., & Svingby, G. (2007). *The use of scoring rubrics: Reliability, validity and educational consequences*. Educational Research Review, 2(2), 130-144.
- Karakaya,İ. (2015). *Comparison of self, peer and instructor assessments in the portfolio assessment by using many facet Rasch model*. Journal of Education and Human Development, 4(2), (182-192).
- Karakaya Özyer, K. (2018). *Açıköğretim sistemindeki açık uçlu soruların çok yüzeyli Rasch modeli analizi yöntemiyle puanlanması*. Açıköğretim Uygulamaları ve Araştırmaları Dergisi, 4(1), 61-77.
- Karasar, N. (2017). *Bilimsel araştırma yöntemi: kavramlar-ilkeler-teknikler*. Nobel Yayın Dağıtım.
- Kaplan, S. (2006). *Performans değerlendirme: Performans görevlerinin hazırlanması ve puanlanması*. Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi.
- Kaptan, A. (2015). *İlköğretim beşinci sınıf öğrencilerinin ikna edici yazma beceri düzeylerinin belirlenmesi*. Yüksek Lisans Tezi, Eğitim Bilimleri Enstitüsü, Sınıf Öğretmenliği ABD. Eskişehir Osmangazi Üniversitesi.
- Kılıç, S. (2009). *Ölçümlerin uyumluluğu ve tıptaki uygulamaları*. Yüksek Lisans Tezi, Sağlık Bilimleri Enstitüsü, Biyoistatistik ABD, Çukurova Üniversitesi, Adana.
- Knoch, U., Read, J., & von Randow, J. (2007). *Re-training writing raters online: How does it compare with face-to-face training?*. Assessing Writing, 12(1), 26-43.
- Kondo-Brown, K. (2002). *A FACETS analysis of rater bias in measuring Japanese second*

language writing performance. *Language Testing*, 19(1), 3-31.

Koons, H.H. (2008). *The reading-writing connection: An investigation of the relationship between reading ability and writing quality across multiple grades and three writing discourse modes*. Doctoral Dissertation. University of North Carolina, ABD.

Korkmaz, H. (2004). *Fen ve teknoloji eğitiminde alternatif değerlendirme yaklaşımları*. Ankara: Yeryüzü Yayınevi.

Korkmaz, Y. (2009). *Fen öğretiminde rubrik kullanma eğitiminin öğretmenlerin ölçme ve değerlendirmeye ilişkin görüş ve uygulamalarına etkisi*. Yüksek Lisans Tezi, İlköğretim Anabilim Dalı, Selçuk Üniversitesi, Konya.

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.

Kubiszyn, T., & Borich, G. (2013). *Educational testing and measurement: Classroom application and practice (10th ed.)*. Hoboken, NJ: John Wiley & Sons, Inc.

Kumar, D. (2005). *Performance appraisal: The importance of rater training*. Journal of the Kuala Lumpur Royal Malaysia Police College, 4, 1-15.

Kurt, S. (2019). *Dört kare yazma yönteminin ortaokul 5. sınıf öğrencilerinin yazılı anlatım becerileri üzerine etkisi*. Yüksek Lisans Tezi, Sosyal Bilimler Enstitüsü, Türkçe Eğitimi Bilim Dalı, Ağrı İbrahim Çeçen Üniversitesi.

Kurudayıoğlu, M., & Yılmaz, E. (2014). *How are we persuaded? Persuasive text, and structure*. Journal of Theory and Practice in Education, 10(1): 75-102. ISSN: 1304-9496

Kutlu, Ö., Doğan, D. C. & Karakaya, İ. (2014). *Öğrenci başarısının belirlenmesi: performansa ve portfolyoya dayalı durum belirleme*. Ankara: Pegem Akademi.

Kutlu, Ö., Yıldırım, Ö. & Bilican, S. (2009). *Öğretmenlerin dereceli puanlama anahtarlarına ilişkin tutum ölçeği geliştirme çalışması*. Yüzüncü Yıl Üniversitesi, Eğitim Fakültesi Dergisi. Cilt:VI, Sayı:II, 76-88.

Landy, R.J., & Farr, J.L. (1980). *Performance rating*. Psychological Bulletin, 87(1), 72-107.

- Lianzhen, H. & Jie, Z. (2008). *Investigating the reliability of CET-SET using multi-facet Rasch model*. Modern Foreign Language Education in China.
- Linacre, J. D. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.
- Linacre, J. M. (1993a). *Rasch-based Generalizability Theory*. Rasch Measurement Transactions, 7:1, p. 283-284. <https://www.rasch.org/rmt/rmt71h.htm> adresinden erişilmiştir.
- Linacre, M. (1993b). *Generalizability theory and many-facet Rasch measurement*. Paper presented at the annual meeting of the American educational research association, Atlanta, GA.
- Linacre, J. M. (1994). *Constructing measurement with a many-facet Rasch model*. In Wilson, M. (Ed.), *Objective measurement: Theory into practice*, Volume 2 (pp. 129-144). Norwood, NJ: Ablex.
- Linacre, J. M. (1996). *Generalizability theory and many-facet Rasch measurement*. In Engelhard, G. & Wilson, M. (Eds.), *Objective measurement: Theory into practice*, Volume 3 (pp. 85-98). Norwood, NJ: Ablex.
- Linacre, J.M. (2002). *Number of person or item strata. Rasch measurement: transactions of the Rasch measurement SIG*, 16, 888. 02.11.2019 tarihinde <https://www.rasch.org/rmt/rmt163f.htm> adresinden erişilmiştir.
- Linacre, J.M. (2018). *A user's guide to FACETS Rasch-model computer programs program manual 3.81. 0 by*. Chicago: MESA Press.
- Linacre, M., Engelhard, G., Tatum, D.S. & Myford, C.M. (1994). *Measurement with judges: Many-faceted conjoint measurement. Chapter 1*. International Journal of Educational Research Volume 21, Issue 6, Pages 569-577.
- Linlin, C. (2020). *Comparison of automatic and expert teachers' rating of computerized english listening-speaking test*. English Language Teaching, 13(1), 18-30.
- Loignon, A. C., Woehr, D. J., Thomas, J. S., Loughry, M. L., Ohland, M. W., & Ferguson, D. M. (2017). *Facilitating peer evaluation in team contexts: The impact of frame-of-reference rater training*. Academy of Management Learning & Education, 16(4), 562-578. <http://dx.doi.org/10.5465/amle.2016.0163>
- Lord, F.M., Novick, M.R., & Birnbaum, A. (2008). *Statistical theories of mental test*

scores. Oxford, England: Addison-Wesley.

- Lonigro, A., Baiocco, R., Baumgartner, E., & Laghi, F. (2017). *Theory of mind, affective empathy, and persuasive strategies in school-aged children*. *Infant and Child Development*, 26(6), e2022.
- Lonigro, A., Baiocco, R., Baumgartner, E., Sette, S., & Laghi, F. (2017). *Persuasion in school-aged children: How does it change if the persuadee is the mother or the peer?*. *Cognitive processing*, 18(1), 67-74.
- Lottridge, S.M., Schulz, E. M. & Mitzel, H.C. (2013). *Using automated scoring to monitor reader performance and detect reader drift in essay scoring*. the handbook of automated essay evaluation: Current applications and new directions, p. 233-250. New York: Routledge Academic.
- Lumley, T. & McNamara, T. F. (1995). *Rater characteristics and rater bias: Implications for training*. *Language Testing*, 12 , 54-71.
- Lunz, M. E., & Linacre, J. M. (1998). *Measurement designs using multifacet Rasch modeling*. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 47–77). Mahwah, NJ: Erlbaum.
- Marzona, R. J. (2001). *Designing a new taxonomy of educational objectives*. Thousand Oak, California: Corwin.
- McDonald, M.B. (1999), *Seed deterioration: physiology repair and assessment*. *Seed Science and Technology*, 27(1), 177-237.
- McNamara, T. (1996). *Measuring second language performance*. London, UK: Longman.
- MEB (2019). *İlköğretim Türkçe dersi (1-8. sınıflar) öğretim programı*. Ankara: Talim ve Terbiye Kurulu Başkanlığı.
- Mertler, C. A. (2001). *Designing scoring rubrics for your classroom*. *Practical Assessment, Research & Evaluation*, 7(25).
- Moore, B.B. (2009). *A consideration of rater effects and rater design via signal detection theory*. Doktora Thesis. <http://proquest.com> sayfasından erişilmiştir.
- Moskal, B.M. (2000). *Scoring rubrics: What, when, how?* *Practical Assessment, Research and Evaluation*, 7(3).
- Mulqueen C., Baker D., & Dismukes, R.K. (2000). *Using multifacet Rasch analysis to*

examine the effectiveness of rater training. Presented at the 15th Annual Conference for the Society for Industrial and Organizational Psychology (SIOP). New Orleans.

Mushquash, C., & O'Connor, B. P. (2006). *SPSS, SAS, and MATLAB programs for generalizability theory analyses.* Behavior Research Methods, 38(3), 542-547.

Myford, C.M., & Wolfe, E.W. (2003). *Detecting and measuring rater effects using many-facet Rasch measurement: Part I.* Journal of Applied Measurement, 4(4), 386-422.

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.

Nakamura, Y. (2000). *Many facet Rasch based analysis of communicative language testing results.* Journal of Communication Students, V12, 3-13.

Nalbantoğlu Yılmaz, F. (2017). *Analysis of the Rater effects on the scoring of diagnostic trees prepared by teacher candidates with the many-facet Rasch model.* Online Submission, 8(18), 174-184.

Newell, G. E., Beach, R., Smith, J., & Van Der Heide, J. (2011). *Teaching and learning argumentative reading and writing: A review of research.* Reading Research Quarterly, 46, 273–304.

Newell, G. E., Bloome, D., Kim, M. Y., & Goff, B. (2019). *Shifting epistemologies during instructional conversations about “good” argumentative writing in a high school English language arts classroom.* Reading and Writing, 32(6), 1359-1382.

Nitko, A. J., & Brookhart, S.M. (2014). *Educational assessment of students (6rd ed.).* Pearson New International Edition.

Okur, A., Göçen, G., & Süğümlü, Ü. (2013). *İkna edici yazma ve karşılaştırmalı bir araştırma/persuasive writing and a comparative study (example of Australian native language teaching materials and Turkish).* Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10(21), 167-197.

Osburn, H. G. (2000). *Coefficient alpha and related internal consistency reliability coefficients.* Psychological methods, 5(3), 343.

Parlak, B., & Doğan, N. (2014). *Dereceli puanlama anahtarı ve puanlama anahtarından elde edilen puanların uyum düzeyleri.* Hacettepe Üniversitesi Eğitim Fakültesi

Dergisi, 29(29-2), 189-197.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). *The hierarchical rater model for rated test items and its application to large-scale educational assessment data*. Journal of Educational and Behavioral Statistics, 27 (4), 341–384.

Perlman, C. C. (2003). *Performance assessment: Designing appropriate performance tasks and scoring rubrics*. In Wall, J.E., & Walz, G.R. (Ed.), *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators* (s. 497-506). Greensboro, NC: CAPS Press. 24 Kasım 2019 tarihinde <https://files.eric.ed.gov/fulltext/ED480070.pdf> adresinden erişilmiştir.

Perloff, M. R. (2010). *The dynamics of persuasion*. Communication and attitudes in the twenty-first century (4th ed. ed.). New York, USA: Routledge.

Polat, M. (2018). *Defining severe graders through many faceted Rasch measurement*. European Journal of Foreign Language Teaching.

Popham, J. W. (1997). *What's wrong and what's right with rubric*. Educational Leadership. 55, (2), 12.

Raymond, M. R., & Viswesvaran, C. (1993). *Least squares models to correct for rater effects in performance assessment*. Journal of Educational Measurement, 30 (3), 253–268.

Russell, M.K., & Airasian, P. W. (2011). *Classroom assessment*. (7 th ed.). New York: McGraw-Hill.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). *Rating the ratings: Assessing the psychometric quality of rating data*. Psychological Bulletin, 88(2), 413-428.

<http://dx.doi.org/10.1037/0033-2909.88.2.413>

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

Semerci, Ç. (2011). *Doktora yeterlikler çerçevesinde öğretim üyesi, akran ve öz değerlendirmelerin Rasch ölçme modeliyle analizi*. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 2(2), 164-171.

Shalvelson, R. J. & Webb, N. M. (1991). *Generalizability theory: a primer*. USA: Sage.

Shalvelson, R. J. & Webb, N. M. (2003). *Generalizability theory*. Editor: Kempf-Leonard

- in, Encyclopdia of Social Measurement. San Diago: Academic Press.
- Shirazi, M. A. (2019). *For a greater good: Bias analysis in writing assessment*. SAGE Open, 9(1), 2158244018822377.
- Shoukri M. M. (2004). *Measures of interobserver agreement*. Chapman & Hall, Newyork.
- Soo Park, Y., & Xing, K. (2019). *Rater model using signal detection theory for latent differential rater functioning*. Multivariate behavioral research, 54(4), 492-504.
- Stark, J. F. (1998). *Measurement and evaluation in education*. Ohio: Great & Great Puplishers.
- 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.
- Şata, M. (2019). *Performans değerlendirme sürecinde puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisinin incelenmesi*. Yayınlanmamış Doktora Tezi. Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara, Türkiye.
- Şata, M. & Karakaya, İ. (2020). *Investigation of the use of electronic portfolios in the determination of student achievement in higher education using the many-facet Rasch measurement model*. Educational Policy Analysis and Strategic Research, 15(7-21).
- Tamanini, K. B. (2008). *Evaluating differential rater functioning in performance ratings: Using a goal-based approach*. Doctoral dissertation, Ohio University.
- Tekindal, S., Alıcı, D., Çakan, M., Kan, A., Karaca, E. Özbek, Y.Ö., & Yaşar, M. (2010). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi.
- Tindal, G., & Haladyna, T. M. (2002). *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*. Routledge.
- Üçpınar, F. K., & Ünaldı, A. (2018). *Validation of writing scales for turkish as a second language through many-facet Rasch measurement*. Boğaziçi Üniversitesi Eğitim Dergisi, 34(1), 23-48.
- Wan, S. (2017). *An application of many-facet Rasch model to the analysis of an implicit associate test: Disentangling implicit gender-science stereotype*. Doctoral dissertation, University of Georgia.

- Wang, J. & Engelhard, G. (2019). *Conceptualizing rater judgments and rating processes for rater-mediated assessments*. Journal of Educational Measurement, Vol. 56, No. 3, pp. 582–609. DOI: 10.1111/jedm.1222
- Weigle, S.C. (1998). *Using FACETS to model rater training effects*. Language Testing, 15(2), 263–287. <https://doi.org/10.1177/026553229801500205>.
- Wesolowski, B. C., Wind, S. A., & Engelhard Jr, G. (2015). *Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning*. Musicae Scientiae, 19(2), 147-170.
- Wind, S.A. (2018a). *Using Guttman errors to explore rater fit in rater-mediated performance assessments*. Methodological Innovations. 1–17. Article reuse guidelines: sagepub.com/journals-permissions.
- Wind, S.A. (2018b). *Examining the impacts of rater effects in performance assessments*. Applied Psychological Measurement, Vol. 43(2) 159–171. Article reuse guidelines: sagepub.com/journals-permissions.
- Wind, S. A., & Guo, W. (2019). *Exploring the combined effects of rater misfit and differential rater functioning in performance assessments*. Educational and Psychological Measurement, 0013164419834613.
- 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 (1994), 67, 189—205.
- Wolf, K., & Stevens, E. (2007). *The role of rubrics in advancing and assessing student learning*. The Journal of Effective Teaching, 7(1), 3-14.
- Wolfe, E. W. (2004). *Identifying rater effects using latent trait models*. Psychology Science, 46(1), 35–51.
- Wolfe, E.W. & McVay, A. (2010). *Rater effects as a function of rater training context*. Using Assessment and Research to Promote Learning, Pearson Assessments.
- Wolfe, E. W., & McVay, A. (2012). *Application of latent trait models to identifying substantively interesting raters*. Educational Measurement: Issues and Practice, 31

(3), 31–37.

- Wolfe, E. W., & Song, T. (2015). *Comparison of models and indices for detecting rater centrality*. Journal of Applied Measurement, 16(3), 228–241.
- Wright, B. D., & Linacre, J. M. (1994). *Reasonable mean-square fit values*. Rasch Measurement Transactions, 8, 370-371.
- Wu, M. (2017). *Some IRT-based analyses for interpreting rater effects*. Psychological Test and Assessment Modeling, 59(4), 453–470.
- Yazıcı, N. (2013). *Başarının ölçülmesinde SOLO taksonomiye dayalı hazırlanan rubrik kullanımının etkisinin karşılaştırmalı olarak incelenmesi*. Yüksek Lisans Tezi, Kahramanmaraş Sütçü İmam Üniversitesi, Sosyal Bilimler Enstitüsü, Kahramanmaraş.
- Youn, S. J. (2007). *Rater bias in assessing the pragmatics of KFL learners using facets analysis*. University of Hawai'i Second Language Studies Paper 26 (1).
- Yurdugül, H. (2005). *Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması*. XIV. Ulusal Eğitim Bilimleri Kongresi (28–30 Eylül 2005, Denizli), Kongre kitabı, s:771-774.
- Yüzüak, A.V., Erten, S. & Kara, Y. (2019). *Analysis of laboratory videos of science teacher candidates with many-facet Rasch measurement model*. Journal of Education in Science, Environment and Health (JESEH), 5(2), 146-155. DOI:10.21891/jeseh.568816
- Yüzüak, A. V., Yüzüak, B., & Kaptan, F. (2015). *Performans görevinin akran gruplar ve öğretmen yaklaşımları doğrultusunda çok-yüzeyle Rasch ölçme modeli ile analizi*. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Cilt 6, Sayı 1, 1-11.
- Zhu, X. (2009). *Assessing fit of item response models for performance assessments using bayesian analysis*. Doctoral Thesis, University of Pittsburgh, Pittsburgh, ABD.

EKLER

EK 1. Evcil Hayvan Besleme



Arkadaşlarınız evlerinde evcil hayvan (kedi, kuş, köpek, balık, vs.) besliyorlar. Sen de evinizde evcil hayvan olsun istiyorsun. Fakat ailenin evde hayvan bakılamayacağı konusunda endişeleri var. Ailenin bu konudaki endişelerini gidermek ve bakacağın evcil hayvanın sorumluluğunu en iyi şekilde üstleneceğin konusunda onları **ikna edecek** bir metin oluştur.

Önemli Not:

1. Yazacağınız metin ikna edici cümleler içermeli, seçtiğiniz görev uygun şekilde gerekçe sunmalı ve **en az 10 cümleden** oluşmalıdır.
2. Yazacağınız metin için **başlık** koymayı unutmayınız.
3. Noktalama işaretlerine ve yazım kurallarına dikkat ediniz.
4. Çalışmanızı **ikna edici** bir slogan cümlesi ile tamamlayabilirsiniz.
5. Metni yazmanız için ideal süre 25 dk'dır.
6. Yazmış olduğunuz metinler "*görüş sunma, içerik, yapı, dil ve anlatım, yazım ve dilbilgisi, sayfa düzeni*" ölçütleri doğrultusunda değerlendirilecektir.
7. Yazma çalışması için sadece size verilen yazma alanını kullanınız.

YAZMA ALANI

[illegible]

EK 2. Eve İnternet Bağlatmak



Annenez eve internet bağlatmak istiyor. Fakat sizin ödev yapmaktan çok oyun oynayacağını, sosyal medyada çok vakit harcayacağını düşünmesi ve bu durumun derslerinizi aksatacağı yönünde endişeleri var. Eve internet bağlatmak konusunda çok kararsız. Annenizi eve internet bağlatması konusunda ikna etmeniz ve endişelerini gidermeniz gerekiyor. Bu konu hakkında annenizi **ikna edecek** bir yazı yazınız.

Önemli Not:

1. Yazacağınız metin ikna edici cümleler içermeli, seçtiğiniz görev uygun şekilde gerekçe sunmalı ve **en az 10 cümleden** oluşmalıdır.
2. Yazacağınız metin için başlık koymayı unutmayınız.
3. Noktalama işaretlerine ve yazım kurallarına dikkat ediniz.
4. Çalışmanızı ikna edici bir slogan cümlesi ile tamamlayabilirsiniz.
5. Metni yazmanız için ideal süre 25 dk'dır.
6. Yazmış olduğunuz metinler "*görüş sunma, içerik, yapı, dil ve anlatım, yazım ve dilbilgisi, sayfa düzeni*" ölçütleri doğrultusunda değerlendirilecektir.
7. Yazma çalışması için sadece size verilen yazma alanını kullanınız.

YAZMA ALANI

[illegible]

EK 3. Çanakkale'ye Gezi



Okulunuz "18 Mart Çanakkale ve Şehitleri Anma Günü" etkinlikleri çerçevesinde Çanakkale'ye gezi organizasyonu düzenledi. Bütün arkadaşlarınız bu geziye katılacak. Sen de bu geziye gitmeyi çok istiyorsun. Fakat ailen bu geziye gitmeni çeşitli sebeplerden dolayı (ücretin çok fazla olması, derslerden geri kalacağınızı düşünmeleri, ... vs.) istemiyor. Bu geziye gitmek için aileni **ikna edecek** bir metin oluşturun.

Önemli Not:

1. Yazacağınız metin ikna edici cümleler içermeli, seçtiğiniz görev uygun şekilde gerekçe sunmalı ve **en az 10 cümleden** oluşmalıdır.
2. Yazacağınız metin için **başlık** koymayı unutmayınız.
3. Noktalama işaretlerine ve yazım kurallarına dikkat ediniz.
4. Çalışmanızı **ikna edici** bir slogan cümlesi ile tamamlayabilirsiniz.
5. Metni yazmanız için ideal süre 25 dk'dır.
6. Yazmış olduğunuz metinler "*görüş sunma, içerik, yapı, dil ve anlatım, yazım ve dilbilgisi, sayfa düzeni*" ölçütleri doğrultusunda değerlendirilecektir.
7. Yazma çalışması için sadece size verilen yazma alanını kullanınız.

YAZMA ALANI

[illegible]

EK 4. Okul Takımı Seçmeleri



Beden eğitimi öğretmeniniz gelecek yıl okul takımlarında (basketbol, futbol, voleybol, hentbol, tenis ...vs.) oynayacak öğrencilerin seçmelerinin yakında yapılacağını söyledi. Siz de bir takımda yer almayı çok istiyorsunuz ancak aileniz son sınıfta derslerinize ve gireceğiniz sınava daha fazla odaklanmanızı istediği için buna karşı çıkıyor. Okul takımına katılmanın derslerinizi ve sınava hazırlığınızı aksatmayacağı konusunda ailenizi **ikna edecek** bir metin oluşturun.

Önemli Not:

1. Yazacağınız metin ikna edici cümleler içermeli, seçtiğiniz görev uygun şekilde gerekçe sunmalı ve **en az 10 cümleden** oluşmalıdır.
2. Yazacağınız metin için başlık koymayı unutmayınız.
3. Noktalama işaretlerine ve yazım kurallarına dikkat ediniz.
4. Çalışmanızı ikna edici bir slogan cümlesi ile tamamlayabilirsiniz.
5. Metni yazmanız için ideal süre 25 dk'dır.
6. Yazmış olduğunuz metinler "*görüş sunma, içerik, yapı, dil ve anlatım, yazım ve dilbilgisi, sayfa düzeni*" ölçütleri doğrultusunda değerlendirilecektir.
7. Yazma çalışması için sadece size verilen yazma alanını kullanınız.

YAZMA ALANI

[illegible]

EK 5. Geri Dönüşüm (Katı Atıklar)



Okulunuzda “**çevre ve temizlik kulübü**” etkinlikleri çerçevesinde katı atıkları toplama kampanyası düzenlenecektir. Bu kulüpte üye olduğunuzu ve katı atık toplama kampanyası için görevli olduğunuzu düşünün. Görevin; arkadaşlarının kağıt, cam, plastik, pil gibi maddeleri yerlere atarak hem çevre kirliliğine hem de ekonomik kayba neden olduğunu fark etmelerini ve bu atıkların okulda belirlenen geri dönüşüm kutularına atılmasını sağlamaktır. Buna dayanarak bir arkadaşlarını bu türden çöpleri okulda belirlenen alanlarındaki geri dönüşüm kutularına atması için onu **ikna edecek** bir metin oluşturun.

Önemli Not:

1. Yazacağınız metin ikna edici cümleler içermeli, seçtiğiniz görev uygun şekilde gerekçe sunmalı ve **en az 10 cümleden** oluşmalıdır.
2. Yazacağınız metin için başlık koymayı unutmayınız.
3. Noktalama işaretlerine ve yazım kurallarına dikkat ediniz.
4. Çalışmanızı ikna edici bir slogan cümlesi ile tamamlayabilirsiniz.
5. Metni yazmanız için ideal süre 25 dk'dır.
6. Yazmış olduğunuz metinler "*görüş sunma, içerik, yapı, dil ve anlatım, yazım ve dilbilgisi, sayfa düzeni*" ölçütleri doğrultusunda değerlendirilecektir.
7. Yazma çalışması için sadece size verilen yazma alanını kullanınız.

YAZMA ALANI

[illegible]

EK 6. Kişisel Temizlik



Sağlıklı yaşam için kişisel temizliğimize özen göstermeliyiz. Kişisel temizlik ile kastedilen şey vücut temizliğini düzenli gerçekleştirmek, gün içerisinde sık sık ellerimizi sabunla yıkamak, ağız ve diş temizliğine dikkat etmektir. Sağlıklı yaşam için temizliğin önemine yönelik sınıftaki bir arkadaşınızı kişisel temizliğini düzenli olarak gerçekleştirmesine yardımcı olacak şekilde onu **ikna edici** bir metin oluşturun.

Önemli Not:

1. Yazacağınız metin **ikna edici** cümleler içermeli, seçtiğiniz görev uygun şekilde gerekçe sunmalı ve **en az 10 cümleden** oluşmalıdır.
2. Yazacağınız metin için **başlık** koymayı unutmayınız.
3. Noktalama işaretlerine ve yazım kurallarına dikkat ediniz.
4. Çalışmanızı **ikna edici** bir slogan cümlesi ile tamamlayabilirsiniz.
5. Metni yazmanız için ideal süre 25 dk'dır.
6. Yazmış olduğunuz metinler "*görüş sunma, içerik, yapı, dil ve anlatım, yazım ve dilbilgisi, sayfa düzeni*" ölçütleri doğrultusunda değerlendirilecektir.
7. Yazma çalışması için sadece size verilen yazma alanını kullanınız.

YAZMA ALANI

[illegible]

EK 7. Benim Kararım



Aileniz sizin için çok önemli olan bir konuda (gideceğiniz okul, ileride seçeceğiniz meslek, sosyal aktivite, sportif faaliyet, hobi çalışması, ... vs.) sizin adınıza karar vermek istiyor. Fakat siz ailenizin seçimini kendiniz için uygun bulmuyorsunuz. Aileniz ise bu konuda çok ısrarcı davranıyor. Ailenizin bu tutumunu değiştirmek ve bazı konularda kendi kararlarınızı kendinizin alabileceği konusunda onları **ikna edecek** bir metin oluşturun.

Önemli Not:

1. Yazacağınız metin **ikna** edici cümleler içermeli, seçtiğiniz görev uygun şekilde gerekçe sunmalı ve **en az 10 cümleden** oluşmalıdır.
2. Yazacağınız metin için **başlık** koymayı unutmayınız.
3. Noktalama işaretlerine ve yazım kurallarına dikkat ediniz.
4. Çalışmanızı **ikna edici** bir slogan cümlesi ile tamamlayabilirsiniz.
5. Metni yazmanız için ideal süre 25 dk'dır.
6. Yazmış olduğunuz metinler "*görüş sunma, içerik, yapı, dil ve anlatım, yazım ve dilbilgisi, sayfa düzeni*" ölçütleri doğrultusunda değerlendirilecektir.
7. Yazma çalışması için sadece size verilen yazma alanını kullanınız.

YAZMA ALANI

[illegible]

EK 8. Yönerge



İKNA EDİCİ METİN YAZMA ÇALIŞMASI

ÖĞRENCİNİN;

Adı-Soyadı :
Sınıfı-Okul No :
Okulu :

Sevgili öğrenciler,

Bu çalışma ikna edici metin yazma becerilerini belirlemek amacıyla oluşturulmuştur. İkna edici yazma; bireyin herhangi bir konu hakkında hedef kitlenin (aile, öğretmeni arkadaş, topluluk, kurum, ..vs.) düşünce yapısında değişiklik yapmayı amaçlayan, herhangi bir şeyin olması gereken veya olmaması gereken şeklinin belirtildiği yazı türüdür. Sizden istenen arka sayfadaki ikna edici metin yazma görevini aşağıdaki yönergeye uygun olarak bir kompozisyon yazmanızdır. Çalışmanıza ait değerlendirmeler alanında uzman kişiler tarafından analitik dereceli puanlama anahtarı ile yapılacaktır. Bu çalışmanın sonuçları bilimsel bir çalışmada kullanılacak olup içtenlikle yazmanız beklenmektedir. Katılımınız için şimdiden teşekkür ederim



EK 9. Öğretmen Değerlendirme Formu

Sevgili puanlayıcı;

Bu form size verilen öğrenci kompozisyonlarının bir listesidir. Bu formda öğrenci kompozisyonlarının üzerindeki numaralar listedeki numaralar ile aynı öğrenciyi temsil etmektedir. Size verilen analitik dereceli puanlama anahtarındaki her bir kritere dayalı olarak verdiğiniz puanı (1-4) tabloda ilgili yerlere yazınız. Çalışmaya verdiğiniz destekten dolayı şimdiden teşekkür eder, kolaylıklar dilerim.

Puanlayıcının;

Adı ve Soyadı:

Branşı :

Kıdem Yılı :

Eğitim düzeyi: () Lisans () Yüksek lisans () Doktora

Çalıştığı okul türü:

() Ortaokul () İmam Hatip Ortaokulu () Diğer *Belirtiniz* :

Cinsiyeti: () Bayan () Bay

İKNA EDİCİ METİN YAZMA PERFORMANSINA YÖNELİK ÖĞRETMEN DEĞERLENDİRME FORMU							
Not: Öğrencilerin ikna metin yazma performansının değerlendirilmesinde size verilen analitik dereceli puanlama anahtarındaki 6 ölçütü dikkate alarak Oldukça Başarısız (1), Orta Düzey (2), İyi (3) ve Oldukça İyi (4) şeklinde puanlamanız gerekmektedir.							
Öğrenci No	Ölçütler						
	Görüş Sunma	İçerik	Yapı	Dil ve Anlatım	Yazım ve Noktalama	Sayfa Düzeni	Görev No
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							
26							
27							
28							
29							
30							
31							
32							

EK 10. Analitik Puanlama Anahtarı

ÖLÇÜTLER	PERFORMANS DÜZEYİ			
	Oldukça Başarısız (1)	Orta Düzey (2)	İyi (3)	Oldukça İyi (4)
Görüş Sunma	Görüş ve tarafını belli etmemiştir. Yazısında hiçbir iddiada bulunmamıştır. İddia varsa da konuya uygun değildir.	Görüş ve tarafını belli etmektedir ancak yeterince açık değildir. İddiasını ilgili kanıtlarla desteklemeye çalışmış ancak ikna edici detaylara yeterince yer vermemiştir.	Konu üzerindeki görüş, taraf ve iddiasını açık ve anlaşılır bir şekilde ele almıştır. Görüşünü tamamen olmasa da ikna edici detaylarla desteklemiştir.	Yazının bütününde belirli bir konuya odaklanmış ve başarılı bir şekilde görüşünü sunmuştur. Konu üzerindeki taraf ve tutumunu açık, anlaşılır ve güçlü bir şekilde ortaya koymuştur. Görüşünü tamamen ikna edici detaylarla desteklemiştir.
İçerik	Öğrenci yazdığı metinde anlatmak istediklerini net bir şekilde ortaya koymamış, düşüncelerini destekleyecek herhangi bir gerekçe sunmamıştır. Veya ortaya koymuş olduğu fikirler okuru ikna edici nitelikte değildir.	Öğrenci yazdığı metinde anlatmak istediklerini net bir şekilde ortaya koymuş, düşüncesini desteklemek için az da olsa gerekçe sunmuş, metni hemen hemen aynı kelimelerle tamamlamış olup fikirlerinin okuru etkilemesi/ikna etmesi zor görünmektedir. Aktarmak istediği mesajı açık, anlaşılır bir şekilde ifade etmek istemiş ancak tümüyle başarılı olamamıştır.	Öğrenci yazdığı metinde anlatmak istediklerini net bir şekilde ortaya koymuş ve düşüncesini desteklemek için yeteri kadar gerekçe sunmuştur. Ayrıca fikirlerini yeteri kadar ikna edici cümle ile açıklamaya çalışmıştır. Doğru kelime seçimi ile aktarmak istediği mesajı iletmıştır.	Öğrenci yazdığı metinde ikna gücü kuvvetli kelimelere, deyimlere ve söz öbeklerine yer vermiştir. Güçlü hareket fiilleri, ikna edici zarf ve sıfatlar kullanmıştır. Öğrenci yazdığı metinde anlatmak istediklerini çok net bir şekilde ortaya koymuş ve düşüncelerini desteklemek için ikna edici gerekçe(ler) sunarak karşısındaki kişiyi etkileyici ifade ve anlatımlara yer vermiştir.
Yapı	Giriş, gelişme ve sonuç bölümlerinden en az bir ya da iki tanesini yapmamıştır. Yazarın oluşturduğu paragraflar arasında konu bütünlüğü sağlanamamıştır.	Giriş, gelişme ve sonuç bölümleri arasındaki bağlantıyı zayıf kurduğundan ve/veya bölümler arasındaki geçiş cümlelerini eksik bıraktığından düzenlemelerde bariz eksiklikler vardır. Yazarın oluşturduğu paragraflar arasında konu bütünlüğü kısmen de olsa sağlanmaya çalışılmıştır.	Giriş, gelişme ve sonuç bölümlerini oluşturmuştur ancak genel olarak bağlantılarda eksiklikler vardır. Yazarın oluşturduğu paragraflar arasında konu bütünlüğü bir şekilde sağlanmaya çalışılmıştır.	Giriş, gelişme ve sonuç bölümlerini belirli bir bütünlük içerisinde iyi düzenlemiş ve iyi sunmuştur. Yazarın oluşturduğu paragraflar arasında konu bütünlüğü çok başarılı bir şekilde sağlanmıştır.
Dil ve Anlatım	Metinde açık ve anlaşılır bir dil kullanılmamıştır, anlatım akıcı ve etkili değildir. (Anlatım; benzetme, atasözü, deyim, ifade ve slogan cümleleri vb. ile etkili hâle getirilmemiştir.)	Metnin bazı cümlelerinde anlaşılır bir dil kullanılmış ancak anlatım akıcı ve etkili değildir. (Anlatım; benzetme, atasözü, deyim, ifade ve slogan cümleleri vb. ile etkili hâle getirilmeye çalışılmış, ancak başarılı değildir.)	Metin açık ve anlaşılır bir dille yazılmıştır, anlatım akıcıdır ancak etkili değildir. (Anlatım; benzetme, atasözü, deyim, ifade ve slogan cümleleri vb. ile istenen düzeyde etkili hâle getirilmiştir.)	Metin açık ve anlaşılır bir şekilde yazılmıştır. Yazar açık, anlaşılır bir dil kullanış olup, anlatım akıcı ve etkilidir. (Anlatım; benzetme, atasözü, deyim, ifade ve slogan cümleleri vb. ile tamamen etkili hâle getirilmiştir.)
Yazım ve Noktalama	Metinde yazım ve noktalama hataları ile okuru rahatsız etmekte ve okurun dikkatini dağıtmaktadır. (8 ve üstü)	Metinde okurun açıkça fark edebileceği yazım ve noktalama hataları yapmıştır. (5-7 hata)	Metinde birkaç cümlede yazım ve noktalama hataları yapmıştır. (3-4 hata)	Metinde yazım ve noktalama kurallarına uyulmuştur. Noktalama işaretlerini doğru ve yerinde kullanmıştır. Ayrıca yazıda çok az yazım yanlış ve noktalama hatası vardır. (0-2 hata)
Sayfa Düzeni	El yazısını okunaklı yazmamıştır. Kâğıdın temizlik ve düzenine hiç özen göstermemiştir. Kâğıt buruşuk, lekeli ve kirlidir.	Yazısını kısmen okunaklı bir el yazısı ile yazmıştır. Kullandığı el yazısındaki hatalar dikkat çekmektedir. Kâğıt temizliğine ve düzenine yeterince özen göstermemiş, kâğıdı biraz buruşuk ve lekeli.	Kâğıt kenarlarında ve satır aralarında bırakılması gereken boşluk uygundur. Yazısını okunaklı bir el yazısı ile yazmıştır. Kâğıdın temizlik ve düzenine yeteri kadar özen göstermiştir.	Kâğıt kenarlarında ve satır aralarında bırakılması gereken boşluk uygundur. Yazı okunaklı düzgün el yazısı ile iyi bir şekilde sunulmuştur. Kâğıt temizliğine dikkat etmiş ve özenle yazmıştır.
Not: Öğrencilerin ikna edici metin yazma performansının değerlendirilmesinde size verilen analitik dereceli puanlama anahtarındaki 6 ölçütü dikkate alarak Oldukça Başarısız (1), Orta Düzey (2), İyi (3) ve Oldukça İyi (4) şeklinde puanlamanız gerekmektedir. Her bir öğrenciye ait puanlamalar size verilen öğretmen değerlendirme formuna işlenecektir. Değerlendirmeleriniz için teşekkür eder, iyi çalışmalar dilerim.				

EK 11. Uzman Değerlendirme Formu

İKNA EDİCİ METİN YAZMA GÖREVLERİ İÇİN UZMAN DEĞERLENDİRME FORMU

Değerli Uzman;

İkna edici metin türündeki “*Performansın Değerlendirilmesinde Farklılaşan Puanlayıcı Davranışlarının Çok Yüzeyle Rasch Ölçme Modeli ile İncelenmesi*”ne yönelik olarak yüksek lisans tez çalışmasını yürütmekteyim. Tez çalışması kapsamında ilköğretim düzeyinde uygulanacak olup, ölçme araçlarının (ikna edici metin yazma görevleri ve ikna edici metin yazma görevlerini değerlendirme formu) uzman görüşleri doğrultusunda geliştirilmesi gerekmektedir. Bu amaç doğrultusunda; ilköğretim öğrencilerinin ikna edici metin yazma becerilerinin geliştirilmesi ve puanlanması amacıyla hazırlanan yazma görevleri ile bu görevlerin puanlayıcılar tarafından analitik dereceli puanlama anahtarı yardımıyla değerlendirilmesi için uzman görüş formu geliştirilmiştir. Geliştirilen bu formların değerli uzmanların görüşleri doğrultusunda son şeklini alacak ve hazırlanan araçlarının psikometrik özelliklerinin gelişmesine önemli katkı sağlayacaktır.

Form içerisindeki ölçütlerin performans görevleri ile dereceli puanlama anahtarlarının geliştirilme sürecini dikkate alarak ölçütlerin amaç doğrultusunda ne derece etkili olduğuna dair görüşlerinizi *tamamen uygunsa (4), uygun ise (3), bazı değişiklikler gerekiyorsa (2)* ve ölçüt tamamen veya *önemli bir kısmı değişmesi gerekiyorsa (1)* işaretleyerek görüşlerinizi belirtiniz. Yapmış olduğunuz değerlendirmeler ve öneriler için şimdiden teşekkür ederiz.

Not: Öneriler kısmında kaç numaralı performans görevi için öneride bulunduğunuzu **ÖZELLİKLE** belirtiniz.

Demografik Bilgiler

Doç. Dr. İsmail KARAKAYA
Gazi Üniversitesi
Gazi Eğitim Fakültesi
Eğitim Bilimleri Bölümü
Eğitimde Ölçme ve Değerlendirme ABD

Cennet TOBAŞ
Gazi Üniversitesi
Gazi Eğitim Fakültesi
Eğitim Bilimleri Bölümü
Eğitimde Ölçme ve Değerlendirme ABD

1. Adı ve Soyadı:
2. Unvan:
3. Deneyim :.....
4. Branş
5. Eğitim düzeyi () Lisans () Yüksek lisans () Doktora

UYGUNLUĞUNA BAKILCAK ÖLÇÜTLER	DERECELER			
	1	2	3	4
YAZMA GÖREVİ				
Yazma görevleri temel dil becerilerinin öğretilmesi ve değerlendirilmesi açısından uygundur. Öneri:				
Yazma görevleri öğrencilerin ikna edicilik, iletişim becerilerini artırma, eleştirel düşünme gibi üst düzey zihinsel süreçlerden bir ya da birkaçına uygundur. Öneri:				
Görev kısmında verilen soru öğrencilerin dil düzeyine uygun, anlaşılır ve ilgi çekicidir. Öneri:				
Yönergede yer alan açıklamaların ayrıntısı ve anlatım biçimi öğrencinin sınıf (ilköğretim düzeyi), dil düzeyine ve donanımına uygundur. Öneri:				
Yönerge kısmında öğrencilerin görevi yaparken dikkat etmesi gereken noktaların hepsi verilmiştir. Öneri:				
PUANLAMA ANAHTARI				
Yazma görevinin değerlendirilmesi amacıyla hazırlanan analitik dereceli puanlama anahtarı (DPA) hazırlanırken yeteri kadar göreceli ifade kullanılmıştır. Öneri:				
DPA içerisindeki ölçütler ayrı ayrı tanımlanmış ve her ölçüt öğrencinin kullandığı bir beceri hakkında bilgi vermiştir. Öneri:				

DPA içerisindeki ölçütler beklenen performansların tamamının puanlanması için yeterlidir. Öneri:				
Görev ve puanlama için verilen süreler kişilerin düzeyine uygundur ve verilen görev için yeterlidir. Öneri:				
Tanımlama, görev, yönerge ve puanlanma kısımlarının tamamında gereksiz kavram ve resimler kullanılmamıştır. Öneri:				

İkna edici metin yazma çalışması için beğendiğiniz görevler (en az 5 adet);

.....
.....
.....
.....
.....

EK 12. MEB İzin Belgesi

Evrak Tarih ve Sayısı: 26.02.2020-E.9352



T.C.
ANKARA VALİLİĞİ
Milli Eğitim Müdürlüğü

Sayı : 14588481-605.99-E.4007269
Konu : Araştırma izni

24.02.2020

GAZİ ÜNİVERSİTESİ REKTÖRLÜĞÜNE

- İlgi a) 13/02/2020 tarihli ve 5459 sayılı yazınız.
b) MEB Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü'nün 2017/25 nolu Genelgesi.

Üniversiteniz Eğitim Bilimleri Enstitüsü Yüksek Lisans Öğrencisi Cennet TOBAŞ'ın **"Performansın Değerlendirilmesinde Farklılaşan Puanlayıcı Davranışlarının Çok Yüzeyle Rasch Ölçme Modeli İle İncelenmesi"** konulu tezi kapsamında Keçiören ve Yenimahalle İlçelerine bağlı ortaokullarda uygulanacak olan veri toplama araçları ilgi (b) Genelge çerçevesinde incelenmiştir.

Yapılan inceleme sonucunda, söz konusu araştırmanın Müdürlüğümüzde muhafaza edilen ölçme araçlarının; Türkiye Cumhuriyeti Anayasası, Milli Eğitim Temel Kanunu ile Türk Milli Eğitiminin genel amaçlarına uygun olarak, ilgili yasal düzenlemelerde belirtilen ilke, esas ve amaçlara aykırılık teşkil etmeyecek, eğitim-öğretim faaliyetlerini aksatmayacak şekilde okul ve kurum yöneticilerinin sorumluluğunda gönüllülük esasına göre uygulanması Müdürlüğümüzce uygun görülmüştür.

Bilgilerinizi ve gereğini rica ederim.

Turan AKPINAR
Vali a.
Milli Eğitim Müdürü

Ek:
Uygulama araçları (12 sayfa)
Dağıtım:
Gereği:
Gazi Üniversitesi Rektörlüğü
Bilgi:
Keçiören-Yenimahalle

Adres: Emniyet Mah. Alparslan Türkeş Cad. 4/A Yenimahalle

Elektronik Ağ: ankara.meb.gov.tr
e-posta: istatistik06@meb.gov.tr

Bilgi için: Emine Konuk

Tel: 0 (312) 306 89 30
Faks: 0 () _ _ _

Bu evrak güvenli elektronik imza ile imzalanmıştır. <https://evraksorgu.meb.gov.tr> adresinden 4ec3-2c4e-3994-91c9-13e5 kodu ile teyit edilebilir.

EK 13. Öğrencilerin Yazma Çalışmalarından Örnekler

YAZMA ALANI

Gurukkle Gezmiş

Anneciğim tarihimi de olan ilgimi biliyorsun. Gerçekten bu geziye gitmeyi çok istiyorum çünkü şehitlerimizin bu vatana ne zorluklarla aldığını görmek istiyorum. Biliyorum ki ücret birazcık fazla ama bunun içinde bir şey düşündüm günlük harçlığımdan birazcık kısarsanız böylece ücret sorunu da hallederiz. Gelelim senin ve benim en büyük sıkıntımızda. Biliyorum derslerimden geri kalacağım ama oraya gitmemde bana büyük katkılar sağlayacak, örneğin vatanın ne zorluklarla bidede kaldığını öğrenirim aynı bir farkındalık bilincinde olurum. Ben bir yana genel derslerimden geri kalacağım ama yazıları gitmeyen arkadaşlarımla tamamlaırım. Anladığım pekiştiremediğim konuları kurslarımda tekrar ederim. O da olmasa konu anlatımlı videolar izler, kaynak kitaplarımdan soru çözerim yazıları tekrar ederim. Yani ne yapar ne eder geri kaldığım konuların hepsini öğrenir geziye gitmeyen arkadaşlarımla aynı seviyeye gelirim. Anneciğim lütfen bu geziye gitmeme izin ver. Sevgilerimle kün...

YAZMA ALANI

Sayın Ailem

Baba senin baban seni okutmadı. Sana yön verdi. Okuma ne gelecektir dedi. Anay ben okuycam. Peki sen anne senin baban sana hatırlım dik diye Okula göndermedi şimdi hayatınız nasıl soruyorum sizlere. Ben polis olacağım dediğinde "Siz baba israf ettiniz öğretmen ol" diye. Benim hobim silahlar ancak Siz baba öğretmen ol diyorsun. Herkes kendi sevdiği işi yapar. Siz baba sevme için işi yapıyorsunuz. Ben kendi kırıncıyı kendim çizerim. Herkes sevdiği işe özen gösterir. Ben polisliğe özen gösteririm, bir baskısı öğretmene, bir baskısı itfaiyeye. Önemli olan isteyecek yaptığın şey. Sizde kendi anne ve babanızdan esinlenerek bir bakın doğru bir şey mi yapıyorsunuz. Hobimiz mesleğimizi mesleğimizi mutluluklarımız mutluluklarımız bizim yapımız, oluşturmamız. Çünkü kendi hayatımızı yapayız. Mutlu olalım. Kendi yapımızı oluşturalım. "Noktayı kışkırtma" demişler. Çünkü cümleyi bitiriyor. Siz çok seviyorum. Anne ve baba ellerinizden geçiriyor. Mutlu olmanız destekliyorum.

YAZMA ALANI

Vicut Temizliği

Merhaba Arkadaşlar bilirsiniz ya da bilmiyorsunuz çokça yağın den bir vircis var. Bu vircis Türkiye de olmadığı için çok rahatız ama her an Türkiye de gelebilir. Bu vircisin ülkenize gelmesi için sık sık ellerimizi yıkayalım. Ellerimiz kirlenmeye çok müsait olduğu için onları ilk suyla en az 20 saniye boyunca yıkayalım. Ayrıca banyo ve duşta temizliğimize de çok dikkat etmeliyiz. Ağzımızı 2,3 gün ara ile tuzlu su ile gar gar yapmalıyız. Bu şekilde ağızımızdaki mikropları öldürebiliriz. Ayrıca ağızımızı günde iki kez fırçalamamızda dişlerimizin gülmemesi için çok önemli. Arkadaşlar ayrıca sanımda tek kullanımlık peçete ve ıslak mendil olması gerek bir kez kullandığımız peçete ve ıslak mendil bir daha asla kullanmamalıyız. Arkadaşlar değinmem gereken son konu ise kışırken ve öksürürken ağızımızı kolumuzun dirsek tarafına kapatmalıyız ve ağızımızdaki kuyafeti şifanana kadar bir daha sıkmamalıyız.

YAZMA ALANI

Spor Seçmelerimiz

Anneciğim biliyorsun ki voleybol oynamayı çok seviyorum. Sen de derslerimi aksatacağımı düşündüğün için seçmelere katılmama izin vermiyorsun. Senin de beni anlamanı istiyorum. Voleybolu bir ilim var ve ben de bunu değerlendirmek istiyorum. Derslerimi aksatmayacağım. Zaten geleceğim çok önemli. Telefondaki oynamak yerine voleybol ile ilgilenebilirim. Hem bu sağığım içinde iyi olacak. Derslerime son derece odaklıyım, bundan emin olabilirsiniz. Seni hayal kırıklığına uğratmayacağım, bana güvenebilirsiniz. Beni anlayışla tavsiyelerinizi düşünüyorum. Şimdiden teşekkür ederim.

YAZMA ALANI

Temiz Çevre, Temiz Hayat

Sevgili arkadaşlar, bizler Göplerimizi geri dönüştürüp Dünya'ya tekrar kazandırıp daha temiz bir Dünya elde ederiz. Böylece hem bizim geleceğimiz, hem de kaden sonra geleceklerin geleceği için bir faydamız olur. Hem onlarda bizleri örnek olarak bizim gibi temiz çevre oluşturmak için çöpleri yere atmak yerine ayrı ayrı çöp kutularına atıp geri dönüşüm katkı sağlarken. Çözümler Göpleri yere atıp, sadece çevreyi kirletip ve çevrenin dengesiyle oynuyoruz sanıyoruz. Unutmayın! Lakin çevreye ne kadar gelse aynı zarar bilede getir. Hastalanırız, salgınlardan kurtuluruz. Ama her kez Göpleri çöp kutularına attığını düşünürsek daha fazla bir çevre elde ederiz. Böylece hastalıklardan korunuruz. Bu Dünya'yı koruyarak bizim elimizdedir.

"Bizler çevreyi korursak
Çevre de bizi
korur."

YAZMA ALANI

Efsa'nın Köpeği Olsa

Canım annem ve babam sizcede ailemizin bir köpeği olsa evimize neşe katmaz mı? Bakımı ben yapacağım. Her şeyle ben ilgileneceğim. Sizin bir şey yapmanıza gerek yok ki. Her şeyini ben yapacağım. mama suyu, tuvalet ihtiyacını ben gidericem. Sadece ben okula olduğumda bir kere dışarı çıkartıp gezdirseniz zaten ihtiyaçlarını gidermiş olursunuz. O da sizin levediniz olarak benden bir farkı olmayarak. Bir canlıya yuva olmak istemesi misiniz? Sizi annesi babası beni ablası yerine kalsın. Hem o mutlu olsun. hem de ben mutlu olayım. Evimize neşe katar hem anne, baba. Çok fazla büyük bir cins seçip denmek zorunda değiliz. Ki küçük, tatlı, minnaş bir şey olsa yeter. Maltese Terrier mesela gerçekten hem çok tatlı, hem de maksimum boyu 20 cm'e kadar uzuyorlar yetişkin hali bile 3 Kg. Lütfen, anne lütfen baba. Köpek niyetine değil, evlat niyetine besliyoruz. Hem arkadaşlarımdan çoğunun köpeği veya kedisi var. Çokta zorluğu yokmuş. Konuştum ben onlar ile lütfen...

EK 14. Öğrenci Yüzeyine İlişkin Ölçüm Raporu

Öğrenci Yüzeyine İlişkin Ölçüm Raporu						
Öğrenci No	Ortalamalar		Model		Uygunluk Ölçütü	
	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Ölçütü	Std. Hata	Uyum-ıçi (INFIT)	Uyum-dışı (OUTFIT)
Ö35	3,79	3,84	2,29	0,29	0,74	0,68
Ö27	3,82	3,81	2,13	0,31	0,85	0,83
Ö26	3,79	3,78	1,95	0,29	1,11	0,76
Ö12	3,65	3,72	1,69	0,23	1,00	0,86
Ö21	3,70	3,66	1,48	0,25	0,85	0,86
Ö43	3,71	3,65	1,45	0,25	0,80	0,58
Ö33	3,33	3,54	1,12	0,18	0,71	0,70
Ö41	3,47	3,49	0,98	0,20	1,35	1,56
Ö40	3,36	3,47	0,93	0,19	0,71	0,69
Ö36	3,33	3,42	0,82	0,18	0,61	0,58
Ö23	3,29	3,36	0,69	0,18	0,88	0,83
Ö39	3,38	3,34	0,66	0,19	0,92	0,85
Ö44	3,18	3,33	0,64	0,17	0,93	0,88
Ö37	3,18	3,29	0,55	0,17	0,90	0,85
Ö46	3,24	3,29	0,54	0,18	0,86	0,82
Ö22	3,38	3,28	0,53	0,19	1,01	0,98
Ö15	3,17	3,27	0,52	0,17	1,44	1,41
Ö4	2,94	3,24	0,47	0,16	0,87	0,82
Ö19	3,38	3,23	0,45	0,19	1,27	1,28
Ö24	3,08	3,14	0,28	0,17	1,14	1,06
Ö3	2,79	3,10	0,22	0,16	0,95	0,93
Ö31	2,97	3,10	0,21	0,16	0,86	0,83
Ö42	3,09	3,07	0,16	0,17	1,75	1,64
Ö45	2,85	3,00	0,05	0,16	1,04	0,99
Ö38	2,97	2,87	-0,14	0,16	1,62	1,54
Ö10	2,80	2,86	-0,16	0,16	0,87	0,88
Ö14	2,74	2,83	-0,21	0,16	1,14	1,22
Ö7	2,52	2,76	-0,31	0,15	0,84	0,86
Ö5	2,64	2,75	-0,33	0,15	1,17	1,16
Ö13	2,85	2,73	-0,35	0,16	0,72	0,72
Ö2	2,83	2,67	-0,45	0,16	0,51	0,51
Ö18	2,73	2,66	-0,45	0,16	0,96	1,00
Ö30	2,52	2,62	-0,52	0,15	1,75	1,78
Ö29	2,42	2,52	-0,66	0,15	0,76	0,75
Ö28	2,59	2,40	-0,83	0,15	0,89	0,90
Ö32	2,17	2,38	-0,85	0,15	1,08	1,08
Ö34	2,17	2,38	-0,85	0,15	0,79	0,78
Ö6	2,26	2,34	-0,92	0,15	1,86	1,73
Ö48	2,36	2,33	-0,93	0,15	1,34	1,35
Ö8	2,24	2,25	-1,04	0,15	1,14	1,18
Ö25	2,23	2,22	-1,09	0,15	0,70	0,76
Ö47	2,24	2,20	-1,12	0,15	0,88	0,85
Ö16	2,15	2,19	-1,13	0,16	0,78	0,89
Ö20	1,95	2,04	-1,37	0,16	0,70	0,81
Ö9	2,02	2,02	-1,40	0,16	0,65	0,70
Ö17	1,97	1,87	-1,64	0,16	0,80	0,93
Ö11	1,76	1,75	-1,86	0,17	1,91	1,82
Ö1	1,76	1,59	-2,19	0,17	0,51	0,50
Ortalama	2,85	2,89	0,00	0,18	1,00	0,98
Ss(Evren)	0,58	0,59	1,05	0,04	0,33	0,33
Ss(Örneklem)	0,58	0,59	1,06	0,04	0,34	0,33
Model, Evren: RMSE= 0,18 Ss= 1,03 Ayırma Oranı= 5,67 Ayırma İndeksi= 7,90 Güvenirlilik= 0,97 Model, Örneklem: RMSE= 0,18 Ss= 1,04 Ayırma Oranı= 5,73 Ayırma İndeksi= 7,98 Güvenirlilik= 0,97 Model, Sabit etkili Ki-kare= 1371,0 sd= 47 p= 0,00 Model, Normal Ki-kare= 45,2 sd= 46 p= 0,51						
Not: sd: serbestlik derecesi, ss: standart sapma, RMSE:hata kareleri ortalamasının karekökü						

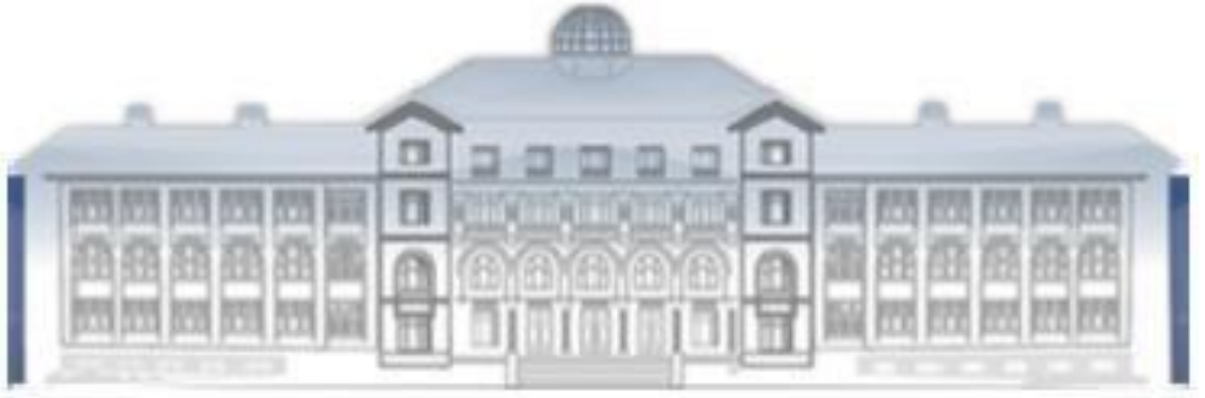
EK 15. “Puanlayıcı x Ölçüt” Etkileşimine Yönelik Ölçüm Raporu

Puanlayıcı	Ölçüt	Gözlenen Puan	Beklenen Puan	Yanlılık	Std. Hata	t-değeri
PÖ1	Yazım ve Noktalama	138	129.42	.35	.20	1.72
PÖ3	Yazım ve Noktalama	126	117.68	.34	.19	1.71
PÖ6	Görüş Sunma	143	134.81	.30	.20	1.60
PU3	Dil ve Anlatım	144	157.16	.28	.23	1.48
PÖ2	Yapı	121	114.77	.23	.19	1.21
PU3	Yapı	137	131.10	.22	.19	1.14
PU1	Sayfa Düzeni	142	137.09	.19	.20	.96
PÖ7	Yapı	106	101.13	.19	.20	.97
PU2	Yapı	144	139.41	.18	.20	.90
PÖ4	Yazım ve Noktalama	144	139.59	.17	.20	.86
PÖ2	Dil ve Anlatım	147	142.98	.16	.20	.80
PU2	Sayfa Düzeni	163	160.19	.14	.23	.62
PÖ3	İçerik	137	133.38	.14	.19	.70
PÖ5	İçerik	151	147.79	.13	.21	.65
PU2	Dil ve Anlatım	166	163.56	.13	.24	.56
PU1	Görüş Sunma	151	147.91	.13	.21	.63
PÖ7	İçerik	135	131.71	.12	.19	.63
PÖ8	Sayfa Düzeni	136	132.85	.12	.19	.61
PÖ8	Yazım ve Noktalama	126	123.19	.10	.19	.54
PÖ5	Görüş Sunma	154	151.77	.10	.21	.46
PÖ3	Dil ve Anlatım	133	131.00	.07	.19	.39
PÖ4	Sayfa Düzeni	150	148.54	.06	.20	.30
PÖ5	Yapı	119	117.60	.05	.19	.27
PU2	İçerik	166	165.23	.04	.24	.18
PU3	Sayfa Düzeni	154	153.39	.03	.21	.13
PU1	Yazım ve Noktalama	128	127.53	.02	.19	.09
PÖ5	Dil ve Anlatım	146	145.60	.02	.20	.08
PÖ8	İçerik	140	139.68	.01	.20	.06
PÖ6	Sayfa Düzeni	124	123.67	.01	.19	.06
PÖ1	Görüş Sunma	149	149.00	.00	.20	.00
PÖ6	Yapı	100	100.19	-.01	.20	-.04
PÖ4	Görüş Sunma	158	158.23	-.01	.22	-.05
PU1	Dil ve Anlatım	141	141.50	-.02	.20	-.10
PÖ4	Dil ve Anlatım	152	152.55	-.02	.21	-.12
PÖ7	Sayfa Düzeni	124	124.69	-.03	.19	-.13
PÖ1	Dil ve Anlatım	142	142.66	-.03	.20	-.13
PÖ6	İçerik	130	130.70	-.03	.19	-.13
PÖ1	İçerik	144	144.90	-.04	.20	-.18
PÖ8	Yapı	108	108.95	-.04	.19	-.18
PÖ8	Görüş Sunma	143	143.96	-.04	.20	-.19
PÖ7	Görüş Sunma	135	136.17	-.04	.19	-.23
PÖ3	Sayfa Düzeni	125	126.39	-.05	.19	-.27
PÖ4	İçerik	153	154.58	-.07	.21	-.33
PÖ5	Yazım ve Noktalama	130	131.92	-.07	.19	-.37
PU3	Görüş Sunma	161	162.45	-.07	.22	-.33
PÖ6	Yazım ve Noktalama	112	113.98	-.07	.19	-.38
PÖ2	İçerik	143	145.22	-.09	.20	-.44
PÖ2	Görüş Sunma	147	149.31	-.10	.20	-.47
PÖ2	Sayfa Düzeni	136	138.61	-.10	.19	-.51
PU1	İçerik	141	143.76	-.11	.20	-.55
PÖ7	Yazım ve Noktalama	112	114.99	-.11	.19	-.57
PÖ2	Yazım ve Noktalama	126	129.11	-.11	.19	-.60
PÖ7	Dil ve Anlatım	126	129.31	-.12	.19	-.64
PÖ1	Sayfa Düzeni	135	138.27	-.12	.19	-.64
PU2	Görüş Sunma	166	168.20	-.12	.24	-.53
PÖ4	Yapı	122	125.50	-.13	.19	-.67
PU3	İçerik	156	159.05	-.14	.21	-.67
PÖ8	Dil ve Anlatım	133	137.36	-.16	.19	-.85
PÖ1	Yapı	110	114.42	-.16	.19	-.85
PÖ3	Görüş Sunma	133	137.82	-.18	.19	-.94
PÖ6	Dil ve Anlatım	123	128.29	-.19	.19	-1.01
PU1	Yapı	108	113.21	-.20	.19	-1.00
PÖ5	Sayfa Düzeni	136	141.31	-.20	.19	-1.05
PU3	Yazım ve Noktalama	137	144.62	-.28	.19	-1.49
PU2	Yazım ve Noktalama	144	151.28	-.25	.20	-1.37
PÖ3	Yapı	94	102.72	-.35	.20	-2.44
Sabit etkili Ki-kare=42.9 sd= 66 p=0.99						

EK 16. “Puanlayıcı x Görev” Etkileşimine Yönelik Ölçüm Raporu

Puanlayıcı	Görev	Gözlenen Puan	Beklenen Puan	Yanlılık	Std. Hata	t-değeri
PÖ6	G6	50	41.93	.78	.31	2.49
PU1	G6	55	47.11	.75	.32	2.34
PU3	G2	137	126.56	.44	.22	2.05
PÖ3	G7	87	82.04	.35	.27	1.30
PÖ5	G4	154	145.82	.32	.20	1.58
PÖ2	G7	94	89.58	.30	.27	1.13
PÖ3	G7	87	82.44	.29	.25	1.13
PU2	G2	140	133.94	.29	.22	1.28
PÖ2	G4	150	143.03	.27	.20	1.33
PÖ6	G5	105	99.96	.27	.23	1.14
PÖ5	G1	148	142.59	.24	.21	1.13
PÖ6	G3	167	159.23	.23	.17	1.34
PÖ3	G5	106	101.88	.22	.23	.94
PÖ8	G1	140	135.11	.21	.21	1.00
PÖ1	G7	92	89.38	.18	.26	.67
PÖ5	G2	118	113.75	.16	.19	.82
PÖ4	G3	196	191.52	.14	.17	.78
PÖ6	G4	131	127.11	.13	.18	.71
PU2	G1	162	159.59	.13	.24	.56
PÖ8	G3	175	170.76	.12	.17	.72
PÖ2	G1	143	140.20	.12	.21	.58
PÖ2	G6	49	47.83	.11	.31	.36
PÖ3	G6	44	43.00	.09	.31	.31
PU3	G5	121	119.82	.08	.27	.31
PU1	G3	179	176.21	.08	.17	.48
PÖ1	G3	180	177.75	.07	.17	.38
PÖ4	G2	123	121.39	.06	.20	.32
PÖ7	G6	43	42.34	.06	.31	.20
PÖ5	G7	92	91.14	.06	.26	.22
PÖ4	G7	96	95.29	.05	.27	.19
PÖ7	G1	129	127.91	.05	.20	.22
PÖ7	G3	162	160.50	.04	.17	.26
PÖ4	G4	154	153.17	.03	.20	.17
PU3	G1	154	153.40	.03	.22	.13
PÖ7	G5	101	100.69	.02	.23	.07
PÖ1	G4	143	142.67	.01	.19	.06
PU2	G7	102	101.92	.01	.29	.02
PÖ7	G2	97	96.91	.00	.19	.02
PÖ8	G2	105	105.04	.00	.19	-.01
PÖ2	G5	110	110.23	-.01	.24	-.06
PÖ8	G7	86	86.24	-.02	.25	-.06
PU1	G4	141	141.44	-.02	.19	-.08
PÖ3	G3	162	162.62	-.02	.17	-.11
PÖ1	G2	110	110.59	-.02	.19	-.11
PÖ3	G2	98	98.58	-.02	.19	-.11
PÖ7	G7	81	81.43	-.03	.25	-.11
PU2	G4	164	164.53	-.03	.23	-.12
PU2	G3	207	208.17	-.04	.18	-.21
PÖ1	G1	139	139.90	-.04	.21	-.19
PU3	G3	197	198.29	-.04	.18	-.23
PÖ4	G1	148	149.04	-.05	.21	-.22
PÖ3	G1	128	129.41	-.06	.20	-.29
PÖ1	G6	47	47.69	-.07	.31	-.21
PÖ7	G4	125	128.22	-.11	.18	-.59
PÖ8	G5	104	106.35	-.13	.23	-.55
PÖ8	G4	133	136.97	-.14	.19	-.74
PÖ2	G2	107	110.94	-.14	.19	-.76
PU3	G7	96	98.05	-.15	.27	-.57
PÖ1	G5	107	110.01	-.17	.23	-.72
PÖ4	G5	114	116.73	-.17	.25	-.69
PU1	G1	134	138.85	-.21	.21	-1.00
PU3	G4	153	157.97	-.21	.20	-1.05
PÖ3	G4	123	130.06	-.24	.18	-1.30
PÖ8	G6	43	45.53	-.24	.31	-.78
PÖ5	G3	173	181.75	-.26	.17	-1.50
PU1	G2	102	109.36	-.27	.19	-1.41
PÖ5	G5	107	112.02	-.29	.23	-1.22
PU1	G7	84	88.70	-.30	.25	-1.20

PU2	G5	120	124.08	-.30	.26	-1.14
PU2	G6	54	56.77	-.30	.32	-.93
PÖ2	G3	167	178.20	-.33	.17	-1.91
PÖ6	G7	75	80.82	-.35	.25	-1.44
PÖ6	G7	75	80.90	-.37	.25	-1.48
PÖ6	G2	86	95.45	-.39	.21	-1.88
PÖ6	G1	118	127.22	-.40	.21	-1.91
PÖ5	G6	44	48.84	-.44	.30	-1.47
Sabit etkili Ki-kare= 71.4 sd=77 p=0.66						
NOT: Görevler G1: Evcil Hayvan Besleme, G2: Eve İnternet Bağlatma, G3: Çanakkale'ye Gezi, G4: Okul Takımı Seçmeleri, G5: Geri Dönüşüm, G6: Kişisel Temizlik, G7: Benim Kararım						



GAZİLİ OLMAK AYRICALIKTIR