
GAZİ ÜNİVERSİTESİ
Gazi
1926

PERFORMANS DEĞERLENDİRME SÜRECİNDE PUANLAYICI EĞİTİMİNİN PUANLAYICI DAVRANIŞLARI ÜZERİNDEKİ ETKİSİNİN İNCELENMESİ

MEHMET ŞATA

DOKTORA TEZİ

EĞİTİM BİLİMLERİ ANABİLİM DALI

GAZİ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

EKİM, 2019

TELİF HAKKI VE TEZ FOTOKOPİ İZİN FORMU

Bu tezin tüm hakları saklıdır. Kaynak göstermek koşuluyla tezin teslim tarihinden itibaren 24 (yirmi dört) ay sonra tezden fotokopi çekilebilir.

YAZARIN

Adı : Mehmet

Soyadı : ŞATA

Bölümü : Eğitimde Ölçme ve Değerlendirme

İmza :

Teslim tarihi :

TEZİN

Türkçe Adı : Performans değerlendirme sürecinde puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisinin incelenmesi

İngilizce Adı : The investigation of the effect of rater training on the rater behaviors in the performance assessment process

ETİK İLKELERE UYGUNLUK BEYANI

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

Yazar Adı Soyadı: Mehmet ŞATA

İmza: ……..

JÜRİ ONAY SAYFASI

Mehmet ŞATA tarafından hazırlanan "Performans Değerlendirme Sürecinde Puanlayıcı Eğitiminin Puanlayıcı Davranışları Üzerindeki Etkisinin İncelenmesi" adlı tez çalışması aşağıdaki jüri tarafından oy birliği ile Gazi Üniversitesi Eğitim Bilimleri Anabilim Dalı'nda Doktora Tezi olarak kabul edilmiştir.

Danışman: Doç. Dr. İsmail KARAKAYA
(Eğitimde Ölçme ve Değerlendirme, Gazi Üniversitesi) ……..

Başkan: Prof. Dr. Ömay ÇOKLUK BÖKEOĞLU
(Eğitimde Ölçme ve Değerlendirme, Ankara Üniversitesi) ……

Üye: Prof. Dr. Mehtap ÇAKAN
(Eğitimde Ölçme ve Değerlendirme, Gazi Üniversitesi) ……

Üye: Doç. Dr. Nilüfer KAHRAMAN
(Eğitimde Ölçme ve Değerlendirme, Gazi Üniversitesi) ……

Üye: Doç. Dr. Celal Deha DOĞAN
(Eğitimde Ölçme ve Değerlendirme, Ankara Üniversitesi) ……

Tez Savunma Tarihi: 04 /10 /2019

Bu tezin Eğitim Bilimleri Anabilim Dalı'nda Doktora tezi olması için şartları yerine getirdiğini onaylıyorum.

Prof. Dr. Selma YEL
Eğitim Bilimleri Enstitüsü Müdürü ……

Benden gidenlere….

TEŞEKKÜR

Zorlu ve yorucu bir süreç içerisinde, kafa karışıklığının çokça yaşandığı bir dönemde yol gösteren, değerli fikir ve düşüncelerini paylaşan, kuramsaldan öte uygulama boyutuna odaklanmamı sağlayan değerli danışman hocam Sayın Doç. Dr. İsmail KARAKAYA'ya çok teşekkür ediyorum.

Bu çalışmanın başından sonuna kadar birçok katkı sağlayan, yeri geldiğinde kritik geri dönütler veren, ne zaman kapılarını çaldığımda her türlü ilgi ve desteklerini esirgemeyen çok değerli hocalarım Sayın Prof. Dr. Ömay ÇOKLUK BÖKEOĞLU ve Sayın Doç. Dr. Nilüfer KAHRAMAN'a çok teşekkür ediyorum.

Savunma jürimde bulanarak yapmış olduğum bu çalışmanın daha nitelikli hale gelmesine katkı sağlayan hocalarım Sayın Prof. Dr. Mehtap ÇAKAN ve Sayın Doç. Dr. Celal Deha DOĞAN'a çok teşekkür ediyorum.

Veri toplama sürecinde ve ölçme aracının hazırlanmasında insanüstü bir çaba ve emeği ile bu çalışmanın bu günlere gelmesinde değerli katkılar sunan arkadaşım Arş. Gör. Fatıma Nur FİŞNE'ye sonsuz teşekkür ediyorum.

Çalıştığınız yerde iyi dostlarınız olduğunda işiniz daha kolay ve daha zevkli olmaktadır. Bana bu kolaylığı ve zevki veren dostlarım ve kardeşlerim Arş. Gör. Ergün Cihat ÇORBACI, Arş. Gör. Fuat ELKONCA, Arş. Gör. Görkem CEYHAN, Arş. Gör. Hikmet ŞEVGİN ve Arş. Gör. Mahmut Sami KOYUNCU'ya çok teşekkür ediyorum.

Anabilim dalı başkanımız Sayın Prof. Dr. Mehtap ÇAKAN başta olmak üzere Gazi Üniversitesi Eğitimde Ölçme ve Değerlendirme anabilim dalı öğretim üyeleri ve elemanlarına çok teşekkür ediyorum.

PERFORMANS DEĞERLENDİRME SÜRECİNDE PUANLAYICI EĞİTİMİNİN PUANLAYICI DAVRANIŞLARI ÜZERİNDEKİ ETKİSİNİN İNCELENMESİ

(Doktora Tezi)

MEHMET ŞATA
GAZİ ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
EKİM 2019

ÖZ

Bu araştırmada, performans değerlendirme sürecinde ölçümlere karışan problemli puanlayıcı davranışlarının belirlenmesi ve puanlayıcı eğitiminin bu problemli davranışlar üzerindeki etkisinin incelenmesi amaçlanmıştır. Araştırma, deney ve kontrol gruplu ön test-son test uygulamalı yarı deneysel desen ile tasarlanmıştır. Araştırmanın çalışma grubu 23 kontrol grubu ve 22 deney grubu olmak üzere toplam 45 puanlayıcıdan oluşmaktadır. Araştırmacı tarafından geliştirilen analitik dereceli puanlama anahtarı, Uluslararası İngilizce Dil Test Sistemi tarafından hazırlanan düşünce temelli yazma görevi ve kişisel bilgi formu kullanılarak araştırmanın verileri toplanmıştır. Araştırma verileri; çok yüzeyli Rasch ölçme modeli, genellenebilirlik kuramı ve bağımsız örneklemler t-testi kullanılarak analiz edilmiştir. Hem çok yüzeyli Rasch hem de genellenebilirlik analizlerinde tamamen çaprazlanmış desen kullanılmıştır. Deneysel işlem olarak puanlayıcı eğitimi verilmiş olup bu eğitim, puanlayıcı hatası eğitimi ve referans çerçevesi eğitimi desenlerinin birleştirilmesi ile oluşturulmuştur. Deney grubuna uygulanan puanlayıcı eğitimi her hafta bir saat olmak üzere toplam dört hafta sürmüştür. Puanlayıcı eğitimi öncesi hem kontrol grubunun hem de deney grubunun problemli puanlayıcı davranışlarının istatistiksel olarak benzer olduğu

belirlenmiştir. Bu araştırmada; puanlayıcı katılığı, puanlayıcı cömertliği, merkeze yönelim, halo, farklılaşan puanlayıcı katılığı ve farklılaşan puanlayıcı cömertliği olmak üzere altı tane problemli puanlayıcı davranışı incelenmiştir. Araştırmanın bulguları incelendiğinde puanlayıcı eğitimi öncesi deney ve kontrol grubunda; puanlayıcı katılığı, puanlayıcı cömertliği, farklılaşan puanlayıcı katılığı ve farklılaşan puanlayıcı cömertliğinin hem grup düzeyinde hem de bireysel düzeyde, halo ve merkeze yönelim davranışlarının ise sadece bireysel düzeyde ortaya çıktığı tespit edilmiştir. Puanlayıcı eğitimi sonrası incelenen problemli puanlayıcı davranışlarının yeniden ortaya çıktığı fakat deney grubundaki bu davranışların ölçümlere karışma düzeylerinin istatistiksel olarak kontrol grubundan görece daha az olduğu belirlenmiştir. Puanlayıcı eğitiminin puanlayıcı katılığı ve puanlayıcı cömertliği davranışı üzerinde büyük bir etkiye sahip olduğu ve katılık ile cömertlik davranışındaki değişimin %15'inin puanlayıcı eğitimi ile açıklanabildiği bulunmuştur. Puanlayıcı eğitiminin puanlayıcı güvenirliği üzerindeki etkisine bakıldığında ise puanlayıcılar arası güvenirliğin arttığı ve bu etkinin orta düzeyde olduğu ayrıca puanlayıcı güvenirliğindeki değişimin %10'unun puanlayıcı eğitimi ile açıklanabildiği tespit edilmiştir. Araştırmanın verileri hem çok yüzeyli Rasch hem de genellenebilirlik kuramı ile analiz edilmiş ve her bir yüzeyden gelen varyans kaynakları incelenmiştir. Elde edilen bulgulara göre her iki analiz yöntemin benzer sonuçlar verdiği bulunmuştur. Araştırmanın ulaştığı sonuçlardan hareketle jüri üyelerinin yer aldığı ve önemli kararların verildiği durumlarda yer alan puanlayıcılara yönelik puanlayıcı eğitiminin verilmesi önerilmektedir. Bu araştırmada akademik yazma becerilerinin değerlendirilmesine yönelik puanlayıcı eğitimi verildiğinden benzer olarak diğer yazma türlerine ve sayısal ders içeriklerine de yönelik puanlayıcı eğitimlerinin verilmesi önerilmektedir.

Anahtar Kelimeler : Puanlayıcı Eğitimi, Problemli Puanlayıcı Davranışları, Geçerlik, Güvenirlik, Çok Yüzeyli Rasch Ölçme Modeli.
Sayfa Adedi : 211
Danışman : Doç. Dr. İsmail KARAKAYA

THE INVESTIGATION OF THE EFFECT OF RATER TRAINING ON THE RATER BEHAVIORS IN THE PERFORMANCE ASSESSMENT PROCESS

(Ph.D Thesis)

MEHMET ŞATA

GAZI UNIVERSITY

GRADUATE SCHOOL OF EDUCATIONAL SCIENCES

OCTOBER, 2019

ABSTRACT

In this study, it is aimed to determine the problematic rater behaviors interfering with the measurements in the performance assessment process and to investigate the effect of rater training on these problematic behaviors. The research was designed with a quasi-experimental design with pre-test and post-test with experimental and control groups. The study group consists of 45 raters, 23 participants in control group and 22 participants in experimental groups. Data were collected by using analytical writing scoring rubric developed by the researcher, opinion-based writing task prepared by International English Language Test System and personal information form. The research data were analyzed by using the many facet Rasch measurement model, generalizability theory and independent samples t-test. Fully-crossed design was used in both many facet Rasch and generalizability analyses. Rater training was provided as an experimental procedure. Besides, this training was prepared by combining rater error training and frame of reference training design. The rater training for the experimental group lasted four weeks, one hour per week. Before the rater training, it was found that problematic rater behaviors of both control and experimental groups were statistically similar. In this study, six problematic rater behaviors, which are

rater stiffness, rater generosity, center orientation, halo, differing rater stiffness, and differing rater generosity, were examined. When the findings of the study were examined, the experimental and control groups before the rater training; rater rigidity, rater generosity, differing rater rigidity and differing rater generosity were determined both at group level and at individual level, and halo and center orientation behaviors appeared only at individual level. The problematic rater behaviors reappeared after rater training. However, it was determined that the level of interference of these behaviors in the experimental group was statistically less than the control group. It was found that rater training had a large effect on rater stiffness and rater generosity behavior, and 15% of the change in stiffness and generosity behavior could be explained by rater training. When the effect of rater training on rater reliability was examined, it was found that reliability between raters increased, and this effect was moderate, and also 10% of the change in rater reliability could be explained by rater training. The data of the research were analyzed with both many facet Rasch and generalizability theory, and the sources of variance from each surface were examined. According to the findings, both analysis methods gave similar results. Based on the results of the research, it is recommended to provide rater training for the raters on the occasions when jury members take part in, and important decisions should be made. In this study, since rater training is provided for the evaluation of academic writing skills, it is suggested that rater trainings for other types of writing and course content should be given in the similar way.

Key Words : Rater Training, Problematic Rater Behaviors, Validity, Reliability, Many Facet Rasch Measurement Model.

Page Number : 211

Supervisor : Associate Professor İsmail KARAKAYA

İÇİNDEKİLER

TELİF HAKKI VE TEZ FOTOKOPİ İZİN FORMU ..... **i**
ETİK İLKELERE UYGUNLUK BEYANI ..... **ii**
JÜRİ ONAY SAYFASI ..... **iii**
TEŞEKKÜR ..... **v**
ÖZ ..... **vi**
ABSTRACT ..... **viii**
İÇİNDEKİLER ..... **x**
TABLOLAR LİSTESİ ..... **xiv**
ŞEKİLLER LİSTESİ ..... **xvi**
SİMGELER VE KISALTMALAR LİSTESİ ..... **xvii**
BÖLÜM 1 ..... **1**
GİRİŞ **1**
Problem İfadesi **7**
Araştırmanın Soruları **7**
Araştırmanın Amacı **8**
Araştırmanın Önemi **8**
Sınırlılıklar **10**
Tanımlar **10**
BÖLÜM 2 ..... **11**
KURAMSAL TEMELLER **11**
Akademik Yazma Becerilerinin Değerlendirilmesi **11**
Performans Değerlendirme **13**
Problemli puanlayıcı Davranışları **15**
Puanlayıcı Katılığı ve Cömertliği **17**
Merkeze Yönelim **20**
Halo Etkisi (Halo Effect) **22**

Farklılaşan Puanlayıcı Katılığı ve Cömertliği (Puanlayıcı Yanlılığı) **23**

Problemli Puanlayıcı Davranışlarını Azaltma veya Kontrol Altına Almaya Yönelik Uygulamalar **24**

Otomatik Puanlama/Değerlendirme **25**

Birden Fazla Puanlayıcının Kullanılması **25**

Dereceli Puanlama Anahtarları **26**

Puanlayıcı Eğitimi **30**

Öz Liderlik Eğitimi ***32***
Davranış Gözlem Eğitimi ***33***
Puanlayıcı Değişkenliği Eğitimi ***34***
Performans Boyutları Eğitimi ***34***
Puanlayıcı Hatası Eğitimi ***35***
Referans Çerçevesi Eğitimi ***36***

Puanlayıcı Güvenirliği **38**

Klasik Test Kuramına Dayalı Puanlayıcı Güvenirliği Yöntemleri **39**

Genellenebilirlik Kuramına Dayalı Puanlayıcı Güvenirliği Yöntemleri **40**

Madde Tepki Kuramına Dayalı Puanlayıcı Güvenirliği Yöntemleri **41**

Çok Yüzeyli Rasch Ölçme Modeli ***43***

İLGİLİ ARAŞTIRMALAR **53**

Puanlayıcı Davranışları ile İlgili Araştırmalar **53**
Puanlayıcı Eğitimi ile İlgili Araştırmalar **64**

BÖLÜM 3 ..... **73**
YÖNTEM **73**

Araştırmanın Deseni **73**
Çalışma Grubu **74**
Veri Toplama Araçları **74**

Yazma Görevi **74**

Analitik Dereceli Puanlama Anahtarı **75**

Kişisel Bilgi Formu **84**

Deneysel İşlem **84**

Puanlayıcı Eğitimi 85

Veri Analizi 87

BÖLÜM 4 91

BULGULAR ve YORUMLAR 91

Ölçüm Raporlarına Ait Uygunluk ve Model Ölçümleri 91

Puanlayıcı Eğitimi Öncesi Problemli Puanlayıcı Davranışların Ölçümlere Karışma Düzeyleri 104

Puanlayıcı Eğitimi Öncesi Deney ve Kontrol Grubundaki Puanlayıcıların Katılık ve Cömertlik Davranışları 104

Puanlayıcı Eğitimi Öncesi Deney ve Kontrol Grubundaki Puanlayıcıların Merkeze Yönelim Davranışları 106

Puanlayıcı Eğitimi Öncesi Deney ve Kontrol Grubundaki Puanlayıcıların Halo Davranışları 109

Puanlayıcı Eğitimi Öncesi Deney ve Kontrol Grubundaki Puanlayıcıların Farklılaşan Puanlayıcı Katılığı ve Cömertliği (Yanlılık) Davranışları 113

Puanlayıcı Eğitimi Sonrası Problemli Puanlayıcı Davranışların Ölçümlere Karışma Düzeyleri 115

Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubundaki Puanlayıcıların Katılık ve Cömertlik Davranışları 116

Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubundaki Puanlayıcıların Merkeze Yönelim Davranışları 119

Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubundaki Puanlayıcıların Halo Davranışları 123

Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubundaki Puanlayıcıların Farklılaşan Puanlayıcı Katılığı ve Cömertliği (Yanlılık) Davranışları 127

Puanlayıcı Eğitiminin Puanlayıcı Güvenirliği Üzerindeki Etkisi 133

Çok Yüzeyli Rasch Ölçme Modelinin ve Genellenebilirlik Kuramının Analiz Sonuçlarının Karşılaştırılması 137

BÖLÜM 5 141

SONUÇLAR **141**

Öneriler **144**

Araştırmacılara Yönelik Öneriler **144**

Uygulayıcılara Yönelik Öneriler **145**

KAYNAKLAR **147**

EKLER **173**

EK 1. Gazi Üniversitesi Etik Komisyonu İzin Belgesi **173**

EK 2. Puanlayıcı Eğitiminin Genel Yapısı **174**

EK 3. Kişisel Bilgi Formu **175**

EK 4. Analitik Dereceli Puanlama Anahtarı (Akademik Yazma Becerileri İçin) **176**

EK 5. Akademik Yazma Görevi **178**

EK 6. Kontrol Grubunun Ön Testi İçin *Puanlayıcı x Birey* Etkileşim Analizi Sonucunda Yanlı Olan Etkileşimler **179**

EK 7. Deney Grubunun Ön Testi İçin Puanlayıcı x Birey Etkileşim Analizi Sonucunda Yanlı Olan Etkileşimler **183**

EK 8. Kontrol Grubunun Son Testi İçin Puanlayıcı x Birey Etkileşim Analizi Sonucunda Yanlı Olan Etkileşimler **186**

EK 9. Deney Grubunun Son Testi İçin *Puanlayıcı x Birey* Etkileşim Analizi Sonucunda Yanlı Olan Etkileşimler **189**

EK 10. Analitik Dereceli Puanlama Anahtarının Türkçe Formu **191**

EK 11. ÇYRÖM Analizi İçin FACETS Sintaksı **192**

TABLOLAR LİSTESİ

Tablo 1. *Problemli Puanlayıcı Davranışlarının Listesi*.....16

Tablo 2. *KGO Değerlerine Göre Kriter Seçiminde Eşik Değerler*.....77

Tablo 3. *Ölçme Aracının Taslak Formunda Bulunan Kriterlere İlişkin Hesaplanan KGO Değerleri*.....78

Tablo 4. *Ölçme Aracına Dâhil Edilen Kriterler ve Puanlanması*.....79

Tablo 5. *Kriterlere İlişkin Tek Değişkenli Normallik Testi*.....82

Tablo 6. *Genelleştirilmiş Kısmi Puan Modeli ile Kestirilen Marjinal Uyum χ^2 ve Standartlaştırılmış LD χ^2 Değerleri*.....88

Tablo 7. *Öğrenci Kompozisyonları Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu*.....92

Tablo 8. *Kriter Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu*.....94

Tablo 9. *Puanlayıcı Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu*.....96

Tablo 10. *Öğrenci Kompozisyonları Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu*.....98

Tablo 11. *Kriter Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu*.....100

Tablo 12. *Puanlayıcı Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu*.....102

Tablo 13. *Deney ve Kontrol Gruplarının Ön Test Puanlamaları Arasındaki Farkla İlgili Bağımsız Örneklemler t-testi Sonucu*.....106

Tablo 14. *Deney ve Kontrol Gruplarının Ön Testi İçin Hesaplanan Kategori İstatistikleri*.....106

Tablo 15. *Deney ve Kontrol Grubundaki Puanlayıcıların Ön Testine İlişkin Her Bir Puanlayıcı İçin Hesaplanan Kategori İstatistikleri* 108
Tablo 16. *Kriter Güçlüğü Eşitlendikten Sonra Puanlayıcı Yüzeyinin Ön Test Ölçümleri için Kestirilen Ölçüm Raporu* 111
Tablo 17. *Deney ve Kontrol Grubunun Son Test Puanlamaları Arasındaki Logit Ölçülerinin Farklarına İlişkin Bağımsız Örneklemler t-testi Sonucu* 118
Tablo 18. *Deney ve Kontrol Gruplarının Son Testi İçin Hesaplanan Kategori İstatistikleri* 119
Tablo 19. *"Introduction-Body-Conclusion" Kriteri İçin Hesaplanan Kategori İstatistikleri* 121
Tablo 20. *Deney ve Kontrol Grubundaki Her Bir Puanlayıcı İçin Kategori İstatistiği* 122
Tablo 21. *Kriter Güçlüğü Eşitlendikten Sonra Son Test Puanlarına İlişkin Puanlayıcı Yüzeyinin Ölçüm Raporu* 125
Tablo 22. *Deney Grubundaki Puanlayıcılara İlişkin Yanlı Olan Puanlayıcı x Öğrenci Kompozisyonu Etkileşimleri* 129
Tablo 23. *Deney ve Kontrol Grubundaki Puanlayıcıların Ön Testlerinden Elde Edilen Puanlayıcı Güvenirliklerine İlişkin Bağımsız Örneklemler t-testi* 134
Tablo 24. *Deney ve Kontrol Grubundaki Puanlayıcıların Son Testlerinden Elde Edilen Puanlayıcı Güvenirliklerine İlişkin Bağımsız Örneklemler t-testi* 136
Tablo 25. *Krippendorff Alfa ve Rasch-Kappa Güvenirlik Katsayılarının Karşılaştırılması* 136
Tablo 26. *Deney ve Kontrol Grubundaki Puanlayıcıların Ön Testlerine İlişkin G-Kuramı ve Çok Yüzeyli Rasch Analizlerinin Karşılaştırılması* 138
Tablo 27. *Deney ve Kontrol Grubundaki Puanlayıcıların Son Testlerine İlişkin G-Kuramı ve Çok Yüzeyli Rasch Analizlerinin Karşılaştırılması* 139

ŞEKİLLER LİSTESİ

Şekil 1. Analitik dereceli puanlama anahtarına ait saçılma grafiği. 83

Şekil 2. Deney ve kontrol grubunun ön testlerinden elde edilen t-değerleri. 105

Şekil 3. Deney ve kontrol grubunun son testlerinden elde edilen t-değerleri 117

Şekil 4. 22 numaralı puanlayıcı için *puanlayıcı x öğrenci kompozisyonu* etkileşimleri. ... 130

Şekil 5. 11 numaralı puanlayıcı için *puanlayıcı x öğrenci kompozisyonu* etkileşimleri. ... 131

Şekil 6. Altı numaralı puanlayıcı için *puanlayıcı x öğrenci kompozisyonu* etkileşimleri .. 131

Şekil 7. Deney grubundaki tüm puanlayıcılar için *puanlayıcı x öğrenci kompozisyonu* etkileşimleri..... 132

Şekil 8. Deney ve kontrol grubundaki puanlayıcıların ön test ölçümlerine ilişkin puanlayıcı güvenirlikleri. 134

Şekil 9. Deney ve Kontrol grubundaki puanlayıcıların son test ölçümlerine ait puanlayıcı güvenirlikleri. 135

SİMGELER VE KISALTMALAR LİSTESİ

ADPA	Analitik Dereceli Puanlama Anahtarı
AFA	Açımlayıcı Faktör Analizi
ANOVA	Varyans Analizi
ÇYRÖM	Çok Yüzeyli Rasch Ölçme Modeli
GK	Genellenebilirlik Kuramı
IELTS	International English Language Testing System
KTK	Klasik Test Kuramı
MEB	Milli Eğitim Bakanlığı
MTK	Madde Tepki Kuramı
PISA	Programme for International Student Assessment
TOEFL	Test of English as a Foreign Language
χ^2	Ki-kare İstatistiği

BÖLÜM 1

GİRİŞ

İçinde bulunduğumuz yüzyılda, öğrencilerin ne kadar bilgi öğrendiklerinden ziyade, öğrendiği bilgiyi gerçek yaşam durumlarında kullanmaları daha önemli hale gelmiştir. Bu nedenle eğitim ve öğretimde; problem çözen, kendini tanıyan, eleştirel düşünen, empati kuran ve özgün bilgiler ortaya koyan bireyleri yetiştirmek kaçınılmaz hale gelmiştir. Genel anlamda bu becerilere 21.yy becerileri adı verilmiş ve okullar bu becerileri öğrencilerine kazandırmak amacıyla yapı değişikliğine gitmiştir. Eğitim kurumlarının önceki yapıları incelendiğinde daha çok alt düzey zihinsel becerilerin ölçülmesine odaklandığı görülmektedir. Yeni öğretim programlarının bu becerileri ölçmeye yönelik hazırlanmasından dolayı daha çok üst düzey zihinsel becerilerin ölçülmesine odaklandığı görülmektedir.

Zihinsel beceriler en genel anlamda alt ve üst düzey zihinsel beceriler olmak üzere ikiye ayrılmaktadır. Alt düzey zihinsel beceriler; bireyden bireye değişmeyen, genellikle hatırlama ve kavrama davranışlarını içeren, geleneksel ölçme araçları (çoktan seçmeli, doğru yanlış, eşleştirme vb.) ile ölçülen, kısa zamanda gelişim gösteren ve süreçten çok sonuca odaklı olan davranışları kapsamaktadır. Üst düzey zihinsel becerilerin özellikleri ise farklıdır. Üst düzey zihinsel beceriler; bireyden bireye değişen, süreç odaklı, daha çok tamamlayıcı ölçme araçları (kompozisyon, portfolyo, performans görevi gibi) ile ölçülen ve

kazanılması uzun zaman alan becerilerdir. Üst düzey zihinsel beceriler, bireyin yeteneğini sergilediğinde kullanmış olduğu zihinsel, duyuşsal ve psikomotor özelliklerinin bir bütünüdür. Alanyazın incelendiğinde üst düzey zihinsel beceriler için; üst düzey zihinsel özellikler, üst düzey düşünme gibi kavramların kullanıldığı görülmektedir (Kutlu, Doğan & Karakaya, 2014, s.15). Haladyna'ya (1997) göre üst düzey zihinsel beceriler başarı gelişiminin bir göstergesi olduğundan geçerli ve güvenilir bir şekilde ölçülmeleri oldukça önemlidir.

İçinde bulunduğumuz yüzyılda bireyin, bilgi edinmesinin ötesinde karar verme, eleştirel düşünme, problem çözme gibi üst düzey zihinsel becerilerini geliştirmiş olması bir zorunluluk haline gelmiştir (Barak, Ben-Chaim, & Zoller, 2007). Bireyin bu becerileri kazanması ya da geliştirmesi öğretim programlarının odak noktası haline gelmiştir (Boddy, Watson & Aubusson, 2003; Watts, Jofili & Bezerra, 1997). Üst düzey zihinsel becerilerin ölçülmesi denildiğinde akla, OECD tarafından üç yılda bir yapılan geniş ölçekli ölçme ve değerlendirme uygulamalarından biri olan Uluslararası Öğrenci Değerlendirme Programı (The Programme for International Student Assessment-PISA) sınavı gelmektedir. Katılan ülkelerin eğitim programlarının durum belirlemesini yapan PISA, ayrıca katılan ülkelerdeki öğrencilerin üst düzey zihinsel becerilerinin gelişip gelişmediğini de belirleyen önemli bir geniş ölçekli ölçme ve değerlendirme uygulamasıdır. PISA 2015 sonuçları incelendiğinde, üst düzey zihinsel becerileri gelişen (5. Düzey ve üstü) öğrenci oranı OECD ülkelerinde %10,70, tüm ülkelerde %8,20 ve Türkiye'de ise %2,01 olduğu bulunmuştur (Taş, Arıcı, Ozarkan, & Özgürlük, 2016, s.33). Bu durum, Türkiye'de üst düzey zihinsel becerileri gelişmiş öğrenci sayısının çok düşük bir orana karşılık geldiğini göstermektedir. Milli Eğitim Bakanlığı (MEB) tarafından hazırlanan öğretim programlarında, öğrencilerin kazanması gereken temel ve üst düzey becerilerin kazanımlarla ilişkilendirildiği vurgulanmıştır (MEB, 2017, s.5). Fakat yapılan çalışmalara göre öğretim programlarının hedeflediği gibi öğrencilerde üst düzey düşünme becerilerini geliştiremediğini ortaya koyan çeşitli çalışmalar bulunmaktadır (Eke, 2018; Söylemez, 2018; Yaz & Kurnaz, 2017).

Günümüz dünyasında üst düzey zihinsel beceriler önemli bir konuma sahip olduğundan birçok ülke eğitim programlarında üst düzey becerilerinin gelişimine katkı sağlayacak bir revizyona gitmiştir. Özellikle uluslararası geniş ölçekli ölçme ve değerlendirme (PISA gibi) faaliyetlerinin önemli bir kısmında üst düzey zihinsel becerilerinin yer aldığı düşünüldüğünde, ülkelerin başarılı sayılabilmeleri için öğrencilerinin üst düzey zihinsel becerileri edinmiş olmaları gerekmektedir. Dolaysıyla üst düzey zihinsel becerilerin geçerli ve güvenilir bir biçimde ölçülmesi önemli bir uygulama problemi olmaktadır. Üst düzey zihinsel becerilerin ölçümünde çoktan seçmeli maddelerin kullanımının uygun olmadığı, bu tür madde formatlarının daha çok bilgi ve alt düzey zihinsel becerileri ölçmek için uygun olduğu vurgulanmıştır (Ebel, 1965; Kutlu vd., 2014, s.27). Bu bağlamda üst düzey zihinsel becerilerin ölçülmesi için yeni madde formatlarının kullanılma gerekliliği ortaya çıkmaktadır.

Bir öğretim programında hedeflenen öğrenme çıktılarına ulaşmak için uygun madde formatının kullanılması önem taşımaktadır (Reiner, Bothell & Sudweeks, 2002, s.14). Hedeflenen öğrenme çıktısı üst düzey zihinsel beceriyi içerdiği bir durumda çoktan seçmeli bir maddenin kullanılması uygun olmayabilir. Bunun için açık uçlu madde formatı (kompozisyon, performans görevi gibi) daha uygun olabilir. Çünkü hedeflenen öğrenme çıktısının birey tarafından elde edilme derecesinin (bireyin öğrenmesinin durum belirlenmesinin yapılması) doğru bir şekilde temsil edilmesi, kullanılmış olan test maddesinin (sınav sorusunun) amaca uygun olması ile sağlanabilir (Reiner vd., 2002, s.15). MEB tarafından hazırlanan öğretim programları incelendiğinde yeterlikler ve beceriler başlığı altında; problem çözme, eleştirel düşünme ve yaratıcılık gibi üst düzey zihinsel becerilerin öğrencilere kazandırılmasının amaçlandığı görülmektedir. Problem çözme, eleştirel düşünme ve yaratıcılık gibi üst düzey zihinsel becerilerin ölçülmesinde geleneksel başarı testlerinin (çoktan seçmeli, doğru-yanlış, kısa cevaplı vb.) kullanımı hem zor olmakta hem de uygunluğu tartışmalı görünmektedir. Bundan dolayı ölçme ve değerlendirme uygulamalarında yeni madde formatlarının kullanılması daha uygun olacaktır. Öğrencinin

cevabını kendisinin yapılandırdığı madde formatları (performans görevi gibi) öğrencilere otantik (gerçek yaşamla ilişkin) değerlendirmeler yapma fırsatı sunduğundan eğitim öğretim faaliyetlerine de katkı sağlamaktadır (Messick, 1994).

Üst düzey zihinsel becerilerin tutarlı ve doğru bir şekilde ölçülmesi ve değerlendirilmesinde, performans değerlendirmenin kullanılması daha uygun görünmektedir (Johnson, Penny & Gordon, 2008). En genel anlamda performans değerlendirme, bireylerin yaptıklarının ve eksiklerinin gözlemlenmesi ve bu gözleme dayanarak güçlü ve eksik taraflarının belirlenmesi, güçlü yönlerinin geliştirilmesi, eksik yönlerinin ise belirlenip iyileştirilme çalışmasıdır (Bennet, 1998, s.232). Performans değerlendirmeyi geleneksel değerlendirmeden ayıran birkaç özellik bulunmaktadır. Performans değerlendirme; gerçek yaşamdan kesitlere dayalı, üründen daha çok sürece odaklanan, bireyin iyi ve kötü becerilerini belirleme ve bireyi daha fazla düşünme ve problem çözmeye sevk etme gibi özelliklere sahipken, geleneksel değerlendirmede bu özellikler bulunmamaktadır (Brown & Hudson, 1998; Khattri, Kane & Reeve, 1995; Moore, 2009).

Performans değerlendirme ile ilgili en önemli kaygının, birey performansının puanlanması ve durum belirlemesinin yapılması sürecinde objektiflik olduğu ifade edilebilir. Çünkü performans değerlendirmenin, geleneksel değerlendirme (sabit tepkili değerlendirme) gibi objektif (nesnel) bir şekilde puanlaması çok zordur (Romagnano, 2001). Alanyazında performans değerlendirmede nesnelliği sağlamak için birçok yöntem önerilmiştir. Bu yöntemler; otomatik puanlama (Attali, Bridgeman & Trapani, 2010; Burstein vd., 1998; Landauer, Laham & Foltz, 2003), birden fazla puanlayıcının kullanılması (Gronlund, 1977, s.85; Kubiszyn & Borich, 2013, s.170), dereceli puanlama anahtarlarının kullanılması (Dunbar, Brooks & Miller, 2006; Ebel & Frisbie, 1991, s. 194; Kutlu vd., 2014, s.51; Oosterhof, 2003, s.81) ve puanlayıcı eğitimi (Bernardin & Buckley, 1981; Haladyna, 1997, s.143; İlhan & Çetin, 2014; Lumley & McNamara, 1995) olarak sıralanabilir. Önerilmiş olan bu yöntemlerin her birinin avantaj ve dezavantajları, birbirlerine göre üstün veya zayıf yönleri mevcuttur. Haladyna (1997, s.137) kullanılan yöntem fark etmeksizin puanlayıcılar

arasında bir tutarlığın sağlanmasının zor olduğunu vurgulamaktadır. Başka bir değişle kullanılan yöntemden bağımsız olarak performans değerlendirmede birey performansının dışındaki diğer bazı dışsal değişkenlerin her zaman için ölçümleri etkileme (karışma) olasılığı bulunmaktadır. Birey performansının değerlendirilme sürecinde meydana gelen bu tutarsızlıklar "puanlayıcı etkileri/davranışları/yanlılığı" olarak tanımlanmaktadır (Farrokhi, Esfandiari & Vaez Dalili, 2011; Haladyna, 1997, s.139; İlhan, 2015, s.3). İlgili araştırmada bu tutarsızlıklar problemli puanlayıcı davranışı olarak adlandırılmıştır.

Birey performansının puanlanması/değerlendirmesi sürecinde puanlayıcı davranışlarından bir veya daha fazlasının ortaya çıkması durumunda, öğrencilerin yetenek düzeylerine ilişkin kestirimlerin hata miktarları yüksek olacaktır. Başka bir değişle elde edilen kestirimler güvenilir ölçümler olmayacaktır. Birey performansının puanlanması sürecinde, ortaya çıkan problemli puanlayıcı davranışları geçerlik üzerinde de olumsuz etkilere sahiptir. Puanlayıcı davranışları, yapıyla ilişkisiz varyansa atfedildiğinden doğrudan bir geçerlik tehditi oluşturmaktadır (Abu Kassim, 2011; Brennan, Gao & Colton, 1995; Congdon & McQueen, 2000; Farrokhi vd., 2011). Puanlayıcılar kullanılarak yapılan bütün durum belirleme çalışmaları için temel geçerlik kanıtı puanlayıcılar arası tutarlılıktır (Downing, 2005). Bundan dolayı puanlanma sürecindeki problemli puanlayıcı davranışlarının ölçümlere karışma düzeylerinin minimum seviyeye getirilmesi veya kontrol altına alınması önem arz etmektedir (Kim, 2009; Linacre, 1994).

Performans değerlendirme, üst düzey zihinsel becerilerin ölçümünde önemli bir avantaj sağlamakla birlikte, birey performansının puanlanmasında nesnellik sorunu önemli bir uygulama problemi olarak ortaya çıkmaktadır. Puanlamadaki nesnelliğin önemli kaynaklarından biri problemli puanlayıcı davranışlarıdır. Bu yüzden problemli puanlayıcı davranışlarının ölçümlere karışma düzeylerinin minimum veya kontrol edilebilir seviyeye getirilmesi geçerlik ve güvenirlik bağlamında önem taşımaktadır. Bunun için alanyazında çeşitli yöntem ve teknikler önerilmiştir. Bu önerilerden biri de puanlayıcı eğitimidir (Bernardin & Buckley, 1981; Feldman, Lazzara, Vanderbilt & DiazGranados, 2012;

Haladyna, 1997, s.143; Hauenstein, & McCusker, 2017; Stamoulis & Hauenstein, 1993; Weigle, 1998; Zedeck & Cascio, 1982). Puanlayıcı eğitimi, ölçümlere karışan problemli puanlayıcı varyansını azalmak için yaygın olarak kullanılmaktadır (Brijmohan, 2016). Alanyazında puanlayıcı eğitimi ile ilgili birçok yöntem/desen önerilmektedir. Bu çalışmada puanlayıcıların eğitiminde, puanlayıcı hatası eğitimi ve referans çerçevesi eğitimi desenleri birleştirilerek kullanılmıştır.

Puanlayıcı eğitiminin asıl amacı, puanlayıcıların öğrenci performansına ve durum belirleme ölçütlerine yönelik ortak bir anlayış geliştirmelerini sağlamaktır (Eckes, 2008; Shale, 1996). Başka bir değişle, birey performansının geçerli ve güvenilir bir şekilde değerlendirilmesinin yapılmasını sağlamaktır (Moser, Kemter, Wachsmann, Köver & Soucek, 2016). Öğrencilerin açık uçlu bir sınavdan aldıkları puanlar, hem öğrencinin performansından hem de puanlayıcının öğrenci performansını yorumlamasından oluştuğu için test sonuçlarında sabit bir geçerlik kaygısı oluşturmaktadır (Ellis, Johnson & Papajohn, 2002; McNamara, 1996, s.117). Test sonuçlarına dayalı alınan kararlar önemli olduğunda, problemli puanlayıcı davranışları belirlenmeli ve bu davranışlar kabul edilebilir bir seviyeye indirilmelidir (Ellis vd., 2002).

Alanyazın incelendiğinde, birey performansının birden fazla puanlayıcı tarafından değerlendirilmesi sonucunda birey performansı ile ilgisi olmayan puanlayıcı kaynaklı bir varyansın ölçümlere karıştığı ve bunun sonucunda ölçümlerin geçerliği ve güvenirliğinin olumsuz yönde etkilendiği görülmektedir. Bu bağlamda mevcut çalışmada alanyazında rapor edildiği gibi en fazla ölçümlere karışan problemli puanlayıcı davranışlarından katılık, cömertlik, halo, merkeze yönelim, farklılaşan katılık ve farklılaşan cömertlik davranışları üzerinde puanlayıcı eğitiminin (referans çerçevesi eğitimi ve puanlayıcı hatası eğitimi birleştirilerek) etkisinin ne olduğu bu araştırmanın problemini oluşturmaktadır.

Problem İfadesi

Puanlayıcı eğitiminin, performans değerlendirme sürecinde ortaya çıkan problemli puanlayıcı davranışları üzerindeki etkisinin incelenmesi.

Araştırmanın Soruları

Araştırmanın problem ifadesi doğrultusunda aşağıdaki sorulara yanıt aranmıştır:

1. Öğrenci kompozisyonu, kriter ve puanlayıcı yüzeylerinin ön ve son test ölçümleri için elde edilen ölçüm raporlarının uygunluk değerleri ve model ölçümleri nasıldır?
2. Puanlayıcı eğitimi öncesi birey performansının durum belirlemesi sürecinde deney ve kontrol grubundaki puanlayıcıların puanlamalarına ilişkin;

a. Katılık ve cömertlik durumları nedir?

b. Merkeze yönelim (eğilim) durumları nedir?

c. Halo durumları nedir?

d. Farklılaşan puanlayıcı katılığı ve cömertliği (yanlılık) durumları nedir?

3. Puanlayıcı eğitimi sonrası birey performansının durum belirlemesi sürecinde deney ve kontrol gruplarının;

a. Puanlayıcı katılığı ve cömertliği davranışındaki değişim miktarları arasındaki fark anlamlı mıdır?

b. Merkeze yönelim davranışındaki değişim miktarları arasındaki fark anlamlı mıdır?

c. Halo davranışındaki değişim miktarları arasındaki fark anlamlı mıdır?

d. Farklılaşan puanlayıcı katılığı ve cömertliği (yanlılık) davranışındaki değişim miktarları arasındaki fark anlamlı mıdır?

4. Deney ve kontrol grubundaki puanlayıcıların puanlayıcı eğitimi öncesi ve sonrası puanlayıcı güvenirlikleri nedir ve anlamlı bir şekilde farklılaşmakta mıdır?
5. Çok yüzeyli Rasch ölçme modeli ile genellenebilirlik kuramı analiz sonuçları benzerlik göstermekte midir?

Araştırmanın Amacı

Bu araştırma, puanlayıcı eğitiminin birey performansının değerlendirilmesi sürecinde ortaya çıkan problemli puanlayıcı davranışları üzerindeki etkisini inceleme amacını taşımaktadır. Bu amaç doğrultusunda, alanyazında önerilen puanlayıcı eğitimi desenlerinden puanlayıcı hatası eğitimi ve referans çerçevesi eğitimi birleştirilerek bir eğitim modülü geliştirilmiştir. Araştırmanın alt amaçlarından biri öğrencilerin akademik yazma becerilerinin değerlendirmesini yapmak olduğundan, bu beceriyi değerlendirmek için analitik dereceli puanlama anahtarı geliştirilmiştir.

Araştırmanın Önemi

Ulusal alanyazın incelendiğinde problemli puanlayıcı davranışlarına yönelik çalışmaların yapıldığı (Atılgan, 2005; Çetin & İlhan, 2017; Karakaya, 2015; Ok, Bilgiç & Bilgiç, 2013) buna karşılık problemli puanlayıcı davranışlarına yönelik herhangi bir puanlayıcı eğitimine ve puanlayıcı eğitimlerinin etkililiğine ilişkin herhangi bir çalışmaya rastlanmadığı görülmektedir. Birey performansının değerlendirilmesi sürecinde problemli puanlayıcı davranışlarının hem geçerlik hem de güvenirlik üzerindeki etkileri dikkate alındığında, bu etkilerin azaltılmasına yönelik önlemlerin alınması önem arz etmektedir.

Hem ulusal hem de uluslararası geniş ölçekli ölçme ve değerlendirme uygulamalarının üst düzey zihinsel becerileri ölçmenin önemine yönelik atıflarda bulunduğu göz önüne alındığında, öğrencilerin ölçme ve değerlendirme uygulamalarında üst düzey zihinsel becerileri kazanmaları son derece önemli görülmektedir. Üst düzey zihinsel becerilerin ölçülmesi için geleneksel değerlendirmeden (çoktan seçmeli gibi) çok performans değerlendirmenin (portfolyo, performans görevi gibi) daha uygun olduğu alanyazında sıklıkla dile getirilmiştir. Performans değerlendirmenin uygulamadaki önemli bir problemi olan puanlama sürecinde nesnelliği sağlamanın zor olduğu düşünüldüğünde, bu tür değerlendirmelerde nesnelliği artıracak yöntem ve tekniklerin etkililiklerinin incelenmesi,

ölçümlerin güvenirliği ve ölçüm sonuçlarına dayalı olarak yapılan çıkarımların geçerliği için önem taşımaktadır.

Uluslararası alanyazın incelendiğinde eğitim bilimleri alanında birey performansının durum belirleme sürecinde problemli puanlayıcı davranış düzeylerini azaltmak veya kontrol altına almak için birçok puanlayıcı eğitimi yöntemine başvurulduğu görülmektedir. Ulusal alanyazında eğitim bilimlerinde kuramsal bir derleme çalışması (İlhan & Çetin, 2014) ve bireyden daha çok kurumların performansına yönelik 360 derece geri bildirim yöntemi dışında herhangi bir puanlayıcı eğitiminin olmadığı görülmektedir. Mevcut çalışma ile ulusal alanyazına puanlayıcı eğitiminin kuramsal ve uygulamalı olarak kazandırılması ve ulusal alanyazına katkı sağlaması önemli görülmektedir.

Problemli puanlayıcı davranışlarının, birey performansının durum belirlemesi aşamasında hem geçerliği hem de güvenirliği olumsuz yönde etkilediği dikkate alındığında, bu davranışların ölçümlere karışma düzeylerinin azaltılması veya kontrol altına alınması önemli görülmektedir. Bu bağlamda ilgili çalışmanın ulusal alanyazına bu yönüyle de katkı sağlayacağı ve birey performansının durum belirleme sürecinde daha geçerli ve güvenilir ölçümler sağlayacağı düşünülmektedir.

Birçok alanda (eğitim, güzel sanatlar, tıp vb.) çoğu zaman birey performansı, jüri veya hakem heyeti tarafından değerlendirilmektedir. Türkiye'de güzel sanatlar alanında yapılan özel yetenek sınavlarında en büyük sorunun uygulanan puan sistemi (yapılan değerlendirme) olduğu belirtilmiştir. Hem güzel sanatlar hem de spor bilimlerinin yetenek sınavları her üniversitenin kendi belirlediği değerlendirme kriterleri doğrultusunda yapılmakta ve bunun sonucunda birçok farklı uygulama problemi ortaya çıkmaktadır (Ayaydın, 2011; Kavuran, 2003). Böyle bir durumda alınan karaların geçerliği ve güvenirliği önemli bir endişe kaynağıdır. Yapılan değerlendirmenin geçerliğe ve güvenirliğe yönelik tehditleri azaltmak için puanlayıcı eğitiminin yapılması vurgulandığından, yapılan bu araştırma ile bu tür uygulamaya sahip alanlara katkı sağlanacağı düşünülmektedir.

Sınırlılıklar

Araştırmanın sahip olduğu sınırlılıklar maddeler halinde aşağıda verilmiştir;

- Alanyazında birey performansının değerlendirmesinde 30'dan fazla problemli puanlayıcı davranışı rapor edilmesine karşın, bu araştırmada, en çok ortaya çıkan; puanlayıcı katılığı ve cömertliği, merkeze yönelim, halo ile farklılaşan puanlayıcı katılığı ve cömertliği davranışları dikkate alınmıştır. Problemli puanlayıcı davranışı denilrken bu altı davranış kast edilmektedir.
- Alanyazında birçok puanlayıcı eğitimi deseni rapor edilirken, mevcut çalışmada sadece iki puanlayıcı deseni (puanlayıcı hataları eğitimi, referans çerçevesi eğitimi) dikkate alınmıştır.
- Mevcut araştırmada hem ön testte hem de son testte puanlayıcılar aynı öğrenci kompozisyonlarını puanlanmışlardır.
- Araştırma kapsamındaki puanlayıcılar, daha önce puanlama deneyimi olmayan 45 İngilizce öğretmen adayı ile yapılmıştır. Ayrıca deney ve kontrol grupları seçkisiz olarak seçilememiştir.
- Mevcut araştırmada değerlendirmesi yapılan performans, öğrencilerin ikinci dil akademik yazma becerileri olarak dikkate alınmış olup diğer yazma türleri incelenmemiştir.

Tanımlar

Problemli Puanlayıcı Davranışları: Performans değerlendirmede birey performansının dışındaki diğer bazı dışsal değişkenlerin her zaman için ölçümleri etkileme (karışma) olasılığı bulunmaktadır. Birey performansının değerlendirilme sürecinde meydana gelen bu tutarsızlıklar problemli puanlayıcı davranışları olarak tanımlanmaktadır. Metin içinde puanlayıcı davranışları geçtiğinde burada problemli puanlayıcı davranışları kast edilmektedir.

BÖLÜM 2

KURAMSAL TEMELLER

Bu bölüm, araştırma problemi ile ilgili kuramsal temeller ve araştırma konusu ile ilgili araştırmalar olmak üzere iki kısımdan oluşmaktadır. Kuramsal temeller kısmında; akademik yazma becerilerinin değerlendirilmesi, performans değerlendirme, problemli puanlayıcı davranışları, problemli puanlayıcı davranışlarının azaltılması veya kontrol altına alınmasına yönelik başvurulan uygulamalar ve puanlayıcı güvenirliğine ilişkin bilgilere yer verilmiştir. İlgili araştırmalar kısmında ise, performans değerlendirmede ortaya çıkan problemli puanlayıcı davranışlarının incelenmesi ve performans değerlendirmede nesnelliği yükseltmek için kullanılan puanlayıcı eğitimi desenlerine ilişkin araştırmalara yer verilmiştir.

Akademik Yazma Becerilerinin Değerlendirilmesi

Akademik yazma, düşüncelerin mantıksal olarak yapılandırılıp gerekçelendirildiği bir metin türü olarak tanılanmaktadır (Bayat, 2014). Bir başka tanıma göre akademik yazma, bireyin düşünce temeline dayalı olarak görüş, fikir ve duygularıyla gözlem, deney ve tecrübelerini, seçilen konuyla ilgisi ölçüsünde planlayarak dilin kurallarına uygun biçimde anlatması olarak ifade edilmektedir (Göçer, 2010). Akademik yazmanın tanımlarına bakıldığında birçok beceriyi gerektirdiği ve karmaşık bir sürece sahip olduğu görülmektedir. Çünkü akademik yazma becerileri; zihinsel, devinimsel ve duyuşsal becerilerinin aynı anda

kullanılmasını gerektiren çoklu bir dil beceri setidir (Çekici, 2018). Akademik yazma, sahip olduğu farklı özelliklerden dolayı diğer metin türlerinden ayrılmaktadır. Bu özellikler; organizasyon, içerik, dil kullanımı ve sözcük seçimidir. Akademik yazılar belli bir organizasyona sahip olmalıdır. Akademik yazının niteliği, mantıklı ve bilimsel dayanaklarla işleyen bir düşünce içermesine bağlıdır. Akademik yazılar, ilgili alanın jargonu dahilinde bir dil ve argodan uzak bir sözcük seçimine sahiptir (Bayat, 2014).

Yükseköğretimdeki öğrencilerin yazmış oldukları kompozisyonlar (essay), tezler ve araştırma raporları akademik yazı türlerine girmektedir (Gillet, Hammond & Martala, 2009). Akademik yazma; karmaşık düşünceleri, soyut kavramları ve üst düzey zihinsel süreçleri aktarma amacını taşımaktadır (Zwiers, 2008). Akademik yazma, üsy düzey zihinsel beceri olarak ele alındığında önemli bir yere sahip olduğu belirtilmiştir (Carter, Bishop ve Kravits, 2002, s.20). Yükseköğretimdeki öğrencilerin akademik yazma becerilerini geliştirmeleri için; zengin kelime hazinesine sahip olma, bol ve dikkatli okuma çalışmaları, duygu ve düşüncelerini belli bir düzende yönetebilme, özgün yazı çalışmaları, gözlem yeteneğini geliştirme, yazma konusunda azimli olma ve belli bir amaca sahip olma gibi birçok özelliğin bir arada bulunması gerektiği belirtilmiştir (Deniz, 2003).

Öğrencilerin akademik yazma becerilerinin geliştirilmesi önemli olmakla birlikte, bu becerilerin güvenilir ve geçerli bir şekilde ölçülüp değerlendirmeleri de önemlidir. Akademik yazma becerilerinin üst düzey zihinsel beceri olduğu göz önüne alındığında, geleneksel değerlendirme yöntemlerinin uygun olmadığı, bunun yerine performans değerlendirme yöntemlerinin daha uygun olduğu araştırmanın problem durumunda daha önce belirtilmiştir. Bu bağlamda mevcut çalışmada yükseköğretimdeki öğrencilerin akademik yazma becerilerinin değerlendirilmesinde, performans değerlendirme yöntemlerine başvurulmuş, böylelikle yapılan ölçümlerin güvenirliği ve ölçümlere dayalı yapılan çıkarımların geçerliğine katkı sağlanması amaçlanmıştır. Aşağıda performans değerlendirme detaylı bir şekilde açıklanmıştır.

Performans Değerlendirme

Üst düzey zihinsel beceri gerektiren, bir görev ya da etkinliğin yapılmasında ortaya konan çaba ve sonucunda ortaya çıkan ürün performans olarak tanımlanmıştır. Öğrenci başarısının gelişimini sağlamak amacıyla yapılan üst düzey zihinsel çaba gerektiren çalışmalar performansa dayalı durum belirleme olarak tanımlanmaktadır (Kutlu vd., 2014, s.30). Geleneksel değerlendirmenin aksine performans değerlendirme, bireyin gerçek koşullardaki problem veya görevi yaparken öğrendiği bilgi ve beceriyi ne kadar iyi kullandığına odaklanmaktadır (Erman-Aslanoğlu ve Kutlu, 2003).

Performans değerlendirmede, birey performansının durum belirleme sürecinde izlenen adımlar şu şekildedir; amaç ve kazanımlar belirleme, performans görevi geliştirme, puanlama için dereceli puanlama anahtarını geliştirilme, performans görevinin uygulanması ve performansın puanlanması, yorumlanması ve sonuçlarının kullanılmasıdır (Moskal, 2003). Performans değerlendirmede amacın belirlenme aşamasında, değerlendirilecek öğenin ne olacağına karar verilmelidir. Başka bir değişle bireyin performansının değerlendirilmesinde sürecin mi, süreç sonunda ortaya çıkan ürünün mü yoksa her ikisinin birlikte mi ele alınacağına karar verilmelidir (Airasian, 1994). Performans değerlendirmede amaç belirlendikten sonra, ilgili performansın ölçülmesi için performans görevi geliştirilmelidir. Performans görevleri en genel anlamda, genişletilmiş yanıtlı ve sınırlandırılmış olmak üzere ikiye ayrılmaktadır (Mc Millan'dan aktaran, Kutlu vd., 2014, s.32). Ölçülecek performansın niteliğine göre, performans görevlerinden biri seçilmelidir. Bu araştırmada, yükseköğretimdeki öğrencilerin akademik yazma becerilerinin ölçülmesi amaçlandığından genişletilmiş yanıtlı performans görevi seçilmiştir. Bir performans görevi; tanımlama, görev, yönerge ve puanlama yöntemi bölümlerinden oluşmaktadır (Kutlu vd., 2014, s.34).

Performans görevi geliştirdikten sonra, ilgili performansın puanlaması için dereceli puanlama anahtarının geliştirilmesi gerekmektedir (detaylı bilgi için dereceli puanlama anahtarı bölümüne bakınız). Dereceli puanlama anahtarı geliştirilirken, durum belirlemesi

yapılan performansın doğası dikkate alınmalıdır. Söz gelimi bu çalışma için dikkate alınan akademik yazma becerisi, doğası gereği birçok farklı beceriden oluştuğu için farklı ölçütlerle puanlaması ölçümlerin geçerliği ve güvenirliği açısından daha uygun görünmektedir. Dereceli puanlama anahtarı geliştirilirken, ilgili performansın bütün boyutlarını kapsaması ölçümlerin geçerliğini tehdit eden yapının eksik temsil edilme durumunu bertaraf edecektir (Messick, 1989). Dereceli puanlama anahtarı geliştirdikten sonra, performans görevi (akademik yazma gibi) bireylere uygulanır. Uygulamadan sonra birey performanslarının değerlendirilmesi yapılıp sonuçlar raporlaştırılır (Moskal, 2003).

Geleneksel değerlendirmeden ziyade performans değerlendirmenin kullanılmasının birçok avantajının olduğu belirtilmiştir (Brown & Hudson, 1998; Khattri, Kane & Reeve, 1995; Moore, 2009). İlk olarak, öğrencinin güçlü ve zayıf yönlerinin belirlenmesi ve öğrendiklerine ilişkin daha detaylı bilgi sahibi olma olanağı tanımaktadır. İkincisi, bireyin üst düzey zihinsel becerilerinin gelişmesine odaklanmaktadır. Üçüncüsü bireyin öğrendiği bilgi ve beceriyi gerçek yaşam durumuna aktarmasına imkân tanımaktadır.

Performans değerlendirmenin avantajları olmakla birlikte dezavantajları da mevcuttur. Geleneksel değerlendirmeye göre performans değerlendirme, sistematik ve rasgele hatalardan daha fazla etkilenmektedir (Yue, 2011). Bundan dolayı performans değerlendirme ile yapılan ölçümlerin güvenirliği ve bu ölçümlere dayalı yapılan çıkarımların geçerliği daha fazla savunmasız kalmaktadır (Abu Kassim, 2007). Performans değerlendirmede diğer bir dezavantaj kapsam geçerliğinin sağlanma zorluğudur. Çünkü performans değerlendirmede verilen görev geleneksel değerlendirme maddelerine göre daha fazla zaman aldığından verilen görev ya da görevler sınırlı olmaktadır (Arias, 2010). Performans değerlendirmenin, uygulanması ve puanlamasının zaman alıcı ve uzmanlık isteyen bir sürece sahip olması diğer bir dezavantajıdır (Kubiszyn & Borich, 2013, s.161; Oosterhof, 2003, s.72; Reiner vd., 2002, s.10).

Performans değerlendirmede önemli bir endişe kaynağı, birey performansının puanlanması sürecinde, başka dışsal değişkenlerin ölçümlere karışma ihtimalinin daha fazla olmasıdır.

Performans değerlendirmede birey performansına karışan değişkenlerden biri problemli puanlayıcı davranışlarıdır. Ölçülen yapı ile ilgili olmayan bu davranışlar genellikle yapıyla ilişkisiz varyansa (geçerliği tehdit eden bir faktör) atfedilmektedir (Abu Kassim, 2011; Brennan, vd., 1995; Congdon & McQueen, 2000; Farrokhi vd., 2011). Birey performansının değerlendirilmesinde ölçümlere karışan problemli puanlayıcı davranışlarının belirlenmesi, ölçümlerin geçerliğine ve güvenirliğine katkı vereceğinden, performans değerlendirme sürecinde bu davranışların belirlenip, azaltılması veya kontrol altına alınması önem arz etmektedir. Problemli puanlayıcı davranışlarına ilişkin detaylı bilgi aşağıda verilmiştir.

Problemli puanlayıcı Davranışları

Alanyazında performans değerlendirme sürecinde, birey performansına puanlayıcılardan dolayı karışan varyans; *problemli puanlayıcı davranışları, puanlayıcı etkisi*, *puanlayıcı hatası* ve *puanlayıcı yanlılığı* gibi farklı isimlerle tanımlanmıştır (Farrokhi vd., 2011; Haladyna, 1997, s.139; Myford & Wolfe, 2004; Royal & Hecker, 2016). Birey performansının değerlendirilmesi sürecinde, bir veya daha fazla problemli puanlayıcı davranışı ortaya çıkabilir. Alanyazın incelendiğinde birey performansının durum belirleme sürecinde 30'dan fazla problemli puanlayıcı davranışının olduğu görülmektedir (Royal & Hecker, 2016). Birey performansının değerlendirilmesi sürecinde ortaya çıkan problemli puanlayıcı davranışlarının bir listesi Tablo 1'de verilmiştir.

Tablo 1

Problemli Puanlayıcı Davranışlarının Listesi

Davranış tipi	Puanlayıcının eğilimi/yönelimi
Asimilasyon etkisi	"Aşırı/ölçüsüz" görünmekten kaçınmak için diğer puanlayıcılara benzeyen puanlamalar yapma
Aktarma etkisi	Bir önceki öğrencinin performansının daha sonraki öğrencinin puanını etkilemesine izin verme (hem pozitif hem de negatif yönde)
Merkeze yönelim	Birey performansına bakılmaksızın dereceli puanlama anhtarının bir kategorisini aşırı kullanma
Amigo etkisi	Performansı ne olursa olsun, öğrencilerin destekçisi görünmek için herkese yüksek puan verme
Çarpışan standartlar	Puanlayıcıların kişisel derecelendirme standartları ölçüm aracında ifade edilen standartlarla çakıştığında daha düşük puan verme
Çarpışan değerler	Performansta ifade edilen değerin puanlayıcının verdiği değerle çakışması sonucunda daha düşük puan verme
Korkaklık	Öğrencinin kafa tutması veya meydan okuması nedeniyle daha yüksek puan verme
Kayma etkisi	Zamanla gittikçe katılaşan veya cömertleşen puanlar verme
Aşırılık etkisi	Sadece derecelendirme ölçeğinin en uç kategorilerini kullanma
Yorgunluk etkisi	Yorgunluktan puanların etkilenmesine izin verme
Taraf tutma	Kurumda çok sevilen bir öğrencinin bazı kusurlarını gözden kaçırma/görmeme
İlk izlenim	Sonraki bilgileri göz ardı ederek ilk izlenimine göre puan verme
Halo etkisi	Öğrenciye yüksek puan verileceği şüphesi uyandırma
Gestalt olgusu	Puanlayıcının, öğrencinin farklı yönleri birbirinden ayırt edemediğinde, genel bir izlenime göre puan verme
Bağlantı etkisi	Öğrencinin bir başkasıyla ilişkisine dayalı bir değerlendirme yapma
Güvensizlik	Puanlayıcı daha daha düşük puanı savunacak donanıma sahip olmadığı zaman yüksek puan verme
Uzunluk	Daha büyük/yüksek performanslara daha yüksek puan verme
Cömertlik	Sürekli olarak şişme puan verme/değerlendirme yapma
Kişilik çatışması	Bazı bireyleri ya da bireyin bazı özelliklerini sevmediğinden dolayı düşük puan verme
Puanlayıcı yeterliği	Puanlama yeterliğinin düşük olmasından dolayı düşük puan verme
Tekrarlama faktörü	Tekrarlanan performansın belirlenmesi durumunda daha düşük puan verme
Ölçme aracının yorumu	Ölçme aracının puanlamasını yanlış anlama ve farklı bir şeye puan vermesi sonucunda oluşan uygunsuz puanlama
Puan ranj sınırlanması	Bireylerin performanslarını birbirinden ayıramama
Öz değerlendirme	Performansından daha fazla puan verme
Benzerlik etkisi	Kendine benzettiği bireylere daha yüksek puan verme
Yüzeysel tarama	Performansın sadece sınırlı bir bölümünü değerlendirme
Ani kayıp	Bir puanlayıcının olumsuz tepkisini kışkırtan performansın bazı yönlerinden dolayı düşük puan verme
Sempati puanı	Puanlayıcının duygularına hitap eden veya çaba sarfetme isteğini gösteren bir performansa yüksek puan verme
Katılık	Sürekli olarak düşük puan verme
Kişisel özellik	Bireyin tek bir özelliğine (iletişim, görünüş, başarı gibi) odaklanarak bu özelliğe daha fazla puan verme

Royal, K. D., & Hecker, K. G. (2016). Rater errors in clinical performance assessments. *Journal of veterinary medical education*, 43(1), 5-8. https://doi.org/10.3138/jvme.0715-112R

Tablo 1'de görüldüğü gibi, performans değerlendirme sürecinde puanlayıcıların gösterdikleri birçok davranış mevcuttur. Myford & Wolfe (2004) problemli puanlayıcı davranışlarını; puanlayıcıların katılığı, puanlayıcıların cömertliği, halo etkisi, ranj sınırlaması, merkeze yönelim etkisi, tutarsızlık ve yanlılık gibi başlıklar altında ele almaktadırlar. Haladyna (1997, s.139) puanlayıcı davranışları terimine karşılık performansın puanlanmasındaki yanlılıklar kavramını kullanmaktadır. Bu yanlılık kaynakları; yanıt seti (response set), cömertlik hatası, katılık hatası, merkeze yönelim hatası, yakınlık hatası (proximity error), halo hatası, mantıksal hata, ilgi eksikliği hatası, duruma/kişiye özgü hata (idiosyncratic error) ve diğer yanlılıklar gibi başlıklar altında ele alınmıştır. Problemli puanlayıcı davranışlarının belirlenmesinde; korelasyonel, grafiksel, regresyon modelleri, gizil sınıf modelleri ve çok düzeyli modeller kullanılmaktadır. Bu yöntemlerden genellenebilirlik kuramı ve çok yüzeyli Rasch modeli her geçen gün artarak en fazla tercih edilen modellerdir (Royal & Hecker, 2016).

Tablo 1'de verilen problemli puanlayıcı davranışlarından bazıları doğası gereği diğer davranışlardan daha fazla ortaya çıkmaktadır. Bu davranışlar; katılık ve cömertlik, merkeze yönelim, halo etkisi, tutarsızlık ve farklılaşan katlık ve cömertliktir (Myford & Wolfe, 2003, 2004). Problemli puanlayıcı davranışları, yapıyla ilişkisiz varyansa atfedildiğinden, ölçümlerin güvenirliği ve geçerliği üzerinde önemli bir etkiye sahiptir. Bu yüzden, performans değerlendirme sürecinde problemli puanlayıcı davranışlarının belirlenerek birey performansının mümkün olduğunca bu davranışlardan arınık hale getirilmesi gerekmektedir (Bachman, 2004; Eckes, 2015, s.189). Performans değerlendirme sürecinde en fazla ortaya çıkan problemli puanlayıcı davranışları aşağıda detaylı bir şekilde anlatılmıştır.

Puanlayıcı Katılığı ve Cömertliği

Puanlayıcı katılığı, herhangi bir puanlayıcının diğer puanlayıcılardan ya da puanlama ölçütlerinden sürekli olarak düşük puanlama yapma eğilimi olarak tanımlanmakta iken,

puanlayıcı cömertliği ise herhangi bir puanlayıcının diğer puanlayıcı veya puanlama ölçütlerinden sürekli olarak daha yüksek puanlama yapma eğilimi olarak tanımlanmaktadır (Myford & Wolfe, 2004). Başka bir tanımda ise puanlayıcı katılığı, bir puanlayıcının puanlama ölçeğinin orta noktasının altında puan verme eğiliminde olması, puanlayıcı cömertliği ise bir puanlayıcının puanlama ölçeğinin orta noktasının üstünde puan verme eğilimi olarak tanımlanmaktadır (Saal, Downey & Lahey, 1980). Puanlayıcı katılığı ve cömertliği kişilik özelliğinden (trait) bağımsız olarak süreklilik gösteren bir eğilimdir (Guilford, 1954, s.278). Puanlayıcı katılığı ve cömertliği davranışının birey performansına karıştığı belirlendiğinde, bireylerin puanlarından puanlayıcı davranışlarını çıkartarak yeniden düzenleme yapılmasını öneren araştırmalar bulunmaktadır (Raymond, Webb & Houston, 1991). Fakat problemli puanlayıcı davranışının kaynağı belirlenmedikçe, bu puanlayıcı davranışları sürekli olarak ortaya çıkmaya devam edeceğinden, bireylerin puanlarını yeniden ayarlamak çok zor görünmektedir (Myford & Wolfe, 2000, s.1). Bir puanlayıcının katılığı ve cömertliği; birey, kriterler veya belirli bir puanlama zamanı gibi değişkenlerle değişebilir (Lynch & McNamara, 1998; Wigglesworth, 1993). Araştırmacılar üç nedenden dolayı puanlayıcı katılığı ve cömertliğine odaklanmaktadır; 1) katılık derecesi yüksek olan puanlayıcılar tarafından puanlanan bireyler için ciddi bir dezavantajdır, 2) puanlayıcıların cömertliği, bireylerin puanlarını şişirerek, ölçümlerin geçerliğini tehdit eder. Örneğin, Wolfe (2004) tarafından yapılan araştırmada, en cömert puanlayıcı tarafından puanlanan kompozisyonlarının %75'i geçer notu alırken, en katı puanlayıcı tarafından puanlanan kompozisyonların sadece %14'ünün geçer notu aldığı bulunmuştur. 3) problemli puanlayıcı davranışları üzerine yapılan çalışmalar puanlayıcı uyumsuzluğunda, katılık ve cömertlik davranışının önemli bir faktör olduğunu göstermiştir (Kane, Bernardin, Villanueva & Peyrefitte, 1995).

Puanlayıcı eğitimi genellikle puanlayıcılar arası uyumu yükseltir fakat, puanlayıcı kaymasından (rater drift) dolayı, puanlayıcı katılığı ve cömertliği davranışını ortaya çıkarabilir (Moore, 2009, s.11). Elder, Barkhuizen, Knoch & von Randow (2007) tarafından

yapılan çalışmada, puanlayıcı eğitiminden sonra, puanlayıcı katılığının devam ettiği ve birey performansındaki varyansın %35'ine tekabül ettiği bulunmuştur. Katılık ve cömertlik davranışının; puanlayıcı güvenirliği, puanlayıcı eğitimi, geçerlik ve güvenirlik üzerinde çok etkili olduğu kanıtlanmıştır (Moore, 2009, s.11).

Katılık ve cömertlik davranışı, birey performansına karıştığı zaman öğrencinin aldığı puan, bireyin yeteneği ve puanlayıcının katılık veya cömertlik davranışının toplamı yapacaktır. Bu durumda özellikle puanları kesme noktasında olan öğrenciler için ölçümlere dayalı yapılan çıkarımların (örneğin geçti-kaldı, başarılı-başarısız gibi) geçerliği daha savunmasız olabilir (McNamara, 1996, s. 10). Örneğin yüksek riskli bir sınavda geçme barajının hemen altında kalan bir öğrencinin katı bir puanlayıcı tarafından puanlandığını ve barajın hemen üstündeki bir öğrencinin ise cömert bir puanlayıcı tarafından değerlendirildiğini düşünelim. Eğer bu iki öğrenci için puanlayıcı davranışlarının benzer olduğunu veya performanslarına puanlayıcı davranışlarının karışmadığı bilinirse, bu iki öğrenci için yapılan değerlendirmenin geçerliğinden söz edilebilir. Fakat problemli puanlayıcı davranışlarının ölçümlere karıştığı böyle bir durumda bu iki öğrenci için alınan kararların geçerliği ciddi bir endişe kaynağıdır. Bu bağlamda performans değerlendirmede problemli puanlayıcı davranışlarının belirlenerek, birey performanslarının durum belirlemesi sürecinde etkili olan puanlamaların düzeltilerek dikkate alınması yapılan ölçümlerin güvenirliğine ve ölçümlerden yapılan çıkarımların geçerliğine (karar geçerliği gibi) katkı sağlayacaktır.

Myford ve Wolfe (2003) tarafından katılık ve cömertlik davranışlarını en aza indirmek için birkaç strateji önerilmiştir:

- Ölçülecek performansın özellikleri açıkça tanıtılmalı ve mümkün ise ölçme aracının kategorileri için ortak tanımlamalar yapılmalıdır. Böylece puanlayıcı, performansın farklı düzeyleri arasındaki ayrımı yapabilecek ve ölçme aracının her bir kategorisinin ne anlama geldiğini bilecektir.

- Puanlayıcıların cömert davranma eğilimlerine karşı koymak için, ölçme aracının kategorilerinin yarısı negatif/olumsuz diğer yarısı ise olumlu/pozitif olacak şekilde tasarlanmalıdır.
- Katılık ve cömertlik davranışının ve bu davranışın sahip olabileceği etkinin farkında olmak için puanlayıcılar eğitilmelidir. Böylece puanlayıcılar bu eğilime karşı koyabilirler.
- Ölçme aracının her bir kategorisinde önceden belirlenen sayıda öğrencinin olması şartıyla, puanlayıcılardan öğrencilerin ölçme aracındaki her bir kategoriye yerleştirmeleri istenmelidir (zorlanmış bir dağılım yapmaları istenir). Böylece puanlayıcılar ölçme aracının her kategorisinde belli bir sayıda öğrenciyi yerleştirmek zorunda kalacaktır.
- Puanlayıcı katılığı ve cömertliğindeki farklılıklar için, öğrenci puanlarını ayarlamak amacıyla istatistiksel teknikler kullanılmalıdır.

Merkeze Yönelim

Merkeze yönelim etkisi, herhangi bir puanlayıcının puanlama ölçeğindeki orta kategoriye diğer kategorilere göre daha fazla eğilim göstermesi olarak tanımlanmaktadır (Royal & Hecker, 2016; Wolfe & McVay, 2010). Merkeze yönelim etkisi, puanlayıcının puanlama ölçeğini nasıl kullandığını yansıtan bir puanlayıcı davranışı olarak tanımlanmaktadır (Engelhard, 1994). Merkeze yönelim etkisi, farklı puanlayıcıların puanlama kategorilerini farklı bir şekilde kullanan, yanıt setinin bir örneği olarak görülebilir. Başka bir değişle, bazı puanlayıcılar uçlardaki kategorileri aşırı kullanmakta iken, diğer puanlayıcılar orta kategorileri aşırı kullanma eğiliminde olabilmektedir (Engelhard, 1994). Merkeze yönelim davranışı, genel olarak çoklu derecelendirmeye sahip ölçme araçlarının kullanılması sonucunda tecrübesiz puanlayıcıların yaptığı puanlamalarda ortaya çıkmaktadır (Baird, Hayes, Johnson & Lamprianou, 2013). Anastasi (1988), puanlayıcıların puanlama

ölçeğindeki aşırı uçları kullanmaktan kaçınmaları, puanlamanın değişkenliğini fark edilir bir şekilde azaltacağından hem güvenirliğin hem de geçerliğin azalmasına neden olacağını belirtir (s. 646). Puanlama verilerinde merkeze yönelim davranışının olup olmadığını belirlemeye çalışan araştırmacılar, ölçülen özelliğin/performansın ortalama puanının puanlama ölçeğinin orta noktasına ne kadar yakın olduğuna bakarlar (DeCotiis, 1977; Landy & Farr, 1983).

Myford ve Wolfe (2003) tarafından birey performansının değerlendirilmesi sürecinde puanlayıcıların merkeze yönelim davranışlarını en aza indirmek için çeşitli stratejiler önerilmiştir:

- Ölçme aracının her kategorisi için ortak tanımlamalar yapılmalı. Böylece puanlayıcının ölçme aracının her kategorisinin ne anlama geldiğini ve performansın farklı seviyeleri arasındaki ayrımı daha iyi ayırt edebilir.
- Ölçeğin sonuna yakın ölçek kategorileri arasındaki anlam farklılıkları, ölçeğin merkezine yakın ölçek kategorileri arasındaki farklılıklardan daha belirgin tanımlanmalıdır.
- Ölçme aracında aşırı kategorileri kullanmaktan kaçınmalı, çünkü az sayıdaki puanlayıcı öğrencileri bu kategorilere atamaktadır.
- Puanlayıcıların merkeze yönelim davranışından ve bu davranışın puanlar üzerindeki etkisinin farkında olması için eğitimler verilmelidir.
- Ölçme aracının her bir kategorisinde önceden belirlenen sayıda öğrencinin olması şartıyla, puanlayıcılardan öğrencilerin ölçme aracındaki her bir kategoriye yerleştirmeleri istenmeli (zorlanmış bir dağılım yapmaları istenir). Böylece puanlayıcılar ölçme aracının her kategorisinde belli bir sayıda öğrenciyi yerleştirmek zorunda kalacaktır.

Halo Etkisi (Halo Effect)

Halo etkisi, herhangi bir puanlayıcının birey performansını değerlendirme sürecinde bireyin farklı özelliklerine aynı puanları verme eğilimi olarak tanımlanmaktadır (Engelhard, 2002; MyFord & Wolfe, 2004). Problemli puanlayıcı davranışları üzerine yapılan çalışmalar incelendiğinde en fazla çalışılan konunun halo etkisi olduğu görülmektedir (Myford & Wolfe, 2003). DeCotiis'e (1977) göre halo etkisi bir puanlayıcının, performansını değerlendirdiği bireyin davranışları arasındaki farklılıkları ortaya koymadaki yetersizliği ya da isteksizliğidir.

Daha önceki araştırmalara göre, performans değerlendirmede halo davranışı veya etkisi, genel izlenim oluşturmada ve birey özelliklerinin algılanması esnasında kendini göstermektedir (Murphy, Jako & Anhalt, 1993). Başka bir değişle puanlayıcı, performansını değerlendirdiği bireyin niteliği ile ilgili yetersiz bilgiye sahip ise, o zaman genel izlenimini daha çok kullanmaktadır (Borman, 1975; Decotiis, 1977). Puanlayıcı ve performansı değerlendirilen bireyin birbirini tanıması, halo davranışının ortaya çıkmasında önemli bir değişken olduğu belirlenmiştir (Jacobs & Kozlowski,1985; Kozlowski, Kirsh & Chao,1986). Bu bağlamda araştırmada, puanlayıcılar ve performansı değerlendirilen bireylerin birbirini tanımayacak şekilde seçilerek halo etkisinin en aza indirilmesi amaçlanmıştır.

Duygusal puanlayıcıların, duygusal olmayan puanlayıcılara göre daha çok halo etkisi gösterdiği belirtilmiştir (Steiner, 1954). Ayrıca kadın puanlayıcıların, erkek puanlayıcılara göre daha fazla halo davranışı gösterdikleri tespit edilmiştir (Miller, 1970). Birey performansının değerlendirilmesi sürecinde ortaya çıkan halo davranışı; zamana, bireye, puanlama boyutlarına ve bu faktörlerin bileşime bağlı olarak değişkenlik gösterebilmektedir (Murhpy & Anhalt, 1992).

Myford ve Wolfe (2003) tarafından birey performansının değerlendirilmesi sürecinde puanlayıcıların göstermiş oldukları halo davranışlarını en aza indirmek için çeşitli stratejiler önerilmiştir:

- Durum belirlemesi yapılacak olan performans veya özelliğinin dikkatli bir şekilde tanımlandığından ve bu performansın/özelliğin arasındaki farkların açıkça belli olduğundan emin olunmalı.
- Eğer puanlayıcılar birden çok özelliği puanlıyorlarsa, puanlama ölçeklerinin yüksek ve düşük uçlarının sıralaması tersine çevrilmelidir. Böylece, bir puanlayıcının farklı özellik veya performanslar arasında aynı puanlama dizisi yapmaya teşebbüs etmesi engellenebilir.
- Puanlayıcıların halo davranışından ve bu davranışın puanlar üzerindeki etkisinin farkında olması için eğitimler verilmelidir.
- Eğer birden fazla performans veya özelliğin değerlendirilmesi yapılıyor ise, her bir puanlayıcı bütün öğrencileri ilk performansa göre puanladıktan sonra ikinci performansa geçmelidir.
- Değerlendirilecek çok sayıda özellik varsa, puanlayıcıların bunalmaması için özelliklerin sayısının azaltılması düşünülebilir. Özelliklerin sayısının azaltılması bir seçenek değilse, aynı puanlayıcıların bütün özellikleri puanlaması yerine birkaç farklı puanlayıcı farklı bir özelliği, diğer puanlayıcılar ise farklı bir özelliği puanlamalıdır. Başka bir değişle tamamen çaprazlanmış desen yerine yuvalanmış desen tercih edilebilir.

Farklılaşan Puanlayıcı Katılığı ve Cömertliği (Puanlayıcı Yanlılığı)

Farklılaşan puanlayıcı katılığı ve cömertliği, herhangi bir puanlayıcının öğrencilerden bazılarına göre daha katı ve cömert davranma eğilimi olarak tanımlanmaktadır (Knoch, Read ve Randow, 2007). Farklılaşan puanlayıcı katılığı ve cömertliği davranışını çalışan araştırmacılar, daha önceden tanımlanmış olan (a priori) bir veya daha fazla yanlılık etkileşim analizi için çok yüzeyli Rasch modellerini kullanmaktadır (Myford & Wolfe, 2004). Başka bir değişle araştırmacılar, kurama dayalı olarak önceden tanımlanmış olan

puanlayıcı yanlılıklarını çalışmaktadır. Linacre (2002) FACETS programını kullanarak farklılaşan puanlayıcı katılığı ve cömertliği davranışını belirlemek amacıyla alternatif bir yaklaşım önermiştir. Linacre (2002), analiz modelini araştırmacının tespit etmek istediği anormal (farklılaşan puanlayıcı katlığı ve cömertliği) davranışı belirlemesi için yeniden tasarlamıştır. Eğer yeni tasarlanan modelde, bir veya daha fazla puanlayıcı mükemmel uyum gösterir ise, bu puanlayıcı veya puanlayıcıların farklılaşan puanlayıcı davranışı gösterdiği kabul edilir.

Problemli Puanlayıcı Davranışlarını Azaltma veya Kontrol Altına Almaya Yönelik Uygulamalar

Performans değerlendirme, üst düzey zihinsel becerilerin ölçülmesinde önemli avantajlar sağlamasına rağmen, uygulamadaki önemli bir dezavantajı ölçümlerin objektifliğini sağlama zorluğudur. Ölçümlerin objekliğini düşüren kaynaklar problemli puanlayıcı davranışı olarak tanımlanmıştır. Alanyazında problemli puanlayıcı davranışlarının ölçümlere karışma düzeylerinin azaltılması veya kontrol altına alınmasına yönelik birçok uygulama önerilmiş olup en fazla kullanılan uygulamalar şunlardır: 1) otomatik puanlama (Attali vd., 2010; Burstein vd., 1998; Landauer vd., 2003), 2) birden fazla puanlayıcının kullanılması (Gronlund, 1977, s.85; Kubiszyn & Borich, 2013, s.170), 3) dereceli puanlama anahtarlarının kullanılması (Dunbar vd., 2006; Ebel & Frisbie, 1991, s. 194; Kutlu vd., 2014, s.51; Oosterhof, 2003, s.81) ve 4) puanlayıcı eğitimi (Bernardin & Buckley, 1981; Haladyna, 1997, s.143; İlhan & Çetin, 2014; Lumley & McNamara, 1995) olarak verilmiştir. Bu araştırma kapsamında ilgili performansın değerlendirilmesi sürecinde puanlayıcı davranışlarını azaltmak veya kontrol altına almak için; birden fazla puanlayıcı, dereceli puanlama anahtarı ve puanlayıcı eğitimi kullanılmıştır. Aşağıda performans değerlendime sürecinde puanlayıcı davranışlarını azaltmak veya kontrol etmek için başvurulan uygulamalar açıklanmıştır.

Otomatik Puanlama/Değerlendirme

Performans değerlendirme sürecinde otomatik puanlama yöntemi, insan kaynaklı puanlama yöntemleri kadar güvenilir duruma gelmiştir (Attali & Burstein, 2006; Rudner, Garcia, & Welch, 2006; Shermis, Burstein, Higgins & Zechner 2010; Valenti, Neri, & Cucchiarelli, 2003). Graesser ve McNamara (2012) açık uçlu maddelerin otomatik puanlama performansı ile iki puanlayıcının puanlama performansları arasındaki ilişkiyi incelemiş ve korelasyon katsayının yüksek olduğunu bulmuştur. Her ne kadar otomatik puanlama sistemlerinin kullanımı avantajlı olsa da ilgili konu alanında uzmanlığı bulunmayan kişilerin kullanması bazı problemleri de beraberinde getirmektedir. Ayrıca bu yöntemde bazen birey performansının değerlendirme sürecinde dikkate alınması gereken bazı ek değişkenleri de dikkate almamaktadır (Calfee, 2000). Ek olarak puanlama sürecinde olası teknik hataların ortaya çıkması bu yöntemi elverişsiz hale getirmektedir. Otomatik puanlama yöntemlerinin geniş ölçekli durum belirleme çalışmaları için daha elverişli olduğu belirtilmiştir (Lopez, 2009; Parshall, Spray, Kalohn & Davey, 2002). Örneğin Eğitim Test Servisi (Education Testing Service-ETS) tarafından tapılan yabancı dil olarak ingilizce sınavı (Test of English as a Foreign Language-TOEFL) hem otomatik puanlama hem de insan puanlayıcılar tarafından değerlendirilen geniş ölçekli bir ölçme ve değerlendirme uygulamasıdır.

Birden Fazla Puanlayıcının Kullanılması

Performans değerlendirme sürecinde, objektliği artırmanın en basit uygulaması birden fazla puanlayıcının kullanılmasıdır (Gronlund, 1977, s.85). Performans değerlendirme sürecinde en fazla kullanılan uygulama birden fazla puanlayıcının kullanılmasıdır. Örneğin, güzel sanatlar bölümüne seçilen öğrencilerin belirlemesinde bir jürinin kullanılması, lisansüstü öğrenci alımlarında bir komitenin kurulması, devlet kurumlarına personel alımında bir heyetin kurulması ve buna benzer birçok durumda birden fazla puanlayıcı/değerlendirici kullanılmasıdır. Örneğin TOEFL sınavını değerlendiren puanlayıcılar, kapsamlı bir şekilde

eğitim aldıktan sonra puanlama yapmaktadır. Puanlayıcılar eğitildikten sonra bir sınava tabi tutulurlar. Her puanlama öncesi tüm puanlayıcılar onay sınavına tabi tutulduktan sonra başarılı olduğunda puanlama yapmaktadır. Puanlama liderlerleri denilen üst puanlayıcılar, diğer puanlayıcıları sürekli izler ve rehberlik hizmeti verirler. TOEFL sınavını değerlendirmek için ETS tarafından büyük bir çeşitliliğe sahip puanlayıcı havuzu kurulmuş olup konuşma ve yazma becerileri birden fazla puanlayıcı tarafından değerlendirilerek nesnelliğin sağlanması amaçlanmıştır (Education Testing Service, 2019).

Otomatik puanlamaya nazaran, bu uygulama sınıf içi durum belirleme çalışmaları için daha uygundur. Öğrenci hakkında önemli kararların alınacağı açık uçlu maddeler (performans) kullanıldığında iki veya daha fazla bağımsız puanlayıcı kullanmak hayati önem taşımaktadır (Gronlund, 1977, s.86). Bu sonuçların genellenebilir ve güvenilir olmasını sağlamaktadır. Ayrıca puanlayıcılardan elde edilen değerler kullanılarak ortalama bir değer elde edildiğinden, puanlama ile ilgisi olmayan yapıların etkisi azaltılarak puanlamanın geçerliği artırılmaktadır (Congdon & McQueen, 2000; Farrokhi vd., 2011).

Dereceli Puanlama Anahtarları

Performans değerlendirme sürecinde puanlayıcı davranışlarını azaltmak veya kontrol etmek için kullanılan uygulamalardan biri de dereceli puanlama anahtarlarıdır (Myford & Wolfe, 2003). Dereceli puanlama anahtarları, öğrenci yanıtlarının hangi ölçütlere göre değerlendirileceği ve öğrencinin performansının kaç puana denk geleceğini gösteren puanlama ölçekleri olarak tanımlanmaktadır (Goodrich, 1997; Kutlu vd., 2014, s.51). Jonsson ve Svingby (2007) dereceli puanlama anahtarının, belirli bir konuya özgü ve iyi bir yönergeye sahip olması durumunda güvenilir puanlamalar sağladığını vurgulamıştır. Ayrıca öğrenci performansının değerlendirilmesinde kapsamlı bir geçerlik çerçevesi kullanılarak, geçerli ölçümler sağlanabileceği belirtilmiştir. Dereceli puanlama anahtarlarının asıl amacının bireyin göstermiş olduğu performansın durum belirlemesini yapmak olduğu

belirtilmiştir (Brookhart, 2013, s.4). Buradaki performanstan kasıt hedeflenen öğrenme çıktılarıdır.

Dereceli puanlama anahtarları geliştirilirken geçerlik ve güvenirlik göz önünde bulundurulmalıdır (Moskal, 2000). Geçerleme(validation)/pratik anlamdaki geçerlik, belirli durum belirleme çalışmaları için öğrenci yanıtlarından yapılan çıkarımların uygunluğunu destekleyen kanıt biriktirme süreci olarak tanımlanmaktadır (American Educational Research Association (AERA), American Psychological Association(APA) & National Council on Measurement in Education(NCME), 1999). Dereceli puanlama anahtarı geliştirilirken geçerlik için kanıtların toplanması gerekmektedir. Bir durum belirleme aracının geçerliliğini desteklemek için yaygın olarak üç tür kanıt incelenmektedir. Bunlar; kapsam, yapı ve ölçüt geçerliğidir (Moskal & Leydens, 2000).

Dereceli puanlama anahtarı, ilgili maddedeki konu alanını yansıttığında (öğrenci yanıtlarının tümünü içermesi) kapsam geçerliğinin yüksek olduğu söylenebilir. Böyle bir dereceli puanlama anahtarından elde edilen puanlamaların kullanımı ve yorumu geçerli olmaktadır (Moskal & Leydens, 2000). Başka bir değişle dereceli puanlama anahtarının kapsam geçerliği düşük olursa (öğrenci yanıtlarının çoğunu yansıtmaz ise) bu durumda elde edilen puanlara dayalı yapılan çıkarımlar ve yorumlamalar hedeflenen amaca ulaşmayı zorlaştırabilir. Dereceli puanlama anahtarının kapsam geçerliğinin düşük olması puanlayıcıların puanlama sürecinde farklı bir puanlama yapmalarına neden olacağından, puanlayıcı davranışlarının ortaya çıkmasına neden olur. Puanlayıcı davranışlarını minimum seviyeye getirme amacıyla kullanılan dereceli puanlama anahtarının, puanlayıcı davranışlarını ortaya çıkarması amacına hizmet etmediğinin bir göstergesi olmaktadır (Moskal, 2000; Moskal & Leydens, 2000). Bu bağlamda üst düzey zihinsel becerilerin durum belirlenmesi için kullanılacak olan dereceli puanlama anahtarının kapsam geçerliğinin yüksek olması önem arz etmektedir. Kapsam geçerliği yüksek olan dereceli puanlama anahtarlarının hazırlanması için ilgili konu alanında uzman kişilerden görüş alınıp bu doğrultuda revize edilmelidir (Kutlu vd., 2014, s.86; Moskal & Leydens, 2000).

Üst düzey zihinsel beceriler birer yapıdır. Bu yapıların durum belirlenmesinin yapılması amacıyla geliştirilen dereceli puanlama anahtarlarının yapı geçerliklerinin yüksek olması gerekir. Dereceli puanlama anahtarları yapının uygulanmasını ve ölçülmesini açıkça belirtiklerinden dolayı önem arz etmektedirler (Crusan, 2010, s.43). Messick (1996) yapı geçerliği için kapsamlı bir çerçeve sunmuştur. Bu çerçeveye göre yapı geçerliğinin altı yönü mevcuttur. Bunlar; İçerik, genellenebilirlik, dışsal, yapısal, asli ve sonuçsal yönleridir. Performansın durum belirlemesinde yapı geçerliğin bu yönlerinin tartışılıp göz ardı edilmemeleri gerekir (Jonsson & Svingby, 2007). Yapı geçerliğinin içerik boyutu/yönü, durum belirlemesi yapılan performansın ya da yapının temsil edilmesi ile ilgilidir. Yapılan çalışmaların üçte birinde yapı geçerliğinin bu boyutu dikkate alınmıştır. Yapı geçerliğinin içerik boyutu için genellikle uzman görüşlerine başvurularak deneysel kanıt sağlanmaktadır (Jonsson & Svingby, 2007). Yapı geçerliğinin genellenebilirlik boyutu için bireylerin puanları ile mezuniyet derecelerinin karşılaştırılması yapılarak geçerliğin bu boyutu için kanıt elde edilebilir(Gearhart, Herman, Novak & Wolf, 1995). Burada yapı geçerliğinin genellenebilirlik boyutu ile dışsal boyutu arasındaki sınır biraz belirsizdir(Baker, 1994). Yapı geçerliğinin dışsal boyutu üzerine yapılan raporlar incelendiğinde, dereceli puanlama anahtarlarının; ders öncesi değerlendirmeler (Roblyer & Wiencke, 2003), ulusal anket maddeleri(Stoering & Lu, 2002), ön bilgi testleri (Waltman, Kahn & Koency, 1998) arasındaki korelasyonlarının .40 ile .60 arasında değiştiği bildirilmiştir (Jonsson & Svingby, 2007). Yapı geçerliğinin dışsal boyutunun diğer bir odak noktası, hedeflediği amaç için dereceli puanlama anahtarının uygunluğu ve faydasıdır (Flowers & Hancock, 2003). Başka bir ifadeyle geliştirilen dereceli puanlama anahtarı hitap ettiği kesim (öğretmenler gibi) tarafından benimsenir ise dışsal geçerliğinin yüksek olduğu söylenebilir.

Messick (1996)'e göre hem performans görevi (kompozisyon gibi) hem de dereceli puanlama anahtarı (puanlama yönergesi) ölçülmek istenen yapının teorisi ile tutarlı olmalıdır. Bu yapı geçerliğinin yapısal boyutunu temsil etmektedir. Yapı geçerliğinin yapısal boyutuna ilişkin kanıt toplamak için genellikle faktör analizine başvurulmaktadır (Baker,

1994; Baker, Abedi, Linn & Niemi, 1995). Yapı geçerliğinin asli boyutu ise alan uzmanları tarafından kullanılan düşünme süreçlerini yansıtan yanıtlarının tutarlılığının teorik gerekçelerini ve deneysel kanıtlarını içerir. Dereceli puanlama anahtarlarının kullanıldığı birçok araştırmada, süreçten ziyade açık uçlu madde, çalışma örnekleri ve laboratuvar raporları gibi ürün odaklı oldukları belirlenmiştir (Jonsson & Svingby, 2007). Dereceli puanlama anahtarının asli geçerliğinin yüksek olması için, sadece bir durum belirleme aracı değil aynı zamanda düşünme sürecinin teorik bir modeli olarak amaca hizmet etmelidir (Osana & Seymour, 2004). Yapı geçerliğinin sonuçsal boyutu ise hem kısa hem de uzun vadede hedeflenen ve hedeflenmeyen sonuçların, puanlarının yorumlanmasını içerir. "Sonuçsal geçerlik" farklı kesme puanlarına dayalı olarak alınan yeterli veya yetersiz kararlarının tutarlılığı ve anlamının yanı sıra puanlayıcıların puanlama kullanışlıklarının yansımalarından elde edilen kanıtlarını kapsamaktadır (Jonsson & Svingby, 2007).

Yapı geçerliği yüksek bir dereceli puanlama anahtarında, birey yanıtlarının puanlanması sürecinde, konu alanı ile ilgili olmayan yapılar puanlamaya dahil edilmeyecektir (Moskal & Leydens, 2000). Dereceli puanlama anahtarının yapı geçerliğinin düşük olması halinde, ilgili yapının eksik temsil edilme durumu ortaya çıkacak (Messick, 1989) ve bunun sonucunda puanlayıcı davranışları ortaya çıkacaktır. Birey performansının durum belirlenmesinin yapılması amacıyla kullanılan dereceli puanlama anahtarının geçerliği şüpheli veya düşük ise bu durumda puanlayıcı davranışlarının ortaya çıkmasında iki önemli etken oluşmaktadır. Bu etkenler yapıyla ilişkisiz varyans ve yapının eksik temsil edilme durumudur (Downing & Haladyna, 2004; Farrokhi vd., 2011; Jonsson & Svingby, 2007; Messick, 1996).

Dereceli puanlama anahtarları geliştirilirken dikkate alınması gereken diğer bir geçerlik türü de kritere dayalı geçerliktir. Kritere dayalı geçerliğe kanıt sağlamak için bireyin dereceli puanlama anahtarından aldığı puan ile aynı performansın başka bir ölçme aracı ile ölçülmesi sonucu aldığı puanın korelasyonu incelenir (Rafilson, 1991). Bu geçerlik türü için iyi bir örnek mühendislik fakültelerindeki bir uygulamadır. Bu uygulamada mühendis adaylarına gerçek çalışma ortamlarına deneyim kazandırmak amacıyla gerçeğine benzer bir kurs

verilir.Mühendis adayının kurstaki başarısının değerlendirilmesinde dereceli puanlama anahtarları kullanılır. Burada kullanılan dereceli puanlama anahtarı mühendislerin gerçek çalışma ortamlarındaki başarılı performanslarını gösteren kriterlerden oluştuğundan, mühendis adaylarının aldıkları puanlar mühendislerin meslekteki başarı performansları ile korelasyonuna bakılarak mühendis adayı değerlendirilir (Moskal & Leydens, 2000).

Puanlayıcı Eğitimi

Performans değerlendirme sürecinde, ölçümlere karışan puanlayıcı davranışlarını azaltmak veya kontrol altına almak için başvurulan yöntemlerden biri de puanlayıcı eğitimidir. Puanlayıcı eğitimi puanlayıcıların varyansını azalmak için yaygın olarak kullanılmaktadır (Brijmohan, 2016). Puanlayıcı eğitiminde asıl amaç, örnek uygulamalar üzerinden ölçme araçları puanlayıcılara açıklanarak puanlayıcılar arasında ortak bir anlayış ve kavramsallaştırma yapılmasıdır (Alderson, Clapham & Wall, 1995; Cellar, Curtis, Kohlepp, Poczapski & Mohiuddin, 1989; Fahim & Bijani, 2011). Puanlayıcı eğitimi genel olarak katılık ve cömertlik davranışı üzerinden, puanlayıcılardan gelen değişkenliği ve tutarsızlığı azaltmayı amaçlamaktadır. Bu hedefin gerçekleştirilmesinin en yaygın yolu puanlayıcılara bir takım kriterlerin tanıtıldığı ve daha sonra bu kriterlere göre değerlendirmenin yapılması istendiği puanlayıcı eğitimi vermektir (Fahim & Bijani, 2011). Puanlayıcı eğitimi puanlayıcı davranışları üzerinden değişkenliği azaltabilir fakat yok edemez. Puanlayıcı eğitimi katılık ve cömertlik açısından aşırı uçlardaki puanları azaltarak, aynı hizaya getirmektedir (McIntyre, 1993).

Puanlayıcı eğitiminin işlevselliği ile ilgili tartışmalara rağmen, puanlayıcı eğitimi sırasında gerçekte ne olduğu ve bireysel puanlayıcıları nasıl etkilediği hakkında çok az şey bilinmektedir (Fahim & Bijani, 2011). Charney (1984), puanlayıcı eğitiminin öncelikle puanlayıcıların kendi yargılarını uygulamalarını engellemeye çalıştıklarını ileri sürmektedir. Puanlayıcı eğitimi, puanlayıcı içi tutarlılığını artırarak, genel tutarlılığı artırabilir (McNamara, 1996). Puanlayıcı eğitiminin puanlayıcı katılığındaki aşırı farklılıkları ortadan

kaldırdığı ve puanlayıcıların bireysel ön yargılarını azaltarak, puanlayıcı içi tutarlılığını artırdığı için etkili olabileceğine dair kanıtların olduğu belirtilmiştir (Weigle, 1994).

Alanyazında puanlayıcı eğitiminin puanlayıcılar arası tutarlılığı yükseltiğine yönelik olan araştırmaların yanı sıra, etkisinin çok az olduğuna yönelik araştırmaların da olduğu görülmektedir (Cook, Dupras, Beckman, Thomas, & Pankratz, 2009; Eckes, 2015; Gingerich, Kogan, Yeates, Govaerts & Holmboe, 2014; Gingerich, Regehr & Eva, 2011; Holmboe, Hawkins, & Huot, 2004; Sebok & Syer, 2015). Davis (2016), puanlayıcı eğitimin etkili olması için puanlayıcıların puanlama tecrübelerinin artırılması gerektiğini belirtmiştir. Bazı yeni/acemi puanlayıcılar, puanlayıcı eğitiminden önce veya hemen sonra mükemmel kabul edilebilir puanlama örüntüleri gösterse de, iç tutarlığı zayıf/kötü ve aşırı katı veya cömert olan yeni/acemi puanlayıcılar, birkaç puanlama dersinden/eğitiminden sonra daha deneyimli akranlarına daha çok benzeyebileceği belirtilmiştir (Lim, 2011; Weigle, 1998).

Puanlayıcı eğitiminin amaçlarından biri, puanlayıcıların deneyim, puanlama stili veya puanlama tercihi gibi faktörelerin gözlemlenerek, puanlayacılara geri dönütler verilip puanlayıcılar arası ve puanlayıcı içi tutarlığın artırılmasıdır (Kim, 2017). Puanlayıcılar eğitildiğinde, puanlama yapıldıktan hemen sonra geri dönüt verilmelidir (Knoch, 2011). Puanlayıcı eğitiminde geri dönütlerin verilmesi sonucunda, puanlayıcı(lar) yapmış oldukları hataları ve bu hataların çözümü için neler yapabileceklerinin farkına varmaktadır (Elder vd., 2007). Geri dönütler, grup tartışmaları veya bireysel olarak verilebilir.

Puanlayıcıların ön yargılarını azaltmak, yapılan değerlendirmenin doğruluğunu artırmak, gözlem becerilerini geliştirmek, davranışsal doğruluğu ve puanlayıcı güvenirliğini artırmak için birçok puanlayıcı eğitimi deseni/modeli önerilmiştir (Woehr & Huffuct, 1994; Zedeck & Cascio, 1982). Bu desenlerden en çok tercih edilenler şunlardır:

- Öz Liderlik Eğitimi (Self-Leadership Training-SLT),
- Davranış Gözlem Eğitimi (Behavioral Observation Training-BOT),
- Puanlayıcı Değişkenliği Eğitimi (Rater Variability Training-RVT),
- Performans Boyutları Eğitimi (Performance Dimension Training-PDT),

- Puanlayıcı Hatası Eğitimi (Rater Error Training-RET),
- Referans Çerçevesi Eğitimi (Frame-of-Reference Training-FORT).

Öz Liderlik Eğitimi

Öz liderlik eğitiminin amacı, değerlendirmeyi yapmak için puanlayıcıların kendilerine olan güvenlerini artırmaktır. Öz liderlik (kendi kendini yönetme), kendini gerçekleştirmek için gerekli olan öz yönetim (self-direction) ve öz motivasyonu (self-motivation) meydana getirmek için kişinin kendini ikna etme süreci olarak tanımlanmaktadır (Neck, 1996). Öz liderlik, sosyal zihinsel teorinin bir türevi olarak tanımlanmaktadır. Bu teori davranışın; kişi, davranış ve çevre arasındaki üçlü karşılılılığın bir fonksiyonu olduğunu belirtmektedir.

Neck, Stewart & Manz (1995) tarafından, öz liderliğin performans değerlendirme sürecini nasıl etkileyebileceğini açıklayan bir model önerilmiştir. Bu model, kendi kendine konuşma, zihinsel hayaller, düşünce kalıpları, inanç ve varsayımlar gibi birçok sayıda zihinsel yapıları kullanmaktadır. Burada puanlayıcı performansı(yapı) birçok yolla kavramsallaştırılabilir. Puanlayıcı performansının bir yönü/boyutu doğruluktur. Doğruluk, önceden belirlenen standartlarla bireysel puanlamalar karşılaştırılarak ölçülebilir. Puanlayıcı performansının başka bir yönü adalettir. Başka bir değişle bu modelde ilk adım, puanlayıcı performansını etkileyen zihinsel yapılar kullanılarak puanlaycının performansını artırmaya yönelik eğitim verilir. İkinci adımda ise, puanlayıcının puanlama örüntüsü incelenir. Puanlama örüntüsü ile ilişkili olan yapılar belirlenerek, spesifik eğitimler verilir. Son adımda ise puanlayıcıların algıladıkları öz yeterlilikleri dikkate alınır. Burada puanlama örüntüsünün öz yeterlik ile doğrudan ilişkili olduğu belirtilir.

Davranış Gözlem Eğitimi

Davranış gözlem eğitiminde, puanlayıcıların değerlendirme davranışlarının aksine puanlayıcıların gözlem yapma davranışlarına odaklanmaktadır (Woehr & Huffuct, 1994). Thornton ve Zorich (1980), gözlemleme süreçleri ile değerlendirme süreçleri arasında bir ayrımın performans değerlendirme için önemli olduğunu öne sürmektedir. Değerlendirme/puanlama süreçlerinin; bilginin sınıflandırılması, bütünleştirilmesi ve değerlendirilmesini içermesine karşın, gözlem süreçleri; belirli davranışsal olayların tespiti, algılanması ve hatırlanması veya tanınmasını içerir. Ayrıca, cömertlik ve halo gibi puanlayıcı davranışlarının, gözlemsel süreçlerdeki sorunlardan kaynaklanan bilgi eksikliğinden kaynaklandığını ileri sürülmektedir. Sonuç olarak, davranışsal gözlemi iyileştirmeye yönelik stratejilere odaklanan puanlayıcı eğitimi, performans değerlendirme süreçlerinin etkinliğini arttırmalıdır.

Davranış gözlem eğitiminin değerlendirilmesi genellikle bir tür bellek ölçümlerini içermektedir. Bu eğitim deseninde, davranışsal bilgilerin iyi bir şekilde gözlemlenmesi sonucu performansı değerlendirilen bireyin hatırlanma ve tanınma düzeyi daha fazla olacağından daha iyi bir değerlendirmeye tabi tutulacağı varsayılmaktadır. Burada gözlemsel doğruluk, puanlama veya değerlendirme doğruluğundan ayırt edilebilir (Lord, 1985). Gözlemsel doğruluk ölçümlerinin, değerlendirme doğruluğu ölçümlerinden ayıran önemli bir özellik, gözlemsel doğruluk ölçümlerini uyaran etmenler/faktörlerin nesnel ölçülebilir özelliğe sahip olmasından kaynaklanmaktadır.

Davranış gözlem eğitiminde puanlayıcıların performanslarını değerlendirdikleri bireyi doğru bir şekilde gözlemleyebilecek bir yeterliliğe ulaşmaları amaçlanmaktadır (Noonan & Sulsky, 2001). Bu eğitimde puanlayıcının gözlem yapma yeteneğini geliştirmek için bireye ilişkin gözlem kayıtları kullanılmaktadır (Antonioni & Woehr, 2001).

Puanlayıcı Değişkenliği Eğitimi

Puanlayıcı değişkenliği eğitimi, referans çerçevesi eğitimi ve puanlayıcı hatası eğitiminden meydana getirilmiş karışık bir desendir. Puanlayıcı hatası eğitimi, puanlayıcı davranışları hakkındaki bilgiyi artırırken değerlendirme doğruluğunu artırmadığı bulunmuştur(Murphy & Balzer, 1989). Diğer yandan, referans çerçevesi eğitiminin değerlendirme doğruluğunun artmasına yardımcı olduğu fakat performansı değerlendirilen bireylerin belirli davranışlarının gözlemlenmesini engellediği bulunmuştur (Sulsky & Day, 1992). Bu bağlamda her iki yöntemin dezavantajlarını gidermek için puanlayıcı değişkenliği eğitimi geliştirilmiştir (Stamoulis & Hauenstein, 1993).

Diğer puanlayıcı eğitimi desenlerine nazaran puanlayıcı değişkenliği eğitimi (PDE) daha az çalışılmıştır. PDE referans çerçevesi temelli, puanlayıcı hatası hedefli bir yaklaşımdır (Hauenstein, 1998). PDE, değişkenliği artıracak olan temel puanlayıcı hatalarını kullanmak yerine, puanlayıcıların performanstaki gerçek değişkenliğini belirlemek için performansı etkili bir şekilde temsil eden standartlar belirlemeye odaklanır. PDE, referans çerçevesi eğitiminde olduğu gibi, her bir boyuttaki performans düzeyini ayırt etmek için davranışsal örnekler üretmek yerine, davranışların temsil ettiği özel performans boyutlarına bakılmaksızın, tüm davranışları temsil eden örnekler üretmeye odaklanır (Hauenstein & McCusker, 2017). PDE, referans çerçevesi eğitimine benzediğinden, eğitim süreci boyunca katılımcılara/puanlayıcılara geri dönütlerin verilmesi gerekir. Özetlemek gerekirse PDE, referans çerçevesi eğitimine benzeyen fakat hedefleri puanlayıcı hatası eğitimi hedeflerine benzeyen karışık bir desendir.

Performans Boyutları Eğitimi

Performans boyutları eğitiminin amacı, değerlendirilmesi amaçlanan performansın her bir boyutu ile ilişkili olan uygun davranışları tanımlamak için puanlayıcıları eğitmektir. Boyut, yeterlilik ya da beceri puanlayıcılara anlatılır ve boyutun davranışsal örnekleri; rol yaparak

öğrenme (role-play), video veya yazılı metin olarak puanlayıcılara verilir (Feldman vd., 2012). Örneğin, ekip çalışmasının değerlendirileceği bir durum dikkate alındığında, burada takım çalışması ve takım çalışmasını oluşturan belirli davranışları (durum izleme, bilgi verme gibi) açıklayan bir ders puanlayıcılara verilebilir.

PBE, puanlayıcı hatası eğitiminin tutarsız sonuçlarına bir tepki olarak geliştirilen desenlerden biridir. PBE deseninde, puanlayıcı eğitiminin başarı anahtarı olarak puanlayıcılara bilginin zihinsel sürecinin vurgulanması ile başlanmaktadır (DeNisi, Cafferty, & Meglino, 1984; Feldman, 1981). PBE, daha doğru puanlamalar için boyutun evrensel değerlendirilmesi yerine uygun değerlendirilmesinin yapılması varsayımına dayanmaktadır(Woehr, 1992). PBE deseninin odak noktası puanlama hatalarının yerine performansın değerlendirilme doğruluğu üzerinedir (Woehr & Huffuct, 1994).

Puanlayıcı Hatası Eğitimi

Puanlayıcı eğitimi programlarının ilk odak noktası öznel performans değerlendirmenin psikometrik özellikleri üzerinde yoğunlaşması ile başlamıştır. İlk olarak ranj sınırlaması, yüksek derecede birbiri ile ilişkili puanlamaların performans değerlendirme sürecinde ortaya çıktığı bulunmuştur. Daha sonra puanlayıcı katılığı ve cömertliği, halo etkisi ve merkeze yönelim gibi puanlayıcı davranışları rapor edilmiştir (Cooper, 1981; Landy & Farr, 1980; Saal vd., 1980). Puanlayıcı davranışlarının raporlanmasından sonra performans değerlendirmenin de nispeten güvenilmez olduğu belirtilmiştir (Pearlman, Schmidt & Hunter, 1980). Bu psikometrik hatalar, performans değerlendirmenin hem sistematik hem de sistematik olmayan hataları önemli miktarlarda içerdiğinin kanıtı olarak yorumlanmıştır. Sonuç olarak, birçok araştırmacı bu problemlerden kaçınmak ve tanımlamak için puanlayıcılara eğitim verilmesiyle bu puanlayıcı hatalarının(davranışlarının) giderilebileceğini savunmuştur (Woehr & Huffuct, 1994).

PHE desenin temel varsayımı, yaygın puanlayıcı davranışlarına aşina olmanın ve bu hatalardan kaçınmak için puanlayıcıları teşvik etmenin puanlayıcı davranışlarının doğrudan azalmasına ve dolaysıyla daha etkili performans değerlendirmeye neden olacağı şeklindedir (Woehr & Huffuct, 1994). Yapılan çalışmalar PHE deseninin puanlayıcılar arası güvenirlik gibi puanlama özellikleri üzerinde herhangi bir pozitif etki yaptığına yönelik bulguya rastlanmamıştır (Bernardin & Pence, 1980; Borman, 1975). PHE deseninde, puanlayıcı katılığı ve cömertliği gibi puanlayıcı davranışları azalmasına rağmen, puanlama doğruluğunun da azaldığına yönelik bulgular rapor edilmiştir (Bernardin & Pence, 1980). Bu sonuçları gerekçe gösteren birçok araştırmacı PHE deseninin uygun olmayan bir yaklaşım olduğunu belirtmiştir.

Referans Çerçevesi Eğitimi

Birçok puanlayıcı eğitimi deseni var olmasına rağmen, en fazla tercih edilen yöntemin referans çerçevesi eğitimi (frame-of-referance traning, FORT) olduğu belirtilmiştir (Roch, Woehr, Mishra & Kieszczynska, 2012). Bunun temel sebebi performans gözlenip ve değerlendirildiğinde puanlayıcılar için performansa yönelik ortak bir kavramsallaştırmanın kullanılmasından kaynaklanmaktadır (Aguinis, Mazurkiewicz & Heggestad, 2009; Athey & McIntyre, 1987; Woehr, 1994). Referans çerçevesi eğitimin etkin bir şekilde kullanılmasının bir sebebi de çeşitli performans boyutlarının açıklaması olan performans teorisini kapsamasından kaynaklanmaktadır. Performans teorisi, puanlayıcı davranışının uygun boyutla nasıl eşleştiğini, puanlayıcı davranışlarının etkililiğinin nasıl değerlendirildiğini ve farklı yargıların performansın puanlama boyutu ile nasıl birleştiğini açıklamaktadır (Sulsky & Day, 1992).

RÇE deseninde, durum belirlemesi yapılan performansın çok boyutlu olduğu temel alınmaktadır (Schmitt, 2012; Selden, Sherrier & Wooters, 2012). Bu nedenle, performansın tüm alt boyutları tanımlanmalı ve bu boyutların temsilcisi olan davranışsal örnekler

puanlayıcılara verilmelidir. RÇE deseninde temel ilke, puanlayıcıların değerlendirilmesi yapılan performans boyutlarının belli standartlara sahip olduğuna yönelik eğitilmesidir. Böylece puanlayıcının verdiği puanlar ile öğrencinin gerçek puanları arasında bir eşleştirme yapılabilecektir (Woehr & Huffuct, 1994). RÇE deseni üzerine yapılan araştırmalar, puanlayıcı doğruluğunu artırmak için RÇE'nin kullanımını önemli bir ölçüde desteklemiştir (Stamoulis & Hauenstein, 1993; Hedge & Kavanagh, 1988)

Referans çerçevesi puanlayıcı eğitiminde puanlayıcı davranışlarını/yanlılıklarını en aza indirmek için izlenen adımlar şunlardır: a) puanlayıcılara her bir puanlama boyutunun bir tanımını yapmak, b) ölçek referans noktası belirlemek, c) her bir boyutun hangi davranışların göstergesi olduğunu belirlemek, d) puanlayıcılara, puanlama becerilerini uygulamalarına izin vermek ve e) uygulama ile ilgili geri bildirim vermektir (Cascio & Aguinis, 2004). Puanlayıcı eğitimi üzerine yapılan meta analiz çalışmalarına göre, puanlayıcı doğruluğunu artırmak için en etkili yöntemin referans çerçevesi eğitimi olduğu belirlenmiştir (Aguinis, 2012, s.180; Roch vd., 2012; Woehr & Huffcutt, 1994).

Alanyazındaki her bir puanlayıcı eğitiminin farklı yaklaşımlara sahip olduğu görülmektedir. Puanlayıcı eğitimi desenlerinden hangisinin kullanıldığına bakılmaksızın, temel hedef eğitimin puanlayıcı güvenirliği ve doğruluğunu artırması ve puanlayıcı davranışlarının azlatılması beklenmektedir. Puanlayıcı eğitimine yönelik yapılan araştırmalar, puanlayıcı eğitimlerinin puanlama doğruluğunu ve güvenirliğini artırdığını ve puanlayıcı davranışlarını azalttığını göstermektedir. Fakat puanlayıcı davranışlarının ve bireysel yanlılıkların tamamen elimine edilmesinde puanlayıcı eğitimlerinin başarısız oldukları görünmektedir (Galati, 2007). Ayrıca puanlayıcı eğitimlerinin kısa ömürlü oldukları (Farrokhi vd., 2011) bu yüzden puanlayıcı eğitimlerinin sürekli olarak verilmeleri gerektiği belirtilmiştir (Wang, 2010).

Puanlayıcı Güvenirliği

Güvenirlik, hata kaynağına göre farklı isimlendirmelere ve farklı anlamlara gelmektedir (Atılgan, 2004, s.3). Örneğin, test tekrar testte güvenirlik kararlılık anlamında ve buradaki hata kaynağı ise iki uygulama arasındaki zaman iken, parelel formda güvenirlik tutarlılık anlamında ve buradaki hata kaynağı ise ölçme aracıdır. Ölçme ve değerlendirmede güvenirliğin en genel tanımı, ölçme aracından elde edilen puanların tutarlılığı olarak tanımlanır. Ölçmenin amacı gerçek değeri elde etmektir. Fizik, Kimya gibi beşeri bilimlerde ölçümün gerçek değeri daha az hatayla başka bir değişle daha yüksek güvenirlikle elde edilirken, psikolojik yapıların ölçülmesine odaklanan sosyal ve eğitim bilimlerinde ölçümün gerçek değeri daha fazla hata oranıyla yani daha düşük güvenirlikle elde edilmektedir. Ölçme sonuçlarına karışan hata; ölçme aracından, ölçme ortamından, ölçülen özellikten, ölçme zamanından, ölçme yönteminden, ölçmeyi yapan kişiden ve bütün kaynakların etkileşiminden kaynaklanabilir (Crocker & Algina, 2008, s.105-106; Lord & Novick, 2008, s.39; Tekin, 2013, s.55; Turgut & Baykul, 2014, s.120).

Güvenirliğin, hata kaynağına göre farklı isimlendirildiği önceden belirtilmişti. Performans değerlendirmede önemli bir hata kaynağı, birey performansının değerlendirilme sürecinde puanlayıcıdan veya değerlendiriciden gelen varyanstır. Performans değerlendirmede güvenirlik, puanlayıcı tutarlılığı anlamında ve hata kaynağı ise puanlayıcılardır. Puanlayıcı güvenirliği ile aynı bireye iki veya daha fazla puanlayıcı tarafından verilen puanların tutarlılığı (puanlayıcılar arası güvenirlik) veya aynı bireye tek puanlayıcı tarafından iki veya daha fazla kez verilen puanların tutarlılığı (puanlayıcı içi güvenirlik) kastedilmektedir (Alharby, 2006; Jonsson & Svingby, 2007; Moskal & Leydens, 2000).

Performans değerlendirme sürecinde puanlayıcılardan gelen değişkenliğin belirlenmesi için kullanılan puanlayıcı güvenirliği yönteminin seçiminde, kullanılan ölçme aracının türü, araştırmanın deseni, elde edilen ölçümlerin kullanım amacı, puanlayıcı sayısı gibi birçok faktör göz önünde bulundurulmalıdır (Hallgren, 2012). Örneğin, iki puanlayıcı için

geliştirilen bir puanlayıcı güvenirliği yöntemi, ikiden fazla puanlayıcının olduğu bir ölçme durumda kullanılması uygun olmayabilir.

Ölçme sonuçlarına karışan hata kaynaklarının doğasına göre güvenirlik belirleme yöntemleri geliştirilmiştir. Bu güvenirlik yöntemleri, üç farklı ölçme kuramı altında toplanmıştır. Bu kuramlar; Klasik Test Kuramı (KTK), Genellenebilirlik Kuramı (GK) ve Madde Tepki Kuramı (MTK) olarak verilmiştir. Bu kuramlar aşağıda detaylı bir şekilde açıklanmıştır.

Klasik Test Kuramına Dayalı Puanlayıcı Güvenirliği Yöntemleri

Performans değerlendirme sürecinde puanlayıcı güvenirliği yöntemleri ilk olarak KTK'ya dayalı olarak geliştirilmiştir. KTK'nın temel varsayımı, bir ölçüme ilişkin gözlenen puan, gerçek puan ve hata puanı olmak üzere iki bileşenden oluşmaktadır. Performans değerlendirmede hata varyansına neden olan faktörlerden biri de puanlayıcı davranışlarıdır. KTK'ya dayalı puanlayıcı güvenirlik katsayıları genellikle 0 ile 1 arasında değer almakta ve 1'e yakın değerler yüksek güvenirlik değerlerine işaret etmektedir. Başka bir değişle puanlayıcıların yaptığı puanlamaların tutarlı olduğu belirtilir (Tan, 2013, s.135).

KTK'ya dayalı farklı ölçme durumları için geliştirilen birçok puanlayıcı güvenirliği yöntemi bulunmaktadır. Bu yöntemler; Pearson momentler çarpımı korelasyon katsayısı, Spearman sıra farkları korelasyon katsayısı, Cohen'in kappa istatistiği, ağırlıklandırılmış kappa (weighted kappa) istatistiği, Fleiss kappa istatistiği, Kendall'ın uyum katsayısı, Krippendorff'un alfa katsayısı, Kendall τ, Goodman ve Kruskal'ın ϒ ve sınıf içi korelasyon (ICC) katsayısı şeklindedir.

Mevcut araştırma kapsamında yukarıda adı geçen yöntemlerden Krippendorff'un alfa katsayısı kullanılarak puanlayıcı güvenirliğine ilişkin kestirimler yapılmıştır. Bu bağlamda Krippendorff'un alfa katsayısının açıklanmasına yer verilmiştir.

Krippendorff'un alfa katsayısı ilk olarak içerik analizinde kodlayıcılar arasındaki uyumu belirlemeye yönelik geliştirilmiştir. Aynı zamanda puanlayıcılar arasındaki uyumu

belirlemek için bir uyum istatistiği olarak da kullanılmaktadır (Krippendorff, 1995). Krippendorff alfa katsayısı, iki veya daha fazla puanlayıcı durumu, herhangi bir ölçek türü ile ölçülmüş ölçümler, farklı büyüklükteki örneklemler(küçük veya büyük) ve kayıp veri durumu gibi birçok ölçme durumuna uygulanabilmektedir (Krippendorff, 1980). Krippendorff alfa, puanlayıcının puanlamasına bakmaksızın, toplam puan dağılımı üzerinden hesaplanmaktadır (Artstein & Poesio, 2008). Krippendorff alfa katsayısı aşağıdaki förmül yardımıyla hesaplanmaktadır:

$$\widehat{\alpha_K} = 1 - \frac{D_o}{D_e} \quad (2.1)$$

$$D_o = \frac{1}{n}\sum_c \sum_k o_{ck\ metric}\delta_{ck}^2 \quad (2.2)$$

$$D_e = \frac{1}{n(n-1)}\sum_c \sum_k n_c . n_{k\ metric}\delta_{ck}^2 \quad (2.3)$$

Burada D_o, gözlenen uyumsuzluğu D_e ise beklenen uyumsuzluğu temsil etmektedir. Krippendorff alfa katsayısı 1 olduğunda puanlayıcılar arasında mükemmel bir uyumun olduğu, sıfır olduğunda ise puanlayıcılar arasında uyumun olmadığı belirtilir. Krippendorff alfa katsayısı, beklenen uyumsuzluk oranı, gözlenen uyumsuzluk oranından büyük olduğunda negatif değer almaktadır. Uygulamada Krippendorff alfa katsayısının pozitif değerleri dikkate alınıp raporlanmaktadır. Krippendorff alfa katsayısı;

$0{,}80 \leq \alpha$ ise puanlayıcılar arasında yüksek düzeyde uyum

$0{,}67 \leq \alpha < 0{,}80$ arasında ise puanlayıcılar arasında orta düzeyde uyum

$\alpha < 0{,}67$ ise puanlayıcılar arasında zayıf düzeyde uyum olduğu belirtilir (Ergai vd., 2016).

Genellenebilirlik Kuramına Dayalı Puanlayıcı Güvenirliği Yöntemleri

Cronbach ve arkadaşları, KTK'ya dayalı güvenirlik katsayılarının farklı değerler üretmesinden dolayı Genellenebilirlik Kuramını (GK) geliştirmiştir. GK; puanlayıcı, zaman,

test formu, görev, madde gibi ölçme durumu içinde yer alan bütün hata kaynaklarını aynı anda değerlendirebilen esnek ve alternatif bir yaklaşımdır (Atılgan, 2005). KTK'da ölçme modeline dâhil edilen olası tüm hata kaynakları için tek bir güvenirlik değer elde edilmekte ve bu kısıtlayıcı bir durum olmaktadır. Genellenebilirlik kuramının geliştirilmesi ile bu sınırlılık ortadan kaldırılmıştır.

KTK'da güvenirlik değeri örneklem için hesaplanmakta iken, GK'da ise güvenirlik değerinin evrenden hesaplandığı kabul edilmektedir (Güler, 2008, s.23). KTK'da ölçme hataları tesadüfi hata olarak ele alınıp tek bir değer şeklinde sunulmaktadır. Fakat GK'da hem sistematik hem de tesdüfi hata kaynakları birlikte ele alınıp farklı katsayılarla üretilmektedir. Örneğin performans görevi, puanlayıcılar ve öğrenciler olmak üzere üç yüzeyli bir ölçme durumunda güvenirlğin belirlenmesi amaçlanmaktadır. Eğer burada KTK'ya dayalı bir güvenirlik belirleme yöntemi kullanılırsa tek bir katsayı elde edilecek ve elde edilen bu değerin; puanlayıcılardan mı, perfromans görevinden mi yoksa öğrencilerden mi kaynaklandığı belirlenmeyecktir. Fakat GK kullanıldığında her bir yüzey ve yüzeyler arasındaki ortak etkileşimler için hesaplanan varyans kaynakları incelenerek elde edilen güvenirlik katsayısı üzerinde hangi yüzeyin veya değişkenin daha etkili olduğu belirlenebilir.

Madde Tepki Kuramına Dayalı Puanlayıcı Güvenirliği Yöntemleri

KTK'da birey ve test özellikleri birbirinden ayrılmaz. Başka bir değişle farklı bir test alan birey ve farklı gruplara uygulanan maddeleri karşılaştırmak zordur. KTK'nın diğer bir sınırlılığı ise güvenirlik ve ölçmenin standart hatası kavramlarına yaptığı tanımlamaya ilişkindir. KTK'da güvenirliği belirlemek için paralel formlar arasındaki korelasyon kullanılmakla birlikte uygulamada paralel formlar oluşturmak oldukça zor bir durumdur. Diğer yandan KTK'da ölçmenin standart hatası bütün bireyler için aynı kabul edilmektedir. KTK'nın son sınırlılığı ise madde temelli olmayıp test temelli olmasıdır (Hambleton,

Swaminathan, & Rogers, 1991; Hambleton & Swaminathan, 1985). KTK'nın bu sınırlılıklarını aşmak için madde tepki kuramı geliştirilmiştir. Madde tepki kuramının bazı özellkleri şu şekildedir (Baker, 2001; Crocker & Algina, 2008; De Ayala, 2009):

- Birey ve madde özellikleri birbirinden bağımsızdır.
- MTK, test düzeyinde olmaktan ziyade madde düzeyindedir. MTK'nın madde düzeyinde olması ölçme araçlarının kullanım amacına uygun hale getirilmesine imkân sağlamaktadır.
- Güvenirliğin belirlenmesinde paralel formalara ihtiyaç duyulmamaktadır.
- Her bir yetenek puanı (birey) için ölçmenin standart hatasını verir.

MTK modelleri, maddeye verilen yanıtların hepsinin altında yatan bir örtük özellik olduğunu varsaymaktadır. Bu örtük özellik ve madde nitelikleri arasındaki ilişki tanımlanmakta ve bir ölçek üzerinde bir maddenin yanıtlanma olasılığı kestirilmektedir. Diğer bir varsayım bireyin maddeye ilişkin performansı ile bu performansın belirleyicisi olan özellik, örtük özellik ya da yetenek arasındaki ilişki madde karakteristik fonksiyonu veya madde karakteristik eğrisi adı verilen monoton olarak artan bir fonksiyonla ifade edilir. Bu fonksiyona göre bireyin yetenek düzeyi arttıkça maddeyi doğru cevaplama olasılığı, bireyin içinde bulunduğu gruptan bağımsız bir şekilde artar (Embretson & Reise, 2000; Hambleton vd., 1991; Hambleton, Swaminathan, Cook, Eignor & Gifford, 1977). MTK modelleri, boyutluluğa, kategori sayısına ve kategorilerin sıralı olup olmamasına göre farklı sınıflandırmalara tabi tutulmaktadır.

MTK'da puanlayıcı güvenirliği veya değişken sayısının birden fazla olduğu durumda güvenirliğin hesaplanmasında kullanılan modellerden biri George Rasch tarafından geliştirilen Rasch modelidir. Rasch modelinde, madde ayırıcılık indeksi tüm maddeler için eşit ve bir değerinde kabul edilmekte olup, madde güçlük indeksi kestirilmektedir. Rasch modelinde maddenin şans ile yanıtlanma olasılığı sıfır olarak kabul edilmektedir. Linacre (1989) Rasch modelinden faydalanarak çok yüzeyli Rasch ölçme modelini geliştirmiştir. Aşağıda çok yüzeyli Rasch ölçme modeli ile ilgili detaylı bilgi verilmiştir.

Çok Yüzeyli Rasch Ölçme Modeli

Temel Rasch modelinde, birey ve test maddeleri veya performans görevi değerlendirilerek bireylerin yetenek farklılıkları ve maddelerin güçlük düzeyleri eşit aralıklı bir ölçek üzerine yerleştirilir. Elde edilen sonuçların örneklemden bağımsız olduğu savunulmaktadır (Sudweeks, Reeve & Bradshaw, 2005). Çok yüzeyli Rasch ölçme modelinde ise birçok değişkenlik kaynağı (puanlayıcı, madde, görev, birey, zaman gibi) tek bir eşit aralıklı ölçek üzerine yerleştirilebilmektedir (Linacre, 1993). Rasch analizinden farklı olarak, çok yüzeyli Rasch ölçme modeli bireylerin test puanlarını etkilediği düşünülen bütün değişkenlik kaynaklarını dikkate alan bir yaklaşıma sahiptir (Baird vd., 2013, s.40; Kim, Park & Kang, 2012; Linacre, 1996). Çok yüzeyli Rasch ölçme modeli aynı zamanda yüzey modelleri (facets models) olarak da bilinmektedir (Eckes, 2015, s.28).

Çok yüzeyli Rasch ölçme modeli tüm değişkenlik kaynaklarını dikkate almakla beraber aynı zamanda bu değişkenlik kaynaklarının birbirleriyle etkileşimlerine de odaklanmaktadır (Abu Kassim, 2007). Çok yüzeyli Rasch ölçme modeli, geleneksel iki kategorili ölçümlerden ziyade çok kategorili ölçme araçlarının değerlendirilmesini mümkün kılan Master (1982) tarafından geliştirilen kısmi kredi modelinin bir uzantısı olarak tanımlanmıştır (Myford, 2002). Çok yüzeyli Rasch ölçme modeli bütün parametreleri kalibre ederek, sıralama ölçeğindeki gözlemleri eşit aralıklı bir logit ölçeğine dönüştüren doğrusal bir modeldir (Bond & Fox, 2015, s.126). Ardışık kategori olasılıkları (log odds) oranlarının lojistik dönüşümü; akran değerlendirme, durum belirleme kriterleri ve açık uçlu madde gibi bağımsız değişkenlerin, bağımlı değişken olarak görülmesine imkân sağlamaktadır (Esfandiari, 2015). Çok yüzeyli Rasch ölçme modelinde bireylerin performans puanlarını etkileyen her bir değişkenlik kaynağına yüzey adı verilmektedir (Sudweeks vd., 2005).

Çok yüzeyli Rasch ölçme modeli üç tanımlayıcı bilgi türü sağlamaktadır. İlk olarak, araştırmacıların akran değerlendirme, durum belirleme kriterleri ve açık uçlu maddeler gibi

yüzeyler için grup düzeyindeki etkiler açısından genel değerlendirmelerinin özetlenmesini ve puanlanan verinin analiz edilmesini sağlayan istatistiksel bir çerçeve sağlamaktadır (Myford & Wolfe, 2003). İkinci olarak, yüzeylerin çeşitli öğelerinin bireysel düzeydeki etkilerini (örneğin, her bir akranın değerlendirmesinin, açık uçlu maddenin ve durum belirleme kriterinin nasıl olacağı gibi) araştırmak için araştırmacılara izin vermektedir (Linacre, 2002). Son olarak, puanlayıcıların puanlamadaki örüntülerini araştırma gibi çeşitli yüzeyler arasındaki etkileşimleri araştırmacılara sağlamaktadır (McNamara, 1996, s.120).

Çok yüzeyli Rasch ölçme modeli analizlerinde, her bir yüzeyin katkısı ayrıştırılmakta ve farklı yüzeylerin ne ölçüde amaçlandığı gibi çalıştığını belirlemek için diğer yüzeylerden bağımsız bir şekilde incelenme imkânı vardır. Örneğin puanlayıcının puanlama işlemini yaptığında yanlılık gösterip göstermediği gibi bir sorunun cevabı çok yüzeyli Rasch analizleri ile elde edilebilmektedir. Çok yüzeyli Rasch yanlılık analizi yapıldığında, araştırmacı puanlayıcının puanlama örüntüsünde kanıt arar (Myford & Wolfe, 2003). Çok yüzeyli Rasch ölçme modeli yaklaşımı kullanılarak, puanlayıcı yanlılıklarının, inançlarının ya da kişisel özelliklerinin puanlama davranışı üzerindeki etkileri çalışılabilir. Benzer bir şekilde puanlayıcının geçmiş deneyimlerinin puanlama davranışı üzerindeki etkileri incelenebilmektedir. Özetlenirse, puanlayıcı etkilerini incelemek için çok yüzeyli Rasch ölçme modeli yaklaşımı, grup düzeyindeki puanlayıcı etkilerini (ana etkiler) ve puanlayıcı-etki etkileşimlerini ANOVA tabanlı bir yaklaşım gibi incelemektedir. Burada unutulmaması gereken bir husus birçok ANOVA uygulamasında etkileşim etkileri olası ana etkilere karıştığı için ana etkilerin yorumlanması zorlaşmaktadır. Fakat çok yüzeyli Rasch ölçme modelinde bu durum söz konusu değildir (Myford & Wolfe, 2003). Çok yüzeyli Rasch ölçme modeli sahip olduğu avantajları sayesinde tıp, psikoloji, eğitim gibi birçok alanda uygulanma imkânı bulmuştur (Eckes, 2015, s.29; Engelhard, 2013, s.56).

Çok yüzeyli Rasch modeli, araştırmacılara klasik test kuramına ve genellenebilirlik kuramına dayalı modellerin sağlayamadığı bilgileri sağlamaktadır (Lunz, Wright & Linacre, 1990). Çok yüzeyli Rasch ölçme modeli araştırmacıya her bir yüzey ile ilgili detaylı bilgi

verebilmektedir. Örneğin bireylerin performansını değerlendiren bir grup puanlayıcıdan hangisinin ne kadar puanlama yaptığı ve ne kadar puanlama yapması gerektiği gibi birçok bilgi elde edilebilmektedir. Çok yüzeyli Rasch ölçme modelinin detaylı geri dönüt sağlaması sonucu hangi puanlayıcının iyi hangisinin kötü olduğu ve buna bağlı olarak nasıl bir müdahale gerektiği belirlenebilmektedir. Bu bağlamda puanlayıcı eğitimi öncesi çok yüzeyli Rasch ölçme modeli ile puanlayıcı etkilerinin belirlenmesi yapılacak eğitimin kalitesini artırmaya hizmet edebilir.

İki kategorili (0-1) puanlamada birey ve maddelere ilişkin Rasch modeli:

$$\log\left(\frac{P_{bm}}{1-P_{bm}}\right) = B_b - D_m \qquad (2.4)$$

P_{bm} :b bireyinin m maddesinde başarılı olma olasılığı

B_b :b bireyinin yeteneği

D_m :m maddesinin güçlüğü

şeklinde tanımlanmaktadır.

Andrich (1978) tarafından denklem (2.4)'deki iki kategorili puanlamaya uygun modelin matematiksel özellikleri korunarak, derecelendirilmiş çok kategorili ölçeklerin puanlamasına uygun hale getirilmiştir. Derecelendirilmiş ölçeklerle ilgili model:

$$\log\left(\frac{P_{bmk}}{P_{bm(k-1)}}\right) = B_b - D_m - F_k \qquad (2.5)$$

P_{bmk} :m maddesinde b bireyi için k kategori gözleminin olasılığı

$P_{bm(k-1)}$:k-1 kategoride gözlemin olasılığı

F_k :k-1 kategoriyle ilgili k kategoride gözlemlenmiş güçlük

eşitliği ile tanımlanmaktadır.

Mevcut araştırmada, bir grup öğrencinin yazmış oldukları akademik yazıları(kompozisyon) bir puanlayıcı grubu tarafından değerlendirildiğinden, araştırmanın modeli:

$$\log\left(\frac{P_{bkpx}}{P_{bkpx-1}}\right) = \theta_b - \beta_k - \alpha_p - \tau_x \qquad (2.6)$$

P_{bkpx} = p. puanlayıcı tarafından b. öğrencinin k. kriterine x puanının verilme olasılığı

P_{bkpx-1} =p. puanlayıcı tarafından b. öğrencinin k. kriterine x-1 puanının verilme olasılığı

θ_b = b. öğrencinin yeterlilik düzeyi,

β_k = k. kriterin zorluk derecesi,

α_p = p. puanlayıcının katılık derecesi,

τ_x = x-1 puanı yerine x puanının alınma zorluğu,

şeklinde tanımlanmıştır.

Model kurulduktan sonra yapılan analizlerde, tüm yüzeyler eş zamanlı olarak analiz edilirken her bir yüzey bağımsız olarak tek bir doğrusal ölçek üzerinde (logit ölçeği) kalibre edilir. Yüzeylerin ortak kalibrasyonu (ölçümleme), aynı ölçek üzerinde madde güçlüğü, birey yeteneği ve puanlayıcı davranışının ölçülmesini mümkün kılmaktadır. Her bir yüzeydeki her bir koşul (öğe) için bir ölçüm (kalibrasyonun logit kestirimi), bir standart hata (logit kestiriminin kesinliği ile ilgili bilgi) ve uyum göstergeleri(öğeye ait verinin ölçme modeline uygunluğu hakkında bilgi) verilmektedir (Myford & Wolfe, 2003). Tüm bu değerler bireysel düzeydeki istatistiksel göstergeler olarak tanımlanmaktadır. Çok yüzeyli Rasch ölçme modeli, aynı zamanda her bir yüzey için grup düzeyinde istatistiksel göstergeler de vermektedir. Bu istatistikler; Sabit etkili ki kare testi (fixed chi-square test), ayırma oranı (separation ratio), ayırma indeksi (separation index) ve ayırma indeksinin güvenirliği (reliability of the rater separation index) şeklindedir. Grup düzeyindeki bu istatistikler tüm yüzeyler için hesaplanmaktadır. Ayrıca puanlayıcı yüzeyi için puanlayıcı güvenirliğine ilişkin istatistik, madde/kriter yüzeyi için de kategori istatistikleri ve yanlılık analizleri için

yanlılık indeksi hesaplanmaktadır. Aşağıda ilk olarak bireysel düzeydeki daha sonra grup düzeyindeki istatistiklere ilişkin detaylı bilgi verilmiştir.

Bireysel düzeydeki ilk istatistik, modeldeki her bir yüzeyin öğeleri için hesaplanan *logit* değeridir. Çok yüzeyli Rasch ölçme modelinde logit değerleri genellikle dikey olan lojistik cetvel üzerinde yer alır. Eğer tüm yüzeyler pozitif olarak tanımlanırsa, yüksek pozitif logit değerlere sahip ölçmeler; birey yeteneğinin yüksek olmasına, madde güçlüğünün kolay olmasına ve puanlayıcıların cömert olmasına işaret etmektedir. Logit değerleri ile toplama, çıkarma, bölme ve çarpma işlemi yapılabilmektedir. Logit değerleri aslında sıralamalı ölçmelerin eşit aralıklı ölçümlere dönüşmüş halidir. Çok yüzeyli Rasch ölçme modeli analizinde kestirilen her logit değeri için *standart hata* değeri de kestirilmektedir. Bu değer, yüzeydeki bir öğe için kestirilen logit değerinin kesinliği hakkında bilgi vermektedir. Bir puanlayıcının yaptığı puanlamaların katı veya cömert olmasına karar vermek için logit ölçüsünün kaç olacağına dair kritik bir değer olmadığından, uygulamada genellikle logit ölçüleri t-değerine dönüştürülür (Myford & Wolfe, 2003). Hesaplanan t-değeri, t dağılım tablosundaki t-değerinden büyük ise t değerinin işaretine göre puanlayıcının katı veya cömert olduğu söylenir. Logit ölçüsünün t-değerine dönüştürmek için aşağıdaki denklem kullanılmaktadır:

$$\text{t} - \text{değeri} = \frac{\text{puanlayıcının logit ölçüsü} - \text{logit değerlerinin ortlaması}}{\text{logit ölçüsünün standart hatası}} \qquad (2.7)$$

Çok yüzeyli Rasch ölçme modeli analizinde, *uyum içi (infit*) ve *uyum dışı (outfit)* uygunluk istatistikleri önemli bireysel düzeydeki istatistiklerdir. Uyum istatistiklerinin temeli, gözlenen puanlar ile beklenen puanlar arasındaki farktır (Atılgan, 2004, s.31). Örneğin düşük yeteneğe sahip bir bireyin zor bir maddeyi yapması veya yüksek yeteneğe sahip bir bireyin kolay bir maddeyi yapmaması düşük bir ihtimaldir. Bu ihtimalin değerlendirilmesi için çok yüzeyli Rasch ölçme modeli analizinde uygunluk istatistikleri kestirilmektedir. Uygunluk değerleri standartlaştırılmış artıkların kareleri alınarak hesaplanmaktadır. Standartlaştırılmış

artıkların karesi, her bir puanlayıcı için, puanlayıcının toplam uyumsuz puanlama miktarını gösteren bir istatistik özetini üretmek amacıyla birey ve madde arasındaki ortalamanın alınmasıdır (Myford & Wolfe, 2003). Standartlaştırılmış artık değerlerin kareleri ağırlıklandırma yapılmadan alınırsa uyum dışı istatistiği, ağırlıklandırma yapılarak alınırsa uyum içi istatistiği elde edilmektedir. P puanlayıcısı için uyum dışı ve uyum içi istatistiklerinin hesaplandığı formüller aşağıda sırasıyla verilmiştir:

$$UD_p = \frac{\sum_{b=1}^{B}\sum_{k=1}^{K} z_{Abkp}^2}{BK} \quad (2.8)$$

$$UI_p = \frac{\sum_{b=1}^{B}\sum_{k=1}^{K} z_{Abkp}^2 v_{B_{bkp}}}{\sum_{b=1}^{B}\sum_{k=1}^{K} v_{B_{bkp}}} \quad (2.9)$$

$$A_{bkp} = Beklenen_{bkp} - Gözlenen_{bkp} \quad (2.10)$$

$$Z_{A_{bkp}} = \frac{A_{bkp}}{\sqrt{v_{B_{bkp}}}} \quad (2.11)$$

$$v_{B_{bkp}} = \sum_{d=0}^{K}\left(d - B_{bkp}\right)^2 P_{bkpx} \quad (2.12)$$

Burada UD uyum dışı istatistiğini, UI uyum içi istatistiğini, B birey sayısını, K kriter sayısını, A beklenen puanlama ile gözlenen puanlama arasındaki farkı, Z; A değerinin standartlaştırılmış halini ve V gözlemin varyansını göstermektedir.

Uyum dışı ve uyum içi istatistiklerinin bire eşit olması mükemmel uyuma işaret etmektedir. Gerçek ölçme durumlarında bu değerlerin bire eşit olması çok zor olduğundan kabul edilebilir bir aralık belirlenmiştir. Buna göre uygunluk değerleri en az 0,50 ve en çok 1,50 olan bireyin, puanlayıcının veya maddenin uygun olduğu kabul edilmekte bu aralığın dışındaki birey, puanlayıcı veya maddenin analizden çıkarılması önerilmektedir (Brentari & Golia, 2008; Myford & Wolfe, 2003; Wright & Linacre, 1994). Bu bağlamda mevcut çalışmada uygunluk istatistikleri için kabul edilebilir aralık 0,50 ile 1,50 olarak benimsenmiştir.

Grup düzeyindeki istatistiklerden biri olan *sabit etkili ki kare testinin* hipotezi, puanlayıcının katılık değerinin anlamlı bir şekilde farklı olmadığı şeklindedir. Başka bir değişle ölçmenin

standart hatası hesaplandıktan sonra tüm puanlayıcıların aynı katılık düzeyine sahip olduğu varsayılır. Buna göre sabit etkili ki kare değerinin istatistiksel olarak anlamlı olması puanlayıcıların birey performansını değerlendirme sürecinde farklı puanlamalar yaptığının göstergesi olarak yorumlanmaktadır. Ki kare testi örneklem büyüklüğüne karşı çok hassas olduğundan, bir yüzeyin grup düzeyindeki istatistikleri raporlandığında diğer istatistiklerin de verilmesi gerekmektedir (Myford & Wolfe, 2004; Weigle, 1998). Ki kare testinin istatistiksel formülü (Linacre, 2017):

$$\chi^2 = \sum(w_i - D_i^2) - (\sum w_i \times D_i)^2 / \sum w_i \quad (2.13)$$

$$w_i = \frac{1}{SH_i^2}\text{ , i=1,2,…L} \quad (2.14)$$

şeklinde verilmektedir. Burada D, madde veya kriterin güçlüğü ve L ise madde veya kriterlerin sayısını temsil etmektedir. Eğer puanlayıcı yüzeyi için ki kare değeri hesaplanacak olursa, D, puanlayıcının katılık ölçüsü ve L ise puanlayıcı sayısını temsil etmektedir.

Diğer bir grup düzeyindeki istatistik *ayırma oranıdır*. Bu oran, puanlayıcı katılık ölçümlerinin kesinliği ile ilgili olan puanlayıcı katılık ölçülerinin dağılımının/yayılımının bir ölçüsüdür. Örneğin puanlayıcı yüzeyi için hesaplanan ayırma oranının 8,45 olduğu varsayılsın. Bunun anlamı, puanlayıcı katılıkları arasındaki farklılıkların bu katılıkların ölçüldüğü hatadan sekiz kat daha fazla olduğu şeklinde yorumlanmaktadır. Ayırma oranı 1 ile +∞ aralığında değerler alır (Sudweeks vd., 2005). Ayırma oranının hesaplanmasında kullanılan formül (Linacre, 2002):

$$\text{G} = \text{Gerçek Standart sapma/RMSE} \quad (2.15)$$

Gerçek standart sapma= gözlenen standart sapma – ölçmenin standart hatasının ortalaması

Ayırma oranının değeri kullanılarak hesaplanan grup düzeyindeki diğer bir istatistik *ayırma indeksidir*. Bu indeks, analize dâhil edilen puanlayıcıların örneklemi arasında puanlayıcı katılığının istatistiksel olarak farklı düzeylerinin sayısını ifade etmektedir. Bu indeks, hata varyans birimlerindeki gerçek varyansı göstermektedir. Örneğin puanlayıcı yüzeyi için hesaplanan ayırma indeksi 12.71 olduğu varsayılsın. Bu değer, analize dâhil edilen puanlayıcı örneklemindeki puanlayıcıların istatistiksel olarak 13 farklı düzeyde/tabakada puanlayıcı katılığının olduğunu göstermektedir. Ayırma indeksinin hesaplanmasında kullanılan formül (Linacre, 2002):

$$H = (4G + 1)/3 \qquad (2.16)$$

Uygulamada, G <3,50 olduğunda H ≈ G + 1 ve G ≥ 3,50 ise H ≈ G + 2 olarak alınmaktadır. Grup düzeyindeki diğer bir istatistik ise *ayırma indeksinin güvenirliğidir*. Bu istatistik, puanlayıcı yüzeyini güvenilir bir şekilde tanımlamak için puanlayıcıların nasıl iyi ayrıldığı konusunda bilgi sağlar. Bu indeks, puanlayıcı katılık ölçülerinin dağılımının bir ölçüsüdür. Birçok durumda puanlayıcı yüzeyi için en çok istenilen sonuç ayırma indeksinin güvenirliğinin sıfıra yakın olmasıdır. Diğer taraftan madde veya kriter yüzeyi için ayırma indeksinin güvenirliği hesaplandığı zaman, bu istatistiğin yüksek olması beklenir (Myford & Wolfe, 2004, 2003). Ayırma indeksinin güvenirliğinin hesaplanmasında kullanılan formül:

$$R = G^2/(1 + G^2) \qquad (2.17)$$

şeklinde verilmektedir. Denklem (2.17)'de G = 1 olduğunda R = 0,50 olacağı görülmektedir. Bunun anlamı tüm puanlayıcıların katılık düzeyleri, tüm maddelerin güçlük düzeyleri ve tüm bireylerin yetenek düzeylerinin birbirine eşit olduğu demektir. Buna göre araştırmalarda

istenen birey ve madde yüzeyi için yüksek, puanlayıcı yüzeyi için ise düşük ayırma indeksi güvenirliği değerleridir (Myford & Wolfe, 2003).

Puanlayıcı yanlılığı araştırıldığında, grup düzeyinde hesaplanan istatistiklerden biri de yanlılık indeksidir (bias index). Çok yüzeyli Rasch analizi çerçevesinde puanlayıcı davranışlarının belirlenmesinde yanlılık indeksi önemli bir yer tutmaktadır (Engelhard, 2002; Linacre, 2017). Denklem (2.6)'da verilmiş olan ölçme modeli için *puanlayıcı x birey* yanlılık analizi yapıldığında, yeni model şu şekilde tanımlanacaktır:

$$\log\left(\frac{P_{bkpx}}{P_{bkpx-1}}\right) = \theta_b - \beta_k - \alpha_p - \tau_x - I_{bp} \qquad (2.18)$$

Denklem (2.18)'de tanımlanan modelde I_{bp}, *puanlayıcı x birey* grup yüzeyine ilişkin yanlılık etkileşim terimi olarak tanımlanmaktadır.

Çok yüzeyli Rasch analizinde puanlayıcılar arası güvenirliğinin, puanlayıcı yüzeyi için belirlenmesi komutu verildiğinde, program hem her bir puanlayıcı için hem de grup düzeyinde beklenen ve gözlenen uyum yüzdelerini hesaplamaktadır. Bir puanlayıcının gözlenen ve beklenen uyum yüzdeleri birbirine eşit olduğunda, bu puanlayıcının birey performansını değerlendirme sürecinde hiçbir puanlayıcı davranışını göstermediği başka bir değişle mükemmel bir puanlayıcı olduğu ifade edilir (Eckes, 2015, s.91-92). Performans değerlendirme üzerine çalışan araştırmacılar genellikle gözlenen ve beklenen uyum yüzdelerinin birbirine yakın olmalarını arzu etmektedir. Linacre (2017) puanlayıcı güvenirliğinin değerlendirilmesinde, kappa indeksinin Rasch modeline göre tasarlanmış versiyonunu önermiştir. Rasch modeline göre tasarlanmış kappa indeksi:

$$\text{Rasch} - \text{kappa} = \left(\frac{\text{gözlenen uyum yüzdesi} - \text{beklenen uyum yüzdesi}}{100 - \text{beklenen uyum yüzdesi}}\right) \qquad (2.19)$$

şeklinde verilmiştir. Rasch modelinde bu indeks 0,00 olmalıdır. Rasch-kappa değeri, hem grup düzeyinde hem de bireysel düzeyde hesaplanmaktadır. Sıfırdan büyük Rasch-kappa değerleri, yüksek düzeyde puanlayıcılar arası uyuma işaret ederken, sıfırdan küçük değerler ise düşük düzeydeki puanlayıcılar arası uyuma işaret etmektedir. Ayrıca Rasch-kappa değerine göre yerel puanlayıcı bağımlılığı da belirlenebilmektedir. Buna göre sıfıra yakın olan Rasch-kappa değerine sahip puanlayıcının yüksek derecede yerel puanlayıcı bağımlığına sahip olduğu belirtilir (Eckes, 2015, s.92).

Performans değerlendirme sürecinde kullanılan puanlama ölçeğine ilişkin kategorilerin/derecelendirmelerin uygunluğuna karar vermek için *kategori istatistikleri* hesaplanır. Çok yüzeyli Rasch analizinde, kategori istatistikleri hem bireysel düzeyde hem de grup düzeyinde hesaplanabilmektedir. Genellikle merkeze yönelim davranışının belirlenmesinde kategori istatistikleri incelenmektedir (Myford & Wolfe, 2004). Kategori istatistiklerinin kalite kontrol kısmında, puanlama ölçeğine ilişkin ortalama ölçüm, beklenen ölçüm ve uyum dışı istatistiği yer almaktadır. Ortalama ölçümlerin monotonik artması ve uyum dışı değerlerinin 0,50 ile 1,50 arasında olması puanlama ölçeğinin uygun olduğu anlamına gelmektedir (Linacre, 2017).

Çok yüzeyli Rasch ölçme modeli, Rasch modelleri ailesine ait olduğundan, Rasch modellerindeki varsayımları karşılaması gerekmektedir (Eckes, 2015, s.124; Farrokhi, Esfandiari & Schaefer, 2012; Farrokhi vd., 2011). Çok yüzeyli Rasch ölçme modeli için karşılanması gereken varsayımlar; tek boyutluluk, yerel bağımsızlık ve model veri uyumudur. Mevcut araştırmanın yöntem bölümünde ilgili varsayımların karşılanma durumu test edilmiş ve sonuçları raporlaştırılmıştır.

İLGİLİ ARAŞTIRMALAR

Bu bölümde ilk olarak performans değerlendirme sürecinde ortaya çıkan puanlayıcı davranışlarının incelenmesine yönelik araştırmalara, daha sonra performans değerlendirmede puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisinin incelendiği araştırmalara yer verilmiştir.

Puanlayıcı Davranışları ile İlgili Araştırmalar

Choi ve Wilson (2018) tarafından genellenebilirlik ve madde tepki kuramı birleştirilerek, karmaşık ölçme durumları için yeni bir model geliştirilmiştir (Lojistik karışık model). Daha sonra geliştirilen bu model örnek bir veri seti üzerinden etkililiği araştırılmıştır. Çalışma kapsamındaki veri setinin üç özelliği bulunmaktadır. Bu özellikler; çok boyutlu olması, çok kategorili olması ve birden fazla puanlayıcıyı kapsaması olarak verilmiştir. Veri seti 869 öğrenciden ve sekiz puanlayıcıdan oluşmakta, ölçme aracı ise 19 maddeden ve 6 gizil yapıdan (faktör) oluşmaktadır. Öğrenci performansının değerlendirilmesinde 4'lü derecelendirmeye sahip dereceli puanlama anahtarı kullanılmıştır. Çalışmanın deseni, yuvalanmış bir desendir. Yapılan analizler neticesinde, lojistik karışık modelin karışık ölçme durumlarında kullanılmasının faydalı olunacağı ve hem bireysel düzeyde hem de grup düzeyinde her bir yüzey için birçok istatistik kestiriminin yapılacağı belirtilmiştir.

Primi, Silvia, Benedek ve Jauk (2018) 21. yüzyıl becerilerinden biri olan yaratıcılığın değerlendirme sürecini çok yüzeyli Rasch ölçme modelini kullanarak incelemiştir. Yaratıcılığın değerlendirilme sürecinde geçerlik ve güvenirlik üzerinde bireysel puanlayıcı

farklılıklarının etkileri araştırılmıştır. Araştırmada iki çalışma grubu üzerinde çalışılmış olup ilk grup 132, ikinci grup ise 298 bireyden oluşmaktadır. Araştırma kapsamında bireyler üç kriter (metafor üretme, ortak nesneler için alternatif kullanımlar, yaratıcı örnekler) üzerinden iki ila dört puanlayıcı tarafından 5'li derecelendirilmiş dereceli puanlama anahtarı kullanılarak değerlendirilmiştir. Yapılan analizler neticesinde, puanlayıcıların birey performansını değerlendirme sürecinde katılık ve cömertlik davranışlarını gösterdikleri bu yüzden puanlayıcı farklılıkları için modellerin yeniden kurulması gerektiği belirtilmiştir. Ayrıca, farklı puanlayıcı kombinasyonlarının kullanımının sistematik hatayı artırdığı ve geçerliği düşürdüğü bulunmuştur. Araştırmacılar performans değerlendirme sürecinde, puanlayıcı davranışlarının modele dâhil edilip belirlenmesi açısından çok yüzeyli Rasch ölçme modelinin güçlü bir araç olduğunu belirtmiştir.

Tseng, Su ve Nix (2018) tarafından, öğrencilerin yabancı dil bağlamında çeviri becerilerinin değerlendirilmesinde çok yüzeyli Rasch ölçme modeli kullanılmıştır. Araştırmanın çalışma grubu 225 lise üçüncü sınıf öğrencisi ve altı puanlayıcıdan (3'ü uzman, 3'ü acemi) oluşmaktadır. Öğrencilerin Çince'den İngilizce'ye ve İngilizce'den Çince'ye çeviri becerilerinin değerlendirilmesinde 10 görevden oluşan bir test kullanılmış olup görevlerin değerlendirilmesinde 5'li derecelendirilmiş puanlama ölçeği kullanılmıştır. 225 öğrenci rasgele numaralandırılmış ve 001 ile 099 arasındaki öğrenciler tüm puanlayıcılar tarafından, 100 ile 164 arasındaki puanlayıcılar iki uzman puanlayıcı tarafından ve 165 ile 225 arasındaki öğrenciler ise iki acemi puanlayıcı tarafından değerlendirilmiştir. Araştırma kapsamında dengeli tamamlanmamış blok desen (balanced incomplete block design) kullanılmıştır. Yapılan analizler neticesinde uzman ve acemi puanlayıcıların benzer katılık ve cömertlik düzeyine sahip oldukları ve puanlayıcılar arası güvenirliğin yüksek olduğu bulunmuştur. Puanlama deneyimi yüzeyi için elde edilen sonuçlar incelendiğinde, puanlama deneyiminin, puanlayıcılar arası güvenirlik üzerinde etkili olduğu bulunmuştur.

Youn (2018) tarafından öğrencilerin konuşma becerilerinin değerlendirilme sürecinde ortaya çıkan puanlayıcı davranışları ÇYRÖM analizi kullanılarak incelenmiştir. Araştırmanın

çalışma grubu, 102 üniversite öğrencisi (İngilizce bölümünde okuyan) ve 12 puanlayıcıdan oluşmaktadır. Tüm puanlayıcılar birbirine benzer puanlama deneyimlerine sahiptir. Araştırmada, eşleştirilmiş rol yaparak öğrenme ve rol yaparak öğrenme kartları ölçme aracı olarak kullanılmış olup öğrenci performansları puanlama kriterleri kullanılarak değerlendirilmiştir. Veri analizi çok yüzeyli Rasch ölçme modeli kullanılarak gerçekleştirilmiştir. Veri analizi sonucunda birey performansının değerlendirilmesi sürecinde puanlayıcı davranışlarının (katılık, cömertlik ve yanlılık) ortaya çıktığı bulunmuştur. Araştırmanın sonucuna göre puanlayıcılar düşük düzeydeki öğrenci performanslarını değerlendirmede zorluklar yaşamakta ve puanlayıcı performansı en çok puanlama ölçeğinin puanlaması sürecinde ortaya çıkmaktadır. Yanlılık analizleri incelendiğinde ise anlamlı bulunan etkileşimler içinde, en fazla varyansın *puanlayıcı x kriter* etkileşiminden (%52) kaynaklandığı tespit edilmiştir. Araştırmacı elde edilen bulgulara dayanarak, karar verme sürecinde puanlama ölçeğinin veya kriterlerinin puanlayıcılar üzerinde önemli bir etkiye sahip olduğunu belirtmiştir.

Çetin ve İlhan (2017) tarafından açık uçlu matematik sorularının puanlanmasında ortaya çıkan puanlayıcı davranışlarının incelenmesi amacıyla çok yüzeyli Rasch ölçme modeli kullanılmıştır. Çalışma grubunu 104 8.sınıf öğrencisi ve yedi puanlayıcı oluşturmaktadır. Araştırmacılar 10 tane açık uçlu maddeden oluşan başarı testini geliştirdikten sonra, çalışma grubunda bulunan öğrencilere uygulamıştır. Çalışmaya katılan puanlayıcılar ilk olarak herhangi bir taksonomiye göre yapılandırılmamış olan dereceli puanlama anahtarlarını kullanarak öğrenci yanıtlarını puanlandırmıştır. Daha sonra SOLO taksonomisine göre hazırlanmış olan dereceli puanlama anahtarları kullanılarak aynı puanlayıcılar öğrenci yanıtlarını tekrar değerlendirmiştir. Veri toplama süreci bittikten sonra çok yüzeyli Rasch ölçme modeli kullanılarak veri analizi yapılmıştır. Veri analizi sonucu herhangi bir taksonomiye dayandırılmadan hazırlanmış olan dereceli puanlama anahtarları öğrenci performansının değerlendirilmesinde kullanıldığında, puanlayıcılar arası uyumun düşük olduğu görülmüştür. Ayrıca puanlamada ortaya sıklıkla çıkan puanlayıcı cömertliği ve

puanlayıcı katılığı gibi puanlayıcı davranışlarının ortaya çıktığı görülmüştür. SOLO taksonomisi kullanılarak hazırlanmış olan dereceli puanlama anahtarları ile öğrenci performansının değerlendirilmesinde puanlayıcılar arası uyumun daha iyi olduğu bulunmuştur. Buna göre, belli bir taksonomiye dayandırılarak hazırlanan dereceli puanlama anahtarlarının öğrenci performansını değerlendirmede daha objektif sonuçlar verdiği söylenebilir.

Zhao, Andersson, Guo ve Xin (2017) öğrencilerin yazma becerilerinin değerlendirilme sürecinde ortaya çıkan puanlayıcı davranışlarını belirlemek için çapraz sınıflandırılmış modelleri kullanmıştır. Puanlayıcı davranışlarının belirlemesinde, 1) ardışık etkilerin varlığı, 2) ardışık etkilerin yönü ve 3) bireysel özelliklerine göre puanlayıcı farklılıkları modelleri kullanılmıştır. Çalışmanın verisi 2011 yılında Çin'de yapılan yüksek riskli bir sınavın verisinden oluşmaktadır. Analizler neticesinde, birey performasnının değerlendirilmesi sürecinde puanlayıcıların ardışık etkisi/davranışı gösterdikleri bulunmuştur. Ayrıca diğer bir puanlayıcı davranışı olan asimilasyon etkisinin(diğer puanlayıcılara benzene eğilimi) de ölçümlere karıştığı belirlenmiştir. Deneyimli, cömert ve nitelikli puanlayıcıların asimilasyon etkisine karşı daha kararlı oldukları bulunmuştur. Ayrıca büyük ölçekli ve yüksek riskli sınavlarda puanlayıcı davranışlarını belirlemek için test edilen modellerin uygun olduğu sonucuna varılmıştır.

Doğan ve Uluman (2017) tarafından derecelendirilmiş kategori derecelendirme ölçeklerinin ve dereceli puanlama anahtarlarının puanlayıcı güvenirliğine katkısının belirlenmesi amacıyla klasik test kuramı, genellenebilirlik kuramı ve çok yüzeyli Rasch ölçme modeli kullanılmıştır. Araştırmanın modeli korelasyonel araştırma türüne girmektedir. Araştırmanın çalışma grubu, 82 ilkokul öğrencisi ve 3 öğretmenden(puanlayıcı) oluşmaktadır. Öğrencilerin performanslarının değerlendirilmesi amacıyla bir performans görevi verilmiştir. Daha sonra öğrenciler rasgele iki gruba ayrılmış ve bir grup derecelendirmiş kategori puanlama ölçeğine göre diğer grup ise dereceli puanlama anahtarına göre aynı puanlayıcılar tarafından değerlendirilmiştir. Araştırmanın sonuçlarına göre,

derecelendirilmiş kategori puanlama ölçeği ile yapılan puanlamaların daha tutarlı olduğu başka bir değişle puanlayıcı güvenirliğinin daha yüksek olduğu bulunmuştur. Birey performansının değerlendirilmesi sürecinde dereceli puanlama anahtarı kullanıldığında puanlayıcıların katılık ve cömertlik düzeylerinin daha yüksek olduğu bulunmuştur. Puanlayıcılarla yapılan görüşmeler neticesinde, puanlayıcıların kısa aralıklarla hızlı puanlama yapmaları gereken durumlarda derecelendirilmiş kategori puanlama ölçeğini kullanmayı tercih ettikleri belirlenmiştir.

Goodwin (2016) öğrencilerin ikinci dil (L2) yazma becerilerinin değerlendirilmesinde ortaya çıkan puanlayıcı davranışlarının belirlenmesi için çok yüzeyli Rasch ölçme modelini kullanmıştır. Araştırmacı aynı dereceli puanlama anahtarının farklı öğrenci gruplarına uygulanması sonucu ortaya çıkan puanlayıcı davranışlarına odaklanmıştır. Araştırmada kullanılan ölçme aracı ABD'nin bir üniversitesinde hem kabul hem de yerleştirme amaçlı kullanılan akademik İngilizce'nin okuma ve yazma testiden oluşmaktadır. Ölçme aracı yedi yarıyıl dönemi boyunca uygulanmış ve her yarıyıl sonunda 43 puanlayıcı tarafından değerlendirilmiştir. Puanlayıcılara kabul veya yerleştirme amacına bakılmaksızın analitik dereceli puanlama anahtarını benzer şekilde kullanmaları için eğitim verilmiştir. Araştırmada; bireyler, puanlayıcılar, test tipi (kabul ve yerleştirme) ve kriterler olmak üzere dört yüzey bulunmaktadır. Analiz sonuçlarına göre 25 puanlayıcından beş tanesi kabul veya yerleştirme testlerinde istatistiksel olarak altı kez puanlayıcı davranışını göstermiştir. Elde edilen bulgular, puanlayıcıların yerleştirme testlerinden çok kabul testlerini daha geniş bir aralıkta puanladığını göstermektedir. Bu beklenmedik bir sonuç çünkü puanlayıcılar eğitilirken test türünün önemli olmadığı aynı yazma becerilerini değerlendirildiği belirtilmiştir. Kabul testinin daha geniş bir aralığa sahip olmasındaki temel etkenlerden biri puanlayıcıların puanlama esnasında göstermiş oldukları puanlayıcı davranışlarından kaynaklandığı belirtilmiştir.

Wu ve Tan (2016) tarafından, öğrenci kompozisyonlarının değerlendirilmesini yapmak ve birey performansına karışan puanlayıcı varyansını belirlemek için çok yüzeyli Rasch ölçme

modeli kullanılmıştır. Araştırmanın çalışma grubu 50 öğrenci ve 40 puanlayıcıdan oluşmaktadır. Araştırma, tamamen çaprazlanmış desen kullanılarak yürütülmüştür. Öğrenci kompozisyonlarının değerlendirilmesinde üç kritere ve 3'lü derecelendirmeye sahip bir dereceli puanlama anahtarı kullanılmıştır. Yapılan analizler neticesinde öğrenci kompozisyonlarının değerlendirilmesinde ölçümlere puanlayıcı davranışlarının karıştığı belirlenmiştir. Puanlayıcılar arası güvenirliğin yüksek olduğu bulunmuştur. Araştırmacılar çalışmanın sonucundan hareketle, kompozisyonların %25'inin tekrar gözden geçirilmesi gerektiğini ve puanlayıcılar arasında daha tutarlı sonuçlar için puanlayıcılara iyi bir eğitimin verilmesi gerektiğini belirtmiştir.

Esfandiari (2015) tarafından ikinci dil yazma becerisinin (kompozisyon) değerlendirme sürecinde akran değerlendirmelerinde sıklıkla ortaya çıkan iki puanlayıcı davranışını (merkeze yönelim ve halo) belirlemek için çok yüzeyli Rasch ölçme modeli kullanılmıştır. Ayrıca akranların değerlendirmelerinin anlaşılması için öğretmenlerinin yapmış olduğu puanlamalarla karşılaştırılması yapılmıştır. Araştırma, tamamen çaprazlanmış desen ile desenlenmiştir. Akran değerlendiriciler 15 kritere ve altılı derecelendirilmeye sahip analitik dereceli puanlama anahtarını kullanarak akranlarını değerlendirmiştir. Yapılan analziler neticesinde grup düzeyinde akran değerlendiricilerin merkeze yönelim ve halo davranışını göstermedikleri fakat bireysel düzeyde akran değerlendiricilerin merkeze yönelim ve halo davranışını gösterdikleri bulunmuştur. Ayrıca akran değerlendiricilerin ve öğretmenlerin puanlaması arasında istatistiksel olarak anlamlı bir farkın olmadığı bulunmuştur.

Güler (2014) tarafından yapılan araştırmada, birey performansının değerlendirilmesi sürecinde ortaya çıkan puanlayıcı davranışları çok yüzeyli Rasch ölçme modeli kullanılarak incelenmiştir. Araştırmanın çalışma grubu 55 üniversite öğrencisi ve üç puanlayıcıdan oluşmaktadır. Veri toplama aracı olarak 10 tane açık uçlu maddeden oluşan istatistik testi kullanılmıştır. Açık uçlu maddelerin değerlendirilmesinde, 11'li derecelendirilmiş dereceli puanlama anahtarı kullanılmıştır. Yapılan analizler neticesinde puanlayıcıların katılık ve cömertlik davranışlarının birbirine çok benzer olduğu ve puanlayıcı güvenirliğinin yüksek

olduğu bulunmuştur. Ayrıca 11 kategoriye sahip dereceli puanlama anahtarının sadece 3 kategorisinin tercih edildiği diğer kategorilerin çok az oranda kullanıldığı sonucundan hareketle dereceli puanlama anahtarının daha sonraki çalışmalarda üçlü derecelendirilmiş olarak kullanılmasının daha uygun olduğu belirtilmiştir.

Ok vd. (2013) tarafından yapılan araştırmada öğrencilerin öğretim elamanının performansını değerlendirme sürecinde ortaya çıkan puanlayıcı davranışlarının belirlenmesi amacıyla açımlayıcı faktör analizi ve ANOVA kullanılmıştır. Araştırmanın çalışma grubu 270 üniversite öğrencisinden ve bir öğretim üyesinden oluşmaktadır. 270 öğrenci öğretim üyesini tanımakta ve ders almaktadır. Veri toplama aracı olarak araştırmacılar tarafından geliştirilen iki ölçek (davranış odaklı değerlendirme ölçeği ve grafik değerlendirme ölçeği) kullanılmıştır. Her iki ölçme aracı da performansın 10 tane boyutunu içermektedir. Araştırmanın ulaştığı sonuçlar incelendiğinde, davranış odaklı değerlendirme ölçeğinin halo etkisine karşı daha savunmasız olduğu bulunmuştur. Diğer yandan grafik değerlendirme ölçeğinde puanlayıcı katılığı veya cömertliğinin ortaya çıkma durumunun daha yüksek olduğu tespit edilmiştir.

Farrokhi vd. (2012) tarafından öz değerlendirme, akran değerlendirme ve puanlayıcı değerlendirmelerin performans değerlendirme sürecinde ortaya çıkan farklılaşan puanlayıcı katılığı ve cömertliği davranışını belirlemek için çok yüzeyli Rasch modeli kullanılmıştır. Çalışmanın grubu 194 puanlayıcıdan (188 öğrenci, 6 öğretmen) oluşmaktadır. Burada öz değerlendirme ve akran değerlendirmeyi 188 öğrenci, puanlayıcı değerlendirmeyi ise altı öğretmen yapmıştır. Öğrencilerin yazma becerilerini değerlendirmek için 15 kritere ve altılı derecelendirmeye sahip dereceli puanlama anahtarı kullanılmıştır. Araştırmada tamamen çaprazlanmış desen kullanılmıştır. *Puanlayıcı tipi x kriter* etkileşimleri incelendiğinde (3x15 = 45 olası etkileşim), toplam 19 etkileşimin anlamlı olduğu bu etkileşimlerden 11 tanesi farklılaşan puanlayıcı katılığı olduğu ve bu puanlamaları yapanların öğretmen ve öğrencinin akranları olduğu bulunmuştur. Anlamlı etkileşimlerin sekiz tanesi ise farklılaşan puanlayıcı cömertliği olduğu ve bu puanlamaların öz değerlendirmeye ait olduğu yani

bireyin kendine verdiği puandan geldiği bulunmuştur. *Puanlayıcı tipi x birey* etkileşimleri (3x188=564) incelendiğinde, 13 etkileşimin anlamlı olduğu (yanlılık içerdiği) ve bu etkileşimlerden 9 tanesinin akran değerlendirmeden geldiği görülmektedir. Bireylerin kendi akranlarını değerlendirirken farklılaşan puanlayıcı cömertliği davranışını gösterdikleri belirlenmiştir.

Wiseman (2012) tarafından öğrencilerin yabancı dil yazma becerilerinin değerlendirilmesi sürecinde ortaya çıkan puanlayıcı davranışlarını belirlemek için çok yüzeyli Rasch ölçme modeli kullanılmıştır. Araştırmada puanlayıcılar bütün bireylerin performanslarını değerlendirmiş olup, tamamen çaprazlanmış desen kullanılmıştır. Araştırmanın çalışma grubu 39 İngilizce üçüncü sınıf öğrencisi ve 8 puanlayıcıdan oluşmaktadır. Çalışma grubundaki öğrencilerden anlatım ve ikna türünde iki kompozisyon yazmaları istenmiştir. Öğrenci kompozisyonlarının değerlendirmesini yapmak için altılı derecelendirilmiş analitik ve holistik dereceli puanlama anahtarları kullanılmıştır. Araştırma kapsamında analitik ve holistik dereceli puanlama anahtarlarının öğrenci performansının değerlendirilmesinde ne kadar başarılı olduklarını belirlemek için karşılaştırılmıştır. Araştırma sonuçlarına göre holistik dereceli puanlama anahtarını kullanan puanlayıcıların daha çok alt derecelendirmeleri veya kategorileri tercih ettikleri bulunmuştur. Diğer bir sonuç ise yazma becerilerinin değerlendirilmesinde analitik dereceli puanlama anahtarlarının daha başarılı olduğu belirlenmiştir. Ayrıca puanlayıcıların birey performansını değerlendirme sürecinde orta düzeyde cömertlik davranışını ve yüksek düzeyde katılık davranışını gösterdiği belirlenmiştir. Puanlayıcı yanlılığı davranışı incelendiğinde birey performansının değerlendirilmesi sürecinde daha çok farklılaşan puanlayıcı cömertliği davranışının ortaya çıktığı bulunmuştur.

Farrokhi ve Esfandiari (2011) tarafından yabancı dil (ikinci dil-L2) yazma becerilerinin değerlendirilmesi sürecinde ortaya çıkan puanlayıcı davranışlarından biri olan halo etkisi çok yüzeyli Rasch ölçme modeli aracığıyla incelenmiştir. Araştırma kapsamında öz değerlendirme, akran değerlendirme ve öğretmen değerlendirmenin performans

değerlendirme sürecinde halo davranışını gösterme durumları karşılaştırılmıştır. Çalışmanın grubu 194 puanlayıcıdan (188 öğrenci, 6 öğretmen) oluşmaktadır. Burada öz değerlendirme ve akran değerlendirmeyi 188 öğrenci, puanlayıcı değerlendirmeyi ise altı öğretmen yapmıştır. Öğrencilerin yazma becerilerini değerlendirmek için 15 kritere ve altılı derecelendirmeye sahip dereceli puanlama anahtarı kullanılmıştır. Araştırmada tamamen çaprazlanmış desen kullanılmıştır. Araştırmanın sonuçlarına göre grup düzeyinde puanlayıcıların halo davranışını göstermedikleri bununla birlikte bireysel düzeyde, tüm puanlayıcı tiplerinde halo davranışının ortaya çıktığı bulunmuştur. Ayrıca diğer iki puanlayıcı tipine (öğretmen değerlendirme, akran değerlendirme) göre öz değerlendirmede daha fazla halo davranışının ortaya çıktığı bulunmuştur.

Farrokhi vd. (2011) tarafından yapılan araştırmada, yabancı dil (ikinci dil-L2) yazma becerilerini değerlendirilme sürecinde ortaya çıkan puanlayıcı davranışlarından biri olan merkeze yönelim davranışı çok yüzeyli Rasch ölçme modeli aracılığıyla incelenmiştir. Araştırma kapsamında öz değerlendirme, akran değerlendirme ve öğretmen değerlendirmenin performans değerlendirme sürecinde merkeze yönelim davranışını gösterme durumları karşılaştırılmıştır. Çalışmanın grubu 194 puanlayıcıdan(188 öğrenci, 6 öğretmen) oluşmaktadır. Burada öz değerlendirme ve akran değerlendirmeyi 188 öğrenci, puanlayıcı değerlendirmeyi ise altı öğretmen yapmıştır. Öğrencilerin yazma becerilerini değerlendirmek için 15 kritere ve altılı derecelendirmeye sahip dereceli puanlama anahtarı kullanılmıştır. Araştırmanın sonuçlarına göre hem grup düzeyinde hem de bireysel düzeyde puanlayıcıların merkeze yönelim davranışını göstermedikleri bulunmuştur.

Leckie ve Baird (2011) tarafından 14 yaş öğrencileri için hazırlanan İngilizce yazma testinin operasyonel izleme sistemi (operational monitoring system) ile değerlendirilmesi sürecinde ortaya çıkan puanlayıcı davranışlarının belirlenmesi amacıyla çok düzeyli modeller kullanılmıştır. Öğrenci kompozisyonlarının (yazma testi) değerlendirilmesi için 30'lu derecelendirmeye sahip bir ölçek kullanılmıştır. Puanlayıcıların puanlamada zorluk yaşamamaları için ölçme aracı üç alt ölçeğe bölünmüştür. Öğrenci kompozisyonlarını 689

puanlayıcı değerlendirmiştir. Bu puanlayıcılardan 135'i takım lideri, 372'si deneyimli ve 182'si ise yeni puanlayıcıdır. Toplam 84 kompozisyon puanlanmıştır. Araştırma kapsamında, 1) zamanla puanlayıcı katılık etkilerindeki kayma, 2) merkeze yönelim davranışı ve 3) puanlayıcıların önceki puanlama deneyimlerinin puanlayıcı katılığı ve merkeze yönelim üzerindeki farklılıkları incelenmiştir. Araştırma sonuçlarına göre, deneyimsiz puanlayıcılar deneyimli puanlayıcılara göre daha katı görünmesine rağmen istatistiksel olarak anlamlı olmadığı bulunmuştur. Performans değerlendirme sürecinde merkeze yönelim davranışının puanlamalara karıştığı bulunmuştur. Ayrıca puanlayıcı katılığının zamanla tutarlı olmadığı başka bir değişle kararsız olduğu istatistiksel olarak anlamlı bulunmuştur.

Sudweeks vd. (2005) tarafından üniversitede öğrenim gören öğrencilerin yazma becerilerini değerlendirme sürecinde ortaya çıkan puanlayıcı davranışlarının belirlenmesi için genellebilirlik kuramı ve çok yüzeyli Rasch ölçme modeli kullanılmıştır. Araştırmanın çalışma grubu 24 üniversite öğrencisi ve 9 puanlayıcıdan oluşmaktadır. Her bir öğrenci iki kompozisyon yazdığından çalışmada 48 kompozisyon 9 puanlayıcı tarafından değerlendirilmiştir. Araştırma, tamamen çaprazlanmış desen ile desenlenmiştir. Öğrenci kompozisyonlarının değerlendirmesini yapmak için dokuzlu derecelendirmeye sahip holistik dereceli puanlama anahtarı kullanılmıştır. Puanlayıcılar 48 yazma görevini iki kez puanlamıştır. Böylece, puanlayıcılar tarafından iki farklı zaman diliminde yapılan değerlendirmeler arasındaki tutarlılık, bir puanlayıcının farklı zamanlarda yaptığı değerlendirmeler arasındaki tutarlılık ve kompozisyonlar arasındaki tutarlılık incelenmiştir. Genellenebilirlik sonuçları incelendiğinde, birey performansının değerlendirilmesinde en büyük varyansın *birey x kriter (görev)* etkilişiminden kaynaklandığı bulunmuştur. ÇYRÖM analizleri incelendiğinde ise öğrenci kompozisyonlarının değerlendirilmesi sürecinde puanlayıcı katılığı ve cömertliği davranışlarının ölçümlere karıştığı, zaman içinde puanlayıcıların puanlama davranışlarının tutarlı olduğu ve yüzey etkilişimlerinden en büyük varyansın *birey x kriter* etkileşiminden kaynaklandığı bulunmuştur. Bu sonuçlara göre

genellenebilirlik kuramı ve ÇYRÖM analizi sonucunda elde edilen bulguların uyumlu oldukları görülmüştür.

Engelhard (1994) tarafından öğrencilerin yazma becerilerinin durum belirlemesinin incelenmesi amacıyla çok yüzeyli Rasch ölçme modeli kullanılmıştır. İlgili çalışma kapsamında öğrenci performansının durum belirlemesi yapıldığında ortaya çıkan puanlayıcı etkilerinin neler olduğu araştırılmıştır. Çalışmada 1990 yılında yapılan sekizinci sınıf yazma testindeki 264 kompozisyon çalışması seçkisiz olarak seçilmiştir. Bu çalışmalar 15 puanlayıcı tarafından değerlendirilmiştir. Analizler neticesinde öğrencilerin yazma becerilerinin durum belirlemesinde puanlayıcı etkilerinin ortaya çıktığı gözlemlenmiştir. Bu etkiler; puanlayıcı katılığı, puanlayıcı cömertliği, halo etkisi, merkeze yönelim etkisi ve ranj sınırlaması etkisinin olduğu bulunmuştur. Halo etkisi ortaya çıkan puanlayıcıların analitik dereceli puanlama anahtarını kullanmaktan ziyade bütünsel dereceli puanlama anahtarlarını kullanma eğiliminde oldukları bulunmuştur. Puanlamaya katılan puanlayıcıların %80'inde merkeze yönelim etkisi ortaya çıktığı ve bu puanlayıcıların dereceli puanlama anahtarının ortadaki iki kategorisine yığıldıkları tespit edilmiştir. Öğrencilerin ham puanları incelendiğinde ranj sınırlamasının olduğu gözlemlenmiş ve bu etkinin düzeltilmiş ham puanlarda daha az etkili olduğu bulunmuştur. Ayrıca puanlayıcı katılığı ve cömertliğinin de istatistiksel olarak anlamlı olduğu gözlemlenmiştir.

Alanyazın incelendiğinde, performans değerlendirme sürecinde ölçümlere karışan puanlayıcı davranışlarının belirlenmesinde en fazla tercih edilen yaklaşımın çok yüzeyli Rasch ölçme modeli olduğu görülmektedir. Özellikle puanlayıcıların madde veya birey gibi diğer yüzeylerle ortak etkileşimlerin incelendiği araştırmalarda çok yüzeyli Rasch ölçme modelinin sıklıkla kullanıldığı bunun temel nedeninin ise çok yüzeyli Rasch ölçme modelinin diğer yöntem ve modellere göre daha detaylı bilgi vermesi ve veri yapısının doğasına daha uygun olmasından kaynakladığı düşünülmektedir. Alanyazından hareketle mevcut araştırmada performans değerlendirme sürecinde puanlamalara karışan puanlayıcı davranışlarının belirlenmesinde çok yüzeyli Rasch ölçme modeli kullanılmış olup elde

edilen varyans oranlarının (yüzdelerinin) geçerliğine kanıt sağlamak amacıyla aynı veri seti üzerinden genellenebilirlik analizi de yapılmıştır.

Puanlayıcı Eğitimi ile İlgili Araştırmalar

Bijani (2018) tarafından öğrencilerin sözlü değerlendirilmeleri sürecinde ortaya çıkan puanlayıcı davranışlarının belirlenmesi ve bu davranışların azaltılmasına yönelik puanlayıcı eğitimin verilmesi amaçlanmıştır. Araştırma karışık desenle (nitel ve nicel) desenlenmiştir. Araştırmanın çalışma grubunu 300 öğrenci ve 21 puanlayıcı oluşturmaktadır. Puanlayıcılar 300 öğrenciyi, puanlayıcı eğitiminden önce ve sonra değerlendirmiş ve puanlayıcıların algıları, tutumları ve beklentileri; anket, görüşme formları ve gözlem yoluyla değerlendirilmiştir. Nitel ve nicel veri analizlerinin bulguları, puanlayıcı eğitimi programlarının puanlayıcıların tutumları, algıları ve değerlendirmeleri üzerinde oldukça yararlı olduğunu göstermiştir. Puanlayıcı eğitimi, puanlayıcıların katılık düzeylerinin ve yanlılıklarının azaltılmasına ve tutarlılık düzeylerinin artmasına neden olmuştur. Ayrıca verilen puanlayıcı eğitimi sonucunda halo davranışının ortaya çıkma durumunun azaldığı görülmüştür. Puanlayıcı eğitimine yönelik olumlu tutuma sahip olan puanlayıcıların daha tutarlı puanlamalar yaptığı ve daha az yanlı davranışlar sergilediği bulunmuştur. Sonuç olarak daha tutarlı değerlendirmeler ve daha az yanlı puanlamalar için puanlayıcı eğitimin verilmesi gerektiği belirtilmiştir.

Knoch, Fairbairn, Myford ve Huisman (2018) tarafından yapılan araştırmada, online ve yüz yüze puanlayıcı eğitimlerinin karşılaştırılması amaçlanmıştır. Online puanlayıcı eğitimi için online eğitim platformu kullanılmıştır. Her iki eğitim programına katılan kişiler yeni/acemi puanlayıcılardır. Puanlayıcılar eğitime tabi tutulmadan önce, verilen bir yazma görevini değerlendirmiş ve puanlayıcı davranışları arasında anlamlı bir farkın olmadığı belirlenmiştir. Daha sonra 11 puanlayıcı online puanlayıcı eğitimi, 13 puanlayıcı ise yüz yüze puanlayıcı eğitimi almıştır. Öğrenci performansları (40 kişi) dört görev üzerinden değerlendirilmiştir. veri analizinde çok yüzeyli Rasch ölçme modeli kullanılmıştır. Analiz sonuçları

incelendiğinde, her iki grubun (online ve yüz yüze puanlayıcı eğitimi alanlar) puanlayıcılar arası uyumlarının istatistiksel olarak benzer oldukları bulunmuştur. Başka bir değişle her iki grubun puanlayıcı tutarlılığı birbirine benzemektedir. Hem online eğitim grubu hem de yüz yüze eğitim grubunun puanlayıcı katılık ve cömertlik düzeylerinin istatistiksel olarak birbirine benzer olduğu bulunmuştur. Araştırmanın bulgularına göre online puanlayıcı eğitiminin yüz yüze puanlayıcı eğitimi kadar etkili olduğu ve performans değerlendirmede online eğitimin kullanılabileceği belirtilmiştir.

Baird, Meadows, Leckie ve Caro (2017) tarafından yapılan araştırmada, İngiltere'deki yüksek riskli sınavlarda uygulanan puanlayıcı izleme verisi ile puanlayıcı doğruluğu incelenmiştir. Araştırmada puanlayıcı doğruluğu, çapraz sınıflandırılmış çok düzeyli modelle kestirilmiştir. Araştırma kapsamında yapılan 22 sınavdan elde edilen 5500 sınav kağıdı 110 takımdan oluşan 567 puanlayıcı (yüz yüze ve online eğitim alan) tarafından değerlendirilmiştir. Araştırmanın sonuçlarına göre puanlayıcı eğitiminin puanlayıcı doğruluğu üzerinde anlamlı bir etkiye sahip olduğu bulunmuştur.

Hauenstein ve McCusker (2017) tarafından yapılan araştırmada üç farklı puanlayıcı eğitiminin puanlayıcı doğruluğu üzerindeki etkisi incelenmiştir. Çalışmada kontrol gruplu deneysel desen kullanılmıştır. Araştırmacılar tarafından kontrol grubuna performans boyutları eğitimi, deney grubuna ise puanlayıcı hataları eğitimi, puanlayıcı değişkenliği eğitimi ve referans çerçevesi eğitimi verilmiştir. Araştırmanın çalıma grubu 89 psikoloji bölümü öğrencisinden oluşmaktadır. 24 katılımcıya referans çerçevesi eğitimi, 22 katılımcıya puanlayıcı hataları eğitimi, 22 katılımcıya puanlayıcı değişkenliği hatası ve 21 katılımcıya (kontrol grubu) performans boyuları eğitimi verilmiştir. Araştırmaya katılan öğrencilere dönem sonunda fazladan puan verilmiştir. Katılımcılar eğitim gruplarına rasgele atanmış olup toplam üç kez ölçüm alınmıştır. Bu ölçümler, ön test, son test ve geri dönütler verdikten sonraki test şeklinde olmuştur. Araştırmanın sonuçlarına göre puanlayıcı doğruluğunu artırmada en etkili eğitim desenin referans çerçevesi eğitimi, daha sonra ise puanlayıcı değişkenliği eğitiminin olduğu bulunmuştur.

Davis (2016) tarafından yapılan araştırmada, puanlayıcı eğitiminin ve puanlama deneyiminin puanlayıcı tutarlılığına katkısı incelenmiştir. Ayrıca puanlayıcı performansı ve değerlendirmede kullanılan dereceli puanlama anahtarları da dikkate alınmıştır. Araştırmanın çalışma grubu daha önce TOEFL konuşma testinde puanlama yapmayan deneyimli 20 İngilizce öğretmeninden oluşmaktadır. Araştırmada TOEFL sınavının konuşma testindeki 100 kayıt puanlanmıştır. İlk puanlama puanlayıcı eğitiminden önce yapılmış olup daha sonra aynı kayıtlar üç defa daha puanlanmıştır. Ses kayıtlarının değerlendirilmesinde altılı derecelendirilmiş dereceli puanlama anahtarı kullanılmıştır. Araştırma, tamamen çaprazlannmış desen ile desenlenmiştir. Sonuçlar incelendiğinde puanlayıcı eğitiminin ve puanlama deneyiminin puanlayıcı katılığındaki değişim üzerinde çok az etkiye sahip olduğu bulunmuştur. Puanlayıcı eğitiminin puanlayıcılar arası güvenirlik üzerindeki etkisi incelendiğinde, puanlayıcı tutarlılığında %13'lük bir artışın olduğu görülmüştür. Sonuç olarak puanlayıcı eğitiminin puanlayıcı katılığı davranışı üzerinde çok az etkili olmasına rağmen puanlayıcı tutarlılığı üzerinde anlamlı bir etkiye sahip olduğu belirlenmiştir.

Hansen, Elholm-Madsen ve Sørensen (2016) geliştirmiş oldukları puanlayıcı eğitiminin puanlama performansı üzerindeki etkisini incelemiştir. Araştırma tek gruplu ön test ve son test desen ile desenlenmiştir. Araştırmanın çalışma grubu üç rehabilitasyon biriminde çalışan 81 uzmandan oluşmaktadır. Araştırmada 36 maddeden oluşan bir ölçme aracı üçlü derecelenmeye sahip bir puanlama ölçeği kullanılarak değerlendirilmiştir. Araştırma kapsamında kullanılan puanlayıcı eğitimi, dört puanlayıcı desenin birleştirilmesi ile oluşturulmuştur. Puanlayıcı eğitimi iki gün sürmüş olup ilk gün; puanlayıcı hataları eğitimi, performans boyutları eğitimi ve davranış gözlem eğitimi verilmiştir. Puanlayıcı eğitiminin ikinci günü ise referans çerçevesi eğitimi verilmiştir. Araştırmanın sonuçlarına göre puanlayıcı eğitiminin puanlayıcıların puanlama performanslarını artırdığı bulunmuştur.

Moser vd. (2016) tarafından referans çerçevesi eğitimi deseninde punlayıcı doğruluğunu artıran faktörlerin neler olduğu incelenmiştir. Araştırmanın çalışma grubu 60 lisansüstü

öğrencisinden oluşmaktadır. Eğitime katılım gönüllük esasına dayalı yapılmış olup katılan öğrencilere ekstra puan verilmiştir. Ön test analizleri sonucunda iki katılımcının uç değerlere (uyumsuz uygunluk istatistikleri) sahip olduğu belirlenmiş ve bu iki katılımcı çalışmadan çıkarılmıştır. Araştırma çift ön test ve tek son test deseni ile desenlenmiştir. İlk olarak performans boyutları ile ilgili sunum yapılmış ve ilk öntest yapılmıştır. Daha sonra referans çerçevesi eğitimi verilmiş ve ikinci ön test yapılmıştır. Referans çerçevesi eğitimi yaklaşık olarak 90 dakika sürmüştür. Son olarak daha özel boyutlara yönelik referans çerçevesi eğitimi verilmiş ve son test yapılmıştır. Referans çerçevesi eğitiminin etkililiğini belirlemek için boylamsal veri setleri için doğrusal karışık modeller kullanılmıştır. Sonuçlar incelendiğinde hem genel amaçlara yönelik hem de özel amaçlar doğrultusunda kullanılan referans çerçevesi eğitiminin puanlayıcı doğruluğu üzerinde etkili olduğu bulunmuştur.

Pufpaff, Clarke ve Jones (2015) tarafından hizmet öncesi öğretmen adayların performanslarının değerlendirilme süreci üzerinde puanlayıcı eğitiminin etkisi araştırılmıştır. Araştırmanın çalışma grubunu 10 puanlayıcı (üç uzman, altı deneyimli bir acemi puanlayıcı) oluşturmaktadır. Araştırma üç aşamalı olarak gerçekleştirilmiştir. İlk olarak puanlayıcı eğitimi öncesi bir eğitim yapılmıştır. Bu aşamada katılımcılardan iki elektronik portfolyoyu, iki durum çalışmasını ve iki makaleyi altı kriter üzerinden ve herhangi bir derecelendirme yapmadan değerlendirmeleri istenmiştir. Ön test öncesi sonuçlara göre puanlayıcılar arası uyumun düşük olduğu bulunmuştur. Daha sonra puanlayıcı eğitimi verilmiş olup bu eğitim kapsamında puanlayıcılara dereceli puanlama anahtarı (dörtlü derecelendirilmiş), dereceli puanlama anahtarının tanımlanması için sunumlar ve el bildirileri verilmiştir. Araştırmanın sonuçlarına göre, verilen puanlayıcı eğitimi sonrası puanlayıcılar arası uyumun çok az artığı bulunmuştur.

Raczynski, Cohen, Engelhard ve Lu (2015) tarafından geniş ölçekli değerlendirme uygulamalarında kullanılan referans çerçevesi eğitiminin iki türü olan işbirlikli ve kendi kendine tempo (self-paced) yöntemlerinin etkililiği incelenmiştir. Araştırma kapsamında geniş ölçekli uygulaması olan bir yazma testi kullanılmıştır. Yazma testi dört kriterden ve

beşli derecelendirmeye sahip bir dereceli puanlama anahtarı kullanılarak değerlendirilmiştir. Araştırmanın çalışma grubunu 66 puanlayıcı oluşturmaktadır. Her iki eğitim yönteminde de aynı eğitim materyalleri kullanılmıştır. Araştırma tamamen çaprazlanmış desen ile desenlenmiştir. Araştırmada genelleştirilmiş doğrusal karışık modeller kullanılarak veri analizi gerçekleştirilmiştir. Araştırmanın sonuçlarına göre her iki puanlayıcı eğitim yöntemi arasında puanlayıcı doğruluğu bağlamında anlamlı bir farkın olmadığı bulunmuştur. Bu sonuç geniş ölçekli performans değerlendirme uygulamalarında her iki yöntemin de kullanılabileceğini göstermektedir.

Weitz vd., (2014) tarafından yapılan araştırmada, fiziksel bir beceri testinin değerlendirilmesinde puanlayıcı eğitiminin puanlayıcı doğruluğu üzerindeki etkisi incelenmiştir. Araştırmada 21 puanlayıcı, 242 üçüncü sınıf tıp öğrencinin performansını değerlendirmiştir. Daha sonra 21 puanlayıcıdan rasgele 11 puanlayıcı seçilip iki gruba ayrılarak puanlayıcı eğitimi verilmiştir. Daha sonra puanlayıcılar tekrar aynı öğrencileri değerlendirmiştir. Araştırmanın sonuçlarına göre puanlayıcı eğitiminin puanlayıcı doğruluğu üzerinde etkili olmadığı bulunmuştur. Ayrıca puanlayıcı eğitimi alan puanlayıcıların eğitim almayan puanlayıcılara göre anlamlı bir şeklide daha katı oldukları bulunmuştur. Anketler eğitimin yapıldığı grupta halo etkisinin farkındalığının daha yüksek olduğunu göstermiştir.

Fahim ve Bijani (2011) tarafından yapılan araştırmada, ikinci dil yazma becerilerinin değerlendirilmesi sürecinde puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisi incelenmiştir. Araştırmada yarı deneysel desen kullanılmıştır. Araştırmanın çalışma grubu 12 puanlayıcıdan oluşmaktadır. Araştırma kapsamında 60 öğrencinin yazmış oldukları kompozisyonlar değerlendirilmiştir. Kompozisyonların değerlendirilmesinde IELTS tarafından hazırlanmış olan puanlama ölçeği kullanılmıştır. Puanlayıcı eğitimi öncesi 60 kompozisyondan 45 tanesi puanlayıcılar tarafından değerlendirilmiştir. Puanlayıcı eğitimi verildikten sonra puanlayıcılar kalan 15 kompozisyonu değerlendirmiştir. Burada ön testte kullanılan 45 kompozisyon ile 15 kompozisyonun denk olması sağlanmış ve böylelikle

verilen eğitime kompozisyonlardan gelecek varyansın azaltılması sağlanmıştır. Araştırmanın sonuçlarına göre, verilen puanlayıcı eğitimi puanlayıcı katılıklarının dağılımını fark edilir bir şekilde azaltmıştır. Başka bir değişle puanlayıcı eğitimi, puanlayıcı tutarlılığı üzerinde olumlu bir etkiye sahiptir. Ayrıca *puanlayıcı x kriter* etkileşimleri incelenmiş ve puanlayıcı eğitimi öncesi anlamlı olan etkileşimlerin (yanlı puanlamaların) puanlayıcı eğitimi sonrası ortaya çıkmadığı görülmüştür.

Joe, Harmes ve Hickerson (2011) tarafından yapılan araştırmada, üniversite öğrencilerinin sözlü iletişim becerilerinin değerlendirilmesinde deneyimli ve deneyimsiz puanlayıcıların dikkat ve öğrencileri sınıflandırmadaki farklılıkları incelenmiştir. Araştırma iki gruptan oluşmakta, deneyimli puanlayıcılar belirlenen kriterlere göre diğer grup ise koşulsuz seçilmiştir. Her iki grubun sözlü iletişim becerisini değerlendirmede benzer oldukları varsayılmıştır. Sekiz deneyimli sekiz de deneyimsiz olmak üzere toplam 16 puanlayıcı çalışmaya katılmıştır. Çalışmada altı ile sekiz dakika uzunluğunda olan sekiz video kaydı kullanılmıştır. Video kayıtlarının değerlendirilmesinde, üçlü derecelendirmeye ve 39 alt kritere sahip analitik dereceli puanlama anahtarı kullanılmıştır. Puanlayıcı eğitimi her iki puanlayıcı grubuna da verilmiş olup puanlayıcı eğitiminin puanlayıcının deneyimi ve zihinsel yeterliği üzerindeki etkisi incelenmiştir. Araştırma tamamen çaprazlanmış desen ile desenlenmiştir. Araştırmanın sonuçlarına göre her iki grubun değerlendirmeleri arasında anlamlı bir fark bulunamamıştır.

Kondo (2010) tarafından öğrencilerin ikinci dil becerilerinin değerlendirilmesi sürecinde puanlayıcı eğitiminin etkililiği incelenmiştir. Puanlayıcıların değerlendirme sürecinde, değerlendirme maddelerinin kriterlerini açıkça anlayabilmeleri için puanlayıcı eğitimi verilmiştir. Puanlayıcı eğitimi Avrupa Ortak Dil Kriterleri Çerçeve Programı temel alınarak yapılmıştır. Puanlayıcı eğitiminde katılımcılara video izletirip Avrupa Ortak Dil Kriterleri Çerçeve Programı kapsamında belirlenen her seviyedeki öğrenci özellikleri üzerine grup tartışmaları yapılmıştır. Araştırmanın çalışma grubunu 73 puanlayıcı oluşturmaktadır. Öğrencilerin ikinci dil becerileri 24 maddeden oluşan ve altılı likert tipi bir ölçek kullanılarak

değerlendirilmiştir. Araştırmada genellenebilirlik kuramı ve çok yüzeyli Rasch ölçme modeli kullanılarak veri seti analiz edilmiştir. Genellenebilirlik sonuçları incelendiğinde, puanlayıcı eğitimi sonrası genellenebilirlik katsayısının yükseldiği ve madde yüzeyine ilişkin varyansın altıda bir oranında azaldığı görülmüştür. ÇYRÖM analizleri incelendiğinde ise eğitim öncesi ve sonrası performans değerlendirmede anlamlı bir farklılığın ortaya çıkmadığı bulunmuştur.

Wang (2010) tarafından puanlayıcı eğitimi ve puanlayıcı doğruluğu üzerine yapılan derleme çalışmasına göre, puanlayıcı faktörleri üzerine yapılan çalışmalar sayesinde; puanlayıcı değişkenliği, puanlayıcı yanlılığı ve puanlayıcı eğitimini daha iyi anlayıp anlamlı çıkarımların yapılacağını belirtmiştir. Puanlayıcı eğitimi, puanlayıcı yanlılıklarını ortadan kaldırmaz, fakat puanlayıcıların daha tutarlı puanlama yapmalarını sağlar. Yapılan çalışmalar incelendiğinde, puanlayıcı eğitimi sonuçlarının bir eğitim dönemine kadar dayanamayacağı görülmüştür. Bu nedenle, her bir test uygulaması öncesi puanlayıcıların yeniden puanlama eğitimine tabi tutulmaları ve böylece puanlayıcının puanlama yaparken diğer puanlayıcılarla ortak bir ölçütler seti oluşturmasına olanak sağlanacaktır. Sonuç olarak puanlayıcılar hizmet öncesi eğitim, hizmet içi eğitim ve görevler üzerine pilot eğitimi alma ihtiyacı içinde oldukları görülmektedir. Başka bir değişle, puanlayıcı eğitimi bir seferlik değil süreklilik arz eden bir iş olmalıdır.

Wolfe ve McVay (2010) tarafından yapılan araştırmada, puanlayıcı eğitiminin daha önce eğitim almış puanlayıcılarda ortaya çıkan puanlayıcı davranışları üzerindeki etkisi incelenmiştir. Araştırmanın çalışma grubunu 400 öğrenci ve 120 puanlayıcı oluşturmaktadır. Puanlayıcılar, 400 öğrenci tarafından yazılan kompozisyonları dörtlü derecelendirilmiş bir puanlama ölçeği kullanarak değerlendirmiştir. Puanlayıcıların hepsi daha önce kompozisyon değerlendirmesini yapma deneyimine sahiptir. Puanlayıcılar 40 kişilik üç gruba yarılmış ve her gruba farklı bir puanlayıcı eğitimi verilmiştir. Araştırmanın sonuçlarına göre verilen puanlayıcı eğitimlerinin puanlayıcı tutarlılığı üzerinde olumlu bir etkiye sahip olduğu

bulunmuştur. Ayrıca online dağıtılmış(online-distributed) puanlayıcı eğitiminin diğer iki eğitimden daha etkili olduğu bulunmuştur.

May (2008) tarafından yapılan araştırmada, akran değerlendirilmesinde sosyal stilli yanlılığın azaltılmasında puanlayıcı eğitiminin etkisi araştırılmıştır. Araştırmanın çalışma grubu işletme fakültesinde okuyan 144 (75 deney ve 69 kontrol) lisans öğrencisinden oluşmaktadır. Araştırma, yarı deneysel denk olmayan kontrol gruplu desen ile desenlenmiştir. Araştırmanın amacına uygun puanlayıcı eğitimini oluşturmak için alanyazında rapor edilen puanlayıcı eğitimlerinden puanlayıcı hatası eğitimi ve performans boyutları eğitimi birleştirilmiştir. Öğrencilerin sosyal stil tercihlerini belirlemek için 10'lu derecelendirilmiş likert tipi bir ölçek kullanılmıştır. Ölçek puanlarına göre katılımcılar üç gruba ayrımıştır. Bu gruplar; katılımcıya benzeyen, katılımcıya yakın ve katılımcının zıttı şeklindedir. Araştırmanın sonuçlarına göre puanlayıcı eğitimi alan gruptaki(deney) puanlayıcılar, kendisine benzeyen, kendisine yakın olan ve kendisinin zıttı olan akranlarını değerlendirirken yanlılık içermeyen ve birbirine yakın puanlamalar yapmakta iken, puanlayıcı eğitimine tabi tutulmayan gruptaki (kontrol) puanlayıcılar ise kendisine benzeyen kişilere yüksek puan kendisinin zıttı olan kişilere ise daha düşük puan vererek yanlılık içeren puanlamalar yaptığı görülmüştür.

Knoch vd. (2007) yüz yüze ve online puanlayıcı eğitimlerinin puanlayıcı davranışları üzerindeki etkililiklerinin incelenmesi amacıyla yaptıkları araştırmada, 16 puanlayıcıyı iki gruba ayırarak bir gruba online puanlayıcı eğitimi diğerine ise yüz yüze puanlayıcı eğitimi verilmiştir. Veri analizi için çok yüzeyli Rasch ölçme modeli kullanılmıştır. Veri analizi sonucunda istatistiksel olarak hem yüz yüze puanlayıcı eğitiminin hem de online puanlayıcı eğitiminin etkili olduğu bulunmuştur. Bireysel puanlayıcı yanlılıkları araştırıldığında, puanlayıcı eğitiminden önce ortaya çıkan puanlayıcı yanlılıklarının yüz yüze puanlayıcı eğitimi sonrasında azaldıkları görülmüştür. Diğer yandan online puanlayıcı eğitiminin puanlayıcı katılığı ve cömertliği etkilerini azalttığı görülmüştür. Fakat yüz yüze yapılan puanlayıcı eğitiminin daha etkili olduğu sonucuna ulaşılmıştır.

Congdon ve McQueen (2000) geniş puanlama periyodu boyunca puanlayıcı katılığının sürekliliğini incelemek amacıyla, 8.285 ilkokul öğrencisinin yazma performanslarını puanlayan 16 puanlayıcı ÇYRÖM aracılığıyla analiz edilmiştir. Her bir bireyin yazma performansı iki farklı puanlayıcı tarafından puanlanmış ve tüm puanlama süreci yedi gün sürmüştür. İlk gün puanlanan bir performans son gün tekrar puanlanmıştır. Puanlama süreci boyunca puanlayıcıların aynı gün ve farklı günleri karşılaştırıldığında istatistiksel olarak manidar bir farkın olduğu gözlenmiştir. 16 puanlayıcıdan 10'unun ilk gün ile son gün yaptıkları puanlama arasında önemli ölçüde farklılıkların olduğu belirlenmiştir. Elde edilen sonuca göre araştırmacılar performansın durum belirlemesi yapılmadan önce puanlayıcıların puanlama eğitimine tabi tutulmalarını ve bunun sürekli olarak tekrarlanması gerektiğini önermiştirler.

Weigle (1998) tarafından hem puanlayıcı eğitimi öncesinde hem de puanlayıcı eğitimi sonrasında deneyimli ve deneyimsiz puanlayıcıların puanlayıcı katılığı ve tutarlılığı incelenmiştir. 16 puanlayıcı (sekiz deneyimli ve sekiz deneyimsiz) merkezi olarak uygulanan bir İngilizce sınavından alınmış olan 60 tane kompozisyon tipindeki öğrenci performanslarını değerlendirmiştir. Birey performansları değerlendirilirken; içerik, dil ve retorik kontrol olmak üzere üç bölümden oluşan bir ölçme aracı kullanılmıştır. Veri analizi için çok yüzeyli Rasch ölçme modeli kullanılmıştır. Yapılan analizler neticesinde puanlayıcı eğitimi öncesinde deneyimsiz puanlayıcıların deneyimli puanlayıcılara göre daha katı puanlama yaptıkları ve daha tutarsız oldukları bulunmuştur. Puanlayıcı eğitiminden sonra, iki puanlayıcı grubu arasındaki farklılıkların azaldığı, ancak birçok puanlayıcının tutarlılığı artmasına rağmen puanlayıcılar arasında puanlayıcı katılığında hala anlamlı farklılıkların olduğu görülmüştür. Bu sonuçlar puanlayıcı eğitiminin daha kesitirlebilir puanları vermek için puanlayıcılara yardımcı olduğunu göstermektedir.

BÖLÜM 3

YÖNTEM

Yöntem bölümünde; araştırmanın deseni, çalışma grubu, veri toplama araçları, puanlayıcı eğitimi (deneysel işlem) ve veri analizi ile ilgili bilgiler sunulmuştur.

Araştırmanın Deseni

Bu araştırmada, puanlayıcı eğitiminin performans değerlendirme sürecinde ortaya çıkan puanlayıcı davranışları üzerindeki etkisinin incelenmesi amaçlandığı için deney ve kontrol gruplu ön ve son test uygulamalı yarı deneysel desen ile desenlenmiştir. Ön test son test kontrol gruplu desen yaygın bir kullanıma sahiptir. Bu desende katılımcılar deneysel işlem uygulanmadan önce ve deneysel işlem uygulandıktan sonra aynı bağımlı değişkenle ilgili olarak ölçülürler (Büyüköztürk, 2011, s.19). Bu desen ilişkili bir desendir. Çünkü aynı kişiler bağımlı değişken üzerinde iki kez ölçülürler. Bununla birlikte, farklı deneklerden oluşan deney ve kontrol gruplarının ölçümlerinin karşılaştırılmasından dolayı bir yönüyle de ilişkisiz bir desen olarak tanımlanmaktadır (Howitt & Cramer, 2008, s.256). Bu iki özelliğinden dolayı ön test son test kontrol gruplu desen karışık desen olarak ifade edilir (Büyüköztürk, 2011, s.19).

Çalışma Grubu

Araştırmanın çalışma grubu İngilizce öğretmen adayı olan 64 puanlayıcıdan oluşmaktadır. Puanlayıcıların yaş ortalaması 21,84 olup daha önce herhangi bir puanlayıcı eğitimine katılmadıkları ve puanlama deneyimlerinin aynı oldukları belirlenmiştir. Araştırmaya katılan 64 puanlayıcı iki şubeden oluştuğu için seçkisiz olmayan bir şeklide iki gruba (33 kontrol ve 31 deney) ayrılmıştır. Araştırmaya katılım gönüllülük esasına dayalı olarak yapılmıştır. Bu nedenle ön teste katılmasına rağmen son teste katılmayan yedi puanlayıcı çalışma kapsamı dışında tutulmuştur. Kalan 57 puanlayıcının (31 kontrol ve 26 deney) ön test puanlamaları çok yüzeyli Rasch ölçme modeli ile analiz edilmiş ve uç değerlere (kötü uygunluk değerlerine) sahip olan 12 puanlayıcının (8 kontrol ve 4 deney grubu olmak üzere), model-veri uyumu üzerinde ciddi bir etkisinin olduğu göz önüne alınarak analiz dışı bırakılmalarına karar verilmiştir. Sonuç olarak kontrol grubundan 23 ve deney grubundan 22 puanlayıcı olmak üzere toplam 45 puanlayıcı üzerinden ilgili araştırma gerçekleştirilmiştir. 45 puanlayıcı hem ön testte hem de son testte 39 öğrenci kompozisyonunu 16 kriter üzerinden değerlendirmiştir.

Veri Toplama Araçları

Veri toplama aracı olarak, yazma görevi (kompozisyon), analitik dereceli puanlama anahtarı ve kişisel bilgi formu kullanılmıştır.

Yazma Görevi

İngilizce bölümü hazırlık öğrencilerinin akademik yazma becerilerinin ölçülmesi amacıyla, Uluslararası İngilizce Dil Test Sistemi (International English Language Testing System, IELTS) tarafından örnek olarak yayınlanan öğrencilerin akademik yazma performanslarını sergiledikleri düşünce temelli (opinion-based) deneme görevi (essay task) (Bakınız EK 5) bu çalışma için tercih edilmiştir (IELTS, 2018). Bu yazma görevinin seçilmesindeki önemli

etkenlerden biri görevin gerçek yaşam durumunu içermesidir. Yazma görevi öğrencilere uygulanmadan önce EK 1'de verilen Gazi Üniversitesi Etik Komisyonu'ndan gerekli izinler alınmıştır. Öğrencilere/bireylere yazma görevinin herhangi bir not verme amacıyla değil akademik araştırma için kullanılacağı, gönüllülük esasına dayalı olduğu ve kişisel hiçbir bilgilerini yazmamaları belirtilmiştir. Yazma görevi için öğrencilerden 40 dakikalık sürede en az 250 kelime içeren bir kompozisyon yazmaları istenmiştir. Yazma görevi araştırmacı tarafından uygulanmış ve toplam 39 öğrenci araştırmaya katılmıştır. Bu öğrenciler B1 seviyesinin üstünde ve "*ileri okuma ve yazma*" dersini almaktadırlar. Öğrenci kompozisyonları toplandıktan sonra birden başlanarak numaralandırılmıştır. Daha sonra öğrenci kompozisyonları çoğaltılmıştır. Öğrenci kompozisyonları 2017-2018 eğitim yılının bahar döneminde yazdırılmıştır.

Analitik Dereceli Puanlama Anahtarı

İngilizce bölümü hazırlık öğrencilerinin yazdıkları kompozisyonlarının değerlendirilmesi amacıyla araştırmacı ve İngilizce bölümünde doktora öğrenimine devam eden ve akademik yazma konusunda deneyime sahip olan bir akademisyen tarafından analitik dereceli puanlama anahtarı (ADPA) geliştirilmiştir.

Dereceli puanlama anahtarları geliştirilirken belli aşamaları içeren sistematik bir süreç takip edilmelidir. Dereceli puanlama anahtarı sistematik bir düzen izlenmeden hazırlanırsa ölçümlerin geçerliği ve güvenirliği olumsuz etkilenebilir. Dereceli puanlama anahtarının geliştirilme sürecinde geçerlik ve güvenirliğin dikkate alınması önem arz etmektedir (Moskal, 2000). Goodrich (1997), Haladyna (1997), Kutlu vd. (2014) ve Moskal'in (2000) önerileri dikkate alındığında, dereceli puanlama anahtarının geliştirme aşamaları aşağıdaki gibi verilebilir:

1. Dereceli puanlama anahtarının amacının belirlenmesi,
2. Durum belirlemesi yapılacak performansa göre kriterlerin/ölçütlerin belirlenmesi,

3. Dereceli puanlama anahtarı taslağının oluşturulması,
4. Uzmanlara gönderilerek düzeltmelerin yapılması,
5. Belirlenen kriterlerin kaç düzeye sahip olacağına karar verilmesi,
6. Amaca uygun olan dereceli puanlama anahtarının (analitik, bütünsel) seçilmesi,
7. Katılımcılara, değerlendirilmeye tabi tutulacakları kriterler bildirilerek açıklanması,

Mevcut çalışmanın amaçlarından biri öğrencilerin akademik yazma becerilerinin değerlendirilmesi olduğundan, dereceli puanlama anahtarının amacı akademik yazma becerisinin değerlendirilmesinin yapılmasıdır. Başka bir değişle ölçülecek ve durum belirlemesi yapılacak olan performans, akademik yazma becerisi olarak belirlenmiştir. Durum belirlemesi yapılacak olan performans belirlendikten sonra, ilgili performansın göstergeleri olacak olan kriterler belirlenmiştir. Kriterlerin belirlenmesinde başlıca Weigle (2002), Hughes (2003), Brown (2004), Brown (2007) ve Brookhart (2013) gibi eserler ve eserlerde yer verilen örnek dereceli puanlama anahtarları detaylı olarak incelenmiştir. Buna ek olarak IELTS tarafından akademik yazma görevinin puanlanması için geliştirilen dereceli puanlama anahtarı da gözden geçirilmiştir. İncelenen dereceli puanlama anahtarları doğrultusunda çalışmaya dâhil edilecek kriterler ve alt kriterler belirlenerek taslak form hazırlanmıştır. Hazırlanmış olan taslak form alan uzmanı olan 11 kişinin görüşüne sunulmuştur. Uzmanlara, akademik yazma becerisinin ölçülmesinde ilgili kriter için; (1) gerekli, (2) gerekli ama düzeltilmeli ve (3) gereksiz şeklinde üçlü derecelendirmeye sahip bir ölçme aracı kullanılarak kriterlerin değerlendirilmesi istenmiştir. Araştırmada, katılımcıların akademik yazma yeterlikleri dikkate alındığından, uzman görüşleri kapsam geçerliği olarak ele alınmıştır. Kriterlerin kapsam geçerliği belirlenirken, Lawshe (1975) yaklaşımı dikkate alınmıştır. Lawshe yaklaşımda kapsam geçerlik oranları belirlenirken aşağıdaki denklem kullanılmaktadır;

$$KGO_i = \frac{N_G - \frac{N}{2}}{\frac{N}{2}} \qquad (3.1)$$

Burada KGO_i i. ölçütün veya maddenin kapsam geçerlik oranını, N_G, i. kriter için "gerekli" diyen uzman sayısını, N ise toplam uzman sayısını göstermektedir. KGO değeri ±1 aralığında değerler almaktadır. KGO değerine göre kriter/madde seçimi için eşik/minimum değerler Tablo 2'de verilmiştir.

Tablo 2

KGO Değerlerine Göre Kriter Seçiminde Eşik Değerler

	Tek yönlü(kuyruklu) hipotezler için anlamlılık düzeyleri					
	,10	,05	,025	,01	,005	,001
	Çift yönlü(kuyruklu) hipotezler için anlamlılık düzeyleri					
Uzman Sayısı (N)	,20	,10	,05	,02	,01	,002
5	0,573	0,736	0,877	0,990	0,990	0,990
6	0,523	0,672	0,800	0,950	0,990	0,990
7	0,485	0,622	0,741	0,879	0,974	0,990
8	0,453	0,582	0,693	0,822	0,911	0,990
9	0,427	0,548	0,653	0,775	0,859	0,990
10	0,405	0,520	0,620	0,736	0,815	0,977
11	0,387	0,496	0,591	0,701	0,777	0,932
12	0,370	0,475	0,566	0,671	0,744	0,892
13	0,356	0,456	0,544	0,645	0,714	0,857
14	0,343	0,440	0,524	0,622	0,688	0,826
15	0,331	0,425	0,506	0,601	0,665	0,798
20	0,287	0,368	0,438	0,520	0,576	0,691
25	0,256	0,329	0,392	0,465	0,515	0,618
30	0,234	0,300	0,358	0,425	0,470	0,564
35	0,217	0,278	0,331	0,393	0,435	0,522
40	0,203	0,260	0,310	0,368	0,407	0,489

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, 45(3), 197-210. https://doi.org/10.1177/0748175612440286

Tablo 2 incelendiğinde, örneğin 20 uzmanın görüş belirttiği bir kriter için KGO değeri 0,05 anlamlılık düzeyi için 0,438 değerine eşit veya daha büyük ise kriterin kapsam geçerliğine sahip olduğu ya da ilgili yapıyı ölçmede gerekli bir kriter olduğuna karar verilip ölçme aracına dâhil edilmektedir. Mevcut araştırmada ölçme aracının taslak formu için 11 alan uzmanından görüş alındığından, 0,05 anlamlılık düzeyi için eşik/minimum değer 0,591 olduğu yukarıdaki tablodan görülmektedir. Buna göre ölçme aracının taslak formunda yer alan kriterlerden KGO değerleri 0,591 değerine eşit veya daha büyük olan kriterler ölçme

aracının esas formuna dâhil edilmiştir. Taslak formda yer alan temel ve alt kriterlere ilişkin kapsam geçerlik oranları Tablo 3'de verilmiştir.

Tablo 3

Ölçme Aracının Taslak Formunda Bulunan Kriterlere İlişkin Hesaplanan KGO Değerleri

			İlgili yapıyı ölçmede kapsam açısından bu kriter			
Kriterler	Temel kriterler	Alt kriter(ler)	Gerekli	Yetersiz	Gereksiz	KGO
1.Kriter	Organization	Title of Essay	10	1	0	0,818
		Introduction-Body-Conclusion	10	1	0	0,818
		Thesis Statement	11	0	0	1,000
		Topic & Supporting Sentences	7	4	0	0,273
		Appropriate Length	9	1	1	0,636
2. Kriter	Content	Topic Relevance	11	0	0	1,000
		Argumentation	8	3	0	0,455
3. Kriter	Vocabulary	Word Choice	11	0	0	1,000
		Lexical Range	9	1	1	0,636
		Word Register	7	1	3	0,273
4. Kriter	Grammar	Accuracy of Grammatical Forms	9	2	0	0,636
		Syntatic Complexity	11	0	0	1,000
5. Kriter	Coherence	Coherence	9	1	1	0,636
6. Kriter	Cohesion	Linking	10	0	1	0,818
		Flow of Sentences	7	3	1	0,273
7. Kriter	Mechanics	Spelling	11	0	0	1,000
		Punctuation	11	0	0	1,000
		Capitalization	7	1	3	0,273
		Appearance	7	1	3	0,273
Kapsam Geçerlik İndeksi (KGİ)						0,750

Not. Analitik dereceli puanlama anahtarının Türkçe formu Ek 10'da verilmiştir.

Tablo 3 incelendiğinde, 19 alt kriterden 6 kriterin KGO değerinin eşik değerinden düşük olduğu bulunmuştur. Bu kriterlerden "*Topic & Supporting Sentences*" kriteri alan uzmanların geri dönütleri sonucunda "*Topic Sentence*" ve "*Supporting Sentence*" şeklinde iki alt kritere bölünerek ölçme aracına dâhil edilmiştir. Ayrıca alan uzmanlarının görüşleri doğrultusunda "*Argumentation*" kriterinin ismi "*Idea Development*" şeklinde değiştirilmiş ve ölçme aracına dâhil edilmiştir. Alan uzmanların geri dönütleri sonucunda yeniden düzenlenen ölçme aracı tekrar alan uzmanlarına gönderilerek görüş alınmıştır. Alan

uzmanları tarafından revize edilmiş formun uygun olduğuna karar verilmiştir. Revize edilen formun son hali ve puanlanması Tablo 4'de verilmiştir.

Tablo 4

Ölçme Aracına Dâhil Edilen Kriterler ve Puanlanması

Kriterler		Puanlama					
Temel kriterler	Alt Kriterler	0 puan	1 puan	2 puan	3 puan	4 puan	Toplam puan
Organization	Title of Essay	( )	( )	( )	( )	( )	( )
	Introduction-Body-Conclusion	( )	( )	( )	( )	( )	( )
	Thesis Statement	( )	( )	( )	( )	( )	( )
	Topic Sentence	( )	( )	( )	( )	( )	( )
	Supporting Sentence	( )	( )	( )	( )	( )	( )
	Appropriate Length	( )	( )	( )	( )	( )	( )
Content	Topic Relevance	( )	( )	( )	( )	( )	( )
	Idea Development	( )	( )	( )	( )	( )	( )
Coherence	Coherence	( )	( )	( )	( )	( )	( )
Cohesion	Linking	( )	( )	( )	( )	( )	( )
Grammar	Accuracy of Grammatical Forms	( )	( )	( )	( )	( )	( )
	Syntatic Complexity	( )	( )	( )	( )	( )	( )
Vocabulary	Word Choice	( )	( )	( )	( )	( )	( )
	Lexical Range	( )	( )	( )	( )	( )	( )
Mechanics	Spelling	( )	( )	( )	( )	( )	( )
	Punctuation	( )	( )	( )	( )	( )	( )

Yazma görevine katılan öğrencilere belirtilmiş olmasına rağmen büyük çoğunluğu yazmış oldukları kompozisyonlarına başlık vermediklerinden, Tablo 4'deki alt ölçütlerden "*Title of Essay*" ölçütü veri yapısını bozduğundan analizler yapılırken kapsam dışı bırakılmıştır. Örneğin 16 numaralı öğrencinin yazmış olduğu kompozisyon "*Title of Essay*" kriteri dışındaki kriterlerden en yüksek puan alıyor iken, kompozisyonun başlığı olmadığı için "*Title of Essay*" kriterinden sıfır puan almaktadır. Bu yüzden ilk kriter analizlerde kapsam dışı bırakılmıştır.

Ölçme aracında bulunan ölçütlere ilişkin kapsam geçerlik oranları hesaplandıktan sonra ölçme aracının bütünü için kapsam geçerlik indeksi (KGİ) elde edilir (Lawshe, 1975). Kapsam geçerlik indeksi, yapı geçerliği sürecinin ön koşuludur (Lawshe, 1985). KGİ, esas uygulamada kullanılacak ölçme aracına dahil edilen ölçütlerin/maddelerin KGO ortalamaları kullanılarak bulunur. Buna göre ölçme aracının KGİ değeri aşağıdaki denklem ile bulunur.

$$\text{KGİ} = \frac{\sum_{i}^{m} \text{KGO}}{m} \qquad (3.2)$$

Burada, m nihai/esas ölçme aracına dâhil edilen ölçüt/madde sayısını göstermektedir. İlgili ölçme aracı için (3.2) numaralı denklem kullanılarak KGİ değeri 0,750 olarak bulunmuştur. Elde edilen bu değer 0,591 değerinden büyük olduğundan, ölçme aracının kapsam geçerliğinin istatistiksel olarak anlamlı olduğu söylenebilir. Başka bir değişle ilgili ölçme aracı akademik yazma becerisini ölçmede yeterli bir kapsama sahip olduğu söylenebilir.

Dereceli puanlama anahtarında yer alacak ölçütler belirlendikten sonra, bu ölçütlerin kaç düzeye sahip olacağına ve her düzeyin kaç puana denk geleceğine karar verilmiştir. Alanyazın taraması ve alan uzmanlarının görüşleri doğrultusunda ölçme aracında yer alan ölçütlerin beş düzeye (4: oldukça yeterli, 0: oldukça yetersiz) ayrılmasına karar verilmiştir. Düzey sayısı belirlendikten sonra her bir düzeyden alınacak puanlamanın analitik tanımlanması (performans göstergeleri) yapılmıştır (bakınız EK 4). Ayrıca düzeyler arasında sistematik geçişleri sağlamak amacıyla analitik tanımlamalarda üst düzeyden alt düzeye doğru sıralanacak biçimde "*highly, comprehensively, extensively, considerably, largely, mostly, adequately, widely, moderately, basically, seemingly, slightly, to some extent, inadequately*" gibi zarf ifadelerine yer verilmiştir. Kutlu vd. (2014), durum belirlemesi yapılacak olan performans alt boyutlara ayrılabiliyorsa, bu durumda analitik dereceli puanlama anahtarının kullanılmasını önermektedir. Akademik yazma becerisi (performansı) alt boyutlara ayrılabilen bir yapıya sahiptir. Bu yüzden mevcut araştırma kapsamında analitik dereceli puanlama anahtarı kullanılmıştır.

Dereceli puanlama anahtarının kapsam geçerliği için kanıtlar toplandıktan sonra yapı geçerliğine ilişkin kanıtlar toplanmıştır. Yapı geçerliği kanıtı için açımlayıcı faktör analizine başvurulmuştur. Açımlayıcı faktör analizine geçilmeden önce açımlayıcı faktör analizi için gerekli olan varsayımlar test edilmiştir. Öncelikle veri setinin faktörleşebilir bir yapıya sahip olup olmadığını belirlemek amacıyla Kaiser-Meyer-Olkin ve Barlett küresellik testi incelenmiştir. Faktör analizi için kabul edilebilir Kaiser-Meyer-Olkin değeri 0,70 ve üstü

olarak belirtilmiştir (Çokluk, Şekercioğlu & Büyüköztürk, 2012, s.207). Mevcut çalışmada Kaiser-Meyer-Olkin değeri 0,875 ve Barlett küresellik testinin anlamlı olduğu bulunmuştur (χ^2 (sd) = 956,427(105), p = 0,000). Açımlayıcı faktör analizinde örneklem büyüklüğünün ne kadar olacağına yönelik alanyazın incelendiğinde, birçok farklı görüş bulunmaktadır. Guadagnoli ve Velicer (1988) tüm bu farklı görüşlerin bir kurama dayalı olmadığını ve deneysel çalışmalarının bulunmadığını belirtmiş ve AFA için gerekli olan örneklem büyüklünün ne olacağına yönelik yaptıkları Monte Carlo simülasyon çalışmasında örneklem sayısından ziyade değişkenlerin faktör yüklerinin önemli olduğunu vurgulamıştır. Buna göre örneklem sayısı 50 kişiden az ve değişken sayısı ne olursa olsun faktör yükü 0,80 ve daha yüksek olan değişkenlerin tutarlı sonuçlar üreteceği belirtilmiştir (Guadagnoli & Velicer, 1988). Mevcut çalışmada örneklem büyüklüğü 50 katılımcının altında olmakla birlikte değişkenlerin üçü hariç diğerlerinin faktör yükü 0,80 değerinden daha büyük olduğundan veri setine AFA yapılması uygun bulunmuştur (ilgili veri seti için faktör yükleri sırasıyla şu şekildedir; 0,842; 0,855; 0,936; 0,968; 0,644; 0,860; 0,960; 0,987; 0,945; 0,605; 0,911; 0,891; 0,899; 0,861 ve 0,622). Veri setinde kayıp ve uç değer bulunmadığından veri seti üzerinde herhangi bir temizleme ve değer atama işlemi yapılmamıştır. Değişkenler arasındaki doğrusallık için ZPRED-ZRESID saçılma diyagramları incelenmiş ve değişkenlerin doğrusal ilişkiye sahip olduğu belirlenmiştir. Değişkenlerin/ölçütlerin normal dağılım gösterip göstermediğini belirlenmesi amacıyla yapılan normallik testi Tablo 5'de verilmiştir.

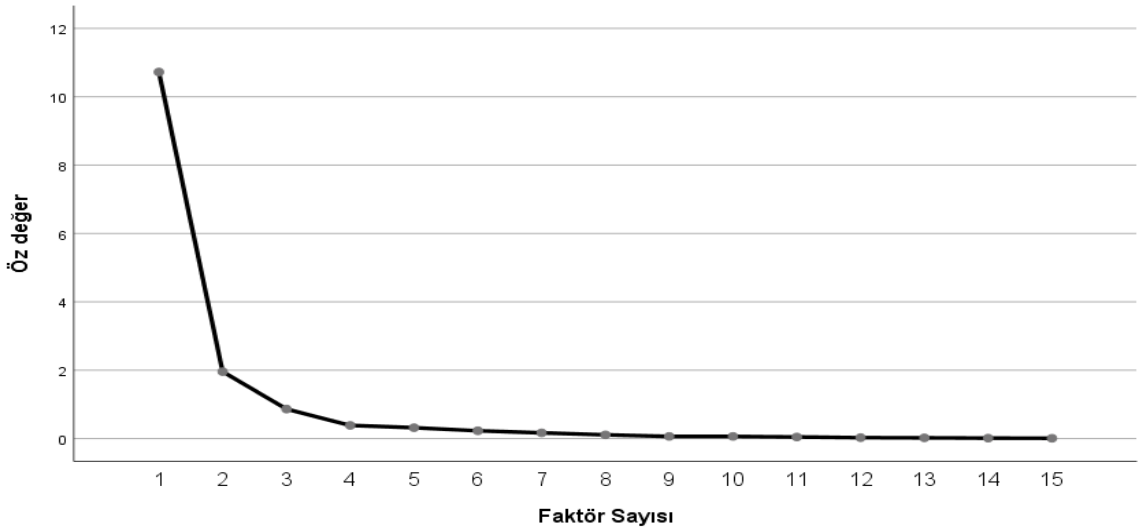
Tablo 5

Kriterlere İlişkin Tek Değişkenli Normallik Testi

	Kolmogorov-Smirnov			Shapiro-Wilk		
	İstatistik	sd	p	İstatistik	sd	p
K1	,226	39	,000	,798	39	,000
K2	,131	39	,088*	,934	39	,024
K3	,089	39	,200*	,948	39	,068*
K4	,091	39	,200*	,963	39	,221*
K5	,232	39	,000	,813	39	,000
K6	,109	39	,200*	,949	39	,078*
K7	,073	39	,200*	,974	39	,484*
K8	,117	39	,198*	,957	39	,141*
K9	,104	39	,200*	,950	39	,082*
K10	,057	39	,200*	,978	39	,635*
K11	,099	39	,200*	,979	39	,673*
K12	,060	39	,200*	,987	39	,936*
K13	,111	39	,200*	,974	39	,481*
K14	,071	39	,200*	,980	39	,715*
K15	,144	39	,039	,925	39	,013

Not. **p* > 0,05 K1: kriter 1, K15: kriter 15

Tablo 5 incelendiğinde, Kolmogorov-Smirnov testine göre iki kriter dışındaki kriterlerin normal dağılım gösterdiği, Shapiro-Wilk testine göre ise dört kriter dışında tüm kriterlerin normal dağılım gösterdiği bulunmuştur. Tek değişkenli normal dağılımın sağlanmamasından dolayı çok değişkenli normallğin test edilmesine karar verilmiş ve ölçme aracından elde edilen toplam puan için normallik testi gerçekleştirilmiştir. Çok değişkenli normallik test edilmiş ve ölçme aracından elde edilen toplam puanının normal dağılım gösterdiği bulunmuştur (ilgili veri için Kolmogrov-Simirnov = 0,100 (39); p = 0,200> 0,05; Shapiro-Wilk = 0,956 (39); p = 0,121> 0,05). Tüm varsayımlar test edildikten sonra AFA yapılmıştır. AFA analizi yapılırken deney ve kontrol grubundaki 45 puanlayıcının 39 katılımcıya verdikleri puanların ortalaması alınmıştır. Yapılan AFA sonucunda ölçütlerin tek bir faktör altında toplandığı ve açıklanan varyansın %70,05 olduğu bulunmuştur. Değişkenlerin tek faktör altında toplandığına ilişkin saçılma grafiği şekil 1'de verilmiştir.



Şekil 1. Analitik dereceli puanlama anahtarına ait saçılma grafiği.

Şekil 1 incelendiğinde, en yüksek öz değere sahip olan birinci faktör ile ikinci en yüksek öz değere sahip olan faktör arasında büyük bir fark olduğundan (yaklaşık beş kat) ve birinci faktör varyansın 2/3'ünü tek başına açıkladığından analitik dereceli puanlama anahtarının tek faktörlü yapıya sahip olduğuna karar verilmiştir (Büyüköztürk, 2013, s.136). Sonuç olarak ilgili ölçme aracının tek faktörlü bir yapıya sahip olduğu bulunmuş ve bu faktöre akademik yazma becerisi adı verilmiştir.

Ölçme aracı kullanılarak elde edilen ölçümlerin geçerliğine ilişkin geçerlik kanıtları toplandıktan sonra, ölçümlerin güvenirliği için McDonald (1999) tarafından önerilen güvenirlik katsayısı (ω) kullanılmıştır. Mevcut çalışmada değişkenlerin faktör yükleri birbirinden farklı olduğundan dolayı (konjenerik ölçümler) bu tür ölçmelerde daha tutarlı kestirimlerin elde edilmesi amaçlandığı (Osburn, 2000) için McDonald ω katsayısının kullanımı tercih edilmiştir. Mplus (versiyon 7) paket programı kullanılarak yapılan analizler neticesinde McDonald ω katsayısının 0,971 (%95 Güven Aralığı: 0,956-0,980) olduğu bulunmuştur. Bu sonuca göre akademik yazma becerisinin ölçülmesi için geliştirilen analitik dereceli puanlama anahtarından elde edilen ölçümlerin geçerli ve güvenilir sonuçlar verdiği savunulabilir.

Kişisel Bilgi Formu

Puanlayıcıların, akademik yazmaya yönelik ilgi, tutum ve kaygıları ve demografik bilgilerinin toplanması amacıyla araştırmacı tarafından kişisel bilgi formu hazırlanmıştır. (Bakınız EK 3). Kişisel bilgi formunun içinde ayrıca puanlayıcıların öğrenci kompozisyonlarına her bir kriter üzerinden verdiği puanı yazacağı puanlama ölçeği eklenmiştir.

Deneysel İşlem

Bu bölümde deney grubuna verilen puanlayıcı eğitimine ilişkin bilgilere yer verilmiştir. Yazma görevlerinin toplanması ve analitik dereceli puanlama anahtarının hazırlanmasından sonra her iki puanlayıcı grubuna yazma görevi ve analitik dereceli puanlama anahtarı hakkında bilgi verilmiştir. Ayrıca deneysel işleme başlamadan önce her iki puanlayıcı grubuna dereceli puanlama anahtarının nasıl hazırlandığı, türleri ve performans değerlendirmeye ilişkin örnek uygulamaları anlatılmıştır. Böylece her iki grubun benzer özellik ve deneyimlere sahip olması sağlanarak ölçülmesi planlanan yapıya farklı varyansların karışması (yapıyla ilişkisiz varyans) azaltılmaya çalışılmıştır. Daha sonra her iki gruba analitik dereceli puanlama anahtarının kriterleri ve kriterlerin derecelendirilmesi detaylı bir şekilde anlatılmıştır. Tüm bu işlemler her hafta bir saat olmak üzere toplam üç saat sürmüştür.

Her iki puanlayıcının ön test puanlamaları incelendiğinde grupların benzer olmasından dolayı bir grup deney diğer grup ise kontrol grubu olarak seçilmiştir. Gruplara deney ve kontrol grubu olarak ayrıldıkları söylenmemiştir. Daha sonra deney ve kontrol grubundaki puanlayıcılara; öğrenci kompozisyonları, analitik dereceli puanlama anahtarı ve puanlayıcı değerlendirme formunu içeren puanlayıcı dosyası (önceden numaralandırılmış) verilmiştir. Puanlayıcılara her bir öğrenci kompozisyonu hazırlanan ADPA'ya göre değerlendirmeleri için bir hafta süre verilmiştir. Bir haftanın sonunda 64 puanlayıcı (33 kontrol ve 31 deney

olmak üzere) performans değerlendirmelerini yapmış oldukları öğrenci kompozisyonlarını içeren puanlayıcı dosyalarını araştırmacıya teslim etmiştir. Her iki grubun yapmış oldukları performans değerlendirmeleri bilgisayar ortamına aktarılmış ve veriler analiz edilmiştir. Yapılan analiz sonucunda her iki grubun benzer puanlayıcı davranışlarına sahip olduğu belirlenmiş ve gruplardan ikincisi deney grubu olarak seçilip puanlayıcı eğitimine geçilmiştir.

Puanlayıcı Eğitimi

Akademik yazma becerisinin değerlendirilmesinde veya puanlanmasında puanlayıcılar arasında ortak bir anlayışın oluşturulması amacıyla ilgili çalışmada referans çerçevesi eğitimi ve puanlayıcı hatası eğitimi birleştirilerek verilmiştir. Bu iki eğitimin birleştirilme nedeni, puanlayıcı hatası eğitiminin puanlayıcı davranışlarının tanımlanmasındaki başarısına rağmen, değerlendirme doğruluğu üzerinde etkili olmamasından dolayı, performans doğruluğunu artıran referans çerçevesi eğitimi ile birleştirilmiştir (Murphy & Balzer, 1989; Sulsky & Day, 1992).

Deney grubuna verilen puanlayıcı eğitimi her hafta bir saat olmak üzere toplam dört hafta sürmüştür. Puanlayıcı eğitimi için araştırmacı tarafından genel yapı şeması EK 2'de verilen bir puanlayıcı modülü geliştirilmiştir. Puanlayıcı eğitiminin ilk haftası, deney grubundaki puanlayıcılara puanlayıcı eğitiminin kapsamı, amacı ve önemi hakkında bilgi verilmiştir. Daha sonra puanlayıcı eğitiminin hedef kitlesinin kimler olduğu ve hangi yöntemlerin kullanıldığı açıklanmıştır. Son olarak puanlayıcıların sorularına cevap verilmiş ve ilk aşama tamamlanmıştır. İlk aşama bittikten sonra ikinci olarak performans değerlendirme sürecinde sıklıkla ortaya çıkan puanlayıcı davranışları anlatılmış ve bu davranışların performans değerlendirme üzerindeki etkisine değinilmiştir. Son olarak örnekler üzerinde puanlayıcı hatalarının neler olduğuna yönelik sınıf içi tartışmalar yapılmış ve eğitimin ilk haftası tamamlanmıştır.

İkinci haftaya başlamadan önce ilk haftanın genel bir tekrarı yapılmıştır. Daha sonra puanlayıcı davranışlarının olası kaynakları hakkında genel bilgiler paylaşılmış ve puanlayıcı davranışlarından kaçınmak için neler yapılacağına yönelik açıklamalar yapılmıştır. Bunun için alanyazında puanlayıcı davranışlarından kaçınmak için uygulanan yöntemlere ilişkin örnekler sunulmuştur. Tüm bu işlemlerden sonra değerlendirmesi yapılmaya çalışılan performans olan "*akademik yazma becerisi*" tanımlanmıştır. Akademik yazmanın tanımı yapıldıktan sonra boyutlarının neler olduğuna yönelik bilgiler verilmiştir. Akademik yazmanın boyutları dikkate alınarak hazırlanmış olan analitik dereceli puanlama anahtarındaki kriterlerin akademik yazmanın hangi boyutuna girdiği belirtilmiştir. Daha sonra puanlayıcıların akademik yazma becerisinin boyutlarına örnekler verilmesi istenmiş ve örnekler sınıf içinde tartışılmıştır. Bu şekilde eğitimin ikinci haftası tamamlanmıştır.

Puanlayıcı eğitiminin üçüncü haftasında ise ikinci haftanın genel tekrarı yapılmış ve akademik yazmanın boyutları ile ilgili birkaç örnek daha paylaşılmıştır. Daha sonra puanlayıcıların ön test puanlamaları incelenerek performans düzeyi yüksek, orta ve düşük olan üç öğrenci kompozisyonu deney grubundaki puanlayıcılara dağıtılmış ve bu öğrencilerin kompozisyonlarını yeniden değerlendirmeleri istenmiştir. Buradaki temel amaç puanlayıcıların akademik yazma becerisine ilişkin göstergeleri belirleyip performans değerlendirme yapmalarını sağlamaktır. Puanlamadan sonra puanlayıcılara verdikleri puanın ne olduğu ve niçin bu puanı verdiği sorulmuştur. Puanlayıcının verdiği puanlama diğer puanlayıcılara sorulmuş ve verilmesi gereken puanın ne olduğu sorulmuştur. Bu şekilde birkaç kriter üzerinde grup tartışmaları yapılmıştır. Buradaki amaç, ilgili performans için (akademik yazma becerisi) ortak bir anlayış oluşturmaktır. Puanlayıcıların yaptıkları puanlamalar hakkında puanlayıcılara geri dönütler verilmiştir. Bunun için puanlayıcıların ön test puanlamalarına ilişkin çok yüzeyli Rasch analizleri kullanılmıştır. Örneğin 1 (bir) numaralı puanlayıcının diğer puanlayıcılara göre daha katı olduğunu ve bunun olası kaynaklarının neler olduğu gibi geri dönütler verilmiştir. Bu şekilde eğitimin üçüncü haftası tamamlanmıştır.

Eğitimin son haftasına gelindiğinde ise üçüncü haftanın devamı şeklinde olmuştur. Puanlayıcıların performans düzeyi yüksek, orta ve düşük olan üç öğrenci kompozisyonlarına verdikleri puanlar ile akademik yazma konusunda uzman olan puanlayıcının verdikleri puanlar karşılaştırılarak puanlayıcıların yaptıkları puanlamalar üzerinde tartışmaları sağlanmıştır. Burada ilk olarak tüm puanlayıcılara örneğin performansı yüksek olan öğrencinin beşinci kriterine kaç puan verilmesi gerektiği sorulup ve grubun ortak bir değer üzerinde anlaşması sağlandıktan sonra uzman puanlayıcının aynı kritere kaç puan verdiği sorulmuştur. Böylece puanlayıcıların ilgili performansın boyutları üzerinde ortak bir anlayışı sürdürmeleri amaçlanmıştır. Son olarak puanlayıcıların analitik dereceli puanlama anahtarının kriterleri üzerinde yeniden tartışmaları istenmiş ve ortak bir fikir birliği oluşturulmaya çalışılmıştır.

Puanlayıcı eğitimi tamamlandıktan sonra, deney ve kontrol grubuna ön test uygulamasında olduğu gibi, öğrenci kompozisyonları, analitik dereceli puanlama anahtarı ve puanlayıcı değerlendirme formunu içeren puanlayıcı dosyası verilmiş ve puanlayıcılara değerlendirmelerini yapmaları için bir haftalık süre verilmiştir. Ön test uygulamasına katılıp son test uygulamasına katılmayan yedi puanlayıcı (3 kontrol ve 4 deney grubundan) araştırma kapsamı dışında tutulmuştur.

Veri Analizi

Veri analizinde; çok yüzeyli Rasch ölçme modeli, genellenebilirlik kuramı ve bağımsız örneklemler t-testi kullanılmıştır. Araştırmada; öğrenci kompozisyonları, puanlayıcılar ve değerlendirme kriterleri olmak üzere üç yüzey bulunmaktadır. Araştırmada puanlayıcılar bütün öğrenci kompozisyonlarını tüm kriterler üzerinden değerlendirdiğinden tamamen çaprazlanmış desen kullanılmıştır. Hem çok yüzeyli Rasch analizleri hem de genellenebilirlik kuramına dayalı analizler bu desen altında yürütülmüştür. Çok yüzeyli Rasch ölçme modeli kullanılarak yapılan analizlerin tutarlı kestirimler yapabilmesi için karşılanması gereken bazı varsayımlar bulunmaktadır. Bu varsayımlar; tek boyutluluk, yerel

bağımsızlık, en az eşit aralık düzeyindeki ölçüm, bir sıralamanın olması ve model veri uyumudur.

Çok yüzeyli Rasch analizinin yapılabilmesi için ölçme aracının tek bir yapıyı ölçmesi gerekmektedir. Bu bağlamda, veri toplama araçları bölümünde gösterildiği gibi araştırmacı tarafından geliştirilen analitik dereceli puanlama anahtarındaki tüm kriterlerin tek bir yapıyı ölçtüğü belirlendiğinden, ilk varsayımın karşılandığı görülmektedir.

Tek boyutluluk varsayımı karşılandıktan sonra yerel bağımsızlık varsayımının karşılanma durumu incelenmiştir. Yerel bağımsızlık varsayımı, Chen ve Thissen (1997) tarafından geliştirilen G^2 istatistiği kullanılarak test edilmiştir. Bu istatistiğe göre, her bir değişken çifti arasında kestirilen standartlaştırılmış LD χ^2 değerlerinin 10'un altında olması ve her bir madde için kestirilen marjinal uyum χ^2 değerlerinin sıfıra yakın olması yerel bağımsızlığın bir göstergesi olarak ele alınmaktadır. Genelleştirilmiş kısmi puan modeline dayalı yapılan kestirimler Tablo 6'da verilmiştir.

Tablo 6

Genelleştirilmiş Kısmi Puan Modeli ile Kestirilen Marjinal Uyum χ^2 ve Standartlaştırılmış LD χ^2 Değerleri

Kriter		Standartlaştırılmış LD χ^2													
		K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12	K13	K14
	Marjinal uyum χ^2														
K1	2,5														
K2	3,1	2,9													
K3	3,3	2,0	3,0												
K4	4,7	4,5	4,0	3,4											
K5	2,3	0,6	0,8	-0,2	0,8										
K6	2,6	-0,2	--	0,6	--	-0,6									
K7	4,3	--	3,6	2,2	--	0,3	--								
K8	3,2	2,3	--	2,1	--	0,7	1,8	--							
K9	1,6	1,6	2,2	1,8	3,5	0,0	0,5	2,1	2,0						
K10	0,2	0,2	0,5	1,3	2,7	0,6	0,8	1,1	1,7	0,8					
K11	0,6	--	--	--	--	0,2	--	--	--	--	1,3				
K12	0,6	-0,1	--	--	--	-0,4	0,3	--	--	--	-0,1	--			
K13	0,6	--	--	--	--	0,2	--	--	--	--	1,3	--	--		
K14	0,0	-0,2	--	0,6	1,9	-0,4	--	0,6	1,6	0,6	--	-0,3	--	-0,3	
K15	0,1	0,4	--	0,8	2,2	0,5	0,1	0,7	--	--	3,5	--	-0,2	--	--

Not. K1: birinci kriter, K15: on beşinci kriter

Tablo 6 incelendiğinde, K4 ve K7 dışındaki kriterlerin marjinal uyum χ^2 değerlerinin sıfıra yakın değerler aldığı ve standartlaştırılmış LD χ^2 değerlerinin ise 10'un altında olduğu görülmektedir. Bu açıdan bakıldığında yerel bağımsızlık varsayımın genel olarak sağlandığı

söylenebilir. Başka bir değişle analitik dereceli puanlama anahtarındaki tüm kriterlerin yerel bağımsızlık varsayımını karşıladığı kabul edilmiştir. Araştırma kapsamındaki analitik dereceli puanlama anahtarı en az eşit aralık düzeyinde ve sıralamalı bir derecelendirmeye (0'dan 4'e kadar) sahip olduğundan bu varsayımların da karşılandığı görülmektedir.

Son olarak standartlaştırılmış artık değerler incelenerek model-veri uyumunun sağlanıp sağlanmadığı test edilmiştir. Model veri uyumu için ±2 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı toplam gözlem sayısının %5'inden fazla olmaması ve ±3 aralığının dışında kalan standartlaştırılmış artık değerlerin de toplam veri sayısının %1'inden fazla olmaması gerektiği belirtilmiştir (Linacre, 2017). Ön test uygulaması için toplam gözlem sayısı 39x45x15 = 26.325 iken ±2 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı 1.067 (%4,05) ve ±3 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı ise 164 (%0,62) olduğundan ön test uygulaması için model-veri uyumunun sağlandığı görülmüştür. Son test uygulaması için toplam gözlem sayısı 26.322 (3 kayıp veri) iken ±2 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı 995 (%3,78) ve ±3 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı ise 186 (%0,71) olduğu belirlenmiş ve son test için model-veri uyumunun sağlandığı kabul edilmiştir. Sonuç olarak tüm varsayımlar karşılanmış olup analizlere geçilmiştir.

BÖLÜM 4

BULGULAR VE YORUMLAR

Bu bölümde, araştırma kapsamında elde edilen bulgular ve bu bulgulara yönelik yorumlara (istatistiksel ve alanyazına dayalı) yer verilmiştir. Çok yüzeyli Rasch ölçme modeli, analize dâhil edilen her bir yüzey için bir ölçüm raporu sunmaktadır. Bu çalışmanın odağı, puanlayıcılar olduğundan puanlayıcı yüzeyine ilişkin ölçüm raporlarının sunulması ile birlikte ele alınan puanlayıcı davranışlarının belirlenmesinde diğer yüzeylere ilişkin ölçüm raporları da incelenmektedir. Bu bağlamda puanlayıcı davranışlarına geçmeden önce ön ve son test ölçümlerine ilişkin ölçüm raporları (öğrenci kompozisyonları, kriter ve puanlayıcı yüzeyleri) verilmiştir. Ölçüm raporları verildikten sonra performans değerlendirmede ortaya çıkan puanlayıcı davranışları ve bu davranışlar üzerinde puanlayıcı eğitiminin etkisi belirlenmeye çalışılmıştır. Daha sonra puanlayıcı eğitiminin puanlayıcı güvenirliği üzerindeki etkisi incelenmiş ve en son çok yüzeyli Rasch analiz sonuçlarının geçerliği için genellenebilirlik analizi yapılmış ve sonuçlar değerlendirilmiştir. Puanlayıcıların hangi puanlayıcı davranışlarını gösterdiklerini belirlemek amacıyla ilk olarak grup düzeyindeki istatistikler daha sonra bireysel düzeydeki istatistikler incelenmiştir.

Ölçüm Raporlarına Ait Uygunluk ve Model Ölçümleri

Birinci araştırma sorusu "Öğrenci kompozisyonu, kriter ve puanlayıcı yüzeylerinin ön ve son test ölçümleri için elde edilen ölçüm raporlarının uygunluk değerleri ve model ölçümleri

nasıldır?" şeklindedir. Bu bağlamda ilk olarak araştırma kapmasında dikkate alınan yüzeylerin ön testlerinden elde edilen ölçüm raporlarına daha sonra ise son testlerden elde edilen ölçüm raporlarına yer verilmiştir. *Öğrenci kompozisyonları (kompozisyonlar)* yüzeyinin ön test ölçümleri için kestirilen ölçüm raporu Tablo 7'de verilmiştir.

Tablo 7

Öğrenci Kompozisyonları Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu

Öğrenci Kompozisyonu Kodu	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit değeri	Standart Hata	Uyum İçi	Uyum Dışı
O01	3,08	3,11	1,39	0,05	0,93	0,94
O02	2,79	2,82	0,94	0,05	0,93	0,91
O03	3,14	3,18	1,51	0,05	1,21	1,22
O04	1,84	1,84	-0,27	0,04	1,20	1,23
O05	2,49	2,52	0,53	0,04	1,03	1,07
O06	2,01	2,03	-0,06	0,04	1,01	1,02
O07	2,82	2,85	0,99	0,05	1,37	1,45
O08	2,05	2,06	-0,02	0,04	0,77	0,77
O09	2,33	2,36	0,32	0,04	0,96	0,96
O10	2,26	2,28	0,23	0,04	1,46	1,49
O11	2,71	2,74	0,83	0,05	0,96	0,96
O12	3,04	3,08	1,34	0,05	1,08	1,12
O13	3,04	3,08	1,34	0,05	0,91	0,90
O14	2,72	2,75	0,84	0,05	0,95	0,99
O15	3,32	3,36	1,86	0,06	0,89	0,98
O16	3,53	3,56	2,35	0,06	1,29	1,21
O17	3,38	3,41	1,98	0,06	1,07	1,12
O18	2,55	2,57	0,60	0,04	0,88	0,88
O19	3,29	3,33	1,79	0,05	0,98	1,02
O20	2,65	2,68	0,75	0,05	0,88	0,89
O21	3,43	3,46	2,10	0,06	1,31	1,31
O22	2,78	2,81	0,93	0,05	0,78	0,79
O23	2,79	2,82	0,93	0,05	0,70	0,70
O24	3,40	3,43	2,03	0,06	0,94	1,01
O25	2,80	2,83	0,95	0,05	0,80	0,81
O26	3,07	3,11	1,39	0,05	0,99	1,02
O27	2,77	2,80	0,92	0,05	0,75	0,74
O28	3,38	3,41	1,98	0,06	1,00	0,99
O29	3,32	3,35	1,85	0,06	0,92	0,91
O30	2,50	2,52	0,54	0,04	0,84	0,85
O31	2,84	2,88	1,02	0,05	0,69	0,68
O32	2,80	2,83	0,95	0,05	0,97	0,96
O33	2,48	2,51	0,52	0,04	0,90	0,90
O34	3,21	3,25	1,64	0,05	0,99	1,03
O35	2,59	2,61	0,65	0,04	1,16	1,15
O36	2,42	2,44	0,43	0,04	0,77	0,78
O37	3,21	3,25	1,64	0,05	1,25	1,21
O38	1,93	1,94	-0,16	0,04	1,34	1,35
O39	1,89	1,90	-0,21	0,04	1,30	1,30
Ortalama	2,79	2,81	1,01	0,05	1,00	1,02
S (Evren)	0,69	0,01	0,19	0,01	0,19	0,20
S (Örneklem)	0,70	0,01	0,20	0,01	0,20	0,20

Model, Evren : RMSE = 0,05 Düzeltilmiş S = 0,69 Ayırma Oranı = 14,16
Ayırma İndeksi = 19,21 Güvenirlik = 1,00
Model, Örneklem: RMSE = 0,05 Düzeltilmiş S = 0,70 Ayırma Oranı = 14,35
Ayırma İndeksi = 19,46 Güvenirlik = 1,00
Model, Ki-kare (Sabit etkili): 7.695,00 sd = 38 p = ,00 Model, Ki-kare (Normal): 37,80 sd = 37 p = ,43

Not. S: standart sapma, sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü, O: öğrenci kompozisyonu

Çok yüzeyli Rasch ölçme modeli, grup düzeyindeki istatistiksel analizlerinde iki farklı model sonucu üretilir. Evren için hesaplanan model, herhangi bir yüzeyin muhtemel bütün

öğelerinin modele dâhil edilmesini dikkate alıyor iken, örneklem için hesaplanan model ise herhangi bir yüzey için seçilen öğeleri dikkate almaktadır (Linacre, 2017, s.202). Bir çalışmadaki yüzeylerin tüm öğelerine ulaşmak maddiyat ve zaman açısından zor olduğundan genellikle yüzeylerden seçilen örneklemler tercih edilmektedir. Dolaysıyla herhangi bir yüzeye ilişkin grup düzeyindeki istatistikler raporlandığında örneklem için hesaplanan model değerlerinin dikkate alınması sonuçların geçerliğine ve güvenirliğine katkı sağlayacaktır. Tablo 7 incelendiğinde, örneklem için hesaplanan ayırma oranının, ayırma indeksinin ve ayırma indeksi güvenirliğinin yüksek olduğu görülmektedir. Bu değerlerin yüksek olması kurulan model tarafından öğrencilerin performanslarına göre birbirinden başarılı bir şekilde ayırt edilebildiğini göstermektedir. Öğrenci performanslarının istatistiksel olarak birbirinden anlamlı bir şekilde farklılaşıp farklılaşmadığı ki-kare istatistiği ile belirlenmektedir. Çok yüzeyli Rasch analizinde her bir yüzey için araştırmacıya iki farklı ki-kare değeri sunulmaktadır. Sabit etkili ki-kare değeri, herhangi bir yüzeyin bileşenleri arasında anlamlı bir farkın olup olmadığını belirlemek amacı ile hesaplanmakta iken, normal ki-kare değeri ise herhangi bir yüzeyin bileşenlerinin normal dağılıma sahip bir evrenden seçkisiz olarak çekilip çekilmediğine karar verme amacı ile hesaplanmaktadır (Linacre, 2017, s.202). Bu bağlamda öğrencilerin performans düzeylerinin istatistiksel olarak birbirinden farklılaşıp farklılaşmadığının belirlenmesi için sabit etkili ki-kare değeri incelenmiştir. Tablo 7'e bakıldığında, birey yüzeyi için hesaplanan sabit etkili ki-kare değerinin istatistiksel olarak manidar olduğu görülmektedir (χ^2(sd) =7.695,00(38), p = 0,00<,01). Bu sonuca göre öğrencilerin performans düzeylerine göre birbirinden başarılı bir şekilde ayırt edilebildiği söylenebilir. Tablo 7 incelendiğinde, birey yüzeyine ilişkin uyum içi ve uyum dışı değerlerinin sırasıyla 0,69 ile 1,49 ve 0,68 ile 1,49 arasında değiştiği ve uygunluk değerlerine ilişkin ortalama değerlerin beklenen değer olan bir(1) değerine yakın olduğu görülmektedir. Bu sonuç, model veri uyumunu bozan bir bireyin olmadığı ve model veri uyumunun iyi olduğunu göstermektedir. Bireylerin logit ölçüleri incelendiğinde, -0,27 ile 2,35 arasında değiştiği ve aradaki logit farkın 2,62 olduğu görülmektedir. Bu değerin

geniş bir aralıkta olması birey performanslarının geniş bir aralığa sahip oluğunun bir göstergesi olarak ele alınmaktadır. Logit değerinin eksi olması birey yeteneğinin (performansın) düşük, artı olması ise yüksek olduğuna işaret etmektedir.

Öğrenci kompozisyonları yüzeyine ilişkin ölçüm raporu incelendikten sonra, kriter yüzeyine ait ölçüm raporu incelenmiştir. Kriter yüzeyinin ön test ölçümleri için kestirlen ölçüm raporu Tablo 8'de verilmiştir.

Tablo 8

Kriter Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu

Kriter Adı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit değeri	Standart Hata	Uyum İçi	Uyum Dışı
Introduction-Body-Conclusion	2,85	2,92	0,07	0,03	1,46	1,32
Thesis Statement	2,58	2,64	-0,32	0,03	1,44	1,42
Topic Sentence	2,58	2,64	-0,32	0,03	1,15	1,13
Supporting Sentence	2,58	2,64	-0,32	0,03	0,82	0,83
Appropriate Length	3,07	3,15	0,44	0,03	1,27	1,19
Topic Relevance	2,99	3,07	0,31	0,03	1,06	1,07
Idea Development	2,53	2,59	-0,39	0,03	0,69	0,70
Coherence	2,78	2,85	-0,02	0,03	0,80	0,81
Linking	2,60	2,66	-0,29	0,03	0,83	0,85
Accuracy of Grammatical Forms	2,82	2,89	0,03	0,03	0,95	0,97
Syntatic Complexity	2,52	2,58	-0,41	0,03	0,72	0,73
Word Choice	2,88	2,95	0,12	0,03	0,83	0,85
Lexical Range	2,60	2,66	-0,29	0,03	0,73	0,75
Spelling	3,25	3,32	0,78	0,03	1,28	1,36
Punctuation	3,16	3,23	0,60	0,03	1,19	1,26
Ortalama	2,79	2,85	0,00	0,03	1,02	1,02
S (Evren)	0,23	0,24	0,37	0,00	0,26	0,24
S (Örneklem)	0,24	0,25	0,38	0,00	0,27	0,25

Model, Evren : RMSE = 0,03 Düzeltilmiş S = 0,37 Ayırma Oranı = 12,46
Ayırma İndeksi = 16,95 Güvenirlik = 0,99
Model, Örneklem: RMSE = 0,03 Düzeltilmiş S = 0,38 Ayırma Oranı = 12,90
Ayırma İndeksi = 17,54 Güvenirlik = 0,99
Model, Ki-kare (Sabit etkili) : 2.151,60 sd = 14 *p* = ,00
Model, Ki-kare (Normal) : 13,90 sd = 13 *p* = ,38

Not. S: Standart Sapma, sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü

Tablo 8 incelendiğinde, kriterlerin güçlük düzeylerinin 0,78 ile -0,41 logit arasında değiştiği görülmektedir. Logit değerinin artı olması kriterin güçlük düzeyinin kolay olduğunu eksi olması ise kriterin güçlük düzeyinin zor olduğunu göstermektedir. Kriterlerin logit değerlerinin ortalaması 0,00 ve standart sapması 0,38 logit olarak bulunmuştur. Kriterler için kestirilen uyum içi ve uyum dışı değerlerinin sırasıyla 0,69 ile 1,46 ve 0,70 ile 1,42 arasında

değiştiği ve ortalama değerlerinin ise 1,02 olduğu bulunmuştur. Bu bulgu, tüm kriterlerin kabul edilebilir uygunluk değerlerine sahip olduğunu ve model veri uyumunu bozmadıklarını göstermektedir. Kriter yüzeyi için hesaplanan; ayırma oranı, ayırma indeksi ve ayırma indeksi güvenirliği değerlerinin yüksek olduğu görülmektedir. Bu değerlerin yüksek olması kriterlerin güçlük düzeylerinin birbirinden farklılık gösterdiğine işaret etmektedir. Kriterlerin güçlük düzeylerinin birbirinden farklılaşıp farklılaşmadığının belirlenmesi için hesaplanan sabit etkili ki-kare değeri incelenmiş ve bu değerin istatistiksel olarak manidar olduğu belirlenmiştir (χ^2(sd) = 2.151,60 (14), p = 0,00<0,01).

Öğrenci kompozisyonları ve kriter yüzeyine ilişkin ölçüm raporları incelendikten sonra son olarak puanlayıcı yüzeyi için çok yüzeyli Rasch analizi sonucunda elde edilen ölçüm raporu incelenmiştir. Puanlayıcı yüzeyinin ön test ölçümleri için kestirilen ölçüm raporu Tablo 9'da verilmiştir.

Tablo 9

Puanlayıcı Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu

Puanlayıcı Kodu	Logit Değeri	Standart Hata	Uyum İçi	Uyum Dışı	Gözlenen Uyum	Beklenen Uyum	Rasch Kappa	t-değeri
OKP01	0,09	0,05	1,26	1,26	41,60	27,60	0,193	1,80
OKP02	-0,51	0,05	1,07	1,09	40,70	27,60	0,181	-10,20
OKP03	-0,04	0,05	0,90	0,93	44,10	27,40	0,230	-0,80
OKP04	0,19	0,05	0,74	0,77	38,40	27,30	0,153	3,80
OKP05	-0,13	0,05	0,74	0,76	42,90	27,80	0,209	-2,60
OKP06	0,15	0,05	0,86	0,87	37,30	26,70	0,145	3,00
OKP07	-0,04	0,05	1,00	1,01	41,40	30,10	0,162	-0,80
OKP08	-0,03	0,05	0,79	0,82	40,80	27,80	0,180	-0,60
OKP09	-0,07	0,05	0,76	0,80	42,00	28,30	0,191	-1,40
OKP10	-0,65	0,05	1,07	1,06	33,20	28,20	0,070	-13,00
OKP11	-0,17	0,05	1,29	1,22	42,00	28,20	0,192	-3,40
OKP12	-0,05	0,05	1,29	1,29	40,50	27,70	0,177	-1,00
OKP13	-0,29	0,05	1,08	1,07	24,50	24,30	0,003	-5,80
OKP14	-0,43	0,05	0,79	0,79	42,00	28,10	0,193	-8,60
OKP15	0,11	0,05	0,84	0,85	41,40	27,80	0,188	2,20
OKP16	0,59	0,06	0,84	0,84	40,10	27,90	0,169	11,80
OKP17	0,18	0,05	1,23	1,20	44,70	28,50	0,227	3,60
OKP18	0,40	0,05	1,18	1,10	42,80	27,90	0,207	8,00
OKP19	0,01	0,05	0,76	0,86	41,50	28,20	0,185	0,20
OKP20	0,40	0,05	1,14	1,11	43,00	28,30	0,205	8,00
OKP21	0,31	0,05	0,78	0,81	43,20	28,70	0,203	6,20
OKP22	0,11	0,05	0,89	0,92	42,50	28,50	0,196	2,20
OKP23	-1,13	0,05	0,93	0,92	42,10	28,00	0,196	-6,60
ODP01	0,26	0,05	0,72	0,75	40,10	31,10	0,131	5,20
ODP02	0,41	0,05	1,36	1,43	39,10	29,30	0,139	8,20
ODP03	0,16	0,05	0,68	0,69	41,20	30,70	0,152	3,20
ODP04	0,69	0,06	1,44	1,44	44,80	30,40	0,207	13,80
ODP05	0,22	0,05	0,74	0,78	41,80	30,00	0,169	4,40
ODP06	-0,65	0,05	1,45	1,49	44,00	30,30	0,197	-13,00
ODP07	0,27	0,05	1,09	1,05	40,10	29,90	0,146	5,40
ODP08	-0,06	0,05	0,60	0,60	43,50	29,90	0,194	-1,20
ODP09	0,03	0,05	1,26	1,31	37,60	28,20	0,131	0,60
ODP10	-0,21	0,05	1,10	1,08	39,70	29,50	0,145	-4,20
ODP11	-0,31	0,05	0,73	0,87	41,30	29,30	0,170	-6,20
ODP12	-1,18	0,05	0,69	0,69	35,80	28,30	0,105	-23,60
ODP13	-0,04	0,05	0,93	0,99	40,10	27,80	0,170	-0,80
ODP14	-0,17	0,05	1,07	1,15	41,50	28,20	0,185	-3,40
ODP15	-0,20	0,05	1,49	1,47	39,70	27,50	0,168	-4,00
ODP16	0,15	0,05	0,72	0,74	41,70	27,60	0,195	3,00
ODP17	-0,19	0,05	0,91	0,92	42,80	26,70	0,220	-3,80
ODP18	0,10	0,05	1,19	1,16	39,60	27,10	0,171	2,00
ODP19	0,50	0,06	1,04	1,08	38,80	27,30	0,158	10,00
ODP20	0,04	0,05	0,86	0,93	44,50	27,60	0,233	0,80
ODP21	0,47	0,06	0,83	0,89	42,30	27,80	0,201	9,40
ODP22	0,70	0,06	1,11	1,10	23,90	24,00	-0,001	14,00
Ortalama	0,00	0,05	1,00	1,02				
S (Evren)	0,40	0,00	0,25	0,25				
S (Örneklem)	0,40	0,00	0,25	0,25				
Model, Evren : RMSE = 0,05 Düzeltilmiş S = 0,39 Ayırma Oranı = 7,63 Ayırma İndeksi = 10,51 Güvenirlik = 0,98								
Model, Örneklem: RMSE = 0,05 Düzeltilmiş S = 0,40 Ayırma Oranı = 7,72 Ayırma İndeksi = 10,63 Güvenirlik = 0,98								
Model, Ki-kare (Sabit etkili) : 2.835,70 sd = 44 *p* = ,00								
Model, Ki-kare (Normal) : 43,30 sd = 43 *p* = ,46								

Not. S: Standart Sapma, sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü, OKP: kontrol grubundan ön testi alan puanlayıcı ODP: deney grubundan ön testi alan puanlayıcı

Tablo 9 incelendiğinde, her bir puanlayıcı için logit ölçüsü, logit ölçüsünün standart hatası (ölçmenin standart hatası), uygunluk değerleri, puanlayıcı güvenirliği için uyum yüzdeleri, t-değeri ve Rasch-kappa değerine ilişkin bilgilerin verildiği görülmektedir. Tablo 9'daki son iki sütun doğrudan çok yüzeyli Rasch analizinden elde edilmeyip denklem (2.7) ve (2.19) kullanılarak elde edilmektedir. Puanlayıcıların logit ölçüleri -1,18 ile 0,70 arasında değişmekte ve puanlayıcıların katılık/cömertliklerine ilişkin aralık ise 1,88 [0,70-(-1,18)] logit olduğu görülmektedir. Burada logit değerinin artı olması cömertlik eksi değerde olması ise katılık davranışının sergilendiğini ifade etmektedir. Puanlayıcıların uyum içi ve uyum dışı değerlerinin ikisi de 0,60 ile 1,49 arasında değiştiği ve ortalamalarının ise sırasıyla 1,00 ve 1,02 olduğu bulunmuştur. Buna göre, tüm puanlayıcıların kabul edilebilir uyuma sahip oldukları ve model veri uyumunu bozmadıkları görülmektedir. Tablo 9 incelendiğinde, örneklem için hesaplanan ayırma oranının, ayırma indeksinin ve ayrıma indeksi güvenirliğinin yüksek olduğu görülmektedir. Bu değerlerin yüksek olması puanlayıcıların birey performansının durum belirleme sürecinde birbirilerinden farklı davranışlar sergilediğini göstermektedir. Puanlayıcıların performans değerlendirme sürecinde bireyden bireye farklı davranışlar göstermesinin grup düzeyindeki istatistiksel anlamlılığı için sabit etkili ki-kare değeri incelenmiştir. Tablo 9'a bakıldığında, puanlayıcı yüzeyi için hesaplanan sabit etkili ki-kare değerinin istatistiksel olarak manidar olduğu görülmektedir (χ^2(sd) = 2.835,70(44), p = 0,00<0,01). Bu sonuca göre deney ve kontrol grubundaki puanlayıcıların grup düzeyinde performans değerlendirme sürecinde farklı davranışlar sergilediği söylenebilir.

Araştırma kapsamındaki tüm yüzeylerin ön test ölçümlerine dayalı ölçüm raporları verildikten sonra son testlere ilişkin ölçüm raporları sunulmuştur. Bu bağlamda ilk olarak öğrenci kompozisyonları yüzeyinin son test ölçümlerine ilişkin ölçüm raporu verilmiştir. Öğrenci kompozisyonları yüzeyinin son test ölçümleri için kestirilen ölçüm raporu Tablo 10'da sunulmuştur.

Tablo 10

Öğrenci Kompozisyonları Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu

Öğrenci Kompozisyonu Kodu	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uyum İçi	Uyum Dışı
O01	3,17	3,21	1,59	0,05	1,05	1,03
O02	2,87	2,90	1,05	0,05	0,84	0,83
O03	3,11	3,14	1,47	0,05	1,31	1,31
O04	1,94	1,96	-0,19	0,04	0,97	1,00
O05	2,50	2,53	0,50	0,04	1,02	1,07
O06	2,15	2,18	0,06	0,04	1,06	1,07
O07	3,00	3,03	1,26	0,05	1,17	1,24
O08	2,23	2,26	0,15	0,04	0,70	0,71
O09	2,46	2,50	0,45	0,04	0,94	0,95
O10	2,23	2,26	0,15	0,04	1,32	1,37
O11	2,74	2,78	0,86	0,05	1,06	1,05
O12	2,93	2,96	1,14	0,05	0,97	1,01
O13	3,03	3,06	1,32	0,05	0,86	0,87
O14	2,99	3,02	1,25	0,05	1,07	1,11
O15	3,42	3,45	2,13	0,06	1,01	1,09
O16	3,55	3,58	2,48	0,07	1,05	1,08
O17	3,39	3,42	2,05	0,06	1,28	1,38
O18	2,56	2,59	0,59	0,05	0,95	0,95
O19	3,33	3,36	1,92	0,06	1,04	1,08
O20	2,68	2,72	0,77	0,05	0,92	0,93
O21	3,37	3,40	2,01	0,06	1,11	1,16
O22	2,82	2,85	0,97	0,05	0,81	0,83
O23	2,85	2,88	1,02	0,05	0,69	0,70
O24	3,44	3,47	2,19	0,06	1,08	1,08
O25	3,07	3,10	1,39	0,05	0,87	0,89
O26	3,05	3,08	1,36	0,05	0,83	0,83
O27	2,91	2,94	1,11	0,05	0,81	0,80
O28	3,40	3,43	2,09	0,06	0,91	0,91
O29	3,40	3,44	2,09	0,06	0,91	0,91
O30	2,77	2,81	0,90	0,05	0,99	1,03
O31	2,98	3,01	1,23	0,05	0,79	0,79
O32	2,83	2,86	0,98	0,05	0,99	0,99
O33	2,45	2,49	0,44	0,04	0,93	0,94
O34	3,29	3,32	1,83	0,06	1,06	1,10
O35	2,74	2,77	0,84	0,05	1,04	1,04
O36	2,65	2,68	0,72	0,05	0,77	0,79
O37	3,29	3,32	1,83	0,06	1,09	1,10
O38	1,78	1,79	-0,39	0,04	1,26	1,27
O39	1,78	1,79	-0,38	0,04	1,38	1,40
Ortalama	2,85	2,88	1,11	0,05	1,00	1,02
S (Evren)	0,46	0,46	0,74	0,01	0,16	0,18
S (Örneklem)	0,47	0,47	0,75	0,01	0,17	0,18

Model, Evren : RMSE = 0,05 Düzeltilmiş S = 0,73 Ayırma Oranı = 14,40
Ayırma İndeksi = 19,53 Güvenirlik = 1,00
Model, Örneklem: RMSE = 0,05 Düzeltilmiş S = 0,74 Ayırma Oranı = 14,59
Ayırma İndeksi = 19,78 Güvenirlik = 1,00
Model, Ki-kare (Sabit etkili) : 8.410,60 sd = 38 p = ,00
Model, Ki-kare (Normal) : 37,80 sd = 37 p = ,43

Not. S: Standart Sapma, sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü, O: öğrenci kompozisyonu

Tablo 10 incelendiğinde, örneklem için hesaplanan ayırma oranının, ayırma indeksinin ve ayrıma indeksi güvenirliğinin yüksek olduğu görülmektedir. Bu değerlerin yüksek olması öğrencilerin/bireylerin performanslarına göre birbirinden başarılı bir şekilde ayırt edilebildiğini göstermektedir. Öğrenci performanslarının istatistiksel olarak birbirinden anlamlı bir şekilde farklılaşıp farklılaşmadığı ki-kare istatistiği ile sınanmış ve sabit etkili ki-kare değerinin istatistiksel olarak manidar olduğu gözlemlenmiştir (χ^2(sd) =8.410,60(38), p = 0,00<,01). Bu sonuca göre öğrencilerin performans düzeylerine göre birbirinden başarılı bir şekilde ayırt edilebildiği söylenebilir. Tablo 10 incelendiğinde, birey yüzeyine ilişkin uyum içi ve uyum dışı değerlerinin sırasıyla 0,69 ile 1,38 ve 0,70 ile 1,40 arasında değiştiği ve uygunluk değerlerine ilişkin ortalama değerlerin beklenen değer olan bire yakın olduğu görülmektedir. Bu sonuç, model veri uyumunu bozan bir bireyin olmadığı ve model veri uyumunun iyi olduğunu göstermektedir. Bireylerin logit ölçüleri incelendiğinde, -0,39 ile 2,48 arasında değiştiği ve aradaki logit farkın 2,87 olduğu görülmektedir. Bu değerin geniş bir aralıkta olması birey performanslarının geniş bir aralığa sahip olduğunun bir göstergesi olarak ele alınmaktadır. Logit değerinin eksi olması performansın düşük, artı olması ise performansın yüksek olduğuna işaret etmektedir.

Öğrenci kompozisyonları yüzeyine ilişkin ölçüm raporu sunulduktan sonra kriter yüzeyine ilişkin ölçüm raporu verilmiştir. Kriter yüzeyinin son test ölçümleri için kestirilen ölçüm raporu Tablo 11'de verilmiştir.

Tablo 11

Kriter Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu

Kriter Adı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uyum İçi	Uyum Dışı
Introduction-Body-Conclusion	2,97	3,05	0,18	0,03	1,88	1,80
Thesis Statement	2,58	2,65	-0,44	0,03	1,38	1,35
Topic Sentence	2,56	2,63	-0,48	0,03	1,05	1,05
Supporting Sentence	2,64	2,72	-0,35	0,03	0,72	0,73
Appropriate Length	3,08	3,15	0,38	0,03	1,23	1,15
Topic Relevance	3,04	3,11	0,29	0,03	1,08	1,07
Idea Development	2,60	2,67	-0,41	0,03	0,67	0,70
Coherence	2,87	2,94	0,01	0,03	0,77	0,78
Linking	2,70	2,77	-0,26	0,03	0,79	0,80
Accuracy of Grammatical Forms	2,91	2,98	0,07	0,03	1,03	1,06
Syntatic Complexity	2,59	2,66	-0,43	0,03	0,70	0,71
Word Choice	2,98	3,05	0,20	0,03	0,90	0,91
Lexical Range	2,66	2,73	-0,32	0,03	0,70	0,71
Spelling	3,33	3,40	0,90	0,04	1,14	1,21
Punctuation	3,22	3,29	0,66	0,03	1,15	1,23
Ortalama	2,85	2,92	0,00	0,03	1,01	1,02
S (Evren)	0,24	0,24	0,42	0,00	0,32	0,30
S (Örneklem)	0,25	0,25	0,43	0,00	0,33	0,31
Model, Evren : RMSE = 0,03 Düzeltilmiş S = 0,42 Ayırma Oranı = 13,43 Ayırma İndeksi = 18,24 Güvenirlik = 0,99						
Model, Örneklem : RMSE = 0,03 Düzeltilmiş S = 0,43 Ayırma Oranı = 13,91 Ayırma İndeksi = 18,87 Güvenirlik = 0,99						
Model, Ki-kare (Sabit etkili) : 2.494,50 sd = 14 *p* = ,00						
Model, Ki-kare (Normal) : 13,90 sd = 13 *p* = ,38						

Not. S: Standart Sapma, sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü

Tablo 11 incelendiğinde, kriterlerin güçlük düzeyleri 0,90 ile -0,48 logit arasında değiştiği görülmektedir. Logit değerinin artı olması kriterin güçlük düzeyinin kolay olduğunu eksi olması ise kriterin güçlük düzeyinin zor olduğunu göstermektedir. Kriterlerin logit değerlerinin ortalaması 0,00 ve standart sapması 0,43 logit olarak bulunmuştur. Kriterler için kestirilen uyum içi ve uyum dışı değerlerinin sırasıyla 0,67 ile 1,88 ve 0,70 ile 1,80 arasında değiştiği ve ortalama değerlerinin ise 1,01 ve 1,02 olduğu bulunmuştur. Bu bulgu, *"Introduction-Body-Conclusion"* kriteri dışındaki diğer kriterlerin kabul edilebilir uygunluk değerlerine sahip olduğunu ve model veri uyumunu bozmadıklarını göstermektedir. Burada *"Introduction-Body-Conclusion"* kriteri model veri uyumunu bozan uygunluk değerlerine sahip olmasına rağmen ön ve son test sonuçlarının karşılaştırılmasının yapılması amaçlandığından veri setinden çıkartılmamıştır. Kriter yüzeyi için hesaplanan; ayırma oranı,

ayırma indeksi ve ayırma indeksi güvenirliği değerlerinin yüksek olduğu görülmektedir. Bu değerlerin yüksek olması kriterlerin güçlük düzeylerinin birbirinden farklılık gösterdiğine işaret etmektedir. Kriterlerin güçlük düzeylerinin birbirinden farklılaşıp farklılaşmadığının belirlenmesi için hesaplanan sabit etkili ki-kare değeri incelenmiş ve bu değerin istatistiksel olarak manidar olduğu belirlenmiştir (χ^2(sd) = 2.494,50 (14), p = 0,00<,01).

Öğrenci kompozisyonları ve kriter yüzeyine ilişkin ölçüm raporları incelendikten sonra, son olarak puanlayıcı yüzeyi için çok yüzeyli Rasch analizi sonucunda elde edilen ölçüm raporu incelenmiştir. Puanlayıcı yüzeyinin son test ölçümleri için kestirilen ölçüm raporu Tablo 12'de verilmiştir.

Tablo 12

Puanlayıcı Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu

Puanlayıcı Kodu	Logit Değeri	Standart Hata	Uyum İçi	Uyum Dışı	Gözlenen Uyum	Beklenen Uyum	Rasch Kappa	t-değeri
SKP01	0,23	0,06	1,18	1,16	46,90	30,20	0,239	4,60
SKP02	-0,71	0,05	1,33	1,34	41,80	30,90	0,158	-14,20
SKP03	0,09	0,05	0,85	0,86	45,90	31,30	0,213	1,80
SKP04	0,18	0,05	0,88	0,90	40,90	29,80	0,158	3,60
SKP05	-0,27	0,05	1,45	1,47	44,10	31,30	0,186	-5,40
SKP06	0,09	0,05	0,82	0,85	44,50	31,00	0,196	1,80
SKP07	0,03	0,05	1,27	1,19	43,30	31,80	0,169	0,60
SKP08	-0,07	0,05	0,91	0,94	45,70	31,20	0,211	-1,40
SKP09	0,03	0,05	1,08	1,11	43,70	31,00	0,184	0,60
SKP10	-0,62	0,05	1,26	1,20	44,40	31,10	0,193	-12,40
SKP11	-0,08	0,05	1,39	1,29	45,20	31,50	0,200	-1,60
SKP12	-0,27	0,05	1,45	1,41	40,90	30,90	0,145	-5,40
SKP13	-0,65	0,05	1,14	1,23	35,60	29,70	0,084	-13,00
SKP14	-0,38	0,05	0,87	0,86	46,80	31,40	0,224	-7,60
SKP15	-0,32	0,05	0,92	0,92	45,60	31,50	0,206	-6,40
SKP16	0,44	0,06	1,45	1,49	42,30	31,40	0,159	8,80
SKP17	0,04	0,05	1,05	1,07	47,20	31,70	0,227	0,80
SKP18	0,27	0,06	0,96	0,96	45,20	31,70	0,198	5,40
SKP19	-0,17	0,05	0,81	0,85	42,10	31,70	0,152	-3,40
SKP20	0,09	0,05	0,96	0,95	45,10	31,40	0,200	1,80
SKP21	0,10	0,05	1,00	0,98	43,70	31,80	0,174	2,00
SKP22	0,00	0,05	0,99	0,96	46,00	31,90	0,207	0,00
SKP23	-1,08	0,05	0,78	0,82	45,10	31,60	0,197	-21,60
SDP01	0,19	0,05	0,61	0,63	42,20	33,20	0,135	2,80
SDP02	0,26	0,06	1,40	1,49	39,70	30,20	0,136	5,20
SDP03	0,29	0,06	0,90	0,89	46,70	32,90	0,206	5,80
SDP04	1,12	0,07	1,23	1,28	44,60	32,30	0,182	22,40
SDP05	0,04	0,05	0,69	0,75	39,70	31,50	0,120	0,80
SDP06	-0,08	0,05	0,94	1,04	44,60	32,30	0,182	-1,60
SDP07	0,12	0,05	0,83	0,89	41,60	31,70	0,145	2,40
SDP08	-0,22	0,05	0,62	0,64	41,00	32,20	0,130	-4,40
SDP09	-0,16	0,05	0,61	0,62	38,90	30,40	0,122	-2,90
SDP10	0,05	0,05	1,08	1,02	42,20	31,80	0,152	1,00
SDP11	-0,35	0,05	0,87	0,96	41,50	31,00	0,152	-7,00
SDP12	-0,72	0,05	0,64	0,67	35,00	29,50	0,078	-14,40
SDP13	0,43	0,06	0,74	0,78	42,60	29,80	0,182	8,60
SDP14	-0,07	0,05	0,70	0,71	41,50	29,10	0,175	-1,40
SDP15	-0,10	0,05	1,28	1,30	39,20	29,40	0,139	-2,00
SDP16	0,20	0,05	0,70	0,73	42,80	29,30	0,191	4,00
SDP17	0,45	0,06	0,93	0,91	46,70	28,90	0,250	9,00
SDP18	0,03	0,05	1,39	1,38	38,70	28,30	0,145	0,60
SDP19	0,38	0,06	1,06	1,07	37,70	31,50	0,120	7,60
SDP20	0,39	0,06	0,84	0,84	43,60	29,40	0,201	7,80
SDP21	0,31	0,06	0,76	0,80	45,30	30,20	0,216	6,20
SDP22	0,49	0,06	1,02	1,00	24,00	23,90	0,001	9,80
Ortalama	0,00	0,05	1,00	1,02				
S (Evren)	0,38	0,00	0,24	0,24				
S (Örneklem)	0,39	0,00	0,24	0,24				

Model, Evren : RMSE = 0,05 Düzeltilmiş S = 0,38 Ayırma Oranı = 7,06
Ayırma İndeksi = 9,75 Güvenirlik = 0,98

Model, Örneklem: RMSE = 0,05 Düzeltilmiş S = 0,38 Ayırma Oranı = 7,14
Ayırma İndeksi = 9,86 Güvenirlik = 0,98

Model, Ki-kare (Sabit etkili) : 2.334,60 sd = 44 *p* = ,00

Model, Ki-kare (Normal) : 43,10 sd = 43 *p* = ,46

Not. S: Standart Sapma, sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü SKP: kontrol grubundan son testi alan puanlayıcı SDP: deney grubundan son testi alan puanlayıcı

Tablo 12 incelendiğinde, puanlayıcıların logit ölçülerinin -1,08 ile 1,12 arasında değişmekte ve puanlayıcıların katılık/cömertliklerine ilişkin aralık ise 2,20 [1,12-(-1,08)] logit olduğu görülmektedir. Burada logit değerinin artı olması cömertlik eksi değerde olması ise katılık davranışının sergilendiğini ifade etmektedir. Puanlayıcıların uyum içi ve uyum dışı değerleri sırasıyla 0,61 ile 1,45 ve 0,62 ile 1,49 arasında değiştiği ve ortalamalarının ise 1,00 ve 1,02 olduğu bulunmuştur. Buna göre, tüm puanlayıcıların kabul edilebilir uyuma sahip oldukları ve model veri uyumunu bozmadıkları görülmektedir. Tablo 12 incelendiğinde, örneklem için hesaplanan ayırma oranının, ayırma indeksinin ve ayrıma indeksi güvenirliğinin yüksek olduğu görülmektedir. Bu değerlerin yüksek olması puanlayıcıların birey performansının durum belirleme sürecinde birbirilerinden farklı davranışlar sergilediğini göstermektedir. Puanlayıcıların performans değerlendirme sürecinde bireyden bireye farklı davranışlar göstermesinin grup düzeyindeki istatistiksel anlamlılığı için sabit etkili ki-kare değeri incelenmiştir. Tablo 12'e bakıldığında, puanlayıcı yüzeyi için hesaplanan sabit etkili ki-kare değerinin istatistiksel olarak manidar olduğu görülmektedir (χ^2(sd) = 2.334,60(44), p = 0,00<0,01). Bu sonuca göre deney ve kontrol grubunda bulunan puanlayıcıların performans değerlendirme sürecinde grup düzeyinde farklı davranışlar sergilediği söylenebilir.

Araştırma kapsamındaki yüzeylerin ön ve son test ölçümlerine ait çok yüzeyli Rasch analizleri incelendiğinde, yüzeylerin hem ön hem de son test ölçümleri için hesaplanan uygunluk değerlerinin ve ölçüm raporlarının kabul edilebilir aralıkta olduğu gözlemlenmiştir. Araştırma kapsamında dikkate alınan yüzeylerin ön ve son testlerine ilişkin ölçüm raporları incelendiğinde hem grup düzeyindeki hem de bireysel düzeydeki istatistiklerin kabul edilebilir ve birbirine benzer sonuçlar aldığı görülmektedir. Benzer olarak ön ve son test ölçümlerine ilişkin her bir yüzey ve yüzeyin öğeleri için kestirilen uygunluk değerlerinin kabul edilebilir aralıkta olduğu gözlemlenmiştir. Bu sonuçlara göre hem ön hem de son teste ilişkin analizlerde model veri uyumunu bozan bir öğenin olmadığı ve böylece her iki analize ilişkin yapılan çıkarımların geçerlik ve güvenirlik düzeylerinin de benzer olduğu söylenebilir.

Puanlayıcı Eğitimi Öncesi Problemli Puanlayıcı Davranışların Ölçümlere Karışma Düzeyleri

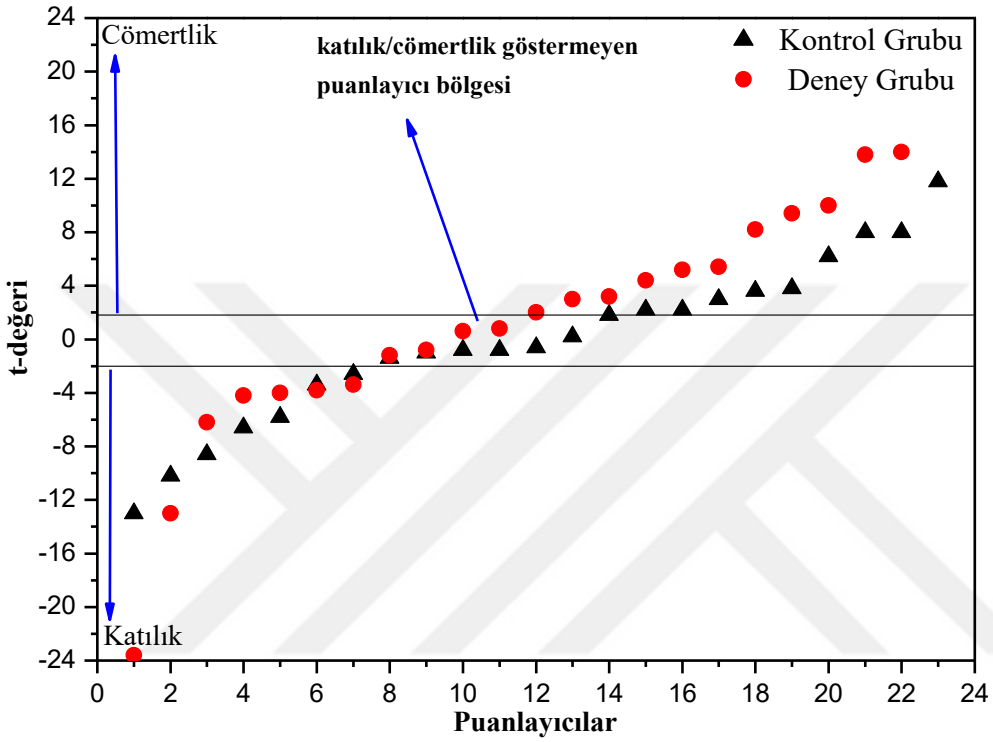
İkinci araştırma sorusu dört alt araştırma sorusundan meydana geldiği için araştırma kapsamında elde edilen bulgular alt başlıklar halinde sunulmuştur.

Puanlayıcı Eğitimi Öncesi Deney ve Kontrol Grubundaki Puanlayıcıların Katılık ve Cömertlik Davranışları

Puanlayıcı eğitimi öncesi, deney ve kontrol gruplarının puanlayıcı katılığı ve cömertliği davranışını gösterip göstermediğini belirlemek amacıyla ilk olarak grup düzeyindeki istatistikler daha sonra bireysel düzeydeki istatistikler incelenmiştir.

Grup düzeyindeki istatistikler incelendiğinde ayırma oranının, ayırma indeksinin ve ayırma indeksi güvenirlik değerlerinin yüksek olduğu görülmektedir. Bu değerlerin yüksek olması puanlayıcıların, birey performansını değerlendirme sürecinde farklı davranışlar sergilediğini göstermektedir. Puanlayıcıların grup düzeyinde farklı davranışlar sergilemesinin istatistiksel anlamlılığı için sabit etkili ki-kare değerine bakılmış ve bu değerin anlamlı olduğu görülmüştür (bakınız Tablo 9). Grup düzeyinde farklı puanlayıcı davranışlarının olduğu belirlendikten sonra bireysel düzeydeki istatistikler incelenmiştir. Bireysel düzeydeki istatistiklerden biri logit ölçüdür. Logit ölçüsü, katılık ve cömertlik için kritik bir değere sahip olmadığından logit değerleri kullanılarak elde edilen t-değeri kullanılmaktadır (bakınız Tablo 9'daki t-değeri sütunu). Bir puanlayıcının t-değeri hesaplanırken, puanlayıcının logit ölçüsünden tüm puanlayıcılara ait logit ölçülerinin ortalaması çıkarılmakta ve elde edilen değer logit ölçüsünün standart hatasına bölünmektedir. Hesaplanan t-değeri, t dağılım tablosundaki kritik t-değeri ile karşılaştırılarak istatistiksel anlamlılığı yapılmaktadır. Araştırmada kontrol grubundan 23 deney grubundan ise 22 puanlayıcı olduğundan küçük grup baz alınarak serbestlik derecesi 21 ve anlamlılık düzeyi de ,05 olarak alınmıştır ($t_{kritik(,05,21)}$ =2,080). Her bir puanlayıcı için hesaplanan t-değeri, kritik t-değeri ile

karşılaştırılarak, bireyin performansı değerlendirilirken hangi puanlayıcı davranışlarının gösterildiği belirlenmeye çalışılmıştır. Deney ve kontrol grubundaki puanlayıcıların ön test puanlamalarında ortaya çıkan katılık ve cömertlik davranışlarının grafiksel gösterimi Şekil 2'de verilmiştir.



Şekil 2. Deney ve kontrol grubunun ön testlerinden elde edilen t-değerleri. (şekildeki her bir nokta bir puanlayıcıyı göstermektedir)

Şekil 2 incelendiğinde, kontrol grubunda yedi puanlayıcının (%30,43) katılık ya da cömertlik davranışını göstermedikleri, yedi puanlayıcının (%30,43) katılık davranışını ve dokuz puanlayıcının (%39,13) ise cömertlik davranışını gösterdikleri görülmektedir. Benzer şekilde, deney grubuna bakıldığında ise beş puanlayıcının (%22,73) katılık veya cömertlik davranışını göstermedikleri, 10 puanlayıcının (%45,46) cömertlik davranışını ve yedi puanlayıcının (%31,81) ise katılık davranışını gösterdikleri görülmektedir.

Deney ve kontrol gruplarının ön testinden elde edilen logit değerlerinin katılık veya cömertlik davranışı açısından birbirinden farklı olup olmadıklarını belirlemek amacıyla bağımsız örneklemler t-testi yapılmış ve sonuçlar Tablo 13'de verilmiştir.

Tablo 13

Deney ve Kontrol Gruplarının Ön Test Puanlamaları Arasındaki Farkla İlgili Bağımsız Örneklemler t-testi Sonucu

Test	Grup	N	$\overline{X}$	S	t	sd	p
Ön test	Kontrol	23	-0,04	0,38	0,735	43	0,466
	Deney	22	0,05	0,43			

Not. *p<,05 Kriterler: "Kontrol=1"; "Deney=2"

Tablo 13 incelendiğinde, deney ve kontrol gruplarının ön testinden elde edilen logit değerlerinin istatistiksel olarak benzer oldukları bulunmuştur (t_{43} =0,735; p = 0,466 > 0,05). Buna göre, puanlayıcı eğitimi öncesi her iki gruptaki puanlayıcıların katılık ve cömertlik düzeylerinin birbirine benzer oldukları söylenebilir. Diğer taraftan her iki grupta da birey performansının değerlendirilmesi sürecinde, puanlayıcı davranışlarının ölçülen yapıya karıştığı görülmektedir. Alanyazın incelendiğinde, performans değerlendirme sürecinde birey performansına mutlaka katılık ve cömertlik davranışının karıştığı görülmektedir (Abu Kassım, 2007; Knoch, vd., 2018). Bu bağlamda puanlayıcı uyumsuzluğunda katılık ve cömertlik davranışının önemli bir davranış olduğu sonucu ortaya çıkmaktadır (Kane vd., 1995).

Puanlayıcı Eğitimi Öncesi Deney ve Kontrol Grubundaki Puanlayıcıların Merkeze Yönelim Davranışları

Puanlayıcıların merkeze yönelim davranışını gösterip göstermediğini belirlemek için ilk olarak grup düzeyindeki istatistikler incelenmiştir. Bu istatistiklerden biri kategori istatistikleridir. Deney ve kontrol gruplarının ön testine ait kategori istatistikleri Tablo 14'de verilmiştir.

Tablo 14

Deney ve Kontrol Gruplarının Ön Testi İçin Hesaplanan Kategori İstatistikleri

Puanlama kategorileri	Frekans	%	Yığmalı %	Ortalama logit ölçüsü	Beklenen logit ölçüsü	Uyum Dışı
0	595	2	2	-0,14	-0,33	1,20
1	2.194	8	11	0,11	0,11	1,00
2	6.435	24	35	0,56	0,59	1,00
3	10.132	38	74	1,10	1,11	1,00
4	6.969	26	100	1,67	1,64	1,00

Tablo 14'e bakıldığında, uçlardaki kategorilerin az orta kategorilerin ise çok kullanıldığı görülmektedir. Bu durumun muhtemel iki nedeni; merkeze yönelim davranışı veya birey performansının orta düzeyde olmasının bir sonucu olabilir. Bu durumun gerçek nedeninin belirlenmesinde grup düzeyindeki istatistiksel göstergelerden biri olan kategori istatistiği tek başına yeterli olmamaktadır. Bu yüzden grup düzeyindeki diğer istatistiksel göstergelerin de incelenmesi gerekir. Deney ve kontrol grubundaki puanlayıcıların ön testine ait birey yüzeyine ilişkin ölçüm raporu incelendiğinde ayırma oranının, ayırma indeksinin ve ayırma indeksi güvenirliğinin yüksek olduğu, başka bir değişle öğrencilerin sahip oldukları performans düzeylerine göre başarılı bir şekilde ayırt edildiği görülmektedir. Ayrıca ki-kare değerinin anlamlı çıkması öğrencilerin performans düzeyine göre anlamlı bir şekilde ayırt edildiğinin istatistiksel kanıtı olarak yorumlanmaktadır (bakınız Tablo 7). Bu bulgulardan hareketle grup düzeyinde merkeze yönelim davranışının olmadığı, kategori istatistiklerindeki mevcut durumun öğrencinin performans düzeyinden kaynaklandığı söylenebilir. Merkeze yönelim davranışını belirlemek için grup düzeyindeki diğer bir istatistiksel gösterge de kriter yüzeyi için hesaplanan ölçüm raporunun incelenmesidir. Kriter yüzeyi için hesaplanan ölçüm raporunda her bir kriterin uygunluk istatistikleri incelenmelidir. Eğer uygunluk istatistikleri (uyum içi ve uyum dışı) kabul edilebilir aralıkta (min=0,50 ve mak=1,50) değilse o zaman merkeze yönelim davranışı ölçümlere karışmış olabilir. Deney ve kontrol gruplarının kriter yüzeyine ilişkin ölçüm raporu incelendiğinde, tüm kriterlerin kabul edilebilir uygunluk istatistiklerine sahip olduğu görülmektedir (bakınız Tablo 8). Sonuç olarak deney ve kontrol grubunda bulunan puanlayıcıların birey performansının durum belirlemesinde grup düzeyinde merkeze yönelim davranışını sergilemedikleri belirlenmiştir.

Deney ve kontrol gruplarının grup düzeyinde merkeze yönelim davranışı incelendikten sonra bireysel düzeydeki merkeze yönelim davranışı incelenmiştir. Bireysel düzeydeki ilk istatistik olarak her bir puanlayıcı için kestirilen uyum içi ve uyum dışı değerlerinin incelenmesidir. Tablo 9 incelendiğinde, tüm puanlayıcıların uyum içi ve uyum dışı

değerlerinin kabul edilebilir aralıkta (0,50 ile 1,50 arasında) olduğu görülmektedir. Merkeze yönelim davranışının bireysel düzeyde olup olmadığının nihai kararı için her bir puanlayıcı için hesaplanan kategori istatistiklerinin incelenmesi gerekmektedir. Deney ve kontrol gruplarının ön test ölçümlerine ait her bir puanlayıcı için hesaplanan kategori istatistikleri Tablo 15'de verilmiştir.

Tablo 15

Deney ve Kontrol Grubundaki Puanlayıcıların Ön Testine İlişkin Her Bir Puanlayıcı İçin Hesaplanan Kategori İstatistikleri

Puanlayıcı	Uyum dışı istatistiği					Merkeze Yönelim
Kodu	Kategori 0	Kategori 1	Kategori 2	Kategori 3	Kategori 4	Etkisi
OKP01	1,10	1,40	1,10	1,00	1,00	Yok
OKP02	1,30	1,20	1,00	1,10	0,90	Yok
OKP03	1,20	0,70	1,20	1,10	1,10	Yok
OKP04	0,70	0,90	0,80	0,90	0,90	Yok
OKP05	1,10	0,80	0,90	1,00	1,00	Yok
OKP06	0,80	0,90	1,00	1,10	1,00	Yok
OKP07	1,20	0,90	0,80	0,80	0,90	Yok
OKP08	--	1,10	0,90	1,10	0,90	Yok
OKP09	1,10	0,90	0,80	0,80	0,90	Yok
OKP10	1,00	0,90	0,90	1,00	1,00	Yok
OKP11	0,80	0,60	0,40	0,70	0,70	Var
OKP12	1,00	1,50	0,80	1,10	1,10	Yok
OKP13	1,10	1,00	1,10	1,10	1,10	Yok
OKP14	0,80	0,90	0,80	0,80	1,00	Yok
OKP15	0,70	0,90	0,90	0,90	0,90	Yok
OKP16	--	0,70	0,90	0,80	0,90	Yok
OKP17	0,80	0,90	1,10	1,20	1,10	Yok
OKP18	1,10	0,90	0,90	0,80	0,90	Yok
OKP19	1,00	0,90	1,30	1,00	1,00	Yok
OKP20	0,90	0,90	1,20	0,90	1,00	Yok
OKP21	0,90	0,50	0,80	0,90	1,00	Yok
OKP22	1,00	1,00	0,80	0,70	0,80	Yok
OKP23	3,50	1,90	1,30	1,70	--	Var
ODP01	--	0,90	0,80	0,90	1,00	Yok
ODP02	0,90	2,00	1,40	1,10	1,10	Var
ODP03	0,80	0,60	0,70	0,80	0,80	Yok
ODP04	1,10	1,40	1,70	1,30	1,20	Var
ODP05	1,10	1,00	0,90	1,00	1,00	Yok
ODP06	2,10	1,20	1,50	1,20	1,00	Var
ODP07	0,90	1,10	0,90	1,00	1,00	Yok
ODP08	0,60	0,80	0,60	0,70	0,90	Yok
ODP09	0,70	1,60	1,50	1,20	1,30	Var
ODP10	1,30	0,80	1,10	1,10	1,00	Yok
ODP11	0,70	1,40	1,10	1,00	0,90	Yok
ODP12	0,90	0,70	0,80	0,70	1,10	Yok
ODP13	1,00	1,10	1,00	1,10	0,80	Yok
ODP14	1,60	1,30	1,10	0,90	0,90	Var
ODP15	1,10	1,50	1,40	1,20	1,10	Yok
ODP16	1,00	0,80	0,90	0,80	0,90	Yok
ODP17	0,90	0,80	0,90	0,90	0,90	Yok
ODP18	0,90	0,70	1,00	1,00	1,00	Yok
ODP19	1,10	0,60	1,20	1,20	1,00	Yok
ODP20	1,10	1,10	1,10	0,90	0,90	Yok
ODP21	1,10	0,90	1,00	1,00	0,90	Yok
ODP22	0,80	0,80	1,10	1,10	0,90	Yok

Not. OKP: kontrol grubundan ön testi alan puanlayıcı ODP: deney grubundan ön testi alan puanlayıcı

Tablo 15 incelendiğinde, kontrol grubundan 11 ve 23 numaralı puanlayıcılar, deney grubunda ise 2, 4, 6, 9 ve 14 numaralı puanlayıcıların analitik dereceli puanlama anahtarının bazı kategorilerinde merkeze yönelim (eğilim) davranışı sergiledikleri görülmektedir. Buna göre deney grubunda daha fazla puanlayıcının birey performansının durum belirleme sürecinde merkeze yönelim davranışı sergiledikleri söylenebilir. Elde edilen bulgulara göre, grup düzeyinde merkeze yönelim etkisi ortaya çıkmaz iken bireysel düzeyde bazı puanlayıcıların merkeze yönelim davranışı sergiledikleri görülmüştür.

Araştırma kapsamında ulaşılan sonuçlar alanyazındaki çalışmalarla uyum göstermektedir. Örneğin Esfandiari (2015) tarafından yapılan çalışmada, akademik yazma becerisinin değerlendirilmesi sürecinde merkeze yönelim davranışını bazı puanlayıcıların gösterdiği fakat grup düzeyinde merkeze yönelim davranışının gösterilmediği bulunmuştur. Benzer bir çalışma ise Engelhard (1994) tarafından yapılmış ve ilgili araştırmada öğrencilerin akademik yazma becerileri değerlendirilirken yapılan puanlamaların %80'ine merkeze yönelim davranışının karıştığı tespit edilmiştir. Ayrıca Leckie ve Baird (2011) tarafından yapılan araştırmada da deneyimsiz puanlayıcıların deneyimli puanlayıcılara göre daha fazla merkeze yönelim davranışını sergiledikleri bulunmuştur. Mevcut çalışmada hem deney hem de kontrol grubundaki puanlayıcıların deneyimsiz olmaları merkeze yönelim davranışının ortaya çıkma olasılığını artıran nedenlerden biri olarak düşünülebilir. Ayrıca katılık ve cömertlik davranışına nispeten merkeze yönelim davranışı performans değerlendirmede daha az ortaya çıkmıştır. Bu sonuç performans değerlendirmede en fazla ortaya çıkan davranışın katılık ve cömertlik olduğunu göstermektedir (Cronbach, 1990).

Puanlayıcı Eğitimi Öncesi Deney ve Kontrol Grubundaki Puanlayıcıların Halo Davranışları

Birey performansının durum belirleme sürecinde puanlayıcıların halo davranışını gösterip göstermediğini belirlemek için ilk olarak grup düzeyindeki istatistiksel göstergeler

incelenmiştir. Kriter yüzeyine ilişkin ölçüm raporu (bakınız Tablo 8) incelendiğinde, ayırma oranının ayırma indeksinin ve ayırma indeksi güvenirliğinin yüksek olduğu görülmektedir. Bu istatistiklerin yüksek olması kriterlerin güçlük düzeylerinin birbirinden farklılık gösterdiği anlamına gelmektedir. Bu farklılığın anlamlılığı için yapılan ki-kare testi istatistiksel olarak manidardır (χ^2(sd) = 2.151,60 (14), p = 0,00<,01). Kriterlerin güçlük düzeylerinin anlamlı çıkması başka bir değişle kriterlerin güçlük düzeylerinin birbirinden başarılı bir şekilde ayırt edilmesi, yapılan puanlamalarda grup düzeyinde halo etkisinin karışmadığını göstermektedir.

Linacre (2017) birey performansının durum belirleme sürecinde halo etkisinin olup olmadığını belirlenmesi için kriter güçlüklerinin birbirine eşitlenerek puanlayıcıların uygunluk istatistiklerinin yeniden incelenmesini önermektedir(s.374). Bu yüzden kriter güçlükleri sıfıra eşitlenerek çok yüzeyli Rasch analizi tekrar edilip puanlayıcılara ait uygunluk istatistikleri incelenmiştir. Kriter güçlükleri eşitlendikten sonra uygunluk istatistiği (uyum içi veya uyum dışı) veri ile mükemmel uyum gösteren (Uİ=1,00 veya UD=1,00 olan) puanlayıcıların halo davranışını gösterdikleri kabul edilmektedir (İlhan, 2015; Linacre, 2017, s.324). Kriter güçlükleri eşitlendikten sonra kontrol grubunun puanlayıcı yüzeyine ait ölçüm raporu Tablo 16'da verilmiştir.

Tablo 16

Kriter Güçlüğü Eşitlendikten Sonra Puanlayıcı Yüzeyinin Ön Test Ölçümleri için Kestirilen Ölçüm Raporu

Puanlayıcı Kodu	Logit değeri	Standart Hata	Uyum İçi	Uyum Dışı	Gözlenen Uyum	Beklenen Uyum	Rasch Kappa	t-değeri
OKP01	0,09	0,05	1,32	1,36	41,40	27,30	0,194	1,80
OKP02	-0,46	0,05	1,00	0,98	41,20	29,40	0,167	-9,20
OKP03	-0,04	0,05	0,93	0,94	42,00	27,60	0,199	-0,80
OKP04	0,18	0,05	0,79	0,80	41,50	27,50	0,193	3,60
OKP05	-0,12	0,05	0,77	0,77	44,00	28,90	0,212	-2,40
OKP06	0,14	0,05	0,89	0,90	40,10	27,10	0,178	2,80
OKP07	-0,04	0,05	1,19	1,19	40,10	29,90	0,146	-0,80
OKP08	-0,03	0,05	0,75	0,78	24,50	24,20	0,004	-0,60
OKP09	-0,07	0,05	0,74	0,76	39,70	27,00	0,174	-1,40
OKP10	-0,59	0,05	1,15	1,14	43,20	28,10	0,210	-11,80
OKP11	-0,15	0,05	1,32	1,26	44,80	29,10	0,221	-3,00
OKP12	-0,05	0,05	1,36	1,32	42,50	27,90	0,202	-1,00
OKP13	-0,27	0,05	1,22	1,19	44,70	27,80	0,234	-5,40
OKP14	-0,39	0,05	0,87	0,86	44,10	27,10	0,233	-7,80
OKP15	0,10	0,05	0,91	0,90	41,80	28,80	0,183	2,00
OKP16	0,54	0,05	0,81	0,81	41,60	27,20	0,198	10,80
OKP17	0,16	0,05	1,35	1,34	40,10	28,60	0,161	3,20
OKP18	0,36	0,05	1,27	1,19	44,50	27,30	0,237	7,20
OKP19	0,01	0,05	0,64	0,67	43,50	28,60	0,209	0,20
OKP20	0,37	0,05	1,07	1,02	42,90	27,30	0,215	7,40
OKP21	0,29	0,05	0,73	0,72	39,60	26,80	0,175	5,80
OKP22	0,10	0,05	0,95	0,96	42,00	27,70	0,198	2,00
OKP23	-1,03	0,04	1,55	1,64	40,50	27,30	0,182	-20,60
ODP01	0,24	0,05	0,75	0,76	23,90	24,20	-0,004	4,80
ODP02	0,38	0,05	1,30	1,28	40,10	27,50	0,174	7,60
ODP03	0,15	0,05	0,77	0,76	41,30	28,10	0,184	3,00
ODP04	0,63	0,06	1,29	1,24	41,70	27,30	0,198	12,60
ODP05	0,20	0,05	0,71	0,72	39,70	28,30	0,159	4,00
ODP06	-0,59	0,05	1,30	1,30	37,30	26,30	0,149	-11,80
ODP07	0,25	0,05	1,11	1,06	40,70	27,20	0,185	5,00
ODP08	-0,06	0,05	0,62	0,62	38,40	26,80	0,158	-1,20
ODP09	0,02	0,05	1,22	1,26	42,80	26,40	0,223	0,40
ODP10	-0,19	0,05	1,17	1,15	35,80	27,40	0,116	-3,80
ODP11	-0,28	0,05	0,56	0,60	41,50	27,60	0,192	-5,60
ODP12	-1,07	0,04	0,86	0,86	33,20	27,60	0,077	-21,40
ODP13	-0,04	0,05	0,91	0,91	41,40	28,80	0,177	-0,80
ODP14	-0,16	0,05	0,95	0,97	42,00	27,60	0,199	-3,20
ODP15	-0,19	0,05	1,50	1,47	37,60	27,20	0,143	-3,80
ODP16	0,14	0,05	0,77	0,76	40,80	27,30	0,186	2,80
ODP17	-0,17	0,05	1,03	1,00	42,10	27,40	0,202	-3,40
ODP18	0,09	0,05	1,22	1,18	38,80	27,00	0,162	1,80
ODP19	0,46	0,05	0,96	0,92	42,80	27,40	0,212	9,20
ODP20	0,04	0,05	0,82	0,83	39,10	28,40	0,149	0,80
ODP21	0,43	0,05	0,78	0,77	43,00	27,70	0,212	8,60
ODP22	0,64	0,06	1,07	1,00	42,30	27,40	0,205	12,80
Ortalama	0,00	0,05	1,01	1,00				
S (Evren)	0,36	0,00	0,25	0,24				
S(Örneklem)	0,37	0,00	0,25	0,25				

Model, Evren : RMSE = 0,05 Düzeltilmiş S = 0,36 Ayırma Oranı = 7,29
Ayırma İndeksi = 10,05 Güvenirlik = 0,98

Model, Örneklem: RMSE = 0,05 Düzeltilmiş S = 0,36 Ayırma Oranı = 7,37
Ayırma İndeksi = 10,17 Güvenirlik = 0,98

Model, Ki-kare (Sabit etkili) : 2.597,00 sd = 44 *p* = ,00

Model, Ki-kare (Normal) : 43,30 sd = 43 *p* = ,46

Not. S: Standart Sapma, sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü, OKP: kontrol grubundan ön testi alan puanlayıcı, ODP: deney grubundan ön testi alan puanlayıcı

Tablo 16 incelendiğinde, grup düzeyindeki istatistiksel göstergeler olan ayırma oranının, ayırma indeksinin ve ayırma indeksi güvenirliğinin yüksek olduğu görülmektedir. Bu değerlerin yüksek olması puanlayıcıların birey performansını değerlendirme sürecinde farklı davranışlar gösterdiğine işaret etmektedir. Puanlayıcıların birey performansını değerlendirme sürecinde farklı davranışlar sergilediğine yönelik yapılan ki-kare testinin manidar olduğu görülmektedir (χ^2(sd) = 2.597,00(44), p = 0,00<0,01). Bireysel düzeydeki istatistiksel göstergeler incelendiğinde Tablo 16'da görüldüğü gibi kriter güçlükleri eşitlendikten sonra uygunluk istatistikleri veri ile mükemmel uyum gösteren kontrol grubundan 2 numaralı puanlayıcı deney grubunda ise 17 ve 22 numaralı puanlayıcıların birey performansını değerlendirirken halo davranışını sergilediği görülmektedir.

Kriter güçlükleri eşitlenmeden yapılan analizlerde ise kriterlerin güçlük düzeyleri arasındaki fark büyük ise puanlayıcılara ait uygunluk istatistikleri 1'den önemli ölçüde büyük olduğunda ve kriterlerin güçlük düzeyleri arasındaki fark küçük ise puanlayıcılara ait uygunluk istatistikleri 1'den önemli ölçüde küçük olduğunda performans değerlendirmeye halo davranışının karıştığı belirtilmiştir (MyFord & Wolfe, 2004). Ön teste ait kriter yüzeyinin ölçüm raporu incelendiğinde (bakınız Tablo 8) kriterlerin güçlük düzeyleri arasındaki fark 1,19 (0,78-[-0,41]) logit olduğu görülmektedir. Bu fark büyük olduğundan uygunluk istatistikleri 1'den önemli derecede büyük olan puanlayıcıların birey performansının durum belirlemesini yaparken halo davranışını gösterdikleri kabul edilmiştir. Buna göre, kontrol grubunda 1, 17 ve 23 numaralı puanlayıcıların, deney grubunda ise 15 numaralı puanlayıcının halo davranışını gösterdikleri belirlenmiştir. Sonuç olarak kontrol grubunda dört puanlayıcı deney grubunda ise üç puanlayıcının birey performansının durum belirleme sürecinde halo etkisi davranışını sergiledikleri tespit edilmiştir.

Merkeze yönelim davranışında olduğu gibi halo davranışının da grup düzeyinde ortaya çıkmadığı fakat bireysel düzeyde ortaya çıktığı bulunmuştur. Alanyazına bakıldığında araştırmanın sonuçlarını destekleyen çalışmaların olduğu görülmektedir. Esfandiari (2015)

ve Engelhard (1994) tarafından yapılan araştırmalarda performans değerlendirmeye halo davranışının karıştığı tespit edilmiştir. Benzer olarak Farrokhi ve Esfandiari (2011) tarafından yapılan araştırmada ise, akran değerlendirme, öz değerlendirme ve öğretmen değerlendirme sürecinde halo davranışının performansa karışma durumu incelenmiş ve her üç değerlendirme türünde de halo davranışının ortaya çıktığı gözlenmiştir. Wu ve Tan (2016) tarafından yapılan çalışmada da puanlayıcıların bir kısmının halo davranışını gösterdiği belirlenmiştir.

Puanlayıcı Eğitimi Öncesi Deney ve Kontrol Grubundaki Puanlayıcıların Farklılaşan Puanlayıcı Katılığı ve Cömertliği (Yanlılık) Davranışları

Farklılaşan puanlayıcı katılığı ve cömertliği davranışını belirlemek için modele dâhil edilen yüzeyler arasındaki etkileşim etkileri incelenmektedir. Bu çalışmada öğrenci kompozisyonları (o), puanlayıcılar (p) ve kriterler (k) olmak üzere üç yüzey bulunduğundan bu yüzeyler arasındaki etkileşimler şu şekildedir; *puanlayıcı x öğrenci kompozisyonu (pxo), puanlayıcı x kriter(pxk), kriter x öğrenci kompozisyonu (kxo)* ve *puanlayıcı x kriter x öğrenci kompozisyonu (pxkxo).* Kriter (madde) yanlılığı mevcut çalışmanın kapsamında olmadığından öğrenci kompozisyonlarının kriterlere göre performans düzeyleri incelenmemiştir. Başka bir değişle *kriter x öğrenci kompozisyonu* etkileşimleri dikkate alınmamıştır. Benzer şekilde puanlayıcıların hangi kritere nasıl yaklaştıkları bu araştırmanın dışında olduğundan *puanlayıcı x kriter (pxk)* etkileşimleri de dikkate alınmamıştır. Ayrıca mevcut çalışmada sadece öğrenci kompozisyonu (performans) ve puanlayıcı etkileşimi odak noktası olduğundan *puanlayıcı x kriter x öğrenci kompozisyonu (pxkxo)* üçlü etkileşimleri de araştırma dışında bırakılmıştır. Herhangi bir puanlayıcının performans değerlendirme sürecinde hangi öğrenci kompozisyonuna daha katı veya cömert davrandığını belirlemek için *puanlayıcı x öğrenci kompozisyonu* etkileşimleri incelenmiştir.

Diğer puanlayıcı etkilerinde olduğu gibi farklılaşan puanlayıcı katılığı ve cömertliği için ilk olarak grup düzeyindeki istatistiksel göstergeler incelenmiştir. Puanlayıcı yanlılığının grup düzeyindeki istatistiksel anlamlılığı için ki-kare istatistiği kullanılmaktadır. Yapılan analiz neticesinde *puanlayıcı x öğrenci kompozisyonu* etkileşimine ilişkin ki-kare değerinin anlamlı olduğu bulunmuştur (χ^2(sd) = 5.298,40(1755), p = 0,00 <0,05). Ki-kare değerinin anlamlı çıkması birey performansının durum belirlemesi yapılırken grup düzeyinde farklılaşan puanlayıcı katılığı ve cömertliği (yanlılık) etkisinin puanlamaya karıştığı anlamına gelmektedir.

Grup düzeyinde farklılaşan puanlayıcı katılığı/cömertliğinin istatistiksel olarak anlamlı çıkması sonucunda bireysel düzeydeki istatistiksel göstergeler incelenmiştir. Çok yüzeyli Rasch analizinde yüzey etkileşimlerinin anlamlılığı için t istatistiği kullanılmaktadır. Herhangi bir etkileşim için hesaplanan t-değeri, t dağılım tablosundaki kritik t-değeri ile karşılaştırılarak istatistiksel anlamlılığı yapılmaktadır. Deney ve kontrol grubundaki puanlayıcıların ön testlerine ilişkin *puanlayıcı x öğrenci kompozisyonu* etkileşimleri incelendiğinde, her bir etkileşim için t-değeri, serbestlik derecesi ve t-değerinin anlamlılık olasılığının verildiği görülmektedir (Bakınız EK 6-7). Burada t değeri ±2 aralığının dışında olan etkileşimlerin puanlayıcı yanlılığına işaret ettiği belirtilir (Linacre, 2017, s.218). Kontrol grubunda bulunan 23 puanlayıcı araştırmaya katılan 39 öğrencinin tamamını puanlandırmış olup toplam 897 (23x39) olası etkileşim meydana getirmiştir. Çok yüzeyli Rasch analizinin çıktıları incelendiğinde 897 olası etkileşimin 203 tanesinin (%22,63) istatistiksel olarak anlamlı olduğu tespit edilmiştir (Bakınız EK 6). Başka bir değişle kontrol grubundaki puanlayıcıların yaptığı 203 puanlamada farklılaşan puanlayıcı katılığı ve cömertliği davranışı ortaya çıkmıştır. Etkileşim raporunda verilen t değerinin işaretine göre farklılaşan puanlayıcı katılığı ve farklılaşan puanlayıcı cömertliği davranışları belirlenebilmektedir. Buna göre t değerleri incelendiğinde 203 puanlamanın 111 tanesi farklılaşan puanlayıcı katılığı iken 92 tanesi farklılaşan puanlayıcı cömertliği davranışıdır. Benzer şekilde deney grubunda bulunan 22 puanlayıcı araştırmaya katılan 39 öğrencinin

tamamını puanlandırmış olup toplam 858 (22x39) olası etkileşim meydana getirmiştir. Çok yüzeyli Rasch analizinin çıktıları incelendiğinde 858 olası etkileşimin 160 tanesinin (%19,00) istatistiksel olarak anlamlı olduğu tespit edilmiştir (Bakınız EK 7). Başka bir değişle deney grubundaki puanlayıcıların yaptığı 160 puanlamada farklılaşan puanlayıcı katılığı ve cömertliği davranışı ortaya çıkmıştır. Etkileşim raporunda verilen t değerinin işaretine göre farklılaşan puanlayıcı katılığı ve farklılaşan puanlayıcı cömertliği davranışları belirlenebilmektedir. Buna göre t değerleri incelendiğinde 160 puanlamanın 92 tanesi farklılaşan puanlayıcı katılığı iken 68 tanesi farklılaşan puanlayıcı cömertliği davranışıdır.

Araştırmadan elde edilen bulgular incelendiğinde, deney ve kontrol grubundaki puanlayıcıların hem grup düzeyinde hem de bireysel düzeyde puanlayıcı yanlılığı davranışı gösterdiği belirlenmiştir. Kontrol grubundaki puanlayıcıların yaptığı puanlamaların yaklaşık %23'ü yanlılık içermekte iken deney grubundaki puanlayıcıların yaptıkları puanlamaların yaklaşık %19'unun yanlılık içerdiği bulunmuştur. Araştırma kapsamında ulaşılan sonuçlar alanyazında yapılan bazı çalışmalar ile uyum göstermektedir. Örneğin, Yan (2014) tarafından yapılan araştırmada, öğrencilerin İngilizce sözlü iletişim becerilerinin değerlendirilmesi sürecinde öğrenci performansına puanlayıcı yanlılığı davranışının karıştığı belirlenmiştir. Başka bir çalışma olan Van Dyke (2008) tarafından yapılan araştırmada, başarılı ve başarısız üniversitelerde çalışan akademisyenlerin meslektaşlarını ve kendilerini değerlendirme sürecinde cömertlik davranışını ve farklılaşan puanlayıcı cömertliği davranışını sergiledikleri bulunmuştur.

Puanlayıcı Eğitimi Sonrası Problemli Puanlayıcı Davranışların Ölçümlere Karışma Düzeyleri

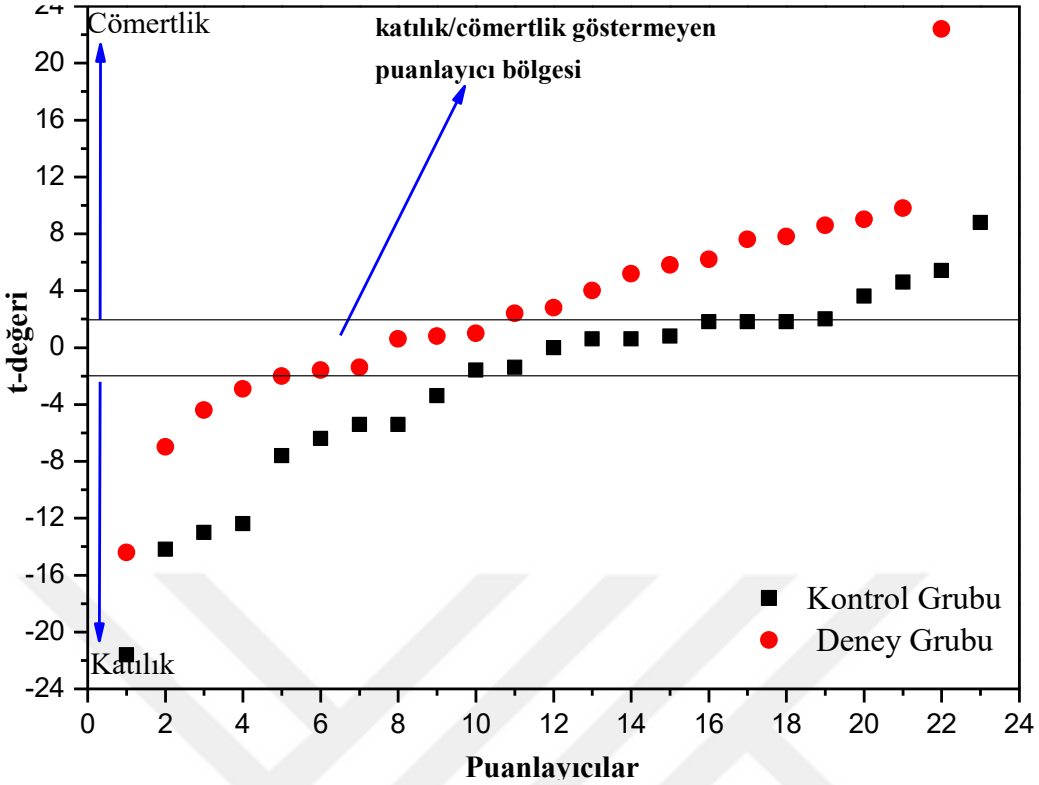
İkinci araştırma sorusu kapsamında deney ve kontrol gruplarının ön test puanlamalarına ilişkin puanlayıcı davranışları belirlendikten sonra, deney grubuna verilen puanlayıcı eğitiminden sonra son test uygulaması kontrol ve deney gruplarına tekrar yapılmış ve yapılan

puanlamadaki problemli puanlayıcı davranışları belirlenmeye çalışılmıştır. Bu bağlamda araştırmanın üçüncü araştırma sorusuna ilişkin bulgular alt başlıklar halinde sunulmuştur.

Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubundaki Puanlayıcıların Katılık ve Cömertlik Davranışları

Deney grubuna verilen puanlayıcı eğitimi sonrasında, her iki gruptaki puanlayıcıların katılık ve cömertlik davranışları sergileme durumları incelenmiştir. Bu bağlamda ilk olarak grup düzeyindeki istatistikler daha sonra ise bireysel düzeydeki istatistikler incelenmiştir.

Grup düzeyindeki istatistikler incelendiğinde ayırma oranının, ayırma indeksinin ve ayırma indeksi güvenirlik değerlerinin yüksek olduğu görülmektedir. Bu değerlerin yüksek olması puanlayıcıların, birey performansını değerlendirme sürecinde farklı davranışlar sergilediğini göstermektedir. Puanlayıcıların grup düzeyinde farklı davranışlar sergilemesini istatistiksel anlamlılığı için sabitlenmiş ki-kare değerine bakılmış ve bu değerin anlamlı olduğu görülmüştür [(χ^2(sd) = 2.334,60(44), p = 0,00<0,01) bakınız Tablo12]. Grup düzeyinde farklı puanlayıcı davranışlarının olduğu belirlendikten sonra bireysel düzeydeki istatistikler incelenmiştir. Bireysel düzeydeki istatistiklerden biri logit ölçüdür. Logit ölçüsü, katılık ve cömertlik için kritik bir değere sahip olmadığından logit değerleri kullanılarak elde edilen t-değeri kullanılmaktadır (bakınız Tablo 12'deki t-değeri sütunu). Deney ve kontrol gruplarına ait t-değerleri için elde edilen grafik Şekil 3'de verilmiştir.



Şekil 3. Deney ve kontrol grubunun son testlerinden elde edilen t-değerleri (şekildeki her bir nokta bir puanlayıcıyı göstermektedir)

Şekil 3 incelendiğinde, kontrol grubunda yedi puanlayıcının (%30,43) katılık ya da cömertlik davranışını göstermedikleri, dokuz puanlayıcının (%39,13) katılık davranışını ve yedi puanlayıcının (%30,43) ise cömertlik davranışını gösterdikleri görülmektedir. Benzer şekilde, deney grubuna bakıldığında ise yedi puanlayıcının (%31,82) katılık veya cömertlik davranışını göstermedikleri, 11 puanlayıcının (%50,00) cömertlik davranışını ve dört puanlayıcının (%18,18) ise katılık davranışını gösterdikleri görülmektedir. Şekil 3 incelendiğinde, deney grubundaki puanlayıcıların katılık ve cömertlik davranışının olmadığı bölgeye yakınlaştığı görülmektedir. Ayrıca katılık ve cömertlik davranışı göstermeyen puanlayıcıların sayısının arttığı bulunmuştur.

Deney ve kontrol gruplarının son testinden elde edilen logit değerlerinin katılık veya cömertlik davranışı açısından birbirinden farklı olup olmadıklarını belirlemek amacıyla bağımsız örneklemler t-testi yapılmıştır. Burada deneysel işlemin etkililiği için puanlayıcıların son testteki logit ölçülerinden ön testteki logit ölçülerinin farkları alınmıştır.

Deney ve kontrol grubundaki puanlayıcıların logit ölçüleri fark puanlarının t-testi istatistiği Tablo 17'de verilmiştir.

Tablo 17

Deney ve Kontrol Grubunun Son Test Puanlamaları Arasındaki Logit Ölçülerinin Farklarına İlişkin Bağımsız Örneklemler t-testi Sonucu

	Grup	N	$\overline{X}$	S	t	sd	p	η^2
Logit fark ölçüleri	Kontrol	23	-0,09	0,16	2,708	43	0,010*	0,15
	Deney	22	0,09	0,28				

Not. *p<,05 Kriterler: "Kontrol=1"; "Deney=2"

Tablo 17 incelendiğinde, deney ve kontrol gruplarının son testinden elde edilen logit değerleri farklarının istatistiksel olarak manidar oldukları bulunmuştur (t_{43} =2,708, p = 0,010 < 0,05, η^2 = 0,15).

Puanlayıcı eğitimi sonrası, deney ve kontrol grubundaki puanlayıcıların katılık ve cömertlik davranışları incelendiğinde hem deney hem de kontrol grubundaki puanlayıcıların grup düzeyinde katılık ve cömertlik davranışı sergiledikleri bulunmuştur. Verilen puanlayıcı eğitimi sonrası deney grubunda katılık ve cömertlik davranışını gösterme durumu %77'ten %68'e düşerken kontrol grubunda ise katılık ve cömertlik davranışının hem ön hem de son testte %30 olduğu bulunmuştur. Puanlayıcı eğitimi, puanlayıcılar arası uyumu yükseltmesine rağmen puanlayıcı kaymasından dolayı katılık ve cömertlik davranışını ortaya çıkarabilmektedir (Moore, 2009, s.11). Mevcut çalışmanın sonuçları incelendiğinde verilen puanlayıcı eğitimi sonrası bazı puanlayıcıların puanlamalarında kaymalar olduğu ve bu yüzden katılık ve cömertlik davranışının ortaya çıktığı gözlemlenmiştir.

Deney ve kontrol grubundaki puanlayıcıların katılık veya cömertlik davranış düzeylerini gösteren logit değerlerinin istatistiksel olarak farklılaşıp farklılaşmadığı test edilmiş ve bu değerlerin anlamlı olduğu bulunmuştur (t_{43} =2,708; p = 0,010 < 0,05; η^2=0,15). Elde edilen etki büyüklüğü dikkate alındığında verilen puanlayıcı eğitiminin puanlayıcı katılığı ve cömertliği üzerinde büyük bir etkiye sahip olduğu ve puanlayıcı katılığı ve cömertliği davranışındaki değişkenliğin %15'inin puanlayıcı eğitimi ile açıklanabildiği belirlenmiştir. Araştırma kapsamında elde edilen bulgular alanyazında yapılan çalışmalar tarafından

desteklenmektedir. Örneğin, Bijani (2018) tarafından yapılan araştırmanın sonuçları incelendiğinde, puanlayıcı eğitiminin puanlayıcı katılık düzeyini azalttığı bulunmuştur. Benzer bir biçimde Fahim ve Bijani (2011) tarafından yapılan araştırmada da öğrencilerin ikinci dil yazma becerilerinin değerlendirilmesi sürecinde puanlamalara karışan puanlayıcı katılılığı ve cömertliği davranışının azaldığı bulunmuştur. Bu sonuçlar mevcut çalışmanın sonuçları ile uyuşmaktadır. Davis (2016) tarafından yapılan araştırmanın sonuçları incelendiğinde ise puanlayıcı eğitiminin puanlayıcı katılığı ve cömertliği üzerinde çok az bir etkiye sahip olduğu fakat puanlayıcı tutarlılığı üzerinde önemli bir etki oluşturduğu bulunmuştur. Diğer yandan Weitz vd. (2014) tarafından yapılan araştırmada, verilen puanlayıcı eğitimi sonrası puanlayıcıların daha katı puanlamalar yaptığı ve Kondo (2010) tarafından yapılan araştırmada ise puanlayıcı eğitimi öncesi ve sonrası puanlayıcı katılık ve cömertlik davranışlarının benzer olduğu bulunmuştur. Alanyazında birçok puanlayıcı eğitimi deseni ve kombinasyonları olduğundan, verilen eğitim sonunda farklı sonuçlar elde edilmektedir.

Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubundaki Puanlayıcıların Merkeze Yönelim Davranışları

Puanlayıcı eğitimi sonrası puanlayıcıların merkeze yönelim davranışını gösterip göstermediğini belirlemek için ilk olarak grup düzeyindeki istatistikler incelenmiştir. Bu istatistiklerden biri kategori istatistikleridir. Deney ve kontrol grubundaki puanlayıcıların son testine ait kategori istatistikleri Tablo 18'de verilmiştir.

Tablo 18

Deney ve Kontrol Gruplarının Son Testi İçin Hesaplanan Kategori İstatistikleri

Puanlama kategorileri	Frekans	%	Yığmalı %	Ortalama logit ölçüsü	Beklenen logit ölçüsü	Uyum dışı
0	554	2	2	-0,14	-0,38	1,40
1	1.836	7	9	0,02	0,08	0,90
2	5.842	22	31	0,59	0,61	1,00
3	10.883	41	73	1,20	1,19	1,00
4	7.207	27	100	1,77	1,77	1,00

Tablo 18'ye bakıldığında, uçlardaki kategorilerin az orta kategorilerin ise çok kullanıldığı görülmektedir. Bu durumun muhtemel iki nedeni; merkeze yönelim etkisi ve birey performansının orta düzeyde olmasının bir sonucu olabilir. Bu durumun gerçek nedeninin belirlenmesinde grup düzeyindeki istatistiksel göstergelerden biri olan kategori istatistiği tek başına yeterli olmamaktadır. Bu yüzden grup düzeyindeki diğer istatistiksel göstergelerin de incelenmesi gerekir. Tablo 10'da verilen son testin öğrenci kompozisyonu yüzeyine ilişkin ölçüm raporu incelendiğinde ayırma oranının, ayırma indeksinin ve ayırma indeksi güvenirliğinin yüksek olduğu, başka bir değişle öğrencilerin sahip oldukları performans düzeylerine göre başarılı bir şekilde ayırt edildiği görülmektedir. Ayrıca ki-kare değerinin anlamlı çıkması öğrencilerin performans düzeyine göre anlamlı bir şekilde ayırt edildiğinin istatistiksel kanıtı olarak yorumlanmaktadır. Bu bulgulardan hareketle grup düzeyinde merkeze yönelim etkisinin olmadığı kategori istatistiklerindeki mevcut durumun öğrencinin performans düzeyinden kaynaklandığı söylenebilir. Merkeze yönelim davranışını belirlemek için grup düzeyindeki diğer bir istatistiksel gösterge de kriter yüzeyi için hesaplanan ölçüm raporunun incelenmesidir. Kriter yüzeyi için hesaplanan ölçüm raporunda her bir kriterin uygunluk istatistikleri incelenmelidir. Eğer uygunluk istatistikleri (uyum içi ve uyum dışı) kabul edilebilir aralıkta (min=0,50 ve mak=1,50) değilse o zaman merkeze yönelim davranışı ölçümlere karışmış olabilir. Deney ve kontrol gruplarının son testlerine ilişkin kriter yüzeyine ilişkin ölçüm raporu (Bakınız Tablo 11) incelendiğinde, birinci kriter olan *"Introduction-Body-Conclusion"* dışındaki kriterlerin kabul edilebilir uygunluk istatistiklerine sahip olduğu görülmektedir. Birinci kriterin kabul edilebilir uygunluk istatistiğine sahip olmasının nedeni, son testi alan puanlayıcılardan bazıları bu kriterin bir veya birden fazla kategorisinde merkeze yönelim davranışını göstermesinden kaynaklanmıştır. Puanlayıcıların ilgili kriterin hangi kategorisinde/derecelendirmesinde merkeze yönelim davranışını gösterdiklerini belirlemek için kategori istatistikleri incelenmiştir. Birinci kriter olan *"Introduction-Body-Conclusion"* için kategori istatistiği Tablo 19'da verilmiştir.

Tablo 19

"Introduction-Body-Conclusion" Kriteri İçin Hesaplanan Kategori İstatistikleri

Puanlama kategorileri	Frekans	%	Yığmalı %	Ortalama logit ölçüsü	Beklenen logit ölçüsü	Uyum dışı
0	162	9	9	-0,14	-0,36	3,30
1	115	7	16	-0,23	0,00	0,50
2	201	11	27	0,26	0,40	0,70
3	405	23	50	0,80	0,83	0,70
4	872	50	100	1,33	1,29	0,90

Tablo 19 incelendiğinde, deney ve kontrol grubundaki puanlayıcılarından bazıları *"Introduction-Body-Conclusion"* kriterinin 0 (sıfır) kategorisine/derecelendirmesine diğer puanlayıcılara göre daha fazla eğilim göstermiştir. Buna göre grup düzeyinde bir merkeze yönelimin söz konusu olduğu görünmektedir. Grup düzeyinde merkeze yönelimin olduğu belirlendikten sonra bireysel düzeydeki istatistikler incelenmiştir.

Bireysel düzeydeki ilk istatistik olarak her bir puanlayıcı için kestirilen uyum içi ve uyum dışı değerlerinin incelenmesidir. Tablo 12 incelendiğinde son testteki tüm puanlayıcıların uyum içi ve uyum dışı değerlerinin kabul edilebilir aralıkta (0,50 ile 1,50 arasında) olduğu görülmektedir. Merkeze yönelim etkisinin bireysel düzeyde olup olmadığının nihai kararı için her bir puanlayıcı için hesaplanan kategori istatistiklerinin incelenmesi gerekmektedir. Deney ve kontrol gruplarının son testine ait her bir puanlayıcı için hesaplanan kategori istatistikleri Tablo 20'de verilmiştir.

Tablo 20

Deney ve Kontrol Grubundaki Her Bir Puanlayıcı İçin Kategori İstatistiği

Puanlayıcı Kodu	Uyum dışı istatistiği					Merkeze Yönelim Etkisi
	Kategori 0	Kategori 1	Kategori 2	Kategori 3	Kategori 4	
SKP01	1,00	1,30	0,80	0,90	0,90	Yok
SKP02	1,70	1,10	1,00	1,80	1,00	Var
SKP03	1,20	0,70	1,00	1,00	0,90	Yok
SKP04	0,80	0,90	0,90	0,90	1,00	Yok
SKP05	3,30	1,10	1,50	1,30	1,30	Var
SKP06	0,70	0,90	0,90	0,90	1,00	Yok
SKP07	1,20	0,60	0,50	1,00	1,00	Yok
SKP08	0,70	0,90	1,20	1,20	1,10	Yok
SKP09	0,90	1,10	0,90	1,20	0,90	Yok
SKP10	1,10	0,80	0,60	0,80	1,30	Yok
SKP11	0,80	0,50	0,60	1,00	0,80	Yok
SKP12	1,40	1,00	1,20	0,90	1,20	Yok
SKP13	1,40	1,50	1,30	1,20	1,20	Yok
SKP14	1,00	1,00	0,80	1,00	1,00	Yok
SKP15	0,90	1,00	1,10	1,00	1,10	Yok
SKP16	5,60	0,50	1,20	1,20	1,20	Var
SKP17	1,30	0,70	1,20	1,00	1,20	Yok
SKP18	1,40	0,50	0,90	0,90	1,00	Yok
SKP19	1,30	1,10	1,20	1,10	1,20	Yok
SKP20	0,70	0,80	1,00	1,00	1,10	Yok
SKP21	0,70	0,80	1,00	0,80	1,10	Yok
SKP22	1,00	0,80	0,60	1,00	0,90	Yok
SKP23	0,90	0,90	0,90	0,80	1,10	Yok
SDP01	--	0,60	0,80	0,90	0,90	Yok
SDP02	0,80	2,50	1,40	1,20	1,10	Var
SDP03	1,30	0,70	0,80	0,80	0,80	Yok
SDP04	0,50	1,00	1,40	1,40	1,10	Yok
SDP05	--	1,10	1,00	1,10	0,90	Yok
SDP06	0,90	1,30	1,30	0,90	1,00	Yok
SDP07	0,60	1,00	1,00	1,00	1,00	Yok
SDP08	1,30	0,60	0,90	0,90	1,00	Yok
SDP09	1,30	0,60	0,70	0,70	0,90	Yok
SDP10	1,30	0,70	0,80	1,00	0,90	Yok
SDP11	1,10	0,70	1,50	1,10	1,20	Yok
SDP12	--	0,80	0,80	0,70	0,90	Yok
SDP13	1,40	0,80	0,80	0,80	0,90	Yok
SDP14	1,00	0,60	0,80	0,80	0,90	Yok
SDP15	1,20	1,10	1,50	0,80	1,10	Yok
SDP16	--	1,00	0,80	0,90	0,90	Yok
SDP17	--	1,00	0,90	1,00	1,00	Yok
SDP18	1,00	1,30	1,30	1,20	1,10	Yok
SDP19	0,90	0,80	1,30	0,90	1,00	Yok
SDP20	--	1,10	0,70	0,90	0,90	Yok
SDP21	1,00	0,70	0,80	1,00	0,90	Yok
SDP22	0,80	0,70	1,10	0,90	0,90	Yok

SKP: kontrol grubundan son testi alan puanlayıcı, SDP: deney grubundan son testi alan puanlayıcı

Tablo 20 incelendiğinde, kontrol grubundan 2, 5 ve 16 numaralı puanlayıcılar, deney grubunda ise 2 numaralı puanlayıcının analitik dereceli puanlama anahtarının bazı kategorilerinde merkeze yönelim davranışı sergiledikleri görülmektedir.

Baird vd. (2013) tarafından merkeze yönelim davranışının genellikle çoklu derecelendirmeye sahip ölçme araçlarını kullanan tecrübesiz puanlayıcılardan dolayı ortaya çıktığı belirtilmiştir. Ön testte deney grubundan beş puanlayıcı (%22,73) merkeze yönelim davranışı sergiler iken, verilen eğitim sonrası sadece bir puanlayıcının (%4,55) merkeze yönelim davranışı sergilediği bulunmuştur. Bu sonuca göre verilen puanlayıcı eğitiminin birey performansını değerlendirme sürecinde sıklıkla ortaya çıkan merkeze yönelim davranışı üzerinde etkili olduğu söylenebilir. Feldman vd. (2012), performans değerlendirme sürecinde merkeze yönelim davranışının ölçümlere karışması durumunda, bireyin performans düzeyinin ayırt edilmesini azaltarak ölçümlerin geçerliğini tehlikeye düşürdüğü belirtilmiştir. Buna göre verilen puanlayıcı eğitiminin ölçümlerin geçerliğine katkı sağladığı savunulabilir. May (2008) merkeze yönelim davranışının azaltılmasında puanlayıcı hatası eğitimi deseninin etkili olduğunu belirtmiştir. Mevcut çalışmada puanlayıcı hatası eğitimi ve referans çerçevesi eğitiminin kombine edildiği dikkate alındığında ulaşılan sonuçların alanyazın ile uyumlu olduğu görülmektedir. Bununla birlikte Bernardin (1978) ve Knoch vd. (2007) tarafından yapılan araştırmada, verilen puanlayıcı eğitimi sonrası puanlayıcıların merkeze yönelim davranışının beklenilenin aksine arttığı bulunmuştur.

Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubundaki Puanlayıcıların Halo Davranışları

Birey performansının durum belirleme sürecinde puanlayıcıların halo davranışını gösterip göstermediğini belirlemek için ilk olarak grup düzeyindeki istatistiksel göstergeler incelenmiştir. Deney ve kontrol grubundaki puanlayıcıların son testinden elde edilen kriter yüzeyine ilişkin ölçüm raporu (Bakınız Tablo 11) incelendiğinde, ayırma oranının, ayırma

indeksinin ve ayırma indeksi güvenirliğinin yüksek olduğu görülmektedir. Bu istatistiklerin yüksek olması kriterlerin güçlük düzeylerinin birbirinden farklılık gösterdiği anlamına gelmektedir. Bu farklılığın anlamlılığı için yapılan ki-kare testi istatistiksel olarak manidardır (χ^2(sd) = 2.494,50 (14), p = 0,00<,01). Kriterlerin güçlük düzeylerinin anlamlı çıkması başka bir değişle kriterlerin güçlük düzeylerinin birbirinden başarılı bir şekilde ayırt edilmesi, yapılan puanlamalarda grup düzeyinde halo davranışının karışmadığını göstermektedir.

Bireysel düzeydeki puanlamalarda halo davranışını belirlemek için kriter güçlükleri birbirine eşitlenerek puanlayıcıların uygunluk istatistikleri incelenmiştir. Kriter güçlükleri eşitlendikten sonra uygunluk istatistiği (uyum içi veya uyum dışı) veri ile mükemmel uyum gösteren (Uİ=1 veya UD=1 olan) puanlayıcıların halo davranışını gösterdikleri kabul edilmektedir (İlhan, 2015; Linacre,2017). Kriter güçlükleri eşitlendikten sonra deney ve kontrol grubundaki puanlayıcıların son testlerine ilişkin puanlayıcı yüzeyine ait ölçüm raporu Tablo 21'de verilmiştir.

Tablo 21

Kriter Güçlüğü Eşitlendikten Sonra Son Test Puanlarına İlişkin Puanlayıcı Yüzeyinin Ölçüm Raporu

Puanlayıcı Kodu	Logit Değeri	Standart Hata	Uyum İçi	Uyum Dışı	Gözlenen Uyum	Beklenen Uyum	Rasch Kappa	t-değeri
SKP01	0,21	0,05	1,28	1,28	45,10	29,60	0,220	4,20
SKP02	-0,64	0,05	1,32	1,35	45,20	29,70	0,220	-12,80
SKP03	0,08	0,05	0,86	0,86	46,70	29,00	0,249	1,60
SKP04	0,16	0,05	0,86	0,86	41,50	28,90	0,177	3,20
SKP05	-0,25	0,05	1,27	1,35	42,80	29,20	0,192	-5,00
SKP06	0,08	0,05	0,84	0,85	45,10	29,30	0,223	1,60
SKP07	0,02	0,05	1,52	1,42	31,10	26,00	0,069	0,40
SKP08	-0,06	0,05	0,82	0,83	43,60	29,00	0,206	-1,20
SKP09	0,03	0,05	0,94	0,94	40,90	28,20	0,177	0,60
SKP10	-0,56	0,05	1,41	1,35	45,30	29,40	0,225	-11,20
SKP11	-0,08	0,05	1,49	1,37	45,20	29,70	0,220	-1,60
SKP12	-0,25	0,05	1,52	1,46	42,30	29,40	0,183	-5,00
SKP13	-0,59	0,05	1,20	1,26	35,00	28,70	0,088	-11,80
SKP14	-0,34	0,05	0,95	0,95	42,20	31,80	0,152	-6,80
SKP15	-0,29	0,05	0,98	0,99	41,80	29,10	0,179	-5,80
SKP16	0,40	0,05	1,57	1,62	42,10	29,40	0,180	8,00
SKP17	0,03	0,05	1,00	0,99	39,70	30,60	0,131	0,60
SKP18	0,24	0,05	1,02	0,98	39,70	29,00	0,151	4,80
SKP19	-0,15	0,05	0,73	0,76	38,90	29,20	0,137	-3,00
SKP20	0,08	0,05	0,96	0,93	42,20	30,50	0,168	1,60
SKP21	0,09	0,05	1,03	1,01	43,30	31,00	0,178	1,80
SKP22	0,00	0,05	1,18	1,13	41,50	30,00	0,164	0,00
SKP23	-0,97	0,04	0,73	0,73	39,20	29,30	0,140	-19,40
SDP01	0,17	0,05	0,68	0,67	43,70	29,20	0,205	3,40
SDP02	0,23	0,05	1,32	1,30	44,40	29,50	0,211	4,60
SDP03	0,26	0,05	0,92	0,88	45,60	29,50	0,228	5,20
SDP04	1,03	0,06	1,09	1,02	46,90	29,00	0,252	20,60
SDP05	0,04	0,05	0,68	0,70	44,10	29,50	0,207	0,80
SDP06	-0,07	0,05	0,86	0,89	46,80	29,70	0,243	-1,40
SDP07	0,11	0,05	0,87	0,88	47,20	29,70	0,249	2,20
SDP08	-0,20	0,05	0,58	0,59	38,70	28,40	0,144	-4,00
SDP09	-0,14	0,05	0,61	0,62	40,90	29,10	0,166	-2,80
SDP10	0,04	0,05	1,22	1,15	46,00	29,90	0,230	0,80
SDP11	-0,31	0,05	0,72	0,77	43,70	29,80	0,198	-6,20
SDP12	-0,65	0,05	0,83	0,83	41,60	30,80	0,156	-13,00
SDP13	0,39	0,05	0,74	0,72	35,60	27,70	0,109	7,80
SDP14	-0,07	0,05	0,66	0,67	46,60	31,20	0,224	-1,40
SDP15	-0,09	0,05	1,26	1,24	46,70	31,60	0,221	-1,80
SDP16	0,18	0,05	0,75	0,75	44,60	31,40	0,192	3,60
SDP17	0,40	0,05	1,01	0,98	45,70	29,30	0,232	8,00
SDP18	0,02	0,05	1,36	1,34	45,90	29,40	0,234	0,40
SDP19	0,34	0,05	1,06	0,99	44,50	28,80	0,221	6,80
SDP20	0,35	0,05	0,81	0,83	43,60	28,70	0,209	7,00
SDP21	0,28	0,05	0,78	0,77	41,00	30,80	0,147	5,60
SDP22	0,44	0,06	1,01	0,95	42,60	29,20	0,189	8,80
Ortalama	0,00	0,05	1,01	1,00				
S (Evren)	0,35	0,00	0,27	0,26				
S (Örneklem)	0,35	0,00	0,27	0,26				

Model, Evren : RMSE = 0,05 Düzeltilmiş S = 0,34 Ayırma Oranı = 6,71
Ayırma İndeksi = 9,28 Güvenirlik = 0,98

Model, Örneklem: RMSE = 0,05 Düzeltilmiş S = 0,35 Ayırma Oranı = 6,78
Ayırma İndeksi = 9,38 Güvenirlik = 0,98

Model, Ki-kare (Sabit etkili) : 2.109,70 sd = 44 *p* = ,00

Model, Ki-kare (Normal) : 43,31 sd = 43 *p* = ,47

Not. S: Standart Sapma, sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü

Tablo 21 incelendiğinde grup düzeyindeki istatistiklerden ayırma oranının, ayırma indeksinin ve ayırma indeksi güvenirliğinin yüksek olduğu görülmektedir. Bu değerlerin yüksek olması puanlayıcıların birey performansının değerlendirme sürecinde farklı davranışlar sergilediğini göstermektedir. Puanlayıcıların birey performansını değerlendirirken göstermiş oldukları bu davranışların grup düzeyinde istatistiksel anlamlılığı için yapılan ki-kare testinin manidar olduğu görülmektedir (χ^2(sd) = 2.109,70(44), p = 0,00<0,01). Kriterlerin güçlükleri eşitlendikten sonra, puanlayıcı yüzeyi için kestirilen ölçüm raporunun uygunluk değerleri kabul edilebilir olduğu görünmektedir. Grup düzeyindeki istatistikler incelendiğinde, puanlayıcıların birey performansını değerlendirme sürecinde farklı davrandıkları belirlenmiştir. Bu davranışlardan halo davranışını gösteren puanlayıcı(ların) belirlenmesi için bireysel düzeydeki istatistikler incelenmiştir. Tablo 21'e bakıldığında, kriter güçlükleri eşitlendikten sonra uygunluk istatistikleri veri ile mükemmel uyum gösteren kontrol grubundan 17 numaralı puanlayıcı birey performansını değerlendirirken halo davranışını sergilediği görülmektedir.

Kriter güçlükleri eşitlenmeden yapılan analizlerde ise kriterlerin güçlük düzeyleri arasındaki fark büyük ise puanlayıcılara ait uygunluk istatistikleri 1'den önemli ölçüde büyük olduğunda ve kriterlerin güçlük düzeyleri arasındaki fark küçük ise puanlayıcılara ait uygunluk istatistikleri 1'den önemli ölçüde küçük olduğunda performans değerlendirmeye halo etkisinin karıştığı belirtilmiştir (MyFord & Wolfe, 2004). Son teste ait kriter yüzeyi ölçüm raporu incelendiğinde (bakınız Tablo 11) kriterlerin güçlük düzeyleri arasındaki fark 1,38 (0,90-[-0,48]) logit olduğu görülmektedir. Bu fark büyük olduğundan uygunluk istatistikleri 1'den önemli derecede büyük olan puanlayıcıların birey performansının durum belirlemesini yaparken halo davranışını gösterdikleri kabul edilmiştir. Buna göre, kontrol grubunda 2, 5, 11, 12 ve 16 numaralı puanlayıcıların, deney grubunda ise 2 ve 18 numaralı puanlayıcının halo davranışını gösterdikleri belirlenmiştir. Sonuç olarak kontrol grubunda 6 puanlayıcı deney grubunda ise 2 puanlayıcının birey performansının durum belirleme sürecinde halo davranışını sergiledikleri tespit edilmiştir.

Kontrol grubunda ön testte halo davranışını gösteren puanlayıcı sayısı üç (%13,04) iken son testte bu sayı altıya (%26,09) çıkmıştır. Deney grubunda ise ön test sonuçlarında halo davranışı gösteren puanlayıcı sayısı 4 (%18,18) iken son testte ikiye (%9,09) düşmüştür. Bu sonuca göre verilen puanlayıcı eğitiminin halo davranışını azaltmada etkili olduğu savunulabilir. Feldman vd. (2012) halo davranışının performans değerlendirmede sistematik hatayı artırdığını, puanlayıcı doğruluğunu azalttığını ve bu yüzden ölçümlerin geçerliği üzerinde önemli bir etkiye sahip olduğunu ifade etmiştir. Bu bağlamda verilen puanlayıcı eğitimi sonrası elde edilen ölçümlerin geçerliğine katkı sağlanmıştır. Bijani (2018) tarafından yapılan araştırmada puanlayıcı eğitiminin halo davranışını azalttığı, Weitz vd. (2014) tarafından yapılan araştırmada ise puanlayıcı eğitiminin puanlayıcıların halo davranışına yönelik farkındalıklarını artırdığı bulunmuştur. Pulakos (1984) tarafından yapılan araştırmada ise puanlayıcı hatası eğitimi deseninin halo davranışını azaltmada etkili olduğu bulunmuştur. Benzer bir şekilde Borman (1975) tarafından yapılan araştırmada da puanlayıcı eğitiminin halo davranışını azalttığı bulunmuştur. Buna göre mevcut çalışmanın ulaştığı sonuçların alanyazın ile desteklendiği görülmektedir.

Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubundaki Puanlayıcıların Farklılaşan Puanlayıcı Katılığı ve Cömertliği (Yanlılık) Davranışları

Diğer puanlayıcı etkilerinde olduğu gibi farklılaşan puanlayıcı katılığı ve cömertliği için ilk olarak grup düzeyindeki istatistiksel göstergeler incelenmiştir. Puanlayıcı yanlılığının grup düzeyindeki istatistiksel anlamlılığı için ki-kare istatistiği kullanılmaktadır. Yapılan analiz neticesinde *puanlayıcı x öğrenci kompozisyonu* etkileşimine ilişkin ki-kare değerinin anlamlı olduğu bulunmuştur (χ^2(sd) = 4.084,90(1755), p = 0,00 <0,01). Ki-kare değerinin anlamlı çıkması birey performansının durum belirlemesi yapılırken grup düzeyinde farklılaşan puanlayıcı katılığı ve cömertliği (yanlılık) etkisinin puanlamaya karıştığı anlamına gelmektedir.

Grup düzeyinde farklılaşan puanlayıcı katılığı ve cömertliğinin istatistiksel olarak anlamlı çıkması sonucunda bireysel düzeydeki istatistikler incelenmiştir. Deney ve kontrol grubundaki puanlayıcıların son testlerine ilişkin *puanlayıcı x öğrenci kompozisyonu* etkileşimleri incelendiğinde, her bir etkileşim için t-değeri, serbestlik derecesi ve t-değerinin anlamlılık olasılığının verildiği görülmektedir (Bakınız EK 8-9). Kontrol grubunda bulunan 23 puanlayıcı araştırmaya katılan 39 öğrencinin tamamını puanlandırmış olup toplam 897 (23x39) olası etkileşim meydana getirmiştir. Çok yüzeyli Rasch analizinin çıktıları incelendiğinde 897 olası etkileşimin 163 (%18,17) tanesinin istatistiksel olarak anlamlı olduğu tespit edilmiştir (Bakınız EK 8). Başka bir değişle kontrol grubundaki puanlayıcıların yaptığı 163 puanlamada farklılaşan puanlayıcı katılığı ve cömertliği davranışı ortaya çıkmıştır. Etkileşim raporunda verilen t değerinin işaretine göre farklılaşan puanlayıcı katılığı ve farklılaşan puanlayıcı cömertliği davranışları belirlenebilmektedir. Buna göre t değerleri incelendiğinde 163 puanlamanın 95 tanesi farklılaşan puanlayıcı katılığı iken 68 tanesi farklılaşan puanlayıcı cömertliği davranışıdır. Benzer şekilde deney grubunda bulunan 22 puanlayıcı araştırmaya katılan 39 öğrencinin tamamını puanlandırmış olup toplam 858 (22x39) olası etkileşim meydana getirmiştir. Çok yüzeyli Rasch analizinin çıktıları incelendiğinde 858 olası etkileşimin 110 (%12,82) tanesinin istatistiksel olarak anlamlı olduğu tespit edilmiştir (Bakınız EK 9). Başka bir değişle deney grubundaki puanlayıcıların yaptığı 110 puanlamada farklılaşan puanlayıcı katılığı ve cömertliği davranışı ortaya çıkmıştır. Etkileşim raporunda verilen t değerinin işaretine göre farklılaşan puanlayıcı katılığı ve farklılaşan puanlayıcı cömertliği davranışları belirlenebilmektedir. Buna göre t değerleri incelendiğinde 110 puanlamanın 59 tanesi farklılaşan puanlayıcı katılığı iken 51 tanesi farklılaşan puanlayıcı cömertliği davranışıdır.

Puanlayıcı eğitiminin puanlayıcı yanlılığı üzerindeki etkisini görmek amacıyla deney grubundaki bazı puanlayıcıların *puanlayıcı x öğrenci kompozisyonu* etkileşimleri karşılaştırılarak betimsel ve grafiksel olarak sunulmuştur. Deney grubundaki puanlayıcılara ait betimsel değerler verilmekle birlikte grafiksel gösterim makul olmadığından belli bir

kritere göre bazı puanlayıcılar seçilerek grafikleri verilmiştir. Puanlayıcılar seçilirken ön ve son test puanlamaları arasında önemli farklılık olmayan (deneysel işlemden etkilenmeyen), ön testi son testinden daha az yanlılık içiren (deneysel işlemden olumsuz etkilenen) ve ön testi son testinden daha fazla yanlılık (deneysel işlemden olumlu etkilenen) içeren puanlayıcılar dikkate alınmıştır. Bu kriterler doğrultusunda üç puanlayıcıya ait grafiksel gösterimler Şekil 4-6'da ve deney grubundaki tüm puanlayıcılara ait betimsel değerler ise Tablo 22'de verilmiştir.

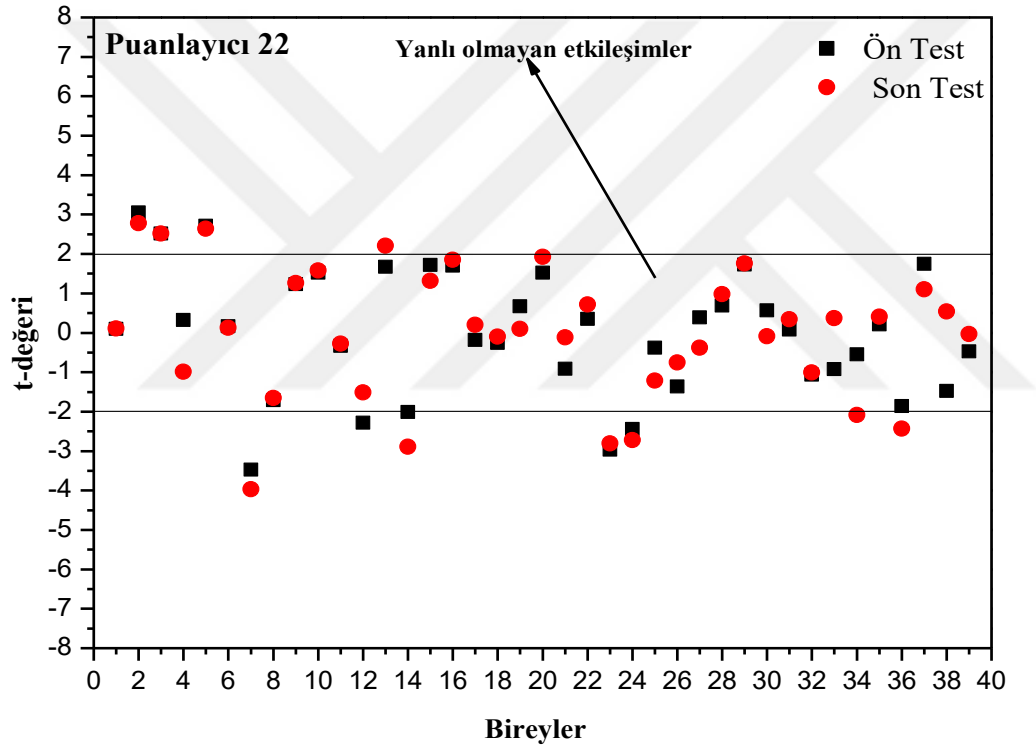
Tablo 22

Deney Grubundaki Puanlayıcılara İlişkin Yanlı Olan Puanlayıcı x Öğrenci Kompozisyonu Etkileşimleri

Puanlayıcı Kodu	Öntest yanlı etkileşimler (%)	Sontest yanlı etkileşimler (%)
SDP01	9 (%23,08)	2 (%5,13)
SDP02	7 (%17,95)	9 (%23,08)
SDP03	7 (%17,95)	10 (%25,64)
SDP04	13 (%33,33)	5 (%12,82)
SDP05	7 (%17,95)	4 (%10,26)
SDP06	11 (%28,21)	2 (%5,13)
SDP07	4 (%10,26)	2 (%5,13)
SDP08	5 (%12,82)	6 (%15,38)
SDP09	13 (%33,33)	6 (%15,38)
SDP10	9 (%23,08)	1 (%2,56)
SDP11	2 (%5,13)	8 (%20,51)
SDP12	3 (%7,69)	4 (%10,26)
SDP13	9 (%23,08)	5 (%12,82)
SDP14	7 (%17,95)	8 (%20,51)
SDP15	13 (%33,33)	8 (%20,51)
SDP16	3 (%7,69)	2 (%5,13)
SDP17	5 (%12,82)	2 (%5,13)
SDP18	7 (%17,95)	6 (%15,38)
SDP19	5 (%12,82)	8 (%20,51)
SDP20	4 (%10,26)	3 (%7,69)
SDP21	6 (%15,38)	3 (%7,69)
SDP22	9 (%23,08)	9 (%23,08)
Toplam (%)	160 (%19,00)	110 (%12,82)

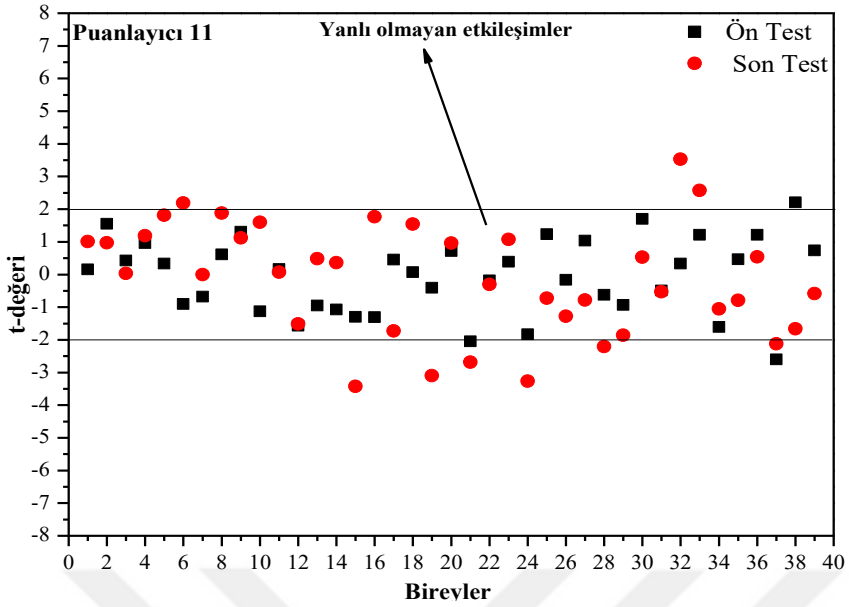
Tablo 22 incelendiğinde, puanlayıcı eğitimi sonrası performans değerlendirme sürecinde; 14 puanlayıcının (1, 4, 5, 6, 7, 9, 10, 13, 15, 16, 17, 18, 20 ve 21 numaralı puanlayıcılar) yanlılık davranışını azalttığını (eğitimden faydalanma), 7 puanlayıcının (2, 3, 8, 11, 12, 14 ve 19 numaralı puanlayıcılar) yanlılık davranışını artırdığını (eğitimden faydalanmama) ve 1

puanlayıcının (22 numaralı puanlayıcı) ise puanlayıcı davranışının sabit kaldığı görülmektedir. Bijani (2018) tarafından yapılan araştırmada, puanlayıcı eğitimine yönelik olumlu tutuma sahip olan puanlayıcılar tutarlılığı yüksek ve yanlılığı az puanlamalar yapmasına rağmen, puanlayıcı eğitimine yönelik olumsuz tutuma sahip olan puanlayıcılar ise tutarlılığı düşük ve yanlılığı fazla olan puanlamalar yaptığı bulunmuştur. Mevcut araştırmada deneysel işlemin yapıldığı süre içerisinde bazı puanlayıcıların olumsuz hal ve tavırlarının olduğu gözlemlenmiş ve bunun sonucunda bu puanlayıcıların eğitim sonrası daha fazla yanlı puanlamalar yapmış olabileceği düşünülmektedir.



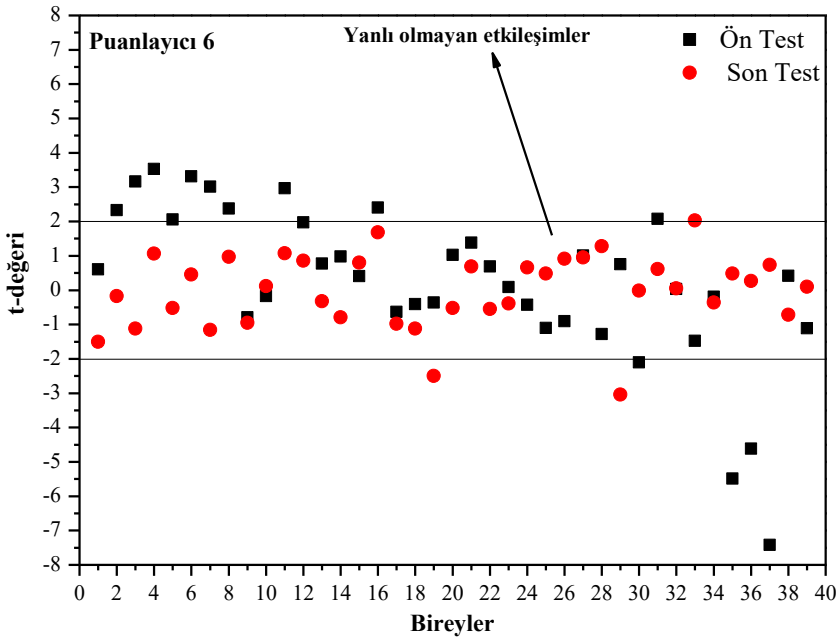
Şekil 4. 22 numaralı puanlayıcı için *puanlayıcı x öğrenci kompozisyonu* etkileşimleri.

Şekil 4 incelendiğinde, yirmi iki numaralı puanlayıcının ön test ve son test puanlamaları arasında önemli bir farklılığın olmadığı görülmektedir. Başka bir değişle verilen puanlayıcı eğitiminin 22 numaralı puanlayıcı üzerinde olumlu veya olumsuz herhangi bir etki meydana getirmediği söylenebilir.



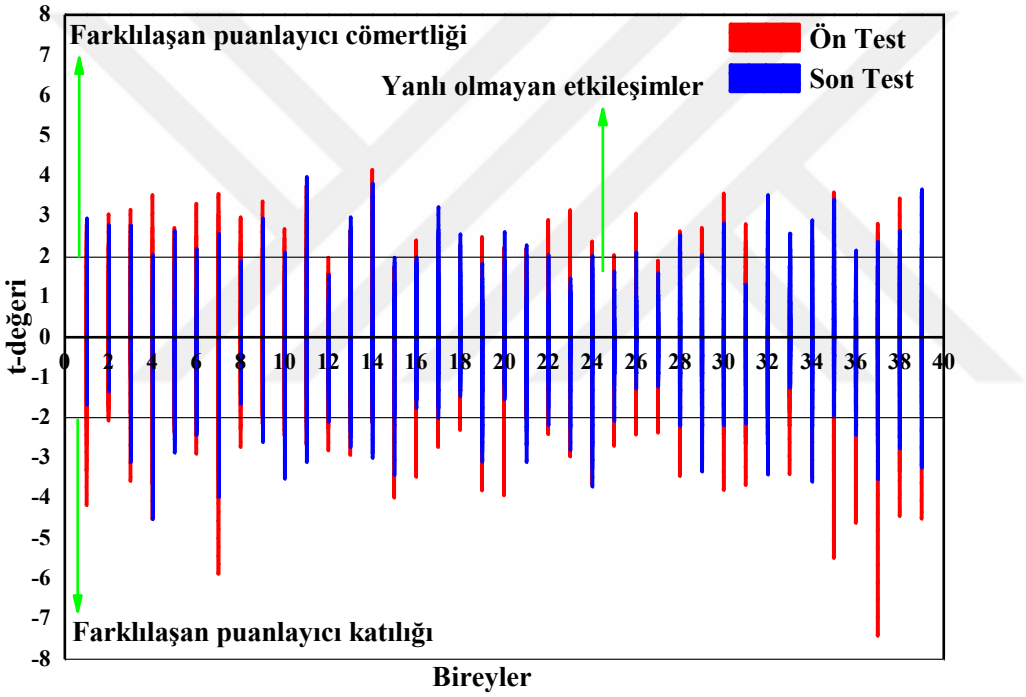
Şekil 5. 11 numaralı puanlayıcı için *puanlayıcı x öğrenci kompozisyonu* etkileşimleri.

Şekil 5 incelendiğinde, 11 numaralı puanlayıcının ön test ve son test puanlamaları arasında önemli farklılıkların olduğu görülmektedir. Örneğin, 11 numaralı puanlayıcı ön testte 39 öğrenci kompozisyonu içinden sadece 2 kompozisyonu yanlı puanlamış iken, puanlayıcı eğitimi sonrası 8 kompozisyonu değerlendirirken yanlı bir davranış sergilediği görülmektedir. Başka bir değişle verilen puanlayıcı eğitiminin 11 numaralı puanlayıcı üzerinde olumsuz bir etki meydana getirdiği söylenebilir.



Şekil 6. Altı numaralı puanlayıcı için *puanlayıcı x öğrenci kompozisyonu* etkileşimleri.

Şekil 6 incelendiğinde, altı numaralı puanlayıcının ön test ve son test puanlamaları arasında önemli farklılıkların olduğu görülmektedir. Örneğin, altı numaralı puanlayıcı ön testte 39 öğrenci kompozisyonu içinden 11 kompozisyonu yanlı puanlamış iken, puanlayıcı eğitimi sonrası sadece 2 kompozisyonu değerlendirirken yanlı bir davranış sergilediği görülmektedir. Başka bir değişle verilen puanlayıcı eğitiminin altı numaralı puanlayıcı üzerinde olumlu bir etki meydana getirdiği söylenebilir. Deney grubundaki tüm puanlayıcıların ön ve son testlerinin *puanlayıcı x öğrenci kompozisyonu* etkileşimlerinin grafiksel gösterimi Şekil 7'de verilmiştir.



Şekil 7. Deney grubundaki tüm puanlayıcılar için *puanlayıcı x öğrenci kompozisyonu* etkileşimleri.

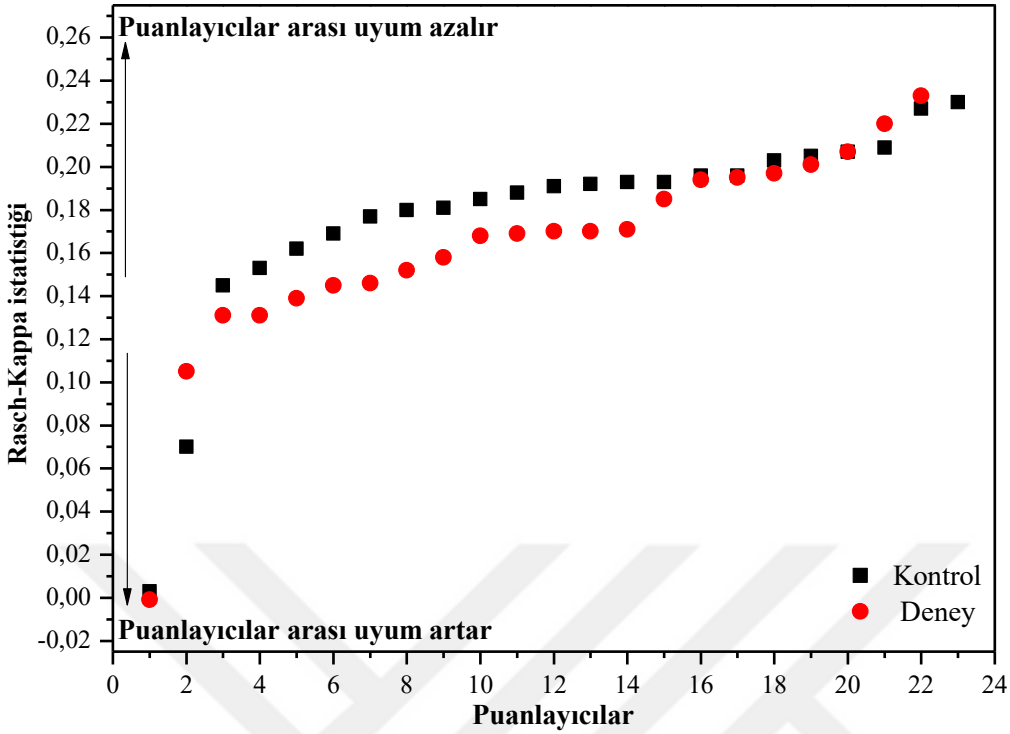
Şekil 7 incelendiğinde, puanlayıcıların ön testini temsil eden kırmızı çizgilerin daha fazla ±2 aralığının dışında olduğu ve puanlayıcıların puanlayıcı eğitiminden sonraki puanlamalarını temsil eden mavi çizgilerin ise daha az ±2 aralığının dışında olduğu gözlemlenmektedir. Şekil 7'e göre bazı bireylerin diğer bireylere göre daha fazla puanlayıcı yanlılığına maruz kaldığı görülmektedir. Örneğin, puanlayıcılar diğer bireylere göre 37 numaralı bireyin performansını değerlendirme sürecinde daha fazla katı davranmıştır. Ayrıca verilen

puanlayıcı eğitiminin genel olarak puanlayıcı davranışı üzerinde olumlu bir etki meydana getirdiği ve bunun sonucunda ölçümlerin geçerliğine katkı sağladığı söylenebilir.

Puanlayıcı eğitimi sonrası, deney ve kontrol grubundaki puanlayıcıların yanlılık davranışları incelendiğinde, grup düzeyinde puanlayıcı yanlılığının performans değerlendirme sürecinde puanlamaya karıştığı belirlenmiştir. Kontrol grubunda ön testte yapılan puanlamaların %23'ü yanlı iken, son testte puanlamaların %18'inin yanlı olduğu bulunmuştur. Deney grubunda ise ön testte yapılan puanlamaların %19'u yanlı iken, son testte puanlamaların %13'ünün yanlı olduğu tespit edilmiştir. Bu sonuca göre hem kontrol grubunda hem de deney grubunda puanlayıcı yanlılığı davranışının azaldığı görülmekle birlikte deney grubunda bu azalışın daha fazla olduğu bulunmuştur. Bijani (2018), Fahim ve Bijani (2011) ve May (2008) tarafından yapılan araştırmalar incelendiğinde, puanlayıcı eğitiminin yanlılık üzerinde etkili olduğu ve yanlılık davranışını azalttığı görülmektedir. Buna göre araştırma sonuçlarının alanyazın tarafından desteklendiği söylenebilir.

Puanlayıcı Eğitiminin Puanlayıcı Güvenirliği Üzerindeki Etkisi

Çok yüzeyli Rasch analizinde her bir puanlayıcı (*puanlayıcı içi*) ve puanlayıcının diğer puanlayıcılar (*puanlayıcılar arası*) ile olan uyumu için iki oran hesaplanmaktadır. Bunlar; gözlenen uyum oranı (yüzdesi) ve beklenen uyum oranıdır. Bu iki oran birbirine yakın ise puanlayıcı içi ve puanlayıcılar arası güvenirliğinin iyi olduğu belirtilir. Çok yüzeyli Rasch analizinden elde edilen bu iki oran kullanılarak her bir puanlayıcının puanlayıcı güvenirliği (puanlayıcı içi) hesaplanmaktadır. Deney ve kontrol grubundaki puanlayıcıların ön testlerine ilişkin puanlayıcı güvenirliklerinin karşılaştırılması (Rasch-Kappa değerleri) Şekil 8'de verilmiştir.



Şekil 8. Deney ve kontrol grubundaki puanlayıcıların ön test ölçümlerine ilişkin puanlayıcı güvenirlikleri.

Şekil 8 incelendiğinde, Deney ve kontrol grubundaki puanlayıcıların güvenirlik değerlerinin birbirine benzer oldukları bununla birlikte deney grubunun biraz daha yüksek güvenirlik değerine sahip olduğu görülmektedir. Her iki grubun puanlayıcı güvenirliklerinin istatistiksel anlamlılığı için bağımsız örneklemler t-testi yapılmıştır. Yapılan t-testi Tablo 23'de verilmiştir.

Tablo 23

Deney ve Kontrol Grubundaki Puanlayıcıların Ön Testlerinden Elde Edilen Puanlayıcı Güvenirliklerine İlişkin Bağımsız Örneklemler t-testi

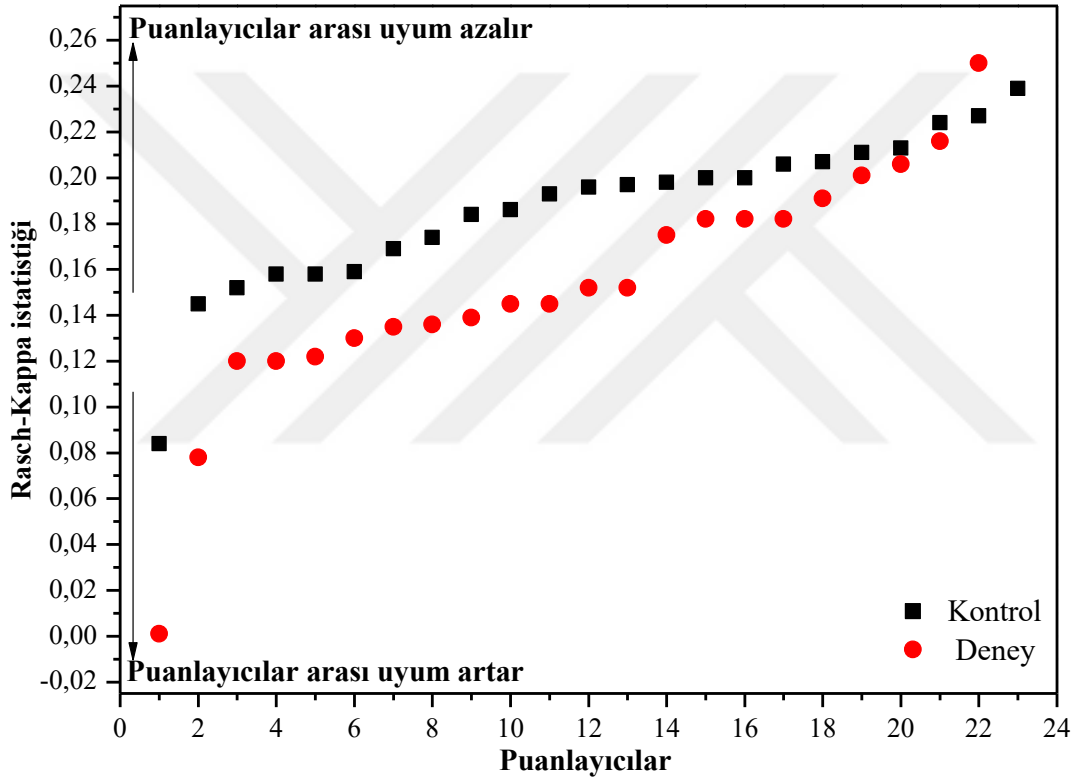
Test	Grup	N	$\overline{X}$	SS	t	sd	p
Ön test	Kontrol	23	0,18	0,05	0,911	43	0,368
	Deney	22	0,16	0,05			

Not. *p<,05 Kriterler: "Kontrol=1"; "Deney=2"

Tablo 23 incelendiğinde, deney ve kontrol gruplarının ön testinden elde edilen puanlayıcı güvenirliklerinin istatistiksel olarak benzer oldukları bulunmuştur (t_{43} =0,911, p = 0,368 > 0,05). Deney ve kontrol grubunun Rasch-Kappa değerlerinin ortalamalarına bakıldığında,

deney grubundaki puanlayıcılar ($\bar{X}$ = 0,16; S=0,05) kontrol grubundaki puanlayıcılara ($\bar{X}$ = 0,18; S=0,05) göre daha yüksek puanlayıcı güvenirlik değerine sahip olmasına rağmen istatistiksel olarak bu farkın anlamlı olmadığı bulunmuştur.

Puanlayıcı eğitimi öncesi deney ve kontrol grubunun puanlayıcı güvenirliklerinin istatistiksel olarak benzer oldukları belirlendikten sonra, puanlayıcı eğitimi sonrası puanlayıcı güvenirliklerine bakılmıştır. Deney ve kontrol grubundaki puanlayıcıların son testine ilişkin güvenirlik değerlerinin karşılaştırılması Şekil 9'da sunulmuştur.



Şekil 9. Deney ve kontrol grubundaki puanlayıcıların son test ölçümlerine ait puanlayıcı güvenirlikleri.

Şekil 9 incelendiğinde, deney grubundaki puanlayıcıların kontrol grubundaki puanlayıcılara göre daha yüksek puanlayıcı güvenirlik değerlerine (daha fazla uyuma) sahip olduğu görünmektedir. Deney ve kontrol grubunun puanlayıcı güvenirliklerine ilişkin farkın istatistiksel anlamlılığı için bağımsız örneklemler t-testi yapılmıştır. Yapılan analizin sonuçları Tablo 24'de verilmiştir.

Tablo 24

Deney ve Kontrol Grubundaki Puanlayıcıların Son Testlerinden Elde Edilen Puanlayıcı Güvenirliklerine İlişkin Bağımsız Örneklemler t-testi

	Grup	N	$\overline{X}$	SS	t	sd	p	η^2
Rasch-Kappa fark	Kontrol	23	0,010	0,04	2,061	43	0,045*	0,09
	Deney	22	-0,010	0,03				

Not. *p<,05 Kriterler: "Kontrol=1"; "Deney=2"

Tablo 24 incelendiğinde, deney ve kontrol gruplarının son testinden elde edilen puanlayıcı güvenirliklerinin istatistiksel olarak farklı oldukları bulunmuştur (t_{43} =2,061, p = 0,045 < 0,05; η^2=0,09). Deney ve kontrol grubunun Rasch-Kappa fark değerlerinin ortalamalarına bakıldığında, deney grubundaki puanlayıcılar ($\overline{X}$ = -0,010; S=0,03) kontrol grubundaki puanlayıcılara ($\overline{X}$ = 0,010; S=0,04) göre daha yüksek puanlayıcı güvenirlik değerine sahip ve bu fark istatistiksel olarak anlamlı bulunmuştur.

KTK'ya dayalı olan Krippendorff alfa katsayısı kullanılarak puanlayıcılar arası güvenirlik katsayısı kestirilmiştir. Krippendorff alfa katsayısı kestirilirken tüm puanlayıcıların her bir öğrencinin her bir kriterine verdiği puanlar dikkate alınmıştır. Araştırmada kullanılan ölçme aracı 15 kritere sahip olduğundan toplam 15 tane Krippendorff katsayısı elde edilmiş ve bu katsayıların ortalaması alınarak tek bir değer elde edilmiştir. Deney ve kontrol grubundaki puanlayıcıların ön ve son testlerine ilişkin Krippendorff katsayısı ve ÇYRÖM analizi ile elde edilen Rasch-Kappa (1 değerinden çıkarılmış) güvenirlik katsayısının karşılaştırılması Tablo 25'de verilmiştir.

Tablo 25

Krippendorff Alfa ve Rasch-Kappa Güvenirlik Katsayılarının Karşılaştırılması

Grup	Krippendorff Alfa		1 – Rasch Kappa	
	Ön Test	Son Test	Ön Test	Son Test
Kontrol	0,2866	0,2864	0,8704	0,8746
Deney	0,3119	0,3439	0,8849	0,9212

Tablo 25 incelendiğinde, Krippendorff alfa katsayısı kullanılarak kestirilen güvenirlik değerlerinin Rasch-Kappa değerlerine göre daha düşük olduğu görülmektedir. Bu durumun temel nedenlerinden biri dayandıkları kuramların ve matematiksel modellerin farklı

olmasından kaynaklanıyor olabilir. Burada dikkat edilmesi gereken husus, her iki katsayının ön ve son testler arasındaki değişim miktarının tutarlı olmasıdır.

Puanlayıcı eğitimi öncesi deney ve kontrol grubundaki puanlayıcıların puanlayıcı güvenirliklerinin istatistiksel olarak benzer olduğu bulunmuştur. Puanlayıcı eğitimi sonrası deney ve kontrol grubundaki puanlayıcıların güvenirlikleri incelendiğinde, istatistiksel olarak anlamlı bir farkın olduğu ve bu farkın deney grubu lehine olduğu belirlenmiştir (t_{43} =2,061, p = 0,045 < 0,05; η2=0,09). Ayrıca verilen puanlayıcı eğitiminin, güvenirlik üzerinde orta düzeyde bir etkiye sahip olduğu ve puanlayıcı güvenirliğindeki değişkenliğin %9,00'nun puanlayıcı eğitimi ile açıklanabildiği belirlenmiştir. Davis (2016) tarafından yapılan araştırmada puanlayıcı eğitiminin puanlayıcı güvenirliğini %13 artırdığı bulunmuştur. Benzer olarak Moser vd. (2016) tarafından yapılan araştırmada ise referans çerçevesi eğitimi deseninin puanlayıcı güvenirliğini artırdığı bulunmuştur. Bu araştırmada kullanılan puanlayıcı eğitiminin bir kısmını referans çerçevesi eğitimi oluşturduğundan, elde edilen sonuçların alanyazın tarafından desteklendiği görülmektedir.

Çok Yüzeyli Rasch Ölçme Modelinin ve Genellenebilirlik Kuramının Analiz Sonuçlarının Karşılaştırılması

Çok yüzeyli Rasch analizi ile elde edilen sonuçların dış geçerliği için aynı veriler genellenebilirlik kuramı kullanılarak da analiz edilmiştir. Burada odaklanan nokta çok yüzeyli Rasch analizinin açıkladığı varyans oranları ile Genellenebilirlik kuramının açıkladığı varyans oranlarının birbiri ile örtüşüp örtüşmediğidir. Bu bağlamda ilk olarak deney ve kontrol gruplarının ön testlerine ilişkin çok yüzeyli Rasch analizi ile genellenebilirlik kuramının varyans bileşenlerinin toplam varyansı açıklama yüzdeleri Tablo 26'da verilmiştir.

Tablo 26

Deney ve Kontrol Grubundaki Puanlayıcıların Ön Testlerine İlişkin G-Kuramı ve Çok Yüzeyli Rasch Analizlerinin Karşılaştırılması

	Kontrol Grubu		Deney Grubu	
Varyans	Açıklanan varyans %		Açıklanan varyans %	
kaynağı	G Kuramı	ÇYRÖM	G Kuramı	ÇYRÖM
O	19,90	21,50	20,60	22,55
P	5,80	6,68	7,80	8,43
K	7,20	7,14	3,60	4,21
PXO	14,00	16,50	11,50	13,41
KXO	10,60	10,60	14,80	14,60
PXK	6,60	6,67	5,90	6,00
PXOXK	35,80	27,24	35,80	27,20
Artık varyans	--	3,67	--	3,60
Toplam (%)	100	100	100	100

Not. O: Öğrenci kompozisyonu, P: Puanlayıcı, K: Kriter, PXO: Puanlayıcı ve öğrenci kompozisyonu etkileşimi, KXO: Kriter ve öğrenci kompozisyonu etkileşimi, PXK: puanlayıcı ve kriter etkileşimi, PXOXK: Puanlayıcı, öğrenci kompozisyonu ve kriter etkileşimi.

Tablo 26 incelendiğinde, her iki ölçme modelinin de birbirine benzer sonuçlar ürettiği görülmektedir. Tablo 26'a göre *puanlayıcı x öğrenci kompozisyonu* etkileşiminin açıkladığı varyans yüzdesi kontrol grubu için G kuramında %14,00, çok yüzeyli Rasch analizinde %16,50 iken, deney grubu için G kuramında %11,50, çok yüzeyli Rasch analizinde ise %13,41 olduğu görülmektedir. Deney ve kontrol gruplarındaki *puanlayıcı x öğrenci kompozisyonu* etkileşimi sonucunda oluşan varyans yüzdelerinin büyük çoğunluğu puanlayıcıların yanlılık analizinde anlamlı çıkan etkileşimlerinden kaynaklanmaktadır.

Deney ve kontrol gruplarındaki puanlayıcıların puanlayıcı eğitimi öncesinde varyans bileşenlerinin toplam varyansı açıklama yüzdeleri incelendikten sonra, puanlayıcı eğitimi sonrası puanlayıcıların varyans bileşenlerinin toplam varyansı açıklama yüzdeleri incelenmiştir. Deney ve kontrol grubundaki puanlayıcıların son testlerine ilişkin çok yüzeyli Rasch analizi ile genellenebilirlik kuramının varyans bileşenlerinin toplam varyansı açıklama yüzdeleri Tablo 27'de verilmiştir.

Tablo 27

Deney ve Kontrol Grubundaki Puanlayıcıların Son Testlerine İlişkin G-Kuramı ve Çok Yüzeyli Rasch Analizlerinin Karşılaştırılması

	Kontrol Grubu		Deney Grubu	
Varyans	Açıklanan varyans %		Açıklanan varyans %	
kaynağı	G Kuramı	ÇYRÖM	G Kuramı	ÇYRÖM
O	19,90	21,44	24,30	28,35
P	4,60	7,34	4,80	4,39
K	6,80	6,34	5,40	5,21
PXO	9,40	11,72	7,10	9,45
KXO	12,80	12,16	15,40	15,09
PXK	8,40	8,61	6,00	6,23
PXOXK	38,10	28,63	37,10	26,89
Artık varyans	--	3,76	--	4,39
Toplam (%)	100	100	100	100

Not. O: Öğrenci kompozisyonu, P: Puanlayıcı, K: Kriter, PXO: Puanlayıcı ve öğrenci kompozisyonu etkileşimi, KXO: Kriter ve öğrenci kompozisyonu etkileşimi, PXK: puanlayıcı ve kriter etkileşimi, PXOXK: Puanlayıcı, öğrenci kompozisyonu ve kriter etkileşimi.

Tablo 27'e bakıldığında hem genellenebilirlik hem de çok yüzeyli Rasch analiz sonuçlarının birbirine yakın değerler ürettikleri görülmektedir. Kontrol grubuna ilişkin kestirilen varyans bileşenlerinin ön ve son test sonuçlarının birbirine yakın olduğu fakat deney grubu için kestirilen varyans bileşenlerinin ön ve son test sonuçları arasında bir farkın olduğu belirlenmiştir.

Ön ve son teste ilişkin çok yüzeyli Rasch analizi sonucunda her bir ana (yüzey) ve ortak etki (yüzey etkileşimleri) için elde edilen varyans bileşenlerinin genellenebilirlik analizi sonucu elde edilen varyans bileşenleri ile benzer oranlara sahip olduğu bulunmuştur. Bu bulguya dayanarak performans değerlendirme sürecinde ölçümlere karışan varyans miktarının belirlenmesinde hem genellenebilirlik hem de çok yüzeyli Rasch analizlerinin kullanımı faydalı görünmektedir. Kim ve Wilson (2009) tarafından yapılan araştırmada, öğrencilerin akademik yazma becerilerinin değerlendirilmesinde genellenebilirlik kuramı ve çok yüzeyli Rasch analizleri kullanılarak incelenmiş ve her iki analiz yöntemin benzer sonuçlar verdiği bulunmuştur. Araştırmacılar tarafından performans değerlendirme sürecinde daha güvenilir ölçümlerin elde edilmesinde hem genellenebilirlik hem de çok yüzeyli Rasch analizinin kullanışlı olduğu belirtilmiştir. Fakat Genellenebilirlik kuramının bütüncül bir bakış açısı

sağladığı fakat daha detaylı incelemeler için çok yüzeyli Rasch analizinin kullanılması gerektiği belirtilmiştir. Iramaneerat, Yudkowsky, Myford ve Downing (2008) tarafından yapılan araştırmada ise, bir yeterlik sınavının değerlendirilmesinde genellenebilirlik kuramı ve çok yüzeyli Rasch analizi kullanılmış olup her iki yöntemin benzer sonuçlar verdiği ve ölçme araçlarının geliştirilmesi ve izlenmesinde her iki yöntemin de kullanılabileceği belirtilmiştir. Buna göre araştırma sonucunda elde edilen bulguların alanyazında desteklendiği görülmektedir.

BÖLÜM 5

SONUÇLAR

Bu araştırma, öğrencilerin kompozisyonlarının değerlendirilmesi sürecinde objektifliği azaltan problemli puanlayıcı davranışlarının belirlenmesi ve bu problemli davranışların ölçümlere karışma düzeylerinin azaltılması veya kontrol edilebilir seviye getirilmesine yönelik uygulamalardan biri olan puanlayıcı eğitiminin uygulanması amacıyla gerçekleştirilmiştir. Bu bölümde araştırmanın sonuçları ve bu sonuçlar doğrultusunda gelecekte yapılacak çalışmalar için öneriler sunulmuştur. Araştırma kapsmında ulaşılan sonuçlar araştırma sorularının sırasına uygun olarak verilmiştir. Araştırmanın sonuçları şu şekildedir;

Bu araştırmada öğrenci kompozisyonları, kriterler ve puanlayıcılar olmak üzere üç yüzey bulunduğundan her bir yüzey için ölçüm değerleri ve uygunluk istatistikleri incelenmiştir. Deney ve kontrol grubundaki puanlayıcıların ön test ölçümlerine ilişkin yapılan çok yüzeyli Rasch analizi sonucunda 12 puanlayıcının model-veri uyumunu bozan uygunluk değerlerine sahip olduğu belirlenmiş ve analiz dışı bırakılmış olup, son test için yapılan çok yüzeyli Rasch analizinde kabul edilebilir aralık dışında herhangi bir puanlayıcının olmadığı görülmüştür. Sonuç olarak uygunluk değerleri uçlarda olan puanlayıcılar çıkarılmış ve her iki test için de kabul edilebilir uygunluk değerleri ve ölçüm değerleri ile analizler gerçekleştirilmiştir. Benzer şekilde kriter yüzeyinin ön ve son test ölçümleri incelenmiş ve uygunluk ölçümlerinin model veri uyumunu bozmadıkları bulunmuştur. öğrenci kompozisyonları yüzeyinin ön ve son test ölçümleri incelendiğinde de model veri uyumunu

bozan kompozisyon olmadığı tespit edilmiştir. Böylece yapılan ölçümlerin güvenirliğine ve ölçümlere dayalı yapılan çıkarımların geçerliğine katkı sağlanmıştır.

Puanlayıcı eğitimi öncesi incelenen problemli puanlayıcı davranışlarının (katılık, cömertlik, merkeze yönelim, halo, farklılaşan katılık ve farklılaşan cömertlik) ölçümlere karışma düzeyleri incelendiğinde, deney ve kontrol grubunudaki puanlayıcıların istatistiksel olarak benzer puanlayıcı davranışları gösterdikleri belirlenmiştir. Ön test ölçümlerine ilişkin çok yüzeyli Rasch analizleri incelendiğinde; katılık, cömertlik, farklılaşan katılık ve farklılaşan cömertlik davranışlarının hem grup düzeyinde hem de bireysel düzeyde ortaya çıkmakta iken merkeze yönelim ve halo davranışının sadece bireysel düzeyde ortaya çıktığı tespit edilmiştir. Bu araştırmada; birden fazla puanlayıcı ve analitik dereceli puanlama anahtarı kullanılmasına rağmen birey performansının değerlendirilmesi sürecinde bireyin yeteneğinin bir veya daha fazla puanlayıcı davranışından etkilendiği görülmüştür.

Araştırma kapsamında durum belirlemesi yapılan performans olarak öğrencilerin ikinci dil akademik yazma becerileri seçilmiştir. Bu bağlamda ingilizce öğretmenliği birinci sınıf öğrencilerine IELTS tarafından yayımlanan yazma görevi verilmiş ve öğrencilerden bir kompozisyon yazmaları istenmiştir. Öğrenci kompozisyonlarının değerlendirilmesi için araştırmacı ve alan uzmanı bir araştırmacı tarafından akademik yazma becerileri için analitik dereceli puanlama anahtarı geliştirilmiştir. Geliştirilen ölçme aracının güvenirlik ve geçerlik çalışması yapılmış ve akademik yazma becerilerinin değerlendirilmesinde ölçme aracının kullanılabileceği bulunmuştur. Sonuç olarak akademik yazma becerilerinin değerlendirilmesinde geçerli ve güvenilir ölçümler veren bir ölçme aracı geliştirilmiştir.

Araştırmaya katılan puanlayıcılara, referans çerçevesi eğitimi ve puanlayıcı hatası eğitimi desenlerinin birleştirilmesi ile oluşturulan bir puanlayıcı eğitimi her hafta birer saat olmak üzere toplam dört hafta verilmiştir. Puanlayıcı eğitimi sonrası araştırma kapsamında incelenen altı puanlayıcı davranışının ölçümlere yeniden karıştığı fakat deney grubundaki ölçümlere karışma düzeylerinin görece daha az olduğu bulunmuştur. Kontrol grubundaki puanlayıcıların ön ve son test ölçümleri arasında istatistiksel olarak bir farklılılığın olmadığı

fakat deney grubundaki puanlayıcıların ön ve son test ölçümleri arasında istatistiksel olarak bir farklılığın olduğu bulunmuştur. Puanlayıcı eğitimi sonrası deney grubundaki katılık, cömertlik, farklılaşan katılık ve farklılaşan cömertlik davranışlarının ölçümlere karışma düzeyleri azalmış olmasına rağmen yok olmamıştır. Bunun temel nedeni alanyazında da belirtildiği gibi (Moore, 2009, s.11) puanlayıcı kaymasından dolayı olduğu düşünülmektedir. Sonuç olarak verilen puanlayıcı eğitiminin puanlayıcı katılığı, puanlayıcı cömertliği, farklılaşan puanlayıcı katılığı ve farklılaşan puanlayıcı cömertliği üzerinde önemli bir etkiye sahip olduğu bulunmuştur.

Deney grubuna verilen puanlayıcı eğitiminin puanlayıcı güvenirliği (puanlayıcı içi) üzerinde orta düzeyde ve istatistiksel olarak anlamlı bir etkiye sahip olduğu bulunmuştur. Bu sonuca göre puanlayıcı güvenirliğinin önemli olduğu ve önemli kararların alındığı durum belirleme çalışmalarında puanlayıcı eğitiminin kullanması faydalı olacaktır. Alanyazın incelendiğinde araştırmanın ulaştığı sonuçların desteklendiği görülmektedir. Çok yüzeyli Rasch her bir puanlayıcıya ilişkin güvenirlik değeri verdiğinden hangi puanlayıcıların daha güvenilir olduğu veya daha güvenilmez olduğu belirlenerek objektifliği daha yüksek olan durum belirleme çalışmaları yapılabilir.

Birey performansının durum belirlemesi sürecinde, öğrenci kompozisyonlarına karışan puanlayıcı kaynaklı varyansı belirlemek amacıyla çok yüzeyli Rasch ve genellenebilirlik kuramı sonuçları incelenmiştir. Yapılan analizler neticesinde her iki analiz yöntemin de benzer sonuçlar ürettiği bulunmuştur. Dolaysıyla birey performansının değerlendirildiği puanlama çalışmalrında puanlayıcı kaynaklı varyansı belirleme amacı varsa her iki analiz yöntemi de tercih edilebilir. Fakat puanlayıcı kaynaklı varyansı oluşturan puanlayıcıların hangileri olduğu ve hangi etkilişiemler sonucunda oluştuğu gibi bir amaç varsa çok yüzeyli Rasch ölçme modelinin tercih edilmesi daha doğru bir yaklaşım olabilir.

Öneriler

Bu araştırma, öğrenci kompozisyonlarının değerlendirilmesi sürecinde, ölçümlere karışan problemli puanlayıcı davranışlarının incelenmesi ve bu davranışların ölçümlere karışma düzeylerinin azaltılması veya kontrol altına alınmasına yönelik uygulamalardan biri olan puanlayıcı eğitiminin uygulaması amacıyla gerçekleştirilmiştir. Araştırma kapsamında ulaşılan sonuçlar doğrultusunda araştırmacılara ve uygulayıcılara bazı öneriler sunulmuştur.

Araştırmacılara Yönelik Öneriler

Bu araştırmada birden fazla puanlayıcı, dereceli puanlama anahtarı ve puanlayıcı eğitimi kullanılmış olmasına rağmen incelenen altı problemli puanlayıcı davranışlarının ölçümlere yeniden karıştığı bulunmuştur. Bu bağlamda performans değerlendirme sürecinde problemli puanlayıcı davranışlarının incelenip belirlenmesi, ölçümlerin güvenirliğine ve ölçümlerden yapılan çıkarımların geçerliğine katkı sağlayacağından, puanlayıcı kaynakları belirlenmelidir.

Verilen puanlayıcı eğitiminin, araştırma kapsamında incelenen altı problemli puanlayıcı davranışı üzerinde etkili olduğu bulunmuştıur. Mevcut çalışmada kapsamında verilen puanlayıcı eğitimi modülünün alanyazında rapor edilen diğer problemli puanlayıcı davranışları üzerindeki etkisi incelenebilir.

Mevcut çalışmada puanlayıcı eğitimi desenlerinden puanlayıcı hatası eğitimi ve referans çerçevesi eğitimi birleştirilerek öğrenci kompozisyonlarının değerledirilmesinde daha objektif sonuçlar vermiştir. Dolaysıyla durum belirlemesi yapılacak performans değiştiğinde bu araştırmada kullaılan puanlayıcı eğitim modülü etkili olmayabilir. Bu bağlamda durum belirlemesi yapılan performansa uygun puanlayıcı eğitim modülleri geliştirilebilir.

Puanlayıcı eğitiminin puanlayıcı güvenirliğini artırdığı tespit edilmiştir. Bu bağlamda puanlayıcı güvenirliğinin yüksek olduğu çalışmalar için puanlayıcı eğitiminin uygulanmasının faydalı olacağı beklenmektedir.

Mevcut araştırmada puanlayıcı eğitimi kalabalık bir gruba (n = 22) uygulanmıştır. Alanyazında daha küçük gruplarla (n =5 veya 6) çalışmanın daha etkili olduğu belirtilmiş olup, gelecek araştırmalarda aynı desenin daha küçük gruplarda uygulaması yapılıp etkililiği incelenebilir.

Bu araştırma kapsamında verilen puanlayıcı eğitimi hazırlık aşaması ile birlikte sekiz saat sürmüştür. Problemli puanlayıcı dvranışlarının ölçümlere karışma düzeylerinin daha fazla azaltılmasının istendiği çalışmalarda daha geniş sürede puanlayıcı eğitimleri verebilir.

Uygulayıcılara Yönelik Öneriler

Milli Eğitim Bakanlığı, merkezi sınavlarda açık uçlu sorulara geçeçeğini duyurmuş ve ilerleyen zamanlarda bunu gerçekleştireceğini belirtmiştir. Milli Eğitim Bakanlığı'nın uyguladığı sınavların hayati öneme sahip olduğu düşünüldüğünde, öğrencilerin performansları değerlendirilirken puanlamaların objektif olması gerekmektedir. Bu bağlamda puanlayıcı eğitiminin problemli puanlayıcı davranışlarının ölçümlere karışma düzeyini azalttığı düşünüldüğünde, puanalyıcılara puanlayıcı eğitiminin verilmesinin faydalı olacağı düşülmektedir.

Jüri üyelerinin yer aldığı ve önemli kararların verildiği ölçme ve değerlendirme uygulamalarında yer alan puanlayıcılara yönelik puanlayıcı eğitiminin verilmesi önerilmektedir. Böylelikle daha adil ve daha objektif değerlendirmeler yapılabilir.

KAYNAKLAR

Abu Kassim, N. L. (2011). Judging behaviour and rater errors: an application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, 11(3), 179-197. http://ejournals.ukm.my/gema/article/view/49 sayfasından erişilmiştir.

Abu Kassim, N.L. (2007). *Exploring rater judging behaviour using the many-facet Rasch model.* Paper Presented in the Second Biennial International Conference on Teaching and Learning of English in Asia: Exploring New Frontiers (TELiA2), Universiti Utara, Malaysia. http://repo.uum.edu.my/3212/ sayfasından erişilmiştir.

Aguinis H. (2012). *Performance management*. Upper Saddle River, NJ: Pearson-Prentice Hall. http://libgen.io/ sayfasından erişilmiştir.

Aguinis, H., Mazurkiewicz, M. D., & Heggestad, E. D. (2009). Using web-based frame-of-reference training to decrease biases in personality-based job analysis: An experimental field study. *Personnel Psychology*, 62(2), 405-438. https://doi.org/10.1111/j.1744-6570.2009.01144.x

Airasian, P. W., (1994). *Classroom Assessment*. New York:McGraw-Hill. http://libgen.io/ sayfasından erişilmiştir.

Alderson, C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press. http://libgen.io/ sayfasından erişilmiştir.

Alharby, E. R. (2006). *A comparison between two scoring methods, holistic vs analytic, using two measurement models, the generalizability theory and the many-facet rasch measurement, within the context of performance assessment.* (Doktora Tezi). http://www.proquest.com/ sayfasından erişilmiştir.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for Educataonal and Psychologlcal Testing.* Washington, DC: American Educational Research Association.

Anastasi, A. (1988). *Psychological testing*. New York: Macmillan. http://libgen.io/ sayfasından erişilmiştir.

Andrich, D. A. (1978). Rating scale formilation for ordered response categories. *Psychometrika*, 43, 561-573. https://doi.org/10.1007/BF02293814

Antonioni, D., & Woehr, D.J. (2001). Improving the quality of multisource rater performance. In D.W. Bracken, C.W. Timmreck and A.H. Church (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and ımplementing MSF processes* (pp. 114- 129). California: Jossey Bass Inc.

Arias, R.M. (2010). Performance assessment. *Papeles del Psicólogo*, 31(1), 85-96.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555-596. https://doi.org/10.1162/coli.07-034-R2

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels–of–processing theory and social facilitation theory perspectives. *Journal of Applied Psychology*, 72, 567–572. https://doi.org/10.1037/0021-9010.72.4.567

Atılgan, H. (2004). *Genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch modelinin karşılaştırılmasına ilişkin bir araştırma.* (Doktora Tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Atılgan, H. (2005). Müzik öğretmenliği özel yetenek seçme sınavının çok-yüzeyli Rasch modeli ile analizi (İnönü üniversitesi örneği). *Eurasian Journal of Educational Measurement*, 20, 62-73. https://www.ebscohost.com/ sayfasından erişilmiştir.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater R V. 2. *Journal of Technology, Learning, and Assessment*, 4, 1–30. https://www.learntechlib.org/p/103244/ sayfasından erişilmiştir.

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *Journal of Technology, Learning, and Assessment*, 10(3),

1-16. https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1603 sayfasından erişilmiştir.

Ayaydın, A. (2011). Güzel sanatlar ve spor liselerinde karşılaşılan sorunlar ve çözüm önerileri: Öğretmenlerin görüşleri. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 8(16), 435-453. http://sbed.mku.edu.tr/article/view/1038000296 sayfasından erişilmiştir.

Bachman, L.F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511667350

Baird, J. A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability. A Comparative exploration from the perspectives of generalisability theory, Rash model and multilevel modelling*. Oxford: University of Oxford for Educational Assessment. http://dera.ioe.ac.uk/17683/1/2013-01-21-marker-effects-and-examination-reliability.pdf sayfasından erişilmiştir.

Baird, J. A., Meadows, M., Leckie, G., & Caro, D. (2017). Rater accuracy and training group effects in expert-and supervisor-based monitoring systems. *Assessment in Education: Principles, Policy & Practice*, 24(1), 44-59. https://doi.org/10.1080/0969594X.2015.1108283

Baker, E. L. (1994). Learning-based assessments of history understanding. *Educational Psychologist,* 29, 97–106. https://doi.org/10.1207/s15326985ep2902_5

Baker, E. L., Abedi, J., Linn, R. L., & Niemi, D. (1995). Dimensionality and generalizability of domain-independent performance assessments. *Journal of Educational Research*, 89, 197–205. https://doi.org/10.1080/00220671.1996.9941205

Baker, F. B. (2001). *The basics of item response theory*. United States of America: ERIC Clearinghouse on Assessment and Evaluation.

Barak, M., Ben-Chaim, D. & Zoller, U. (2007). Purposely teaching for the promotion of higher-order thinking skills: A case of critical thinking. *ResSciEduc*, 37, 353–369. https://doi.org/10.1007/s11165-006-9029-2

Bayat, N. (2014). Öğretmen adaylarının eleştirel düşünme düzeyleri ile akademik yazma başarıları arasındaki ilişki. *Eğitim ve Bilim*, 39(173), 155-168. http://eb.ted.org.tr/index.php/EB/article/view/2333 sayfasından erişilmiştir.

Bennet, J. (1998). *Human resources management*. Singapore: Prentice Hall.

Bernardin, H. J. & Pence, E. C. (1980). Effects of rater training: New response sets and decreasing accuracy. *Journal ofApplied Psychology*, 65, 60-66. https://doi.org/10.1037/0021-9010.65.1.60

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology*, 63(3), 301-308. http://dx.doi.org/10.1037/0021-9010.63.3.301

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, 6(2), 205-212. https://journals.aom.org/doi/abs/10.5465/amr.1981.4287782 sayfasından erişilmiştir.

Bijani, H. (2018). Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training. *Cogent Education*, 5(1), 1-20. https://doi.org/10.1080/2331186X.2018.1460901

Boddy, N., Watson, K., & Aubusson, P. (2003). A trial of the five Es: A referent model for constructivist teaching and learning. *Research in Science Education*, 33, 27–42. https://doi.org/10.1023/A:1023606425452

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York and London: Routledge. https://doi.org/10.4324/9781315814698

Borman, W .C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology*, 60(5), 556-560. https://doi.org/10.1037/0021-9010.60.5.556

Brennan, R.L., Gao, X., & Colton, D.A. (1995). Generalizability analyses of work key listening and writing tests. *Educational and Psychological Measurement*, 55(2), 157-176. https://doi.org/10.1177/0013164495055002001

Brentari, E., & Golia, S. (2008). Measuring job satisfaction in the social services sector with the Rasch model. *Journal of Applied Measurement*, 9(1), 45-56. https://europepmc.org/abstract/med/18180549 sayfasından erişilmiştir.

Brijmohan, A. (2016). *A many-facet Rasch measurement analysis to explore rater effects and rater training in medical school admissions.* (Doktora Tezi). http://www.proquest.com/ sayfasından erişilmiştir.

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

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York: Pearson Education.

Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy.* New York: Pearson Education.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL quarterly*, 32(4), 653-675. https://doi.org/10.2307/3587999

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). *Automated scoring using a hybrid feature identification technique.* In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, Montreal, Quebec, Canada. https://doi.org/10.3115/980845.980879

Büyüköztürk, Ş. (2011). *Deneysel desenler- öntest-sontest kontrol grubu desen ve veri analizi*. Ankara: Pegem Akademi Yayıncılık.

Büyüköztürk, Ş. (2013). *Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum*. Ankara: Pegem Akademi Yayıncılık.

Calfee, R. (2000). To grade or not to grade. *IEEE Intelligent Systems*, 15, 35–37. https://ci.nii.ac.jp/naid/10025900437/ sayfasından erişilmiştir.

Carter, C., Bishop, J. L., & Kravits, S. L. (2002). *Keys to college studying: becoming a lifelong learner.* New Jersey: Printice Hall.

Cascio WF, Aguinis H. (2004). *Applied psychology in human resource management*. Upper Saddle River, NJ: Pearson-Prentice Hall.

Cellar, D.F., Curtis, J.R., Kohlepp, K., Poczapski, P., & Mohiuddin, S. (1989). The effects of rater training, job analysis format and congruence of training on job evaluation ratings. *Journal of Business & Psychology*, 3(4), 387-401. https://doi.org/10.1007/BF01020707

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, 18, 65-81. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.472.9284 sayfasından erişilmiştir.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265-289. https://doi.org/10.3102/10769986022003265

Choi, J., & Wilson, M. R. (2018). Modeling rater effects using a combination of generalizability theory and IRT. *Psychological Test and Assessment Modelin*g, 60(1), 53-80.http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2018_20180323/4_PTAM_Choi___Wilson__2018-03-10__2032.pdf sayfasından erişilmiştir.

Congdon, P., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37(2), 163-178. https://doi.org/10.1111/j.1745-3984.2000.tb01081.x

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, V. S. (2009). Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized, controlled trial. *Journal of General Internal Medicine*, 24(1), 74–79. http://doi.org/10.1007/s11606-008-0842-3

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218-244. https://doi.org/10.1037/0033-2909.90.2.218

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, Ohio: Cengage Learning.

Cronbach, L.I. (1990). *Essentials of psychological testing*. New York: Harper and Row.

Crusan, D. (2010). *Assessment in the second language writing classroom*. Ann Arbor, MI: The University of Michigan Press. https://doi.org/10.3998/mpub.770334

Çekici, Y. E. (2018). Türkçe'nin yabancı dil olarak öğretiminde kullanılan ders kitaplarında yazma görevleri: Yedi iklim ve İstanbul üzerine karşılaştırmalı bir inceleme. *Gaziantep Üniversitesi Eğitim Bilimleri Dergisi*, 2(1), 1-10. http://dergipark.gov.tr/http-dergipark-gov-tr-journal-1517-dashboard/issue/36422/367409 sayfasından erişilmiştir.

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), 217-247. http://eb.ted.org.tr/index.php/EB/article/view/5082 sayfasından erişilmiştir.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları*. Ankara: Pegem Akademi Yayıncılık

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, 33(1), 117-135. https://doi.org/10.1177/0265532215582282

De Ayala, R. J. (2009). *The theory and practice of item response tTheory. Methodology in the social sciences*. New York: Guildford.

DeCotiis, T. A. (1977). An analysis of the external validity and applied relevance of three rating formats. *Organizational Behavior and Human Performance*, 19, 247-266. https://doi.org/10.1016/0030-5073(77)90064-2

DeNisi, A. S., Cafferty, T., & Meglino, B. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Decision Processes*, 33, 360-396. https://doi.org/10.1016/0030-5073(84)90029-1

Deniz, K. (2003). Yazılı anlatım becerileri bakımından köy ve kent beşinci sınıf öğrencilerinin durumu[Özel sayı]. *Türklük Bilimi Araştırmaları*, 13, 233- 255. http://dergipark.gov.tr/download/article-file/156770 sayfasından erişilmiştir.

Doğan, C. D., & Uluman, M. (2017). A comparison of rubrics and graded category rating scales with various methods regarding raters' reliability. *Educational Sciences: Theory and Practice,* 17(2), 631-651. https://doi.org/10.12738/estp.2017.2.0321

Downing, S. M. (2005). Threats to the validity of clinical teaching assessments: what about rater error?. *Medical education,* 39(4), 353-355. https://doi.org/10.1111/j.1365-2929.2005.02138.x

Downing, S. M., & Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38(3), 327-333. https://doi.org/10.1046/j.1365-2923.2004.01777.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), 115-128. https://doi.org/10.1007/s10755-006-9012-x

Ebel, R. L. (1965). *Measuring educational achievement.* New Jersey: Prentice- Hall Press.

Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of educational measurement.* New Jersey: Prentice Hall Press.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155–185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt: Peter Lang.

Education Testing Service (2019). *Rating is kept apart from test administration.* https://www.ets.org/toefl/ibt/scores/ sayfasından erişilmiştir.

Eke, C. (2018). Ortaöğretim fizik dersi öğretim programındaki kazanımların yenilenmiş Bloom taksonomisine göre analizi. *Sosyal Araştırmalar ve Davranış Bilimleri Dergisi*, 4(6), 69-84. http://dergipark.gov.tr/sadab/issue/38907/454687 sayfasından erişilmiştir.

Elder, C, Barkhuizen, G., Knoch, U. & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 24, 37-64. https://doi.org/10.1177/0265532207071511

Ellis, R. O. D., Johnson, K. E., & Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly*, 36(2), 219-233. https://doi.org/10.2307/3588333

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31(2), 93-112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal and T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261-287). Mahway, NJ: Lawrence Erlbaum Associates.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge. https://doi.org/10.4324/9780203073636

Ergai, A., Cohen, T., Sharp, J., Wiegmann, D., Gramopadhye, A., & Shappell, S. (2016). Assessment of the human factors analysis and classification system (HFACS): Intra-rater and inter-rater reliability. *Safety science*, 82, 393-398. https://doi.org/10.1016/j.ssci.2015.09.028

Erman-Aslanoğlu, A. & Kutlu, Ö. (2003). Öğretimde sunu becerilerinin değerlendirilmesinde dereceli puanlama anahtarı (rubric) kullanılmasına ilişkin bir araştırma. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 36 (1-2), 25-36. http://www.acarindex.com/dosyalar/makale/acarindex-1423870052.pdf sayfasından erişilmiştir.

Esfandiari, R. (2015). Rater errors among peer-assessors: applying the many-facet Rasch measurement model. *Iranian Journal of Applied Linguistics*, 18(2), 77-107. https://doi.org/10.18869/acadpub.ijal.18.2.77

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing*, 1(1), 1-16. http://www.ijlt.ir/portal/files/401-2011-01-01.pdf sayfasından erişilmiştir.

Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of raters. *Theory & Practice in Language Studies*, 1(11), 1531-1540. https://doi.org/10.4304/tpls.1.11.1531-1540

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, 34(1), 79-101. https://jalt-publications.org/files/pdf-article/jj2012a-art4.pdf sayfasından erişilmiştir.

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), 76-83. https://pdfs.semanticscholar.org/dd21/ba5683dde8b616374876b0c53da376c10ca9.pdf sayfasından erişilmiştir.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, 66, 127-148. https://doi.org/10.1037/0021-9010.66.2.127

Feldman, M., Lazzara, E. H., Vanderbilt, A. A., & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions*, 32(4), 279-286. https://doi.org/10.1002/chp.21156

Flowers, C. P., & Hancock, D. R. (2003). An interview protocol and scoring rubric for evaluating teacher performance. *Assessment in Education: Principles, Policy and Practice*, 10, 161–168. https://doi.org/10.1080/0969594032000121261

Galati, D.E. (2007). *A g-theory based adjustment process for calibrating marker scores*. (Doktora Tezi). http://www.proquest.com/ sayfasından erişilmiştir.

Gearhart, M., Herman, J. L., Novak, J. R., & Wolf, S. A. (1995). Toward the instructional utility of large-scale writing assessment: Validation of a new narrative rubric. *Assessing Writing*, 2, 207–242. https://doi.org/10.1016/1075-2935(95)90013-6

Gillet, A., Hammond, A. & Martala, M. (2009). *Successful academic writing*. New York: Pearson Longman.

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the "black box" differently: Assessor cognition from three research perspectives. *Medical Education*, 48(11), 1055–1068. http://doi.org/10.1111/medu.12546

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-based assessments as social judgments: Rethinking the etiology of rater errors. *Academic Medicine*, 86(10), 1-7. http://doi.org/10.1097/ACM.0b013e31822a6cf8

Goodrich, H. (1997). Understanding Rubrics: The dictionary may define" rubric," but these models provide more clarity. *Educational Leadership*, 54(4), 14-17.

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

Göçer, A. (2010). Türkçe öğretiminde yazma eğitimi. *Uluslararası Sosyal Araştırmalar Dergisi*, 12 (3), 178-195. http://www.sosyalarastirmalar.com/cilt3/sayi12pdf/gocer_ali.pdf sayfasından erişilmiştir.

Graesser, A. C., & McNamara, D. S. (2012). Automated analysis of essays and open-ended verbal responses. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (pp. 307–325). Washington, DC: American Psychological Association. https://doi.org/10.1037/13619-017

Gronlund, N. E. (1977). *Constructing achievement test.* New Jersey: Prentice-Hall Press.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological bulletin*, 103(2), 265-275. https://doi.org/10.1037/0033-2909.103.2.265

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Güler, N. (2008). *Klasik test kuramı genellenebilirlik kuramı ve Rasch modeli üzerine bir araştırma.* (Doktora Tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

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

Haladyna, T. M. (1997). *Writing test items in order to evaluate higher order thinking.* USA: Allyn & Bacon.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor Quant Methods Psychol*., 8(1), 23-34. https://doi.org/10.20982/tqmp.08.1.p023

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Nijhoff. https://doi.org/10.1007/978-94-017-1988-9

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

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1977). Developments in latent trait theory: A review of models, technical issues, and applications. *Review of Educational Research*, 48, 467–510. https://doi.org/10.3102/00346543048004467

Hansen, T., Elholm-Madsen, E., & Sørensen, A. (2016). The effect of rater training on scoring performance and scale-specific expertise amongst occupational therapists participating in a multicentre study: A single-group pre-post-test study. *Disability and rehabilitation*, 38(12), 1216-1226. https://doi.org/10.3109/09638288.2015.1076069

Hauenstein, N. M. A. (1998). Training raters to increase the accuracy and usefulness of appraisals. In J. Smither (Ed.), *Performance appraisal: State of the art in practice* (pp. 404–444). San Francisco, CA: Josey Bass.

Hauenstein, N. M., & McCusker, M. E. (2017). Rater training: Understanding effects of training content, practice ratings, and feedback. *International Journal of Selection and Assessment*, 25(3), 253-266. https://doi.org/10.1111/ijsa.12177

Holmboe, E., Hawkins, R. E., & Huot, S. J. (2004). Effect of training in direct observation of medical residents' clinical competence: a randomized trial. *Annals of Internal Medicine*, 140, 874–881. https://doi.org/10.7326/0003-4819-140-11-200406010-00008

Howitt, D., & Cramer, D. (2008). *Introduction to statistics in psychology*. Harlow: Pearson Education.

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

IELTS (t.y). *Prepare for IELTS*. https://takeielts.britishcouncil.org/prepare-test/free-sample-tests/writing-sample-test-1-academic/writing-task-2 sayfasından erişilmiştir.

Iramaneerat, C., Yudkowsky, R., Myford, C. M., & Downing, S. M. (2008). Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances in Health Sciences Education*, 13(4), 479-493. https://doi.org/10.1007/s10459-007-9060-8

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). https://tez.yok.gov.tr sayfasından erişilmiştir.

İlhan, M., & Çetin, B. (2014). Performans değerlendirmeye karışan puanlayıcı etkilerini azaltmanın yollarından biri olarak puanlayıcı eğitimleri: Kuramsal bir analiz. *Journal of European Education*, 4(2), 29-38. https://doi.org/10.18656/jee.77087

Jacobs, R. & Kozlowski, S. W. (1985). A closer look at halo error in performance ratings. *Academy of Management Journal*, 28(1), 201-212. https://journals.aom.org/doi/abs/10.5465/256068 sayfasından erişilmiştir.

Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: A mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy & Practice*, 18(3), 239-258. https://doi.org/10.1080/0969594X.2011.577408

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: Guilford Press.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, 2(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kane, J., Bernardin, H., Villanueva, J., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal*, 38, 1036-1051. https://journals.aom.org/doi/abs/10.5465/256619 sayfasından erişilmiştir.

Karakaya, İ. (2015). Comparison of self, peer and instructor assessments in the portfolio assessment by using many facet Rasch model. J*ournal of Education and Human Development*, 4(2), 182-192. https://doi.org/10.15640/jehd.v4n2a22

Kavuran, T. (2003). Resim-iş öğretmenliği anabilim dallarına özel yetenek sınavları ile öğrenci alınırken karşılaşılan sorunlar. *Sosyal Bilimler Dergisi*, 159-180. http://www.acarindex.com/dosyalar/makale/acarindex-1423869755.pdf sayfasından erişilmiştir.

Khaatri, N., Kane, M.B., & Reeve, A.L. (1995). How performance assessments affect teaching and learning. *Educational Leadership*, 53(3), 80-83. https://eric.ed.gov/?id=EJ514709 sayfasından erişilmiştir.

Kim, C. (2017). Rater training for scoring rubrics: Rater-centered bottom-up approach. *Writing*, 33, 1-14. http://minnetesoljournal.org/journal-acrhive/mtj-2015-2/rater-training-for-scoring-rubrics-rater-centered-bottom-up-approach/ sayfasından erişilmiştir.

Kim, S. C., & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model. *Journal of applied measurement*, 10(4), 403-423. https://europepmc.org/abstract/med/19934528 sayfasından erişilmiştir.

Kim, Y., Park, I., & Kang, M. (2012). Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapted Physical Activity Quarterly*, 29(4), 346-365. https://doi.org/10.1123/apaq.29.4.346

Kim, Y.K. (2009). *Combining constructed response items and multiple choice items using a hierarchical rater model* (Doktora Tezi). http://www.proquest.com/ sayfasından erişilmiştir.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior-a longitudinal study. *Language Testing*, 28, 179200. https://doi.org/10.1177/0265532210384252

Knoch, U., Fairbairn, J., Myford, C., & Huisman, A. (2018). Evaluating the relative effectiveness of online and face-to-face training for new writing raters. *Papers in Language Testing and Assessment*, 7(1), 61-86. https://arts.unimelb.edu.au/__data/assets/pdf_file/0009/2748285/4.-PLTA-7.1-Knoch-et-al..pdf sayfasından erişilmiştir.

Knoch, U., J. Read & T. von Randow, (2007). Re-training writing raters online: How does compare with face-to-face training?, *Assessing Writing*, 12(2), 26-43. https://doi.org/10.1016/j.asw.2007.04.001

Kondo, Y. (2010). Examination of rater training effect and rater eligibility in L2 performance assessment. *Journal of Pan-Pacific Association of Applied Linguistics*, 14(2), 1-23. https://eric.ed.gov/?id=EJ920513 sayfasından erişilmiştir.

Kozlowski, S. W. J., Kirsch, M. P., & Chao, G. T. (1986). Job knowledge, ratee familiarity and halo error: An exploration. *Journal of Applied* Psychology, 71, 45-49. https://doi.org/10.1037/0021-9010.71.1.45

Krippendorff, K. (1995). On the reliability of unitizing continuous data. *Sociological Methodology*, 25, 47-76. https://doi.org/10.2307/271061

Kubiszyn, T., & Borich, G. (2013). *Educational testing and measurement*. New Jersey: John Wiley & Sons Incorporated.

Kumar, P., Ghildiyal, J. P., Rajpoot, R. S., & Kumar, S. (2017). Comparison of performance in physiology subject assessed by short essay type questions and multiple choice questions of first year medical students. *Asian Journal of Medical Sciences*, 8(1), 82-84. https://doi.org/10.3126/ajms.v8i1.15977

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

Landauer, T.K., Laham, D., & Foltz, P.W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Landy, E J. & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87, 72-107. https://doi.org/10.1037/0033-2909.87.1.72

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications.* San Diego, CA: Academic Press.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology*, 28(4), 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lawshe, C. H. (1985). Inferences from personnel tests and their validity. *Journal of Applied Psychology*, 70(1), 237-238. https://doi.org/10.1037/0021-9010.70.1.237

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4), 399-418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, 28, 543–560. https://doi.org/10.1177/0265532211406422

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1993). Rasch-based generalizability theory. *Rasch Measurement Transaction*, 7(1), 283-284. https://www.rasch.org/rmt/rmt71h.htm sayfasından erişilmiştir.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago: Mesa Press.

Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. *Objective measurement: Theory into practice*, 3, 85-98.https://files.eric.ed.gov/fulltext/ED364573.pdf sayfasından erişilmiştir.

Linacre, J. M. (2002). Number of person or item strata. Rasch Measurement: *Transactions of the Rasch Measurement SIG*, 16, 888. https://ci.nii.ac.jp/naid/10028294907/ sayfasından erişilmiştir.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *J Appl Meas*, 3(1), 85-106. https://www.ncbi.nlm.nih.gov/pubmed/11997586 sayfasından erişilmiştir.

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

Lopez, L. (2009). *Effects of delayed and immediate feedback in the computer-based testing.* (Doktora Tezi). http://www.proquest.com/ sayfasından erişilmiştir.

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. New York: Information Age Publishing.

Lord, R. C. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology*, 70, 66-71. https://doi.org/10.1037/0021-9010.70.1.66

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54-71. https://doi.org/10.1177/026553229501200104

Lunz, M. E., Wright, B. D. & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3(4), 331-345. https://doi.org/10.1207/s15324818ame0304_3

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15(2), 158-180. https://doi.org/10.1177/026553229801500202

May, G. L. (2008). The effect of rater training on reducing social style bias in peer evaluation. *Business Communication Quarterly*, 71(3), 297-313. https://doi.org/10.1177/1080569908321431

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.

McIntyre, P. N. (1993). *The importance and effectiveness of moderation training on the reliability of teachers' assessment of ESL writing samples*. (Yüksek Lisans Tezi). http://www.proquest.com/ sayfasından erişilmiştir.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Messick, S. (1989). Validity. In: R.L. Linn (Ed.), *Educational measurement*. (pp. 13-104). Macmillan, New York: American Council on Education.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23. https://doi.org/10.3102/0013189X023002013

Messick, S. (1996). Validity of performance assessments. In G. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). Washington, DC: National Center for Education Statistics.

Miller, A. G. (1970). Role of physical attractiveness in impression formation. *Psychonomic Science*, 19(4), 241-243. https://doi.org/10.3758/BF03328797

Milli Eğitim Bakanlığı (2017). *Fen bilimleri dersi öğretim programı (ilkokul ve ortaokul 3, 4, 5, 6, 7 ve 8 .sınıflar).* http://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=143 sayfasından erişilmiştir.

Moore, B.B. (2009). *Consideration of rater effects and rater design via signal detection theory*. (Doktora Tezi). http://www.proquest.com/ sayfasından erişilmiştir.

Moser, K., Kemter, V., Wachsmann, K., Köver, N. Z., & Soucek, R. (2016). Evaluating rater training with double-pretest one-posttest designs: an analysis of testing effects and the moderating role of rater self-efficacy. *The International Journal of Human Resource Management*, 1-23. https://doi.org/10.1080/09585192.2016.1254102

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical assessment, research & evaluation*, 7(10), 71-81. http://pareonline.net/getvn.asp?v=7&n=10 sayfasından erişilmiştir.

Moskal, B. M., (2003). Developing classroom performance assessments and scoring rubrics - part I. ERIC Digest. *Clearinghouse on Assessment and Evaluation College Park MD*., 1-7. https://eric.ed.gov/?id=ED481714 sayfasından erişilmiştir.

Moskal, B.M. (2000). *Scoring rubrics: What, when and how?*. http://pareonline.net/htm/v7n3.htm sayfasından erişilmiştir.

Murphy, K. R., Jako, R. A. ve Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology*, 78, 218-225. https://doi.org/10.1037/0021-9010.78.2.218

Murphy, K.R. & Balzer, W.K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, 74, 619-624. https://doi.org/10.1037/0021-9010.74.4.619

Myford, C. M. (2002). Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education*, 15(2), 187–215. https://doi.org/10.1207/S15324818AME1502_04

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. http://psycnet.apa.org/record/2003-09517-007 sayfasından erişilmiştir.

Myford, C.M., & Wolfe, E.W. (2000). *Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs*. Princeton, NJ: Educational Testing Service.

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. https://europepmc.org/abstract/med/15064538 sayfasından erişilmiştir.

Neck, C. P. (1996). Thought self-leadership: A regulatory approach towards overcoming resistance to organizational change. *International Journal of Organizational Analysis*, 4, 202-216. https://doi.org/10.1108/eb028849

Neck, C. P., Stewart, G. L., & Manz, C. C. (1995). Thought self-leadership as a framework for enhancing the performance of performance appraisers. *The Journal of Applied Behavioral Science*, 31, 278-294. https://doi.org/10.1177/0021886395313004

Noonan, L.E., & Sulsky, L.M. (2001) Impact of frame-of- reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance*, 14(1), 3-26. https://doi.org/10.1207/S15327043HUP1401_02

Ok, A. B., Bilgiç, H. C., & Bilgiç, R. (2013). Öğretim elemanı değerlendirmesinde kullanılan formatların hale etkisi, cömertlik etkisi ve kullanıcı tepkileri açısından karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 44, 237-248. http://dergipark.gov.tr/download/article-file/87250 sayfasından erişilmiştir.

Oosterhof, A. (2003). *Developing and using classroom assessments*. New Jersey: Merrill-Prentice Hall Press.

Osana, H. P., & Seymour, J. R. (2004). Critical thinking in preservice teachers: A rubric for evaluating argumentation and statistical reasoning. *Educational research and evaluation*, 10, 473–498. https://doi.org/10.1080/13803610512331383529

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological methods*, 5(3), 343. http://dx.doi.org/10.1037/1082-989X.5.3.343

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer. https://doi.org/10.1007/978-1-4613-0083-0

Primi, R., Silvia, P., Benedek, M., & Jauk, E. (2018). Applying many-facet Rasch modeling in the assessment of creativity. *PsyArXiv*, 29, 1-57. https://doi.org/10.31234/osf.io/kgzvj

Pufpaff, L. A., Clarke, L., & Jones, R. E. (2015). The effects of rater training on inter-rater agreement. *Mid-Western Educational Researcher*, 27(2), 117-141. https://eric.ed.gov/?id=EJ1059779 sayfasından erişilmiştir.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology*, 69(4), 581-588. http://psycnet.apa.org/doi/10.1037/0021-9010.69.4.581

Raczynski, K. R., Cohen, A. S., Engelhard Jr, G., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement*, 52(3), 301-318. https://doi.org/10.1111/jedm.12079

Rafilson, F. (1991). The case for validity generalization. *Practical Assessment, Research & Evaluation*, 2(13), 1-3. http://ericae.net/pare/99~getvn.html sayfasından erişilmiştir.

Raymond, M. R., Webb, L. C., & Houston, W. M. (1991). Correcting performance-rating errors in oral examinations. *Evaluation & the health professions*, 14(1), 100-122. https://doi.org/10.1177/016327879101400107

Reiner, C. M., Bothell, T. W., & Sudweeks, R. R. (2002). *Preparing effective essay questions: A self-directed workbook for educators.* Oklahoma: New Forums Press.

Roblyer, M. D., & Wiencke, W. R. (2003). Design and use of a rubric to assess and encourage interactive qualities in distance courses. *The American journal of distance education*, 17(2), 77-98. https://doi.org/10.1207/S15389286AJDE1702_2

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, 85(2), 370-395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher*, 94(1), 31-37. http://peterliljedahl.com/wp-content/uploads/Myth-of-Objectivity2.pdf sayfasından erişilmiştir.

Royal, K. D., & Hecker, K. G. (2016). Rater errors in clinical performance assessments. *Journal of veterinary medical education*, 43(1), 5-8. https://doi.org/10.3138/jvme.0715-112R

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric essay scoring system. *Journal of Technology, Learning, and Assessment*, 4, 1–22. https://ci.nii.ac.jp/naid/10027364967/ sayfasından erişilmiştir.

Saal, F. E. , Downey, R. G. & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413-428. https://doi.org/10.1037/0033-2909.88.2.413

Sebok, S. S., & Syer, M. D. (2015). Seeing things differently or seeing different things? Exploring raters' associations of noncognitive attributes. *Academic Medicine: Journal of the Association of American Medical Colleges*, 90(11), 50–55. http://doi.org/10.1097/ACM.0000000000000902

Shale, D. (1996). Essay reliability: Form and meaning. In: White, E. Lutz, W. & Kamusikiri S. (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 76–96). New York: MLAA.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw, & N.S. Petersen (Eds.), *International encyclopedia of education* (3rd ed., pp. 75–80). Oxford, England: Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00233-5

Söylemez, Y. (2018). 2018 Türkçe dersi öğretim programındaki kazanımların üst düzey düşünme becerileri açısından değerlendirilmesi. *Atatürk Üniversitesi Türkiyat Araştırmaları Enstitüsü Dergisi*, (63), 345-384. http://dx.doi.org/10.14222/Turkiyat3991

Stamoulis, D.T. & Hauenstein, N.M.A. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology*, 78(6), 994-1003. https://doi.org/10.1037/0021-9010.78.6.994

Steiner, I. D. (1954). Ethnocentric and tolerance of trait inconsistency. *Journal of Abnormal and Social Psychology*, 49, 25-29. https://doi.org/10.1037/h0054849

Stoering, J. M., & Lu, L. (2002). *Combining the national survey of student engagement with student portfolio assessment.* Toronto: Forum Paper.

Sudweeks, R. R., Reeve, S. & Bradshaw, W.S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9, 239-261. https://doi.org/10.1016/j.asw.2004.11.001

Sulsky, L.M., & Day, D.V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology*, 77(4), 501-510. https://doi.org/10.1037/0021-9010.77.4.501

Tan, Ş. (2013). *Öğretimde ölçme ve değerlendirme KPSS el kitabı.* Ankara: Pegem Akademi.

Taş, U. E., Arıcı, Ö., Ozarkan, H. B., & Özgürlük, B. (2016). *PISA 2015 ulusal raporu*. Ankara: Milli Eğitim Bakanlığı.

Tekin, H. (2013). *Eğitimde ölçme ve değerlendirme*. Ankara: Yargı Yayınları.

Thornton, G. C., & Zorich, S. (1980). Training to improve observer accuracy. *Journal of Applied Psychology*, 65(3), 351-354. https://doi.org/10.1037/0021-9010.65.3.351

Tseng, W. T., Su, T. Y., & Nix, J. M. L. (2018). Validating translation test items via the many-facet Rasch model. *Psychological reports*, 122(2), 748-772. https://doi.org/10.1177/0033294118768664

Turgut, M.F., & Baykul, Y. (2014). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayıncılık.

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2, 319–330. https://doi.org/10.28945/331

Van Dyke, N. (2008). Self-and peer-assessment disparities in university ranking schemes. *Higher Education in Europe*, 33(2/3), 285-293. https://doi.org/10.1080/03797720802254114

Waltman, K., Kahn, A., & Koency, G. (1998). *Alternative approaches to scoring: the effects of using different scoring methods on the validity of scores from a performance assessment*. CSE Technical Report (488). Los Angeles: CREEST. https://cresst.org/wp-content/uploads/TECH488.pdf sayfasından erişilmiştir.

Wang, B. (2010). On rater agreement and rater training. *English Language Teaching*, 3(1), 108-113. https://doi.org/10.5539/elt.v3n1p108

Watts, M., Jofili, Z., & Bezerra, R. (1997). A case for critical constructivism and critical thinking in science education. *Research in Science Education*, 27(2), 309–322. https://doi.org/10.1007/BF02461323

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11, 197-223. https://doi.org/10.1177/026553229401100206

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263-287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Weitz, G., Vinzentius, C., Twesten, C., Lehnert, H., Bonnemeier, H., & König, I. R. (2014). Effects of a rater training on rating accuracy in a physical examination skills assessment. *GMS Zeitschrift für Medizinische Ausbildung*, 31(4), 1-17. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4259060/ sayfasından erişilmiştir.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10, 305-323. https://doi.org/10.1177/026553229301000306

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, 45(3), 197-210. https://doi.org/10.1177/0748175612440286

Wiseman, C.S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, 17(3), 150-173. https://doi.org/10.1016/j.asw.2011.12.001

Woehr, D. J. (1992). Performance dimension accessibility: Implications for rating accuracy. *Journal of Organizational Behavior*, 13(4), 357-367. https://doi.org/10.1002/job.4030130404

Woehr, D. J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology*, 79, 525–534. https://doi.org/10.1037/0021-9010.79.4.525

Woehr, D.J., & Huffuct, A.I. (1994). Rater training for performance appraisal. A qantitative review. *Journal of Occupational and Organizational Psychology*, 67(3), 189-205. https://doi.org/10.1111/j.2044-8325.1994.tb00562.x

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, 46, 35-51. http://psycnet.apa.org/record/2004-19990-003 sayfasından erişilmiştir.

Wolfe, E. W., & McVay, A. (2010). *Rater effects as a function of rater training contex.* http://images.pearsonassessments.com/images/tmrs/tmrs_rg/RaterEffects_101510 . pdf. sayfasından erişilmiştir.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 8(3), 370. https://ci.nii.ac.jp/naid/10017021263/ sayfasından erişilmiştir.

Wu, S. M., & Tan, S. (2016). Managing rater effects through the use of FACETS analysis: the case of a university placement test. *Higher Education Research & Development*, 35(2), 380-394. https://doi.org/10.1080/07294360.2015.1087381

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, 31(4), 501-527. https://doi.org/10.1177/0265532214536171

Yaz, Ö., & Kurnaz, M. (2017). 2013 Fen bilimleri öğretim programının incelenmesi. *Uluslararası Türk Eğitim Bilimleri Dergisi*, 8, 173-184. http://dergipark.gov.tr/goputeb/issue/34591/382211 sayfasından erişilmiştir.

Youn, S. J. (2018). Rater variability across examinees and rating criteria in paired speaking assessment. *Papers in Language Testing and Assessment*, 7(1), 32-60. http://www.altaanz.org/uploads/5/9/0/8/5908292/7_1_youn.pdf sayfasından erişilmiştir.

Yue, X. (2011). *Detecting rater centrality effect using simulation methods and Rasch measurement analysis.* (Doktora Tezi). http://www.proquest.com/ sayfasından erişilmiştir.

Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology*, 67(6), 752-758. https://doi.org/10.1037/0021-9010.67.6.752

Zhao, H., Andersson, B., Guo, B., & Xin, T. (2017). Sequential effects in essay ratings: evidence of assimilation effects using cross-classified models. *Frontiers in psychology*, 8, 1-10. https://doi.org/10.3389/fpsyg.2017.00933

Zwiers, J. (2008). *Building academic language: Essential practices for content classrooms*. San Francisco: Jossey-Bass.

EKLER

EK 1. Gazi Üniversitesi Etik Komisyonu İzin Belgesi

Evrak Tarih ve Sayısı: 10/05/2018-E.75221

T.C.
GAZİ ÜNİVERSİTESİ
Etik Komisyonu

Sayı : 77082166-302.08.01-
Konu : Bilimsel ve Eğitim Amaçlı

EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

İlgi : 03/04/2018 tarihli ve 80287700-302.08.01- 54466 sayılı yazı.

İlgi yazınız ile göndermiş olduğunuz, Eğitim Bilimleri Anabilim Dalı Eğitimde Ölçme ve Değerlendirme Bilim Dalı **Doktora Öğrencisi Mehmet ŞATA'nın, Doç.Dr.İsmail KARAKAYA'nın** danışmanlığında yürüttüğü ***"Performans Değerlendirilmesine İlişkin Puanlayıcı Eğitiminin Puanlayıcıların Davranışları Üzerindeki Etkisinin İncelenmesi"*** adlı tez çalışması ile ilgili konu Komisyonumuzun **08.05.2018** tarih ve **04** sayılı toplantısında görüşülmüş olup,

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

Bilgilerinizi ve gereğini rica ederim.

e-imzalıdır
Prof. Dr. Alper CEYLAN
Komisyon Başkanı

Araştırma Kod No: 2018-165

Ek:1 Liste

Ankara
Tel:0 (312) 202 20 57 - 0 (312) 2 Faks:0 (312) 202 38 76
İnternet Adresi :http://etikkomisyon.gazi.edu.tr/
Bilgi için :Esengül BOŞNAK
Genel Evrak Sorumlusu
Telefon No 03122022666

Bu belge 5070 sayılı Elektronik İmza Kanununun 5. Maddesi gereğince güvenli elektronik imza ile imzalanmıştır.

Süre (dk)	Odak ve Kapsam	Öğrenme Hedefleri	Yöntem
10 dk	Puanlayıcı eğitimi hakkında genel giriş	Katılımcı(lar), puanlayıcı eğitiminin içeriğini, amacını ve önemini bilir(ler). Katılımcı(lar), puanlayıcı eğitiminde hedef kitleyi ve kullanılan yöntemleri bilir(ler).	Anlatım Soru-Cevap
20 dk	Puanlayıcı Hatası Eğitimi (PHE 1) Yaygın puanlayıcı hatalarının açıklanması	Katılımcı(lar) yaygın olarak yapılan puanlayıcı hatalarını bilir ve bu hataların puanlama/değerlendirme üzerindeki olası etki(lerini) açıklar(lar).	Anlatım Sunum yapma
30 dk	Puanlayıcı Hatası Eğitimi (PHE 2) Yaygın puanlayıcı hatalrının tanımlanması	Katılımcı(lar) puanlama davranış(larını) analiz edip hangi hata veya hatalara eğimli olduklarının farkına varır(lar).	Soru-Cevap Grup içi tartışma
20 dk	Puanlayıcı Hatası Eğitimi (PHE 3) Olası hata kaynakları Tipik yanlılıklar Puanlama türleri	Katılımcı(lar) puanlayıcı hatalarından kaçınmanın yolunu bilir(ler).	Anlatım Sunum yapma Literatür Özeti
40 dk	Referans Çerçevesi Eğitimi (RÇE 1) Performans boyutu eğitimi	Katılımcı(lar) ilgili yeterliğin/performansın boyutlarını, göstergelerini ve esas/temel kavramlarını eksiksiz bir şekilde öğrenir(ler). Katılımcı(lar) ilgili performansa ilişkin örnek davranışlar söyleyebilir(ler).	Sınıf içi tartışma Örnekler verme Anlatım
30 dk	Referans Çerçevesi Eğitimi (RÇE 2) Temsili örnek(ler) verme, Uygulamalar yapma, Geri dönütler verme, Sınıflandırma	Katılımcı(lar) ilgili performansa ilişkin basit davranışları düzgün bir biçimde sıralayarak durum belirlemeyi/değerlendirmeyi yapabilir(ler).	Bireysel puanlamalar Grup tartışmaları Geri dönüt verme
30 dk	Referans Çerçevesi Eğitimi (RÇE 3) Temsili örnek(ler) verme, Uygulamalar yapma, Geri dönütler verme, Sınıflandırma, Ağırlıklandırma, Değerlendirme	Katılımcı(lar) kritik olayların durum belirlemesini doğru bir şekilde yapabilir(ler). Katılımcı(lar) kritik olayların sınıflandırmasını yapabilir(ler).	Bireysel puanlamalar Grup tartışmaları Geri dönüt verme
30 dk	Referans Çerçevesi Eğitimi (RÇE 4) Temsili örnek(ler) verme, Uygulamalar yapma, Geri dönütler verme, Sınıflandırma, Ağırlıklandırma, Değerlendirme	Katılımcı(lar) ilgili performansa ilişkin basit davranışları düzgün bir biçimde sıralayarak durum belirlemeyi/değerlendirmeyi yapabilir(ler). Katılımcı(lar) grup tartışmları yaparak bir fikir birliğine ulaşır(lar).	Bireysel puanlamalar Grup tartışmaları Geri dönüt verme
30 dk	Referans Çerçevesi Eğitimi (RÇE 5) Biçimlendirme	Katılımcı(lar) ilgili performansa ilişkin karmaşık/ kompleks davranışları düzgün bir biçimde sıralayarak durum belirlemeyi/değerlendirmeyi yapabilir(ler). Katılımcı(lar) grup tartışmları yaparak bir fikir birliğine ulaşır(lar).	Bireysel puanlamalar Grup tartışmaları Geri dönüt verme

EK 3. Kişisel Bilgi Formu

Sevgili Puanlayıcı;

Aşağıda verilen liste size verilen öğrenci kompozisyonlarının bir listesidir. Öğrenci kompozisyonlarının üzerindeki numara ile aşağıdaki numaralar aynı öğrenciyi temsil etmektedir. Size verilen analitik dereceli puanlama anahtarındaki her bir kritere göre verdiğiniz puanı tabloda ilgili yere yazınız. Çalışmaya verdiğiniz destekten dolayı teşekkür eder kolaylıklar dilerim.

Arş.Gör. Mehmet ŞATA

Adınız ve Soyadınız :

Cinsiyetiniz :

Yaşınız :

1. Hangi sıklıkla akademik yazı/metin yazarsınız.

( ) Her zaman ( ) Çoğunlukla ( ) Bazen ( ) Nadiren ( ) Hiç

2. Akademik yazmaya (örneğin kompozisyon gibi) yönelik kaygınız var mıdır?

( ) Her zaman ( ) Çoğunlukla ( ) Bazen ( ) Nadiren ( ) Hiç

3. Akademik yazmaya yönelik ilginiz var mıdır?

( ) Her zaman ( ) Çoğunlukla ( ) Bazen ( ) Nadiren ( ) Hiç

4. Akademik yazmaya yönelik tutumunuz nasıldır?

( ) Çok severim ( ) Severim ( ) Ne severim nede sevmem ( ) Sevmem ( ) Hiç sevmem

5. Sizce akademik yazma becerisinin değerlendirilmesi için hazırlanan analitik dereceli puanlama anahtarı ve derecelendirmesi (0'dan 4'e kadar) anlaşılır mıdır?

( ) Tamamen anlaşılır ( ) Anlaşılır ( ) Ne anlaşılır nede anlaşılmaz ( ) Anlaşılmaz ( ) Tamamen anlaşılmaz

PUANLAMA ÖLÇEĞİ (RATING SCALE)

Ana Kriterler	Organization						Content		Coherence	Cohesion	Grammar		Vocabulary		Mechanics	
Alt Kriterler Öğr. No	Title of Essay	Introduction-Body-Conclusion	Thesis Statement	Topic Sentence	Supporting Sentences	Appropriate Length	Topic Relevance	Idea Development	Coherence	Linking	Accuracy of Grammatical Forms	Syntactic Complexity	Word Choice	Lexical Range	Spelling	Punctuation
01	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
02	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
03	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
04	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
05	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
06	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
07	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
08	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
09	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
10	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
11	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
12	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
13	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
14	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
15	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
16	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
17	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
18	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
19	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
20	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
21	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
22	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
23	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
24	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
25	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
26	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
27	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
28	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
29	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
30	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
31	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
32	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
33	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
34	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
35	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
36	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
37	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
38	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )
39	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )

EK 4. Analitik Dereceli Puanlama Anahtarı (Akademik Yazma Becerileri İçin)

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

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

EK 5. Akademik Yazma Görevi

ACADEMIC WRITING SAMPLE TASK 2A

You should spend about 40 minutes on this task.

Write about the following topic:

The first car appeared on British roads in 1888. By the year 2000 there may be as many as 29 million vehicles on British roads.

Alternative forms of transport should be encouraged and international laws introduced to control car ownership and use.

To what extent do you agree or disagree?

Give reasons for your answer and include any relevant examples from your knowledge or experience.

Write at least 250 words.

EK 6. Kontrol Grubunun Ön Testi İçin *Puanlayıcı x Birey* Etkileşim Analizi Sonucunda Yanlı Olan Etkileşimler

Sıra No	Gözlenen Puan	Beklenen Puan	Yanlılık Büyüklüğü	Model SH	t değeri	sd	p	Infit	Outfit	Öğrenci	Puanlayıcı
1	14	44,83	-2,59	0,31	-8,41	14	0,000	1,40	1,30	O16	OKP23
2	15	42,38	-2,24	0,30	-7,45	14	0,000	1,30	1,30	O21	OKP23
3	25	46,38	-1,85	0,28	-6,72	14	0,000	1,00	1,00	O07	OKP20
4	35	49,74	-1,49	0,29	-5,21	14	0,000	0,40	0,50	O29	OKP03
5	24	40,83	-1,36	0,28	-4,93	14	0,000	0,50	0,50	O10	OKP16
6	38	50,95	-1,41	0,29	-4,79	14	0,000	0,50	0,50	O26	OKP16
7	39	51,39	-1,39	0,30	-4,67	14	0,000	1,90	2,00	O21	OKP12
8	36	49,08	-1,32	0,29	-4,57	14	0,000	0,40	0,40	O29	OKP05
9	20	36,47	-1,29	0,28	-4,56	14	0,000	1,10	1,10	O18	OKP11
10	34	47,54	-1,29	0,28	-4,55	14	0,001	0,60	0,60	O03	OKP19
11	17	33,72	-1,31	0,29	-4,48	14	0,001	0,30	0,30	O13	OKP23
12	19	35,34	-1,27	0,29	-4,47	14	0,001	0,60	0,60	O06	OKP20
13	19	35,08	-1,25	0,29	-4,40	14	0,001	1,20	1,20	O11	OKP02
14	16	31,89	-1,25	0,30	-4,22	14	0,001	1,30	1,20	O10	OKP11
15	41	51,85	-1,28	0,31	-4,19	14	0,001	0,80	0,80	O24	OKP01
16	13	28,74	-1,28	0,31	-4,08	14	0,001	1,20	1,10	O25	OKP23
17	18	33,08	-1,17	0,29	-4,07	14	0,001	1,60	1,60	O09	OKP11
18	38	49,40	-1,19	0,29	-4,05	14	0,001	0,30	0,30	O19	OKP08
19	26	39,74	-1,11	0,28	-4,03	14	0,001	0,60	0,60	O29	OKP23
20	19	33,53	-1,13	0,29	-3,95	14	0,001	0,30	0,30	O39	OKP20
21	13	28,11	-1,24	0,31	-3,93	14	0,002	0,90	0,90	O06	OKP11
22	28	41,13	-1,09	0,28	-3,93	14	0,002	0,90	0,90	O28	OKP23
23	26	39,23	-1,06	0,28	-3,86	14	0,002	0,60	0,60	O20	OKP09
24	24	37,36	-1,05	0,28	-3,81	14	0,002	0,80	0,80	O37	OKP23
25	31	43,21	-1,06	0,28	-3,79	14	0,002	0,90	0,90	O02	OKP22
26	24	37,01	-1,02	0,28	-3,70	14	0,002	0,90	0,90	O07	OKP02
27	21	34,35	-1,03	0,28	-3,69	14	0,002	0,70	0,70	O26	OKP23
28	38	48,55	-1,08	0,29	-3,68	14	0,003	0,50	0,50	O34	OKP19
29	27	39,46	-1,01	0,28	-3,66	14	0,003	0,40	0,40	O07	OKP13
30	20	33,33	-1,03	0,28	-3,65	14	0,003	0,60	0,60	O11	OKP10
31	41	50,70	-1,11	0,31	-3,61	14	0,003	0,70	0,70	O17	OKP07
32	15	28,64	-1,09	0,30	-3,61	14	0,003	0,70	0,70	O08	OKP11
33	35	45,99	-1,03	0,29	-3,60	14	0,003	0,40	0,50	O26	OKP03
34	37	47,49	-1,04	0,29	-3,56	14	0,003	0,30	0,30	O22	OKP16
35	24	36,37	-0,97	0,28	-3,50	14	0,004	0,50	0,50	O38	OKP16
36	25	37,17	-0,96	0,28	-3,48	14	0,004	0,40	0,40	O05	OKP08
37	46	53,94	-1,15	0,33	-3,44	14	0,004	0,90	0,90	O16	OKP06
38	45	53,24	-1,12	0,33	-3,43	14	0,004	0,40	0,40	O16	OKP19
39	37	47,10	-0,99	0,29	-3,40	14	0,004	2,10	2,00	O03	OKP07
40	20	32,40	-0,96	0,28	-3,39	14	0,004	0,80	0,80	O39	OKP21
41	30	41,17	-0,94	0,28	-3,37	14	0,005	0,70	0,70	O17	OKP23
42	37	46,90	-0,97	0,29	-3,32	14	0,005	0,60	0,60	O12	OKP15
43	30	40,96	-0,92	0,28	-3,30	14	0,005	0,30	0,30	O13	OKP02
44	44	52,18	-1,04	0,32	-3,24	14	0,006	0,50	0,50	O28	OKP04
45	28	39,05	-0,90	0,28	-3,24	14	0,006	1,20	1,20	O11	OKP11
46	31	41,63	-0,90	0,28	-3,23	14	0,006	0,80	0,80	O24	OKP23
47	24	35,44	-0,89	0,28	-3,22	14	0,006	0,20	0,30	O33	OKP11
48	42	50,61	-1,00	0,31	-3,21	14	0,006	0,60	0,70	O13	OKP16
49	28	38,82	-0,88	0,28	-3,17	14	0,007	0,70	0,70	O05	OKP15
50	37	46,51	-0,92	0,29	-3,17	14	0,007	0,20	0,20	O01	OKP19
51	42	50,48	-0,98	0,31	-3,15	14	0,007	1,00	1,00	O17	OKP09
52	21	32,30	-0,87	0,28	-3,10	14	0,008	0,80	0,80	O18	OKP02
53	23	34,10	-0,86	0,28	-3,09	14	0,008	0,60	0,60	O38	OKP20
54	32	42,03	-0,86	0,28	-3,08	14	0,008	1,70	1,80	O07	OKP12

55	25	35,83	-0,85	0,28	-3,07	14	0,008	0,50	0,50	O39	OKP16
56	27	37,58	-0,84	0,28	-3,06	14	0,009	0,50	0,50	O18	OKP09
57	39	47,88	-0,91	0,30	-3,06	14	0,009	0,40	0,40	O15	OKP13
58	28	38,41	-0,84	0,28	-3,04	14	0,009	1,10	1,10	O35	OKP12
59	22	32,97	-0,84	0,28	-3,03	14	0,009	0,60	0,60	O38	OKP21
60	16	27,34	-0,90	0,30	-3,03	14	0,009	1,10	1,20	O14	OKP23
61	42	50,19	-0,94	0,31	-3,02	14	0,009	0,40	0,50	O15	OKP19
62	39	47,78	-0,90	0,30	-3,01	14	0,009	0,90	0,90	O28	OKP14
63	31	40,91	-0,83	0,28	-2,99	14	0,010	0,70	0,70	O33	OKP21
64	22	32,77	-0,83	0,28	-2,97	14	0,010	0,30	0,30	O04	OKP20
65	32	41,67	-0,83	0,28	-2,95	14	0,011	0,70	0,70	O32	OKP12
66	31	40,72	-0,82	0,28	-2,93	14	0,011	0,40	0,40	O14	OKP08
67	45	52,29	-0,96	0,33	-2,92	14	0,011	1,30	1,30	O21	OKP01
68	28	38,05	-0,81	0,28	-2,92	14	0,011	0,40	0,40	O18	OKP08
69	27	37,08	-0,80	0,28	-2,90	14	0,012	1,20	1,20	O35	OKP11
70	32	41,45	-0,81	0,28	-2,88	14	0,012	0,70	0,70	O32	OKP09
71	46	52,93	-0,96	0,33	-2,87	14	0,012	0,70	0,60	O16	OKP03
72	42	49,83	-0,89	0,31	-2,86	14	0,013	0,60	0,70	O29	OKP08
73	44	51,37	-0,91	0,32	-2,84	14	0,013	0,90	1,00	O37	OKP20
74	45	52,10	-0,92	0,33	-2,82	14	0,014	1,00	0,90	O17	OKP17
75	28	37,73	-0,78	0,28	-2,82	14	0,014	0,70	0,70	O10	OKP21
76	24	34,04	-0,78	0,28	-2,81	14	0,014	0,90	0,90	O05	OKP13
77	42	49,73	-0,87	0,31	-2,81	14	0,014	0,50	0,60	O19	OKP19
78	43	50,44	-0,88	0,32	-2,77	14	0,015	0,80	0,90	O19	OKP15
79	33	41,63	-0,75	0,28	-2,64	14	0,019	0,60	0,60	O02	OKP03
80	33	41,63	-0,75	0,28	-2,64	14	0,019	0,30	0,30	O22	OKP08
81	46	52,43	-0,87	0,33	-2,60	14	0,021	0,90	0,90	O16	OKP05
82	32	40,57	-0,72	0,28	-2,58	14	0,022	0,50	0,50	O18	OKP04
83	35	43,23	-0,74	0,29	-2,58	14	0,022	0,70	0,80	O02	OKP15
84	40	47,34	-0,76	0,30	-2,51	14	0,025	1,00	1,00	O19	OKP13
85	34	42,04	-0,70	0,28	-2,48	14	0,027	0,60	0,60	O11	OKP22
86	35	42,92	-0,71	0,29	-2,47	14	0,027	0,60	0,60	O22	OKP01
87	51	55,82	-0,95	0,39	-2,46	14	0,028	0,40	0,40	O16	OKP16
88	26	34,66	-0,67	0,28	-2,45	14	0,028	1,10	1,10	O09	OKP07
89	26	34,51	-0,66	0,28	-2,40	14	0,031	0,70	0,70	O09	OKP12
90	35	42,69	-0,69	0,29	-2,39	14	0,031	0,50	0,50	O07	OKP19
91	8	16,40	-0,89	0,38	-2,38	14	0,032	1,30	1,10	O08	OKP23
92	40	46,99	-0,72	0,30	-2,37	14	0,033	2,00	2,00	O03	OKP12
93	35	42,58	-0,68	0,29	-2,36	14	0,034	0,20	0,20	O31	OKP08
94	38	45,19	-0,69	0,29	-2,34	14	0,035	0,30	0,30	O26	OKP05
95	17	25,63	-0,68	0,29	-2,34	14	0,035	0,90	0,90	O10	OKP10
96	33	40,69	-0,66	0,28	-2,33	14	0,035	0,60	0,60	O02	OKP05
97	46	51,89	-0,78	0,33	-2,33	14	0,035	0,90	0,90	O28	OKP06
98	20	28,54	-0,66	0,28	-2,33	14	0,036	0,80	0,80	O23	OKP23
99	45	51,05	-0,76	0,33	-2,32	14	0,036	0,70	0,80	O17	OKP19
100	28	36,08	-0,64	0,28	-2,31	14	0,036	0,20	0,20	O36	OKP08
101	27	35,09	-0,63	0,28	-2,30	14	0,038	0,40	0,40	O04	OKP16
102	41	47,61	-0,70	0,31	-2,27	14	0,039	0,20	0,20	O13	OKP04
103	39	45,87	-0,68	0,30	-2,27	14	0,040	0,30	0,30	O28	OKP10
104	33	40,44	-0,63	0,28	-2,24	14	0,042	0,20	0,20	O27	OKP05
105	35	42,24	-0,64	0,29	-2,24	14	0,042	0,70	0,70	O07	OKP08
106	18	26,17	-0,64	0,29	-2,22	14	0,044	0,70	0,70	O39	OKP11
107	32	39,46	-0,62	0,28	-2,22	14	0,043	0,60	0,60	O20	OKP12
108	23	30,89	-0,60	0,28	-2,18	14	0,047	0,90	0,90	O39	OKP04
109	29	36,57	-0,60	0,28	-2,18	14	0,047	0,40	0,40	O33	OKP09
110	41	47,37	-0,67	0,31	-2,18	14	0,047	1,00	1,00	O01	OKP15
111	44	49,88	-0,69	0,32	-2,16	14	0,049	0,30	0,30	O15	OKP08
112	54	47,67	0,97	0,45	2,14	14	0,050	0,70	0,80	O25	OKP16
113	47	39,81	0,73	0,34	2,14	14	0,050	1,00	1,00	O09	OKP20
114	32	24,18	0,60	0,28	2,14	14	0,051	0,40	0,40	O18	OKP23
115	52	45,28	0,88	0,40	2,17	14	0,048	0,60	0,60	O13	OKP09

116	48	40,81	0,76	0,35	2,17	14	0,048	1,80	2,30	O07	OKP11
117	57	51,00	1,33	0,61	2,18	14	0,047	0,80	0,60	O24	OKP07
118	60	50,12	3,70	1,70	2,18	14	0,047	0,00	0,00	O24	OKP11
119	56	49,80	1,18	0,54	2,19	14	0,046	0,80	1,00	O17	OKP11
120	56	49,78	1,18	0,54	2,20	14	0,045	0,80	0,80	O28	OKP11
121	44	36,31	0,71	0,32	2,21	14	0,044	0,70	0,70	O22	OKP02
122	59	52,52	2,26	1,02	2,22	14	0,044	0,90	0,80	O37	OKP16
123	36	27,79	0,65	0,29	2,25	14	0,041	0,90	1,00	O39	OKP03
124	58	51,74	1,66	0,73	2,26	14	0,040	0,90	0,90	O03	OKP16
125	36	27,69	0,66	0,29	2,27	14	0,039	2,10	2,20	O39	OKP12
126	59	52,11	2,32	1,02	2,28	14	0,039	0,90	0,80	O29	OKP21
127	58	51,58	1,68	0,73	2,30	14	0,038	1,00	1,60	O17	OKP01
128	55	48,23	1,12	0,49	2,30	14	0,038	0,80	0,90	O37	OKP08
129	51	43,66	0,89	0,39	2,30	14	0,038	0,40	0,40	O07	OKP22
130	56	49,35	1,24	0,54	2,31	14	0,037	0,90	1,10	O37	OKP15
131	53	45,87	0,98	0,43	2,31	14	0,037	0,60	0,60	O31	OKP21
132	45	36,90	0,76	0,33	2,34	14	0,035	0,60	0,70	O33	OKP03
133	53	45,74	1,00	0,43	2,35	14	0,034	0,60	0,60	O26	OKP09
134	32	23,33	0,66	0,28	2,37	14	0,033	0,60	0,50	O30	OKP23
135	51	43,32	0,93	0,39	2,39	14	0,032	0,90	0,90	O32	OKP22
136	45	36,60	0,79	0,33	2,41	14	0,030	0,30	0,30	O36	OKP19
137	38	29,22	0,71	0,29	2,41	14	0,031	0,40	0,40	O07	OKP23
138	56	48,86	1,30	0,54	2,43	14	0,029	0,90	1,00	O15	OKP11
139	46	37,63	0,81	0,33	2,43	14	0,029	0,80	0,80	O11	OKP13
140	43	34,38	0,77	0,32	2,43	14	0,029	0,30	0,30	O01	OKP23
141	52	44,23	0,99	0,40	2,44	14	0,028	0,70	0,70	O14	OKP21
142	60	47,08	4,00	1,63	2,46	14	0,028	0,00	0,00	O34	OKP11
143	55	47,60	1,20	0,49	2,46	14	0,028	0,80	0,80	O02	OKP16
144	50	41,86	0,92	0,37	2,47	14	0,027	0,60	0,60	O33	OKP20
145	50	41,80	0,93	0,37	2,49	14	0,026	1,00	0,90	O32	OKP07
146	57	49,58	1,53	0,61	2,52	14	0,025	0,90	1,00	O15	OKP09
147	33	23,74	0,71	0,28	2,52	14	0,024	1,50	1,60	O04	OKP13
148	55	47,20	1,25	0,49	2,55	14	0,023	1,40	1,70	O12	OKP06
149	44	34,98	0,82	0,32	2,55	14	0,023	0,70	0,70	O18	OKP13
150	54	46,03	1,16	0,45	2,56	14	0,023	1,20	1,40	O12	OKP19
151	55	47,11	1,26	0,49	2,58	14	0,022	0,80	0,70	O28	OKP02
152	57	49,24	1,58	0,61	2,59	14	0,021	0,90	1,00	O12	OKP20
153	30	20,60	0,72	0,28	2,60	14	0,021	0,40	0,40	O09	OKP23
154	50	41,20	0,99	0,37	2,64	14	0,019	0,80	0,90	O22	OKP09
155	57	48,96	1,61	0,61	2,65	14	0,019	1,00	1,30	O26	OKP21
156	49	40,02	0,96	0,36	2,66	14	0,019	0,70	0,80	O36	OKP21
157	35	25,14	0,77	0,29	2,68	14	0,018	1,60	1,60	O38	OKP13
158	60	45,07	4,00	1,46	2,74	14	0,016	0,00	0,00	O35	OKP16
159	53	44,18	1,17	0,43	2,74	14	0,016	0,70	0,70	O20	OKP20
160	54	45,23	1,25	0,45	2,76	14	0,016	0,80	0,90	O25	OKP21
161	39	28,77	0,83	0,30	2,79	14	0,015	0,90	0,90	O32	OKP23
162	57	48,24	1,70	0,61	2,80	14	0,014	0,90	0,90	O31	OKP16
163	57	48,20	1,71	0,61	2,81	14	0,014	0,80	0,80	O03	OKP01
164	58	48,81	2,07	0,73	2,83	14	0,013	0,90	0,70	O29	OKP11
165	55	45,98	1,39	0,49	2,84	14	0,013	0,80	0,80	O02	OKP20
166	50	40,24	1,08	0,37	2,88	14	0,012	0,80	0,80	O35	OKP15
167	58	48,36	2,13	0,73	2,91	14	0,011	1,00	1,60	O19	OKP11
168	41	30,30	0,90	0,31	2,93	14	0,011	2,10	2,00	O10	OKP13
169	55	45,51	1,44	0,49	2,95	14	0,011	1,30	1,60	O12	OKP03
170	49	38,77	1,07	0,36	2,97	14	0,010	1,50	1,70	O10	OKP20
171	53	43,05	1,28	0,43	3,01	14	0,009	0,60	0,60	O02	OKP01
172	58	47,69	2,22	0,73	3,02	14	0,009	0,90	0,80	O32	OKP16
173	33	21,75	0,87	0,28	3,07	14	0,008	0,80	0,80	O39	OKP02
174	57	46,88	1,87	0,61	3,07	14	0,008	1,70	2,30	O12	OKP22
175	56	45,76	1,67	0,54	3,12	14	0,008	0,80	0,90	O01	OKP09
176	58	47,11	2,28	0,73	3,12	14	0,008	0,90	0,70	O37	OKP11

177	55	44,69	1,53	0,49	3,13	14	0,007	1,40	1,70	O12	OKP05
178	56	45,63	1,69	0,54	3,14	14	0,007	1,60	2,10	O12	OKP08
179	39	27,22	0,95	0,30	3,18	14	0,007	0,50	0,50	O38	OKP05
180	59	44,86	3,27	1,02	3,21	14	0,006	0,90	0,80	O26	OKP11
181	57	45,99	1,97	0,61	3,24	14	0,006	0,90	0,90	O29	OKP02
182	57	45,98	1,97	0,61	3,24	14	0,006	0,80	0,70	O03	OKP11
183	34	21,99	0,93	0,28	3,26	14	0,006	0,90	0,90	O36	OKP23
184	58	45,65	2,45	0,73	3,34	14	0,005	1,10	1,70	O13	OKP08
185	41	28,38	1,04	0,31	3,41	14	0,004	0,70	0,70	O38	OKP03
186	46	33,60	1,15	0,33	3,43	14	0,004	1,10	1,20	O10	OKP08
187	57	44,92	2,09	0,61	3,43	14	0,004	0,90	1,10	O11	OKP20
188	52	40,06	1,39	0,40	3,44	14	0,004	0,50	0,50	O27	OKP11
189	47	34,60	1,18	0,34	3,45	14	0,004	0,70	0,70	O09	OKP03
190	56	44,00	1,86	0,54	3,47	14	0,004	1,10	1,40	O34	OKP02
191	50	37,45	1,32	0,37	3,54	14	0,003	0,40	0,40	O09	OKP04
192	48	34,76	1,29	0,35	3,67	14	0,003	0,20	0,30	O09	OKP08
193	48	34,14	1,34	0,35	3,81	14	0,002	2,20	2,30	O10	OKP19
194	54	40,25	1,74	0,45	3,85	14	0,002	0,70	0,70	O23	OKP11
195	50	35,31	1,50	0,37	4,03	14	0,001	0,80	0,80	O07	OKP10
196	44	28,63	1,32	0,32	4,11	14	0,001	0,20	0,20	O02	OKP23
197	40	24,54	1,24	0,30	4,12	14	0,001	1,80	1,80	O39	OKP13
198	42	26,08	1,32	0,31	4,23	14	0,001	0,20	0,20	O20	OKP23
199	30	14,80	1,21	0,28	4,35	14	0,001	0,90	0,90	O38	OKP23
200	40	23,06	1,36	0,30	4,50	14	0,001	0,80	0,70	O33	OKP23
201	36	15,93	1,59	0,29	5,49	14	0,000	0,20	0,20	O06	OKP23
202	41	19,42	1,74	0,31	5,67	14	0,000	0,50	0,50	O10	OKP23
203	40	13,66	2,14	0,30	7,10	14	0,000	0,40	0,50	O04	OKP23

EK 7. Deney Grubunun Ön Testi İçin Puanlayıcı x Birey Etkileşim Analizi Sonucunda Yanlı Olan Etkileşimler

Sıra No	Gözlenen Puan	Beklenen Puan	Yanlılık Büyüklüğü	Model SH	t değeri	sd	p	Infit	Outfit	Öğrenci	Puanlayıcı
1	4	42,62	-3,78	0,51	-7,42	14	0,000	3,70	2,60	O37	ODP06
2	22	42,11	-1,64	0,28	-5,88	14	0,000	0,80	0,80	O07	ODP13
3	4	31,15	-2,80	0,51	-5,49	14	0,000	3,70	2,60	O35	ODP06
4	2	28,30	-3,29	0,71	-4,62	14	0,000	1,80	1,30	O36	ODP06
5	20	36,35	-1,28	0,28	-4,52	14	0,001	1,00	1,00	O04	ODP22
6	12	29,66	-1,46	0,32	-4,51	14	0,001	0,40	0,40	O39	ODP18
7	13	30,25	-1,40	0,31	-4,45	14	0,001	0,50	0,50	O38	ODP18
8	33	45,88	-1,18	0,28	-4,18	14	0,001	0,70	0,70	O01	ODP08
9	30	42,98	-1,11	0,28	-4,00	14	0,001	1,00	1,10	O07	ODP20
10	37	48,52	-1,17	0,29	-3,99	14	0,001	0,80	0,80	O15	ODP10
11	24	37,76	-1,09	0,28	-3,93	14	0,002	1,20	1,20	O20	ODP15
12	37	48,08	-1,11	0,29	-3,81	14	0,002	1,50	1,50	O19	ODP15
13	33	44,87	-1,07	0,28	-3,80	14	0,002	0,80	0,80	O30	ODP04
14	34	45,38	-1,05	0,28	-3,68	14	0,003	0,30	0,30	O31	ODP01
15	47	54,92	-1,25	0,34	-3,65	14	0,003	1,00	0,90	O24	ODP04
16	14	27,90	-1,12	0,31	-3,65	14	0,003	0,90	1,00	O04	ODP09
17	13	26,79	-1,14	0,31	-3,61	14	0,003	0,60	0,50	O04	ODP08
18	35	45,94	-1,03	0,29	-3,58	14	0,003	1,10	1,20	O03	ODP14
19	36	46,46	-1,00	0,29	-3,47	14	0,004	2,00	2,10	O07	ODP02
20	45	53,31	-1,14	0,33	-3,47	14	0,004	1,00	1,00	O16	ODP09
21	44	52,58	-1,11	0,32	-3,46	14	0,004	0,90	0,90	O24	ODP05
22	42	51,11	-1,07	0,31	-3,45	14	0,004	0,80	0,80	O28	ODP09
23	40	49,59	-1,04	0,30	-3,45	14	0,004	0,70	0,70	O34	ODP16
24	34	44,67	-0,97	0,28	-3,41	14	0,004	0,90	0,90	O33	ODP04
25	26	37,60	-0,92	0,28	-3,34	14	0,005	1,00	1,00	O38	ODP22
26	43	51,58	-1,05	0,32	-3,32	14	0,005	0,60	0,60	O01	ODP04
27	45	53,04	-1,09	0,33	-3,32	14	0,005	1,40	1,40	O34	ODP04
28	25	36,67	-0,92	0,28	-3,32	14	0,005	1,60	1,60	O35	ODP15
29	43	51,56	-1,04	0,32	-3,30	14	0,005	0,40	0,40	O15	ODP05
30	35	45,21	-0,95	0,29	-3,30	14	0,005	1,30	1,40	O07	ODP07
31	41	49,82	-0,98	0,31	-3,20	14	0,006	1,20	1,20	O19	ODP09
32	26	37,07	-0,88	0,28	-3,18	14	0,007	1,00	1,00	O39	ODP22
33	27	37,65	-0,85	0,28	-3,08	14	0,008	0,80	0,80	O20	ODP10
34	37	46,00	-0,87	0,29	-2,97	14	0,010	0,60	0,60	O23	ODP02
35	35	44,20	-0,84	0,29	-2,93	14	0,011	0,60	0,60	O13	ODP17
36	19	29,71	-0,83	0,29	-2,90	14	0,012	0,30	0,30	O06	ODP13
37	31	40,60	-0,81	0,28	-2,89	14	0,012	1,60	1,60	O07	ODP17
38	40	48,12	-0,85	0,30	-2,82	14	0,014	0,80	0,80	O12	ODP01
39	24	33,94	-0,77	0,28	-2,78	14	0,015	0,40	0,40	O04	ODP19
40	44	51,14	-0,88	0,32	-2,73	14	0,016	1,40	1,30	O17	ODP09
41	30	39,22	-0,76	0,28	-2,73	14	0,016	0,90	0,90	O08	ODP22
42	31	40,04	-0,75	0,28	-2,70	14	0,017	0,60	0,60	O25	ODP15
43	40	47,80	-0,81	0,30	-2,69	14	0,018	0,50	0,50	O01	ODP03
44	31	39,96	-0,75	0,28	-2,67	14	0,018	1,30	1,30	O32	ODP10
45	32	40,80	-0,75	0,28	-2,66	14	0,019	0,60	0,60	O31	ODP15
46	30	38,99	-0,74	0,28	-2,65	14	0,019	0,80	0,80	O11	ODP14
47	45	51,75	-0,87	0,33	-2,65	14	0,019	0,40	0,40	O34	ODP21
48	38	45,88	-0,77	0,29	-2,60	14	0,021	1,20	1,30	O37	ODP11
49	22	31,09	-0,70	0,28	-2,50	14	0,026	0,90	0,90	O38	ODP03
50	42	48,98	-0,77	0,31	-2,49	14	0,026	0,40	0,30	O31	ODP04
51	31	39,29	-0,69	0,28	-2,46	14	0,028	0,30	0,30	O15	ODP12
52	43	49,73	-0,77	0,32	-2,45	14	0,028	0,40	0,40	O37	ODP03
53	48	53,64	-0,86	0,35	-2,44	14	0,029	1,60	1,80	O24	ODP02
54	34	41,94	-0,69	0,28	-2,44	14	0,029	1,30	1,30	O10	ODP22

55	32	40,10	-0,68	0,28	-2,43	14	0,029	0,50	0,50	O30	ODP05
56	28	36,47	-0,67	0,28	-2,43	14	0,029	0,70	0,70	O06	ODP19
57	32	40,06	-0,68	0,28	-2,42	14	0,030	1,10	1,20	O32	ODP15
58	37	44,54	-0,71	0,29	-2,42	14	0,029	1,10	1,10	O26	ODP15
59	36	43,61	-0,70	0,29	-2,41	14	0,030	0,50	0,50	O22	ODP03
60	37	44,42	-0,70	0,29	-2,38	14	0,032	0,30	0,30	O27	ODP01
61	30	38,10	-0,66	0,28	-2,37	14	0,033	0,70	0,70	O20	ODP14
62	35	42,61	-0,68	0,29	-2,37	14	0,033	0,40	0,40	O05	ODP21
63	42	48,66	-0,73	0,31	-2,35	14	0,034	0,80	0,80	O34	ODP09
64	44	50,28	-0,75	0,32	-2,34	14	0,035	0,60	0,60	O15	ODP09
65	32	39,81	-0,65	0,28	-2,33	14	0,035	1,50	1,50	O10	ODP19
66	28	36,06	-0,64	0,28	-2,31	14	0,037	1,90	1,90	O18	ODP15
67	43	49,31	-0,72	0,32	-2,28	14	0,039	0,80	0,80	O12	ODP02
68	43	49,31	-0,72	0,32	-2,27	14	0,039	0,50	0,50	O19	ODP13
69	32	39,57	-0,63	0,28	-2,26	14	0,040	0,80	0,70	O27	ODP10
70	22	30,24	-0,63	0,28	-2,26	14	0,040	0,40	0,50	O08	ODP13
71	44	50,08	-0,72	0,32	-2,25	14	0,041	0,30	0,30	O34	ODP05
72	38	44,86	-0,66	0,29	-2,22	14	0,043	1,00	1,10	O25	ODP07
73	40	46,61	-0,67	0,30	-2,22	14	0,044	1,40	1,40	O26	ODP09
74	40	46,61	-0,67	0,30	-2,22	14	0,043	0,40	0,50	O25	ODP21
75	38	44,85	-0,65	0,29	-2,22	14	0,043	1,30	1,40	O05	ODP22
76	46	51,65	-0,74	0,33	-2,22	14	0,044	0,70	0,70	O01	ODP22
77	33	40,33	-0,62	0,28	-2,21	14	0,045	0,20	0,20	O33	ODP01
78	50	54,72	-0,82	0,37	-2,20	14	0,045	0,80	1,00	O28	ODP04
79	38	44,76	-0,64	0,29	-2,19	14	0,046	1,20	1,20	O05	ODP04
80	28	35,63	-0,60	0,28	-2,18	14	0,047	0,70	0,80	O30	ODP14
81	27	34,60	-0,59	0,28	-2,15	14	0,049	0,40	0,40	O09	ODP13
82	42	48,14	-0,67	0,31	-2,14	14	0,050	0,50	0,50	O13	ODP01
83	35	41,94	-0,61	0,29	-2,14	14	0,050	0,70	0,70	O35	ODP07
84	59	52,93	2,18	1,02	2,14	14	0,050	0,90	0,70	O16	ODP13
85	58	52,02	1,61	0,73	2,20	14	0,045	2,20	4,40	O16	ODP15
86	59	52,61	2,24	1,02	2,20	14	0,045	1,00	1,10	O21	ODP16
87	50	42,90	0,82	0,37	2,20	14	0,045	1,00	1,20	O09	ODP22
88	33	24,92	0,62	0,28	2,21	14	0,045	0,70	0,80	O38	ODP11
89	56	49,73	1,19	0,54	2,21	14	0,044	1,00	1,20	O13	ODP21
90	55	48,54	1,08	0,49	2,22	14	0,044	0,70	0,70	O26	ODP01
91	52	45,08	0,90	0,40	2,22	14	0,043	0,60	0,60	O07	ODP01
92	52	45,06	0,90	0,40	2,22	14	0,043	0,90	0,90	O20	ODP19
93	55	48,53	1,08	0,49	2,22	14	0,044	0,70	0,70	O32	ODP22
94	59	52,31	2,29	1,02	2,25	14	0,041	1,00	1,20	O28	ODP05
95	55	48,36	1,10	0,49	2,26	14	0,040	0,90	1,00	O02	ODP04
96	57	50,65	1,38	0,61	2,27	14	0,040	0,90	1,00	O28	ODP13
97	56	49,45	1,22	0,54	2,28	14	0,039	0,80	0,80	O28	ODP10
98	55	48,25	1,12	0,49	2,29	14	0,038	0,80	0,80	O13	ODP07
99	34	25,59	0,65	0,28	2,29	14	0,038	2,00	2,00	O39	ODP10
100	51	43,68	0,89	0,39	2,29	14	0,038	0,50	0,50	O18	ODP19
101	31	22,58	0,64	0,28	2,31	14	0,037	0,50	0,50	O05	ODP12
102	47	39,11	0,80	0,34	2,33	14	0,036	0,90	0,90	O08	ODP04
103	43	34,76	0,74	0,32	2,33	14	0,035	0,60	0,70	O02	ODP06
104	31	22,34	0,66	0,28	2,38	14	0,032	0,50	0,50	O08	ODP06
105	52	44,47	0,96	0,40	2,38	14	0,032	0,70	0,70	O01	ODP10
106	59	51,52	2,42	1,02	2,38	14	0,032	1,00	1,10	O24	ODP20
107	51	43,31	0,93	0,39	2,39	14	0,032	1,10	1,00	O31	ODP20
108	56	48,95	1,29	0,54	2,40	14	0,031	0,90	0,90	O16	ODP06
109	35	26,19	0,69	0,29	2,40	14	0,031	2,20	2,20	O38	ODP10
110	57	50,09	1,46	0,61	2,40	14	0,031	0,90	1,00	O26	ODP21
111	34	25,11	0,69	0,28	2,42	14	0,030	0,60	0,70	O04	ODP17
112	43	34,30	0,77	0,32	2,45	14	0,028	1,10	1,10	O10	ODP09
113	56	48,72	1,32	0,54	2,46	14	0,028	0,90	1,20	O07	ODP04
114	50	41,93	0,92	0,37	2,46	14	0,028	0,50	0,50	O11	ODP18
115	59	50,79	2,53	1,02	2,49	14	0,026	0,90	0,80	O19	ODP03

116	59	50,57	2,56	1,02	2,52	14	0,025	1,00	1,40	O03	ODP02
117	59	50,56	2,56	1,02	2,52	14	0,025	0,90	0,80	O17	ODP08
118	24	15,41	0,70	0,28	2,52	14	0,025	0,80	0,80	O06	ODP12
119	44	35,02	0,82	0,32	2,54	14	0,023	1,30	1,40	O05	ODP10
120	49	40,44	0,92	0,36	2,55	14	0,023	1,70	2,20	O07	ODP15
121	53	44,84	1,10	0,43	2,57	14	0,022	0,70	0,80	O01	ODP14
122	51	42,60	1,00	0,39	2,57	14	0,022	0,40	0,40	O14	ODP16
123	59	50,10	2,63	1,02	2,58	14	0,022	1,00	1,10	O37	ODP05
124	47	37,96	0,90	0,34	2,62	14	0,020	1,00	1,10	O30	ODP09
125	59	49,74	2,68	1,02	2,63	14	0,020	1,00	1,90	O28	ODP14
126	55	46,81	1,29	0,49	2,65	14	0,019	0,80	0,90	O13	ODP18
127	46	36,61	0,90	0,33	2,69	14	0,018	1,50	1,60	O10	ODP05
128	51	42,06	1,05	0,39	2,71	14	0,017	1,20	1,30	O05	ODP02
129	57	48,66	1,65	0,61	2,71	14	0,017	0,80	0,80	O29	ODP17
130	41	31,11	0,83	0,31	2,72	14	0,017	1,50	1,40	O04	ODP07
131	49	39,80	0,98	0,36	2,72	14	0,017	1,00	1,10	O22	ODP15
132	52	42,80	1,13	0,40	2,80	14	0,014	1,10	1,10	O09	ODP04
133	57	48,24	1,70	0,61	2,80	14	0,014	1,00	1,20	O03	ODP18
134	56	47,26	1,50	0,54	2,80	14	0,014	0,80	0,70	O01	ODP18
135	53	43,88	1,20	0,43	2,81	14	0,014	0,80	0,80	O31	ODP18
136	53	43,89	1,20	0,43	2,81	14	0,014	1,00	1,00	O35	ODP21
137	57	48,13	1,72	0,61	2,82	14	0,014	0,80	0,80	O37	ODP13
138	53	43,70	1,22	0,43	2,85	14	0,013	0,60	0,60	O14	ODP01
139	59	47,78	2,93	1,02	2,88	14	0,012	0,90	0,70	O26	ODP03
140	58	48,45	2,12	0,73	2,90	14	0,012	0,90	0,90	O32	ODP04
141	47	36,84	0,99	0,34	2,90	14	0,012	0,70	0,70	O35	ODP17
142	50	40,12	1,09	0,37	2,91	14	0,011	0,80	0,90	O22	ODP14
143	39	28,19	0,88	0,30	2,94	14	0,011	0,90	0,90	O08	ODP15
144	44	33,33	0,95	0,32	2,97	14	0,010	1,20	1,20	O11	ODP06
145	42	31,14	0,93	0,31	2,98	14	0,010	0,90	1,00	O08	ODP09
146	46	35,31	1,01	0,33	3,01	14	0,009	1,90	2,40	O07	ODP06
147	56	46,06	1,64	0,54	3,05	14	0,009	1,00	1,20	O02	ODP02
148	56	45,99	1,65	0,54	3,07	14	0,008	1,50	1,60	O26	ODP13
149	53	42,45	1,34	0,43	3,15	14	0,007	1,00	1,00	O23	ODP20
150	52	41,29	1,28	0,40	3,16	14	0,007	1,60	1,80	O03	ODP06
151	34	21,82	0,94	0,28	3,31	14	0,005	0,70	0,70	O06	ODP06
152	50	38,19	1,26	0,37	3,37	14	0,005	0,40	0,40	O09	ODP01
153	42	29,29	1,07	0,31	3,44	14	0,004	1,40	1,40	O38	ODP09
154	32	19,19	0,99	0,28	3,53	14	0,003	0,90	0,90	O04	ODP06
155	57	44,19	2,16	0,61	3,56	14	0,003	0,80	0,70	O07	ODP03
156	53	40,56	1,52	0,43	3,57	14	0,003	0,50	0,50	O30	ODP01
157	51	38,32	1,39	0,39	3,59	14	0,003	0,40	0,50	O35	ODP08
158	42	28,70	1,12	0,31	3,59	14	0,003	1,40	1,40	O39	ODP09
159	52	38,66	1,52	0,40	3,75	14	0,002	1,80	1,70	O11	ODP15
160	57	40,42	2,53	0,61	4,16	14	0,001	0,80	0,60	O14	ODP08

EK 8. Kontrol Grubunun Son Testi İçin Puanlayıcı x Birey Etkileşim Analizi Sonucunda Yanlı Olan Etkileşimler

Sıra No	Gözlenen Puan	Beklenen Puan	Yanlılık Büyüklüğü	Model SH	t değeri	sd	p	Infit	Outfit	Öğrenci	Puanlayıcı
1	34	49,56	-1,64	0,29	-5,72	14	0,000	0,60	0,60	O16	SKP13
2	31	47,43	-1,59	0,28	-5,69	14	0,000	1,50	1,50	O24	SKP13
3	31	46,31	-1,45	0,28	-5,19	14	0,000	0,70	0,70	O17	SKP13
4	24	40,22	-1,33	0,27	-4,83	14	0,000	0,20	0,30	O25	SKP13
5	32	45,18	-1,24	0,28	-4,39	14	0,001	0,60	0,60	O19	SKP13
6	25	39,64	-1,20	0,27	-4,36	14	0,001	1,00	1,00	O20	SKP11
7	42	52,54	-1,40	0,32	-4,36	14	0,001	0,30	0,30	O16	SKP19
8	21	36,19	-1,19	0,28	-4,27	14	0,001	0,70	0,70	O09	SKP11
9	25	39,05	-1,14	0,27	-4,16	14	0,001	1,90	1,90	O30	SKP12
10	25	38,98	-1,13	0,27	-4,13	14	0,001	0,30	0,30	O18	SKP09
11	37	48,37	-1,22	0,30	-4,10	14	0,001	0,20	0,20	O37	SKP19
12	41	51,23	-1,27	0,32	-4,03	14	0,001	1,20	1,20	O17	SKP09
13	22	36,06	-1,10	0,28	-3,98	14	0,001	1,70	1,70	O32	SKP10
14	29	41,73	-1,10	0,28	-3,97	14	0,001	0,70	0,70	O11	SKP09
15	22	36,03	-1,10	0,28	-3,97	14	0,001	1,20	1,20	O33	SKP11
16	25	38,45	-1,08	0,27	-3,95	14	0,001	1,20	1,20	O35	SKP12
17	13	28,03	-1,21	0,31	-3,89	14	0,002	0,30	0,30	O39	SKP21
18	20	34,00	-1,08	0,28	-3,86	14	0,002	1,50	1,50	O30	SKP02
19	13	27,84	-1,19	0,31	-3,84	14	0,002	0,30	0,30	O39	SKP20
20	13	27,81	-1,19	0,31	-3,83	14	0,002	0,30	0,30	O38	SKP20
21	44	52,77	-1,23	0,33	-3,69	14	0,002	1,30	1,20	O15	SKP01
22	39	48,90	-1,11	0,31	-3,64	14	0,003	1,60	1,70	O03	SKP18
23	41	50,42	-1,14	0,32	-3,62	14	0,003	0,50	0,50	O15	SKP19
24	21	33,95	-1,00	0,28	-3,59	14	0,003	0,80	0,80	O09	SKP12
25	48	55,34	-1,31	0,36	-3,59	14	0,003	2,20	2,30	O16	SKP16
26	22	34,59	-0,97	0,28	-3,53	14	0,003	0,60	0,60	O11	SKP10
27	41	50,16	-1,10	0,32	-3,50	14	0,004	0,40	0,40	O28	SKP19
28	42	50,88	-1,11	0,32	-3,47	14	0,004	0,70	0,70	O29	SKP08
29	20	32,54	-0,96	0,28	-3,44	14	0,004	0,70	0,70	O08	SKP11
30	10	23,34	-1,17	0,34	-3,44	14	0,004	0,70	0,60	O33	SKP23
31	44	52,31	-1,14	0,33	-3,43	14	0,004	0,60	0,60	O28	SKP04
32	26	37,70	-0,94	0,27	-3,43	14	0,004	1,70	1,70	O18	SKP11
33	15	28,00	-1,02	0,30	-3,43	14	0,004	0,60	0,60	O38	SKP21
34	44	52,26	-1,13	0,33	-3,40	14	0,004	1,80	1,80	O34	SKP16
35	38	47,55	-1,02	0,30	-3,39	14	0,004	0,90	1,00	O03	SKP20
36	34	44,12	-0,96	0,29	-3,33	14	0,005	0,50	0,60	O02	SKP03
37	31	41,62	-0,93	0,28	-3,33	14	0,005	1,00	1,00	O35	SKP09
38	39	48,04	-0,99	0,31	-3,26	14	0,006	0,70	0,70	O01	SKP07
39	46	53,45	-1,13	0,35	-3,25	14	0,006	2,10	2,10	O17	SKP16
40	34	43,88	-0,93	0,29	-3,24	14	0,006	1,60	1,70	O37	SKP02
41	36	45,46	-0,94	0,29	-3,21	14	0,006	0,60	0,60	O07	SKP09
42	36	45,48	-0,94	0,29	-3,21	14	0,006	0,80	0,80	O03	SKP19
43	41	49,52	-1,01	0,32	-3,19	14	0,007	1,20	1,30	O29	SKP05
44	35	44,59	-0,93	0,29	-3,19	14	0,007	1,40	1,50	O03	SKP12
45	39	47,86	-0,97	0,31	-3,17	14	0,007	0,90	0,90	O19	SKP15
46	44	51,79	-1,05	0,33	-3,16	14	0,007	0,50	0,50	O28	SKP03
47	42	50,17	-1,00	0,32	-3,13	14	0,007	1,00	1,00	O24	SKP05
48	27	37,60	-0,86	0,27	-3,12	14	0,008	2,00	2,00	O20	SKP12
49	33	42,55	-0,87	0,28	-3,05	14	0,009	0,70	0,80	O17	SKP23
50	47	53,82	-1,08	0,35	-3,04	14	0,009	1,90	2,00	O15	SKP16
51	34	43,29	-0,87	0,29	-3,02	14	0,009	0,20	0,30	O23	SKP09
52	40	48,32	-0,94	0,31	-3,02	14	0,009	0,20	0,20	O34	SKP19
53	39	47,45	-0,92	0,31	-3,00	14	0,010	0,60	0,60	O19	SKP14
54	34	43,22	-0,86	0,29	-2,99	14	0,010	1,00	1,00	O13	SKP05

55	34	43,23	-0,86	0,29	-2,99	14	0,010	0,30	0,40	O27	SKP08
56	16	27,20	-0,87	0,29	-2,98	14	0,010	0,40	0,40	O39	SKP17
57	31	40,56	-0,83	0,28	-2,95	14	0,011	1,20	1,20	O11	SKP11
58	33	42,22	-0,84	0,28	-2,94	14	0,011	0,70	0,60	O13	SKP14
59	22	32,56	-0,81	0,28	-2,93	14	0,011	1,20	1,20	O10	SKP11
60	39	47,17	-0,88	0,31	-2,88	14	0,012	1,60	1,70	O03	SKP17
61	28	37,52	-0,77	0,28	-2,81	14	0,014	0,20	0,20	O09	SKP09
62	34	42,73	-0,81	0,29	-2,81	14	0,014	0,50	0,50	O13	SKP15
63	45	51,80	-0,94	0,34	-2,77	14	0,015	0,90	0,90	O28	SKP06
64	44	50,98	-0,92	0,33	-2,76	14	0,016	1,00	1,10	O21	SKP17
65	32	40,85	-0,78	0,28	-2,75	14	0,016	0,70	0,80	O20	SKP09
66	44	50,87	-0,90	0,33	-2,70	14	0,017	1,40	1,50	O01	SKP16
67	45	51,62	-0,91	0,34	-2,68	14	0,018	1,00	1,00	O17	SKP21
68	43	49,92	-0,86	0,33	-2,64	14	0,020	0,30	0,30	O17	SKP19
69	28	36,93	-0,72	0,28	-2,62	14	0,020	0,90	0,90	O05	SKP08
70	30	38,70	-0,73	0,28	-2,62	14	0,020	0,80	0,80	O14	SKP13
71	36	43,86	-0,76	0,29	-2,59	14	0,021	1,00	1,00	O27	SKP22
72	43	49,74	-0,83	0,33	-2,56	14	0,023	1,70	1,80	O15	SKP05
73	37	44,59	-0,75	0,30	-2,54	14	0,024	0,90	0,90	O07	SKP08
74	25	33,98	-0,70	0,27	-2,54	14	0,024	0,30	0,30	O08	SKP09
75	18	27,17	-0,70	0,28	-2,47	14	0,027	0,70	0,70	O38	SKP17
76	40	46,95	-0,76	0,31	-2,45	14	0,028	1,00	0,90	O14	SKP01
77	34	41,65	-0,69	0,29	-2,42	14	0,030	1,00	1,00	O11	SKP07
78	41	47,66	-0,75	0,32	-2,39	14	0,032	1,00	1,00	O03	SKP21
79	31	38,89	-0,67	0,28	-2,38	14	0,032	1,60	1,70	O18	SKP07
80	40	46,71	-0,73	0,31	-2,35	14	0,034	0,40	0,40	O26	SKP20
81	37	44,07	-0,70	0,30	-2,34	14	0,035	1,20	1,20	O27	SKP07
82	34	41,42	-0,67	0,29	-2,34	14	0,035	1,00	1,10	O11	SKP22
83	30	37,81	-0,65	0,28	-2,32	14	0,036	0,80	0,80	O31	SKP02
84	17	25,56	-0,66	0,29	-2,30	14	0,037	0,80	0,80	O08	SKP10
85	35	42,17	-0,66	0,29	-2,29	14	0,038	1,30	1,40	O30	SKP09
86	33	40,37	-0,65	0,28	-2,28	14	0,039	0,80	0,90	O37	SKP23
87	26	34,00	-0,62	0,27	-2,27	14	0,040	0,40	0,40	O10	SKP09
88	24	32,12	-0,62	0,27	-2,26	14	0,040	0,90	0,90	O04	SKP01
89	43	49,02	-0,73	0,33	-2,24	14	0,042	1,50	1,60	O13	SKP16
90	44	49,83	-0,74	0,33	-2,22	14	0,043	0,80	0,90	O37	SKP17
91	39	45,46	-0,67	0,31	-2,20	14	0,045	0,40	0,40	O26	SKP08
92	20	28,12	-0,62	0,28	-2,20	14	0,045	1,70	1,70	O04	SKP11
93	26	33,62	-0,59	0,27	-2,16	14	0,049	1,70	1,80	O10	SKP22
94	31	38,17	-0,60	0,28	-2,15	14	0,050	1,20	1,20	O07	SKP02
95	41	47,08	-0,68	0,32	-2,15	14	0,050	1,80	1,80	O03	SKP07
96	57	51,39	1,33	0,62	2,15	14	0,050	0,80	0,70	O28	SKP07
97	56	50,17	1,18	0,55	2,17	14	0,048	0,70	0,60	O19	SKP22
98	53	46,66	0,96	0,44	2,19	14	0,046	0,60	0,60	O29	SKP13
99	51	44,37	0,88	0,40	2,19	14	0,046	0,90	1,00	O34	SKP13
100	50	43,16	0,86	0,39	2,21	14	0,044	0,60	0,70	O11	SKP04
101	37	28,93	0,66	0,30	2,21	14	0,044	1,20	1,20	O06	SKP05
102	60	50,49	3,72	1,69	2,21	14	0,045	0,00	0,00	O17	SKP11
103	57	51,07	1,38	0,62	2,24	14	0,042	0,80	0,70	O37	SKP01
104	57	50,97	1,40	0,62	2,26	14	0,040	1,10	1,80	O21	SKP09
105	42	33,95	0,73	0,32	2,29	14	0,038	1,60	1,70	O09	SKP05
106	58	51,86	1,70	0,74	2,30	14	0,037	1,00	1,00	O21	SKP04
107	53	46,23	1,01	0,44	2,31	14	0,037	0,80	1,00	O21	SKP10
108	58	51,82	1,71	0,74	2,31	14	0,037	0,90	0,70	O19	SKP18
109	42	33,82	0,75	0,32	2,33	14	0,036	0,60	0,50	O14	SKP23
110	46	38,16	0,81	0,35	2,35	14	0,034	0,50	0,50	O09	SKP03
111	56	49,52	1,28	0,55	2,35	14	0,034	0,90	1,10	O29	SKP12
112	37	28,34	0,70	0,30	2,36	14	0,033	1,00	1,00	O04	SKP08
113	50	42,58	0,92	0,39	2,37	14	0,033	1,40	1,50	O14	SKP05
114	58	51,44	1,77	0,74	2,39	14	0,031	1,10	2,20	O28	SKP09
115	53	45,92	1,05	0,44	2,39	14	0,031	1,00	1,10	O13	SKP09

116	49	41,33	0,90	0,37	2,40	14	0,031	0,40	0,50	O23	SKP19
117	57	50,38	1,49	0,62	2,41	14	0,030	1,00	1,30	O19	SKP09
118	59	51,49	2,50	1,02	2,45	14	0,028	1,10	2,00	O29	SKP09
119	58	50,97	1,85	0,74	2,50	14	0,026	1,00	1,50	O15	SKP11
120	56	48,95	1,36	0,55	2,50	14	0,026	0,90	1,00	O34	SKP11
121	52	44,33	1,05	0,42	2,51	14	0,025	1,10	1,00	O14	SKP11
122	57	49,77	1,58	0,62	2,56	14	0,023	1,10	1,80	O34	SKP09
123	58	50,71	1,89	0,74	2,56	14	0,023	0,90	0,90	O19	SKP20
124	30	20,59	0,72	0,28	2,57	14	0,022	0,30	0,30	O04	SKP13
125	44	35,01	0,86	0,33	2,58	14	0,022	0,40	0,40	O33	SKP19
126	46	37,18	0,90	0,35	2,60	14	0,021	1,40	1,30	O27	SKP13
127	50	41,70	1,01	0,39	2,61	14	0,021	0,70	0,80	O22	SKP11
128	57	49,47	1,62	0,62	2,63	14	0,020	0,90	1,10	O28	SKP12
129	58	50,21	1,97	0,74	2,66	14	0,019	1,10	1,80	O21	SKP11
130	35	25,12	0,77	0,29	2,66	14	0,019	0,50	0,50	O08	SKP13
131	42	32,47	0,85	0,32	2,67	14	0,018	0,50	0,50	O36	SKP13
132	46	36,87	0,93	0,35	2,68	14	0,018	0,40	0,40	O02	SKP10
133	59	49,81	2,77	1,02	2,70	14	0,017	1,10	2,00	O37	SKP09
134	50	41,24	1,05	0,39	2,73	14	0,016	0,40	0,50	O27	SKP12
135	58	49,59	2,06	0,74	2,78	14	0,015	0,80	0,60	O19	SKP11
136	52	43,07	1,19	0,42	2,84	14	0,013	0,70	0,70	O27	SKP11
137	35	24,43	0,82	0,29	2,84	14	0,013	0,20	0,20	O38	SKP19
138	55	46,37	1,43	0,50	2,86	14	0,013	1,10	1,30	O16	SKP23
139	58	49,00	2,14	0,74	2,90	14	0,012	0,90	0,90	O37	SKP11
140	57	48,11	1,81	0,62	2,94	14	0,011	1,10	1,80	O01	SKP09
141	42	31,28	0,95	0,32	2,96	14	0,010	1,30	1,30	O18	SKP10
142	29	18,19	0,83	0,28	2,99	14	0,010	1,10	1,10	O39	SKP13
143	54	44,72	1,39	0,46	3,00	14	0,010	0,60	0,60	O37	SKP10
144	55	45,58	1,52	0,50	3,05	14	0,009	0,80	0,70	O25	SKP11
145	53	43,37	1,34	0,44	3,05	14	0,009	1,10	1,20	O12	SKP11
146	51	40,88	1,24	0,40	3,11	14	0,008	1,30	1,30	O20	SKP17
147	56	46,29	1,71	0,55	3,13	14	0,007	0,90	1,00	O26	SKP09
148	46	34,63	1,12	0,35	3,23	14	0,006	0,60	0,60	O13	SKP23
149	50	39,19	1,25	0,39	3,24	14	0,006	0,60	0,60	O07	SKP10
150	41	28,95	1,03	0,32	3,27	14	0,006	0,50	0,50	O33	SKP13
151	55	44,43	1,65	0,50	3,32	14	0,005	1,40	1,70	O37	SKP13
152	56	45,35	1,82	0,55	3,34	14	0,005	0,90	1,00	O14	SKP09
153	50	38,48	1,32	0,39	3,41	14	0,004	0,70	0,60	O31	SKP13
154	49	36,39	1,36	0,37	3,63	14	0,003	0,60	0,60	O09	SKP08
155	46	32,77	1,27	0,35	3,67	14	0,003	1,00	1,10	O10	SKP08
156	31	17,41	1,05	0,28	3,73	14	0,002	1,60	1,60	O38	SKP02
157	51	38,19	1,50	0,40	3,74	14	0,002	0,50	0,60	O09	SKP06
158	52	39,14	1,58	0,42	3,78	14	0,002	0,70	0,70	O09	SKP04
159	39	24,46	1,18	0,31	3,85	14	0,002	0,20	0,20	O39	SKP19
160	59	39,68	3,94	1,02	3,86	14	0,002	1,00	1,10	O18	SKP21
161	48	33,97	1,43	0,36	3,92	14	0,002	0,70	0,90	O07	SKP23
162	33	17,44	1,20	0,28	4,23	14	0,001	1,60	1,60	O39	SKP02
163	48	32,52	1,55	0,36	4,25	14	0,001	0,50	0,60	O12	SKP23

EK 9. Deney Grubunun Son Testi İçin *Puanlayıcı x Birey* Etkileşim Analizi Sonucunda Yanlı Olan Etkileşimler

Sıra No	Gözlenen Puan	Beklenen Puan	Yanlılık Büyüklüğü	Model SH	t değeri	sd	p	Infit	Outfit	Öğrenci	Puanlayıcı
1	18	34,59	-1,29	0,28	-4,52	14	0,001	0,20	0,20	O04	SDP13
2	28	42,12	-1,21	0,28	-4,41	14	0,001	0,60	0,60	O04	SDP04
3	36	47,29	-1,17	0,29	-3,97	14	0,001	2,20	2,30	O07	SDP02
4	43	52,10	-1,21	0,33	-3,72	14	0,002	0,90	0,90	O24	SDP05
5	41	50,36	-1,13	0,32	-3,59	14	0,003	0,70	0,70	O34	SDP07
6	33	43,82	-1,01	0,28	-3,53	14	0,003	0,40	0,40	O37	SDP12
7	26	37,98	-0,97	0,27	-3,52	14	0,003	1,40	1,50	O10	SDP19
8	35	45,28	-1,01	0,29	-3,46	14	0,004	0,80	0,80	O10	SDP04
9	40	49,23	-1,06	0,31	-3,43	14	0,004	0,50	0,50	O15	SDP11
10	36	45,99	-1,00	0,29	-3,42	14	0,004	0,60	0,60	O32	SDP19
11	41	49,84	-1,05	0,32	-3,34	14	0,005	0,40	0,40	O29	SDP08
12	41	49,67	-1,03	0,32	-3,27	14	0,006	1,00	1,00	O24	SDP11
13	13	25,39	-1,01	0,31	-3,25	14	0,006	0,50	0,60	O39	SDP15
14	32	41,84	-0,88	0,28	-3,11	14	0,008	1,10	1,10	O11	SDP10
15	42	50,13	-1,00	0,32	-3,11	14	0,008	1,20	1,20	O03	SDP17
16	43	50,93	-1,02	0,33	-3,11	14	0,008	2,20	2,10	O21	SDP18
17	39	47,68	-0,95	0,31	-3,10	14	0,008	0,80	0,80	O19	SDP11
18	43	50,78	-0,99	0,33	-3,04	14	0,009	1,30	1,30	O29	SDP06
19	39	47,46	-0,92	0,31	-3,00	14	0,010	0,90	1,00	O14	SDP03
20	39	47,19	-0,88	0,31	-2,89	14	0,012	1,40	1,60	O14	SDP02
21	27	36,83	-0,79	0,27	-2,87	14	0,012	0,50	0,50	O05	SDP14
22	36	44,58	-0,84	0,29	-2,86	14	0,013	0,30	0,40	O19	SDP12
23	37	45,29	-0,83	0,30	-2,81	14	0,014	0,60	0,60	O23	SDP02
24	23	32,94	-0,76	0,28	-2,77	14	0,015	1,00	1,00	O38	SDP22
25	36	44,25	-0,80	0,29	-2,73	14	0,016	0,20	0,20	O13	SDP09
26	47	53,26	-0,97	0,35	-2,72	14	0,017	1,40	1,70	O24	SDP02
27	36	44,16	-0,79	0,29	-2,70	14	0,017	0,90	0,90	O11	SDP03
28	41	48,37	-0,85	0,32	-2,69	14	0,018	0,90	0,90	O21	SDP11
29	44	50,82	-0,89	0,33	-2,68	14	0,018	0,50	0,50	O37	SDP01
30	27	36,00	-0,72	0,27	-2,61	14	0,021	1,30	1,30	O09	SDP15
31	30	38,55	-0,71	0,28	-2,57	14	0,022	1,00	1,00	O10	SDP13
32	32	40,29	-0,72	0,28	-2,56	14	0,023	0,60	0,60	O09	SDP03
33	27	35,87	-0,70	0,27	-2,56	14	0,023	0,50	0,50	O05	SDP09
34	36	43,65	-0,74	0,29	-2,51	14	0,025	0,20	0,20	O13	SDP08
35	43	49,61	-0,81	0,33	-2,50	14	0,026	1,20	1,30	O19	SDP06
36	29	37,46	-0,69	0,28	-2,50	14	0,025	1,40	1,40	O09	SDP18
37	43	49,48	-0,80	0,33	-2,44	14	0,029	1,50	1,60	O19	SDP15
38	35	42,55	-0,70	0,29	-2,43	14	0,029	0,80	0,90	O36	SDP02
39	30	38,12	-0,68	0,28	-2,43	14	0,029	1,40	1,40	O06	SDP22
40	44	50,28	-0,81	0,33	-2,42	14	0,030	0,20	0,20	O29	SDP09
41	38	44,99	-0,71	0,30	-2,36	14	0,034	0,40	0,40	O13	SDP14
42	43	49,25	-0,76	0,33	-2,34	14	0,035	0,50	0,50	O21	SDP08
43	29	36,91	-0,64	0,28	-2,33	14	0,036	1,10	1,10	O06	SDP19
44	38	44,92	-0,70	0,30	-2,33	14	0,035	0,90	0,80	O11	SDP19
45	42	48,27	-0,73	0,32	-2,29	14	0,038	0,90	0,80	O07	SDP20
46	43	49,09	-0,74	0,33	-2,27	14	0,039	0,40	0,40	O13	SDP17
47	46	51,56	-0,78	0,35	-2,25	14	0,041	0,50	0,50	O23	SDP04
48	17	25,36	-0,65	0,29	-2,25	14	0,041	0,60	0,60	O38	SDP15
49	51	55,45	-0,90	0,40	-2,24	14	0,042	1,30	1,30	O34	SDP04
50	46	51,54	-0,78	0,35	-2,24	14	0,042	0,40	0,40	O29	SDP05
51	25	32,97	-0,61	0,27	-2,24	14	0,042	1,50	1,50	O39	SDP22
52	36	42,87	-0,65	0,29	-2,22	14	0,043	0,60	0,60	O36	SDP03
53	45	50,66	-0,75	0,34	-2,22	14	0,044	0,80	0,90	O29	SDP15
54	43	48,94	-0,72	0,33	-2,21	14	0,045	0,70	0,70	O28	SDP11

55	38	44,56	-0,66	0,30	-2,20	14	0,046	0,50	0,60	O30	SDP03
56	44	49,79	-0,73	0,33	-2,20	14	0,045	0,30	0,30	O28	SDP08
57	40	46,32	-0,68	0,31	-2,19	14	0,046	0,40	0,40	O22	SDP13
58	25	32,76	-0,60	0,27	-2,18	14	0,047	1,30	1,30	O06	SDP18
59	32	39,15	-0,61	0,28	-2,17	14	0,047	1,20	1,30	O10	SDP22
60	53	46,65	0,96	0,44	2,19	14	0,046	0,80	0,90	O13	SDP07
61	42	34,35	0,70	0,32	2,19	14	0,046	0,30	0,30	O33	SDP08
62	36	27,95	0,64	0,29	2,19	14	0,046	0,60	0,60	O06	SDP11
63	51	44,33	0,88	0,40	2,20	14	0,045	1,20	1,20	O11	SDP21
64	54	47,71	1,03	0,46	2,21	14	0,044	0,80	0,90	O13	SDP02
65	50	42,96	0,88	0,39	2,27	14	0,040	0,40	0,40	O18	SDP13
66	59	52,53	2,32	1,02	2,27	14	0,040	1,00	0,80	O34	SDP22
67	59	52,48	2,33	1,02	2,28	14	0,039	1,10	2,60	O21	SDP03
68	48	40,53	0,84	0,36	2,31	14	0,036	0,60	0,60	O35	SDP14
69	49	41,68	0,87	0,37	2,31	14	0,037	0,40	0,40	O11	SDP18
70	55	48,39	1,17	0,50	2,34	14	0,035	0,80	0,80	O03	SDP16
71	60	49,08	3,88	1,65	2,36	14	0,034	0,00	0,00	O03	SDP03
72	59	51,96	2,42	1,02	2,37	14	0,033	0,90	0,80	O37	SDP19
73	58	51,42	1,77	0,74	2,40	14	0,031	0,90	0,90	O34	SDP03
74	52	44,73	1,00	0,42	2,40	14	0,031	0,70	0,80	O30	SDP21
75	56	49,21	1,33	0,55	2,43	14	0,029	1,00	1,10	O03	SDP21
76	54	46,87	1,13	0,46	2,44	14	0,028	0,60	0,60	O32	SDP22
77	56	48,83	1,38	0,55	2,52	14	0,024	0,80	0,70	O03	SDP02
78	58	50,77	1,88	0,74	2,54	14	0,024	0,90	0,90	O28	SDP14
79	56	48,72	1,39	0,55	2,55	14	0,023	1,00	1,30	O18	SDP04
80	55	47,56	1,28	0,50	2,56	14	0,023	0,90	1,10	O07	SDP03
81	42	32,87	0,82	0,32	2,57	14	0,022	0,30	0,30	O33	SDP11
82	54	46,41	1,19	0,46	2,57	14	0,022	0,70	0,70	O32	SDP13
83	56	48,60	1,41	0,55	2,58	14	0,022	0,90	1,20	O13	SDP19
84	58	50,54	1,92	0,74	2,59	14	0,021	0,90	0,90	O17	SDP14
85	53	45,09	1,15	0,44	2,61	14	0,020	0,60	0,60	O20	SDP22
86	54	46,21	1,22	0,46	2,62	14	0,020	1,20	1,40	O30	SDP22
87	49	40,44	0,99	0,37	2,64	14	0,020	1,10	1,20	O05	SDP02
88	37	27,25	0,79	0,30	2,64	14	0,019	0,90	1,00	O38	SDP05
89	50	41,55	1,02	0,39	2,65	14	0,019	0,50	0,50	O01	SDP12
90	55	47,24	1,32	0,50	2,65	14	0,019	0,80	0,80	O13	SDP16
91	59	49,99	2,74	1,02	2,68	14	0,018	0,90	0,80	O17	SDP09
92	55	47,08	1,34	0,50	2,69	14	0,018	0,70	0,80	O01	SDP15
93	56	47,97	1,50	0,55	2,74	14	0,016	0,90	0,90	O13	SDP03
94	58	49,74	2,04	0,74	2,76	14	0,016	0,90	0,70	O03	SDP20
95	58	49,67	2,05	0,74	2,77	14	0,015	0,90	1,00	O03	SDP19
96	54	45,58	1,29	0,46	2,78	14	0,015	0,80	0,90	O02	SDP02
97	54	45,39	1,31	0,46	2,83	14	0,013	0,80	0,90	O30	SDP20
98	58	49,01	2,14	0,74	2,90	14	0,012	0,90	0,90	O34	SDP14
99	49	39,21	1,11	0,37	2,95	14	0,011	0,30	0,40	O09	SDP01
100	57	48,06	1,82	0,62	2,95	14	0,011	0,90	0,80	O01	SDP18
101	55	45,87	1,49	0,50	2,98	14	0,010	0,90	0,90	O13	SDP18
102	59	45,74	3,30	1,02	3,23	14	0,006	1,00	1,10	O17	SDP12
103	51	39,66	1,36	0,40	3,40	14	0,004	0,50	0,50	O35	SDP09
104	57	45,75	2,11	0,62	3,42	14	0,004	0,90	0,80	O35	SDP22
105	51	39,12	1,41	0,40	3,53	14	0,003	0,80	0,70	O32	SDP11
106	56	44,40	1,93	0,55	3,53	14	0,003	0,80	0,70	O14	SDP14
107	41	27,27	1,16	0,32	3,67	14	0,003	1,10	1,20	O39	SDP05
108	56	43,03	2,08	0,55	3,81	14	0,002	0,80	0,80	O14	SDP08
109	57	43,64	2,35	0,62	3,81	14	0,002	0,80	0,60	O14	SDP09
110	54	40,39	1,85	0,46	3,98	14	0,001	1,30	1,30	O11	SDP15

EK 10. Analitik Dereceli Puanlama Anahtarının Türkçe Formu

Kriterler		Puanlama					
Temel kriterler	Alt Kriterler	0 puan	1 puan	2 puan	3 puan	4 puan	Toplam puan
Organizasyon	Kompozisyonun Başlığı	( )	( )	( )	( )	( )	( )
	Giriş-Gelişme-Sonuç	( )	( )	( )	( )	( )	( )
	Problem Durumu	( )	( )	( )	( )	( )	( )
	Konu Cümlesi	( )	( )	( )	( )	( )	( )
	Cümleleri Destekleme	( )	( )	( )	( )	( )	( )
	Uygun Kelime Uzunluğu	( )	( )	( )	( )	( )	( )
İçerik	Konu Uygunluğu	( )	( )	( )	( )	( )	( )
	Fikir Geliştirme	( )	( )	( )	( )	( )	( )
Anlam Bütünlüğü	Anlam Bütünlüğü	( )	( )	( )	( )	( )	( )
Bağdaşlık	Bağdaşlık	( )	( )	( )	( )	( )	( )
Dil Bilgisi	Gramer Uygunluğu	( )	( )	( )	( )	( )	( )
	Kullanılan Sözdizimleri	( )	( )	( )	( )	( )	( )
Kelime Haznesi	Kelime Seçimi	( )	( )	( )	( )	( )	( )
	Sözcük Aralığı	( )	( )	( )	( )	( )	( )
Noktalama Bilgisi	Yazım Bilgisi	( )	( )	( )	( )	( )	( )
	Noktalama Bilgisi	( )	( )	( )	( )	( )	( )

EK 11. ÇYRÖM Analizi İçin FACETS Sintaksı

Title=Puanlayıcı Eğitimi

Facets=3

Positive=1,2,3

Non-centered=1

Model=?,?,?,R5

Rating scale=R5,R

Missing=9

Unexpected=2

Usort=1,2,3

Vertical=1A,2A,3A

Pt-biserial=Yes

Arrange=mN

inter-rater=2

Convergence = .01

Iterations (maximum) = 1000

Labels=

1,ÖĞRENCİ/BİREY

1-39

2,PUANLAYICI

1-45

3,KRİTER

1-15

Data=



GAZİLİ OLMAK AYRICALIKTIR...