



**AÇIK UÇLU MADDELERİN PUANLANMASINDA ÖĞRENCİLERE
VERİLEN PUANLAYICI EĞİTİMİNİN PUANLAYICI
DAVRANIŞLARINA ETKİSİNİN ÇOK YÜZEYLİ RASCH ÖLÇME
MODELİ İLE İNCELENMESİ**

Burçin Görgülü Öztürk

YÜKSEK LİSANS TEZİ

EĞİTİM BİLİMLERİ ANA BİLİM DALI

EĞİTİMDE ÖLÇME VE DEĞERLENDİRME BİLİM DALI

GAZİ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

ARALIK, 2023

TELİF HAKKI VE TEZ FOTOKOPİ İZİN FORMU

Bu tezin tüm hakları saklıdır. Kaynak göstermek koşulu ile tezin teslim tarihinden itibaren (...) ay sonra tezden fotokopi çekilebilir.

YAZARIN

Adı : Burçin
Soyadı : GÖRGÜLÜ ÖZTÜRK
Bölümü : Eğitimde Ölçme ve Değerlendirme
İmza :
Teslim Tarihi :

TEZİN

Türkçe Adı : Açık Uçlu Maddelerin Puanlanmasında Öğrencilere Verilen Puanlayıcı Eğitiminin Puanlayıcı Davranışlarına Etkisinin Çok Yüzeyle Rasch Ölçme Modeli ile İncelenmesi
İngilizce Adı : Investigation of the Effect of Rater Training Given to Students on Rater Behavior in Scoring Open-Ended Items Using The Multi-Facet Rasch Measuring Model

ETİK İLKELERE UYGUNLUK BEYANI

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

Yazar Adı Soyadı: Burçin GÖRGÜLÜ ÖZTÜRK

İmza:

JÜRİ ONAY SAYFASI

Burçin GÖRGÜLÜ ÖZTÜRK tarafından hazırlanan “Açık Uçlu Maddelerin Puanlanmasında Öğrencilere Verilen Puanlayıcı Eğitiminin Puanlayıcı Davranışlarına Etkisinin Çok Yüzeyle Rasch Ölçme Modeli ile İncelenmesi” adlı tez çalışması aşağıdaki jüri tarafından oy birliği / oy çokluğu ile Gazi Üniversitesi Eğitim Bilimleri Anabilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalı’nda Yüksek Lisans tezi olarak kabul edilmiştir.

Danışman: Prof. Dr. İsmail KARAKAYA

(Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı, Gazi Üniversitesi)

Başkan: Doç. Dr. Melek Gülşah ŞAHİN

(Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı, Gazi Üniversitesi)

Üye: Prof. Dr. Bayram ÇETİN

(Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı, Gaziantep Üniversitesi)

Tez Savunma Tarihi: 09/11/2023

Bu tezin Eğitim Bilimleri Anabilim Dalı’nda Yüksek Lisans tezi olması için şartları yerine getirdiğini onaylıyorum.

Prof. Dr. Şaban ÇETİN

Eğitim Bilimleri Enstitü Müdürü



Kızım Ezgi'ye...

TEŞEKKÜR

Çalışmamda benden desteğini esirgemeyen ve umutsuzluğa her kapıldığımda beni yüreklendiren ve destekleyen, bana her konuda yardımcı olan değerli hocam Sayın Prof. Dr. İsmail KARAKAYA'ya çok teşekkür ederim.

Tez savunma jürimde yer alıp bana değerli görüşleri ile katkılar sağlayan Sayın Prof. Dr. Bayram ÇETİN ve Sayın Doç. Dr. Melek Gülşah ŞAHİN hocalarıma teşekkürlerimi sunarım.

Tez yazma sürecinde bana destek veren Umur ÖÇ ve Nazira TURSYNBAY'a teşekkür ederim.

Yüksek lisans eğitimim boyunca birbirimize hep destek olduğumuz, yol arkadaşım Serpil KOCAOĞLU'na çok teşekkür ederim.

Ders ve tez döneminde edindiğim bilgilerle bana katkı sağlayan Gazi Üniversitesi Eğitimde Ölçme ve Değerlendirme Bilim Dalı'nda görev yapan tüm hocalarıma teşekkür ederim.

Son olarak beni bugünlere getiren, yetiştiren hayatım boyunca her zaman yanımda olan anneme, babama, abime; elimi hep tutan eşim Erkan'a ve bu zorlu süreç boyunca oyun zamanlarından çaldığım canım kızım Ezgi'ye çok teşekkür ederim.

**AÇIK UÇLU MADDELERİN PUANLANMASINDA ÖĞRENCİLERE
VERİLEN PUANLAYICI EĞİTİMİNİN PUANLAYICI
DAVRANIŞLARINA ETKİSİNİN ÇOK YÜZEYLİ RASCH ÖLÇME
MODELİ İLE İNCELENMESİ**

(Yüksek Lisans Tezi)

Burçin Görgülü Öztürk

GAZİ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Aralık 2023

ÖZ

Bu araştırmada, lise 10.sınıf öğrencilerinin açık uçlu matematik maddelerine verdikleri yanıtların değerlendirilmesinde öğrencilere verilen puanlayıcı eğitiminin puanlayıcı davranışlarına etkisi incelenmiştir. Öğrencilere akran değerlendirmesi yaptırılmıştır ve ilgili analizler Çok Yüzeysel Rasch Ölçme Modeli (ÇYRÖM) ve genellenebilirlik kuramı ile yapılmıştır. Araştırmanın çalışma grubunu 2022-2023 eğitim öğretim yılı ikinci döneminde Ankara ilinde bir devlet okulunda onuncu sınıfta öğrenim gören 65 lise öğrencisi oluşturmaktadır. Araştırmada deney ve kontrol gruplu ön-test son test yarı deneysel desen kullanılmıştır. Deney grubunda 32, kontrol grubunda 33 öğrenci bulunmaktadır. Araştırma öncesinde her iki gruba da 4 açık uçlu matematik maddesinden oluşan başarı testi uygulanmıştır. Araştırmacı tarafından tüm yanıtlar dereceli puanlama anahtarı (DPA) yardımıyla değerlendirilmiş ve içlerinden başarılı, orta başarılı ve daha az başarılı yanıtlar

olmak üzere toplamda üç öğrencinin kağıtları seçilmiştir. Buradaki amaç puanlayıcıların her yanıt tarzını görmelerini sağlamaktır. Deneysel işlem (puanlayıcı eğitimi) uygulanmadan önce hem deney hem de kontrol grubundaki öğrencilerin puanlayıcı davranışlarının istatistiksel olarak benzer olduğu bulunmuştur. Ön test olarak her iki gruptaki öğrencilerin seçilen bu üç öğrencinin kağıdını DPA kullanarak değerlendirmeleri istenmiştir. Deney grubuna puanlayıcı eğitimi verildikten sonra aynı kağıtlar her iki gruba da tekrar uygulanmış ve değerlendirmeleri istenmiştir. Ön test ve son test için deney ve kontrol grubundaki öğrencilerin katılık/cömertlik ve farklılaşan puanlayıcı katılığı /cömertliği (yanlılık) davranışları incelenmiştir. Puanlayıcı eğitimi öncesinde hem grup düzeyinde hem de bireysel düzeyde katılık/cömertlik davranışının ortaya çıktığı görülmüştür. Puanlayıcı eğitimi sonrası deney grubunda puanlayıcıların katılık cömertlik davranışlarının yüzde olarak değişmediği bulunmuştur (%12,5). Kontrol grubunda ise son testteki oranın (%18) ön teste (%33) göre azaldığı bulunmuştur.

Anahtar Kelimeler: Açık Uçlu Madde, Akran Değerlendirme, Puanlayıcı Davranışları, Puanlayıcı Eğitimi

Sayfa Adedi: xvi +102

Danışman: Prof. Dr. İsmail KARAKAYA

**INVESTIGATION OF THE EFFECT OF RATER TRAINING GIVEN
TO STUDENTS ON RATER BEHAVIOR IN SCORING OPEN-ENDED
ITEMS USING THE MULTI-FACET RASCH MEASURING MODEL**

(M.S.Thesis)

Burçin Görgülü Öztürk

GAZI UNIVERSITY

GRADUATE SCHOOL OF EDUCATIONAL SCIENCES

December 2023

ABSTRACT

In this study, the effect of rater training given to 10th grade high school students in evaluating their responses to open-ended mathematics items was examined. The students were peer-evaluated and the relevant analyzes were carried out using the multi-facet Rasch measurement model and generalizability theory. The study group of research consists of 65 high school students studying in the tenth grade at a public school in Ankara in the second semester of the 2022-2023 academic year. In the research, a pre-test post-test quasi-experimental design with experimental and control groups was used. There are 32 students in the experimental group and 33 students in the control group. Before the research, an achievement test consisting of 4 open-ended mathematics items was applied to both groups. All answers were evaluated by the researcher with the help of a rubric, and a total of three students' papers were selected, including successful, moderately successful and less

successful answers. The goal here is to ensure that raters see each response style. Before the experimental procedure (rater training) was implemented, it was found that the rater behaviors of the students in both the experimental and control groups were statistically similar. As a pre-test, students in both groups were asked to evaluate the papers of these three selected students using the rubric. After rater training was given to the experimental group, the same papers were shown again to both groups and they asked to evaluate them. For the pre-test and post-test, the strictness/generosity and differing rater strictness/generosity (bias) behaviors of the students in the experimental and control groups were examined. It was observed that strictness and generosity behavior emerged both at the group level and at the individual level before the rater training. After rater training, it was found that the raters' strictness and generosity behaviors did not change as a percentage in the experimental group (%12,5). In the control group, it was found that the rate in the post-test (%18) decreased compared to the pre-test (%33).

Key Words: Open-Ended Item, Peer Evaluation, Rater Behaviors, Rater Training

Page Number: xvi + 102

Supervisor: 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
BÖLÜM I.....	1
GİRİŞ	1
Problem Durumu	1
Araştırmanın Amacı	5
Araştırmanın Önemi.....	6
Problem İfadesi	8
Araştırma Soruları.....	8
Araştırmanın Sınırlılıkları.....	9

Araştırmanın Sayıtları.....	9
BÖLÜM II	10
KURAMSAL TEMELLER ve İLGİLİ ARAŞTIRMALAR.....	10
Kuramsal Temeller	10
Açık Uçlu Madde	10
Akran Değerlendirme	13
Puanlayıcı Davranışları ve Puanlayıcı Eğitimi	15
Puanlayıcı Güvenirliği	19
Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM)	21
Puanlayıcı Eğitimi ve Akran Değerlendirme ile İlgili Araştırmalar	27
BÖLÜM III.....	33
YÖNTEM	33
Araştırma Deseni.....	33
Çalışma Grubu	33
Veri Toplama Araçları.....	34
Açık Uçlu Maddelerin Hazırlanması	34
Dereceli Puanlama Anahtarının Oluşturulması	35
Verilerin Toplanması.....	40
Verilerin Analizi.....	42
BÖLÜM IV	44
BULGULAR VE YORUMLAR	44
Birinci Araştırma Sorusuna İlişkin Bulgular ve Yorumlar	44
İkinci Araştırma Sorusuna İlişkin Bulgular ve Yorumlar	46
Üçüncü Araştırma Sorusuna İlişkin Bulgular ve Yorumlar	60

Puanlayıcı Eğitimi Öncesi Deney ve Kontrol Grubundaki Puanlayıcıların Katılıp/Cömertlik Durumları.....	61
Puanlayıcı Eğitimi Öncesi Deney ve Kontrol Grubundaki Puanlayıcıların Farklılaşan Puanlayıcı Katılıp/Cömertliği (Yanlılık) Durumları	63
Dördüncü Araştırma Sorusuna İlişkin Bulgular ve Yorumlar	68
Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubundaki Puanlayıcıların Katılıp/Cömertlik Durumları.....	69
Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubundaki Puanlayıcıların Farklılaşan Puanlayıcı Katılıp/Cömertliği (Yanlılık) Durumları	71
BÖLÜM V	76
SONUÇ VE ÖNERİLER	76
SONUÇLAR.....	76
Birinci Araştırma Sorusuna Yönelik Sonuçlar.....	76
İkinci Araştırma Sorusuna Yönelik Sonuçlar	77
Üçüncü ve Dördüncü Araştırma Sorularına Yönelik Sonuçlar	78
ÖNERİLER.....	80
KAYNAKLAR	82
EKLER	92
EK 1. Gazi Üniversitesi Etik Komisyonu İzin Belgesi	93
EK 2. MEB Araştırma İzni.....	95
EK 3. Puanlayıcı Eğitimi Planı	96
EK 4. Araştırmada Kullanılan Matematik Soruları.....	97
EK 5. Dereceli Puanlama Anahtarları	99
Yemek Kartı.....	99
Yakıt Zammı.....	100

Kitap Satışı.....	100
Çark Çevirme.....	101



TABLÖLAR LİSTESİ

Tablo 1. <i>Yaygın Puanlayıcı Hatalarının Listesi</i>	16
Tablo 2. <i>Kriter Seçimi için Geçerlik Oranları</i>	38
Tablo 3. <i>Ölçme Aracının Maddelerine Yönelik Hesaplanan KGO Değerleri</i>	39
Tablo 4. <i>Dereceli Puanlama Anahtarı Kriterlerinin KGO Değerleri</i>	39
Tablo 5. <i>Ön Test Sonucunda Deney ve Kontrol Grubuna Ait Varyans Bileşenleri ve Toplam Varyans Yüzdeleri</i>	44
Tablo 6. <i>Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubun Ait Varyans Bileşenleri ve Toplam Varyans Yüzdeleri</i>	46
Tablo 7. <i>Öğrenci Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu</i>	50
Tablo 8. <i>Puanlayıcı Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu</i>	51
Tablo 9. <i>Cinsiyet Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu</i>	53
Tablo 10. <i>Başarı Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu</i>	54
Tablo 11. <i>Madde Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu</i>	55
Tablo 12. <i>Öğrenci Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu</i>	56
Tablo 13. <i>Puanlayıcı Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu</i>	56
Tablo 14. <i>Cinsiyet Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu</i>	58
Tablo 15. <i>Başarı Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu</i>	59
Tablo 16. <i>Madde Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu</i>	60
Tablo 17. <i>Deney ve Kontrol Grubunun Ön Test Puanlamalarının Bağımsız Örneklem t-testi</i>	63

Tablo 18. Deney ve Kontrol Grubuna Uygulanan Ön Testin Puanlayıcı x Öğrenci Etkileşim Analizinde Yanlı Olan Etkileşimleri.....	64
Tablo 19. Deney ve Kontrol Grubuna Uygulanan Ön Testin Puanlayıcı x Madde Etkileşim Tablosu.....	65
Tablo 20. Deney ve Kontrol Grubuna Uygulanan Ön Testin Öğrenci x Başarı Etkileşim Tablosu.....	66
Tablo 21. Deney ve Kontrol Grubuna Uygulanan Ön Testin Puanlayıcı x Öğrenci x Madde Etkileşim Analizinde Yanlı Olan Etkileşimleri.....	67
Tablo 22. Deney ve Kontrol Grubunun Son Test Puanlamalarının Bağımsız Örneklem t-testi (Logit Ölçüleri Farkı Alınmış).....	70
Tablo 23. Deney ve Kontrol Grubuna Uygulanan Son Testin Puanlayıcı x Öğrenci Etkileşim Analizinde Yanlı Olan Etkileşimler.....	71
Tablo 24. Deney ve Kontrol Grubuna Uygulanan Son Testin Puanlayıcı x Madde Etkileşim Tablosu.....	72
Tablo 25. Deney ve Kontrol Grubuna Uygulanan Son Testin Öğrenci x Başarı Etkileşim Tablosu.....	73
Tablo 26. Deney ve Kontrol Grubuna Uygulanan Son Testin Puanlayıcı x Öğrenci x Madde Etkileşim Analizinde Yanlı Olan Etkileşimleri.....	74

ŞEKİLLER LİSTESİ

<i>Şekil 1.</i> Ön teste ait değişken haritası.....	47
<i>Şekil 2.</i> Son teste ait değişken haritası.....	49
<i>Şekil 3.</i> Deney ve kontrol grubunun ön test t-değerleri.....	62
<i>Şekil 4.</i> Deney ve kontrol grubunun son test t-değerleri.....	69

BÖLÜM I

GİRİŞ

Bu bölümde araştırma ile ilgili problem durumuna, araştırmanın amacına, önemine, sınırlılıklarına yer verilmiştir. Araştırma ile ilgili olarak açık uçlu soru kavramı ve akran değerlendirme yöntemi açıklanmıştır.

Problem Durumu

Öğrenme sürecinde öğrencilerin hangi düzeyde ne kadar bilgi öğrendiklerinden çok bu bilgiyi gerçek yaşama aktarabilmeleri önem kazanmıştır. Eleştirel ve yaratıcı düşünen, problem çözerken değişik bakış açıları kullanan, özgün fikirler bulan, diğer kişilerin görüşlerine saygı duyan bireyler yetiştirmek öğretme sürecinin amacı olmuştur. 21.yy becerileri olarak adlandırılan bu kazanımlar son yıllarda eğitim öğretim süreci için daha da önemli hale gelmiştir. Önceden alt düzey zihinsel becerilerin ölçülmesi daha yaygınken artık üst düzey becerilerin ölçülmesi önem kazanmıştır. İnsanların, çalışanların iş hayatında karşılaştıkları problemleri çözmek, özgün fikirler bulmak için bu becerilere sahip olmaları beklenmektedir. Bu sebeple eğitim kurumlarında bu becerilerin kazandırılması amaçlanmıştır. Üst düzey zihinsel beceriler Bloom taksonomisine göre analiz, sentez ve değerlendirme basamağına karşılık gelmektedir (Brookhart, 2010). Diğer bir bilişsel sınıflama Haladyna'nın sınıflamasıdır. Haladyna, sınıflamasında problem çözme, eleştirel

düşünme ve yaratıcılık basamaklarını üst düzey zihinsel beceriler olarak belirlemiştir. İşte bu üst düzey zihinsel becerileri ölçme konusunda sonuç odaklı olan geleneksel ölçme yöntemleri (çoktan seçmeli, doğru yanlış, kısa cevaplı vb.) yetersiz kalmaktadır. Çünkü geleneksel yöntemle yapılan ölçmelerde öğrencinin soruyu çözme süreci değerlendirmeye alınmaz. Sonuca odaklanılır. Öğrencinin bir becerinin ne kadarını öğrendiği, eksik yönlerinin bulunarak düzeltme yapılması gibi bir durum söz konusu değildir. Bu nedenle ölçme ve değerlendirme sürecinde yeni yaklaşımların kullanılması ihtiyacı doğmuştur. Böylece açık uçlu madde formatı ve performans değerlendirme yaklaşımı önem kazanmıştır. Açık uçlu madde formatında öğrenci cevabı kendisi inşa eder. Düşüncelerini yazıya döker. Eleştirel düşünme, düşünceleri özgürce ifade edebilme, hipotez kurma, problem çözme, farklı çözüm yolları görebilme, neden-sonuç ilişkisi kurabilme, yargıya varabilme gibi üst düzey beceriler için açık uçlu maddeler kullanılabilir (Tan ve Erdoğan, 2004). Son yıllarda açık uçlu maddelerin kullanımı yaygınlaşmıştır. OECD tarafından uygulanan Uluslararası Öğrenci Değerlendirme Programı (The Programme for International Student Assessment- PISA) sınavında açık uçlu maddelere yer verilmektedir. Üst düzey becerilerin ölçüldüğü bu sınavda eğitim programında ancak bu becerilere yer verebilen ülkeler başarılı olmaktadır. Yine bir başka uygulama olan Uluslararası Matematik ve Fen Eğilimleri araştırması (Trends in International Mathematics and Science Study- TIMSS) öğrencilerin fen ve matematik performanslarını ölçmek amacıyla yapılır. Uluslararası sınavlarda açık uçlu maddelerin kullanımı oldukça yaygındır.

Ülkemizde 8. ve 12. sınıfın sonunda uygulanan LGS, YKS gibi geniş ölçekli sınavlarda çoktan seçmeli madde formatı kullanılmaktadır. Öğrencilerin bu sınavlardan başarılı olabilmesi için öğretmenler ve aileler tarafından öğrencilere çoktan seçmeli madde pratikleri yaptırılmaktadır. Ancak bu madde türünde öğrencinin bilgiyi ne kadar öğrendiğinin düzeyi belli olmamaktadır. Öğrenci seçenekleri eleyerek doğru cevaba gidebilmektedir. Bu durum ölçmenin güvenilirliğini önemli ölçüde düşürmektedir. Meslek hayatlarında bu bireylere problem çözebilme, eleştirel düşünebilme, örgütsel çalışabilme gibi üst düzey kazanımlar

lazım olacaktır. Bu kazanımlar için eğitim sürecinde açık uçlu maddelere yer verilmesi önemlidir. Açık uçlu maddeler öğrenmenin derinlemesine gerçekleşmesine yardımcı olur. Öğrenci, alt düzey becerilerde olduğu gibi bilgiyi ezberlemez; yorum yapmayı, farklı bakış açıları ile düşünmeyi, fikirleri bir araya getirmeyi öğrenir. Kısaca bu kazanımlar kişiye hayatı boyunca yarayacak kazanımlardır.

Açık uçlu maddelerin yukarıda sayılan bazı avantajlarının yanı sıra dezavantajları da vardır. Bunlardan en önemlisi puanlayıcı hatasından kaynaklanan ölçme hatalarıdır. Eşleştirmeli, doğru yanlış, çoktan seçmeli maddelerde puanlayıcı hataları oldukça düşük olmasına rağmen açık uçlu soruları değerlendirirken puanlama sürecine hata miktarı daha fazla karışabilir. Geçerli ve güvenilir bir ölçme yapılmak isteniyorsa yazılan açık uçlu maddelerin niteliği yeterli olmayabilir. Puanlamayı yapan puanlayıcının objektif, yansız ve alanında uzman olması gerekmektedir. Bu ölçmenin doğru yapılabilmesi için puanlama anahtarının önemi de oldukça büyüktür. Geçerli ve güvenilir, uzman görüşü alınmış puanlama anahtarları kullanılmalıdır (Karakaya ve Şata, 2022). Ayrıca bu maddeleri değerlendirirken tek bir kişinin, genelde öğretmenin, değerlendirmesi de ölçmeye daha fazla hata karışmasına yol açabilir. Öğretmen gereğinden fazla yumuşak ya da gereğinden fazla katı olabilir. Tek bir puanlayıcı ile ölçme yapmak yerine akran değerlendirme sürece katılarak alternatif bir yol olarak kullanılabilir. Akran değerlendirme ile öğretmenin iş yükü azalabilir. Öğrenen odaklı, öğrenciyi değerlendirme sürecine dahil eden bir süreç gerçekleşir.

Akran değerlendirmede öğrenciler kendi akranlarını değerlendirirler. Değerlendirme sürecine aktif katılırlar. Tabi bu, sadece bir değerlendirme olarak sınırlanmamalıdır. Biçimlendirici değerlendirmenin içinde yer alan akran değerlendirmede öğrenciler bu süreçte eleştirel düşünme becerilerini geliştirirler. Değerlendirme sorumluluğunu alırlar (Falchikov ve Goldfinch, 1995). Topping (1998) akran değerlendirmeyi “bireylerin, benzer statüdeki akranlarının öğrenme işi ile ilgili ürünlerine başarı, nitelik, değer, düzey açısından eleştirel gözle bakmalarını sağlayan bir planlama işi” olarak belirtmiştir. Temizkan (2009) ise akran değerlendirme için hem geri bildirim vermek amacıyla gözleme dayalı

biçimlendirici değerlendirmeye uyduğunu hem de başarıyı belirleme açısından yerleştirmeye dayalı özetleyici değerlendirmeye uyduğunu söylemiştir. Açık uçlu sorularda akran değerlendirme kullanılması öğrenciye farklı çözüm yollarını sunar. Kendi çözüm yöntemine alternatif çözümler görmesi konuyu daha iyi kavramasına yardımcı olur.

Açık uçlu maddeleri değerlendirirken puanlayıcıların ne kadar objektif olabileceği tartışma konusudur. Bir öğrencinin puanı puanlayıcıdan puanlayıcıya değişebilir. Puanlayıcı davranışı (cömertlik, katılık, yazı güzelliğine bakma, düzene göre puan verme, vb.) açık uçlu maddelerde ölçümün geçerlik ve güvenilirliğini etkiler. Bu nedenle bu davranışların kontrol altında tutulması ve puanlayıcı güvenirliliğinin sağlanması gerekmektedir. Bunun için alan yazında DPA'nın özenli bir şekilde hazırlanması ve puanlayıcılara DPA'yı nasıl kullanacakları ile ilgili puanlayıcı eğitimi verilmesi önerilmiştir (Haladyna, 1997; İlhan ve Çetin, 2014). Etkili bir değerlendirme yapılabilmesi için gerekli eğitimin uygulamalı olarak verilmesi puanlayıcıların objektif olmasını sağlar (Uysal, 2008). Açık uçlu maddelerin kullanılması öğrencilerin yanıtlarının değerlendirilmesinde nesnellik sorununu ortaya koymaktadır. Birden fazla puanlayıcı olduğunda puanlama davranışlarında ortaya çıkan farklılık ölçümün geçerlik ve güvenilirliğini etkileyecektir (Eckes, 2008). Bu nedenle bu davranışların olabildiğince kabul edilebilir ölçüte getirilmesi gerekir (Ellis vd., 2002). Puanlayıcı eğitimleri bu ölçümlere karışan puanlayıcı hatalarını azaltmak için kullanılmaktadır. Bu çalışmada alan yazında en sık karşılaşılan davranışlardan katılık, cömertlik ve farklılaşan puanlayıcı katılığı/cömertliği davranışları incelenmiştir.

Puanlayıcı güvenirliliğini hesaplamak için klasik test kuramına dayalı istatistiksel yöntemler, madde tepki kuramına dayalı çok yüzeyli rasch ölçme modeli ve genellenebilirlik kuramının kullanıldığı çalışmalar literatürde mevcuttur. Klasik test kuramında Pearson Momentler Çarpımı, Spearman- Brown formülü, Uzlaşma Katsayısı, Uyum Oranı gibi yöntemlerden yararlanılır. Eğer tüm puanlayıcılar bir öğrenciye aynı puanı verirse istenen uyuma ulaşılabilir. Ancak puanlayıcılar farklı değerlendirmeler yapabilirler. Elde edilen sonuçlar tam ideal sonuçları göstermeyebilir. Klasik yöntemin bu eksik tarafını örtmek için

Genellenebilirlik Kuramı (GK) ve Çok Yüzeyle Rasch Ölçme Modeli (ÇYRÖM) kullanılır. Genellenebilirlik kuramında puanlayıcı, madde, zaman gibi hatalar göz önünde bulundurulur. Hata varyansları dikkate alınır. Karar çalışmaları yapılarak daha yüksek güvenilirlik elde edilebilir (Shavelson ve Webb, 1991). Madde tepki kuramının bir modeli olan Rasch modelinde genellenebilirlik kuramından daha yüksek güvenilirlik elde edilebilir. Çünkü bu model hata varyanslarını madde ve puanlayıcı varyanslarına dahil etmez (Linacre, 1993). Çok yüzeyle rasch ölçme modeli bireylerin gözlenen puanlarının logaritmik olarak bir lojit cetveline dönüştürülmesine olanak veren bir modeldir. (Nakamura, 2002). Facet denilen her bir yüzeyle için yüzeyleler arası doğrusal bağıntılar kurar. ÇYRÖM’de öğrenciler de bir değişkenlik kaynağı olarak ele alınır. Bir öğrencinin akran değerlendirmesini ne kadar etkili yapabildiğini gösterebilir. Genellenebilirlik Kuramında bir güvenilirlik katsayısı kestirilir. ÇYRÖM’de ise her bir değişkenlik kaynağı için ayrı ayrı güvenilirlik katsayısı kestirilir (Allharby, 2006).

Bu çalışmada öğrencilerin açık uçlu matematik maddelerini puanlarken ortaya çıkan puanlayıcı davranışları ÇYRÖM ile incelenmiştir.

Araştırmanın Amacı

Bu araştırmada açık uçlu soruların öğrenciler tarafından etkili ve doğru puanlanabilmesi için verilen puanlayıcı eğitiminin öncesinde ve sonrasında puanlayıcı davranışlarının incelenmesi amaçlanmıştır. Çalışmada elde edilen veriler FACETS paket programı kullanılarak Çok Yüzeyle Rasch Ölçme Modeli’ne göre analiz edilmiştir. Ayrıca Genellenebilirlik Kuramı için EduG programı kullanılarak ÇYRÖM ile elde edilen veriler desteklenmiştir. Bu analizde puanlayıcı katılığı, cömertliği ve farklılaşan puanlayıcı davranışları (yanlılık) incelenmiştir.

Araştırmanın Önemi

Geniş ölçekli sınavlarda öğrenciler çoktan seçmeli maddelerle karşılaşmaktadır. Bazen seçenekleri eleyerek doğru cevaba ulaşabilmeleri, hatta tesadüfi olarak seçeneklerden doğruyu seçebilmeleri öğrencileri çaba göstermeden bir işi yapabileceklerine inandırmaktadır. Bu durum öğrencinin yaşamına da yansımaktadır. Artık insanlar kısa yoldan başarıya ulaşabilme, çaba göstermeden yüksek hedeflere erişebilmeye çalışmaktadırlar. Halbuki kişilerin çok yönlü düşünmeyi, analiz etmeyi, yansıtıcı düşünmeyi öğrenmesi gerekir. Bunun yeri de okullardır. Okulda öğrenilmesi gereken bu üst düzey bilişsel kazanımlar kişilerin okuldan sonraki hayatlarına da yansıyacaktır. İş hayatında örgütsel çalışabilmek, takım olabilmek, üretken olabilmek okullarda kazanılan bu kazanımlarla mümkündür.

Son yıllarda MEB ve ÖSYM tarafından daha geçerli ve güvenilir ölçmenin yapılabileceği bir sınav oluşturmak için bazı değişiklikler yapılmıştır. Üst düzey bilişsel kazanımların ölçülebildiği Türkiye'nin de katıldığı PISA sınavı gibi bir sınav hazırlanabileceği düşünülmüştür ve bu amaçla ABİDE (Akademik Becerilerin İncelenmesi ve Değerlendirilmesi) projesi planlanmıştır. Projedeki amaç, çoktan seçmeli sorular öğrencilerin kendilerini ifade etmelerine, yanıtları özgün bir şekilde oluşturmasına imkan vermediği için, problem çözme, eleştirel düşünme gibi üst düzey zihinsel özellikleri ölçmek için açık uçlu maddeler kullanmaktır. Böylece öğrencilerin üst düzey zihinsel becerilere sahip olma derecesi belirlenmek istenmektedir. Bu projenin pilot uygulaması yapılmış ve açık uçlu maddelerin ulusal sınavlarda kullanılması öngörülmüştür (MEB, 2018). Bu sınavlarda kısa yanıtlı, yanıtı sınırlandırılmış ve yanıtı sınırlandırılmamış madde türü kullanmıştır. Kısa yanıtlı maddeler, yanıtı bir sözcük ya da bir cümle olan maddelerdir. Yanıtı sınırlandırılmış maddeler, maddenin yanıtı birden fazla olduğu ancak sonsuz sayıda olmadığı maddelerdir. Sınırlı sayıda doğru yanıt vardır. Yanıtı sınırlandırılmamış maddeler ise kompozisyon tipi maddelerdir. Bireylere göre farklılık gösteren yanıtlara sahiptir (Karakaya ve Şata, 2022).

Öğrencilerin bu şekilde bir sınava girmeleri aslında tüm okul yaşantısını da değiştirecektir. Okullarda açık uçlu maddelerin kullanımı artacak ve üst düzey zihinsel beceriler önemli hale gelecektir.

Açık uçlu maddelerin avantajlarının yanında sınırlılıkları da mevcuttur. En önemli sınırlılığı puanlayıcı hatalarının ölçme sonuçlarına daha çok karışabilmesidir. Ölçümlerin geçerli ve güvenilir olması için puanlayıcı davranışlarının belirlenmesi ve hata kaynaklarının olabildiğince engellenmesi gerekmektedir. Puanlamayı yapan kişilerin alanlarında uzman bireyler olması önemlidir. Puanlayıcı eğitimi almış olmaları gerekir. Tabi ki burada açık uçlu maddelerin niteliği ve bu maddelerin puanlanması için geliştirilen DPA da büyük önem taşımaktadır. Açık uçlu maddeleri değerlendirirken ölçmeye karışan puanlayıcıdan kaynaklanan hatalar genellikle sistematik hatalardır. Sabit ve tesadüfi hata azdır. İlgili literatür incelendiğinde 30'dan fazla açıklanmış hata türünün olduğu görülmüştür. (Haladyna, 1997; Kutlu vd., 2014; Royal ve Hecker, 2016). Puanlayıcıdan kaynaklı bu hatalar bireylere verilen eğitimlerle azaltılabilir. Mevcut çalışmada öğrencilere puanlayıcı eğitimi verilip akran değerlendirme yapmaları sağlanmıştır. Puanlayıcı eğitimi sırasında öğrenciler birçok açık uçlu soru ile karşılaşmış ve ortak bir paydada buluşana dek puanlama yapmayı öğrenmişlerdir. Öğrenciye arkadaşlarının yanıtlarına objektif ve eleştirel gözle bakma şansı sunulmuştur. Değerlendirme sürecinde öğrenci farklı yanıtlarla öğrenme açığını kapatma ve yeni çözüm yolları görme şansına sahip olmuştur. Puanlama yaparken ne kadar titiz olunması gerektiği, katılık cömertlik davranışının aslında ne kadar da ölçmenin güvenilirliğini etkilediğini fark etmişlerdir. Açık uçlu soruları yanıtlamak ve değerlendirme sürecinde yer almak öğrencilerin yukarıda bahsedilen üst düzey bilişsel kazanımları edinmelerini için önemlidir. ÖSYM sınavlarına hazırlanan ortaöğretim öğrencilerinin üst düzey zihinsel becerileri kazanmalarında akranlarının yanıtlarını puanlamaları ve puanlayıcı davranışları Çok Yüzeysel Rasch Ölçme Modeli ile incelenmiştir ve çalışmanın literatüre katkı sağlayacağı düşünülmektedir.

Problem İfadesi

Açık uçlu matematik maddelerinin puanlanmasında puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisi nasıldır?

Araştırma Soruları

Yukarıdaki amaç ve problem ifadesi doğrultusunda aşağıdaki sorulara yanıt aranmıştır.

1. Yüzeylere ait kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri nasıldır?
 - a) Ön testte yüzeylere ait kestirilen varyans bileşenleri ve yüzeylerin toplam varyansı açıklama yüzdeleri nasıldır?
 - b) Puanlayıcı eğitimi sonrası yapılan son testte yüzeylere ait kestirilen varyans bileşenleri ve yüzeylerin toplam varyansı açıklama yüzdeleri nasıldır?
2. Öğrenci, puanlayıcı, cinsiyet, başarı ve madde yüzeylerinin ön test ve son test ölçüm raporlarının değişken haritası, uygunluk değerleri ve model ölçümleri nasıldır?
3. Puanlayıcı eğitimi öncesinde deney ve kontrol grubundaki puanlayıcıların puanlamalarındaki;
 - a. Katılık ve cömertlik durumları nasıldır?
 - b. Farklılaşan puanlayıcı katılığı ve cömertliği yani yanlılık durumları nasıldır?
4. Puanlayıcı eğitimi sonrasında deney ve kontrol grubundaki puanlayıcıların puanlamalarındaki;
 - a. Katılık ve cömertlik durumlarındaki değişim anlamlı mıdır?
 - b. Farklılaşan puanlayıcı davranışındaki (yanlılık) değişim anlamlı mıdır?

Araştırmanın Sınırlılıkları

Bu çalışmanın sınırlılıkları şu şekildedir.

- 1) Lise 10.sınıfa devam eden öğrencilerden (puanlayıcılar) deney ve kontrol grupları seçkisiz olarak seçilmemiştir.
- 2) Puanlayıcı davranışları, alan yazında daha çok ortaya çıkan puanlayıcı katılığı, cömertliği ve farklılaşan puanlayıcı katılığı/cömertliği (yanlılık) ile sınırlıdır.
- 3) Araştırmadaki 65 puanlayıcı 3 öğrencinin açık uçlu 4 matematik sorusunu yanıtlamıştır. Yukarıdaki veriler ile Rasch analizi yapılmıştır.
- 4) Araştırma öncesi öğrencilerin açık uçlu maddeler hakkında ön bilgileri çok sınırlıdır.

Araştırmanın Sayıtları

Araştırmaya katılan puanlayıcılar, öğrencilerin açık uçlu matematik maddelerine verdikleri cevapları ciddiyle puanlamışlardır.

BÖLÜM II

KURAMSAL TEMELLER VE İLGİLİ ARAŞTIRMALAR

Kuramsal Temeller

Açık Uçlu Madde

Açık uçlu madde, öğrencilerin üst düzey düşünme, akıl yürütme ve bilgiyi bu bağlamda uygulama becerilerine odaklanan madde türüdür. Üst düzey düşünme insan yaşamının bir parçasıdır ve bu düşünmenin gelişebilmesi için açık uçlu maddeler kullanılabilir (Brookhart, 2014, s.54).

Son yıllarda öğrenme ve öğretmeye olan bakış açısının değişmesi araştırmacıların bilginin karmaşıklığını kanıtlamasıyla olmuştur (Badger ve Thomas, 1991). Yöneticiler, iş verenler, içeriği en iyi bilen insanların yanı sıra bilgiyi yönetebilen, sorunları farklı bakış açılarıyla çözebilen insanlarla çalışmanın önemini anlamışlardır. Bilginin tanım olarak öğrenilmesinden çok yorumlanması önem kazanmıştır. Bilgiyi ezberlemek yerine bilgiyi içselleştirip farklı alanlara transfer edebilmek, bilgiye ulaşabilmek için çok yönlü düşünmek gerekli bir beceri haline gelmiştir. Bu becerilerin kazanılması için açık uçlu madde formatı kullanılabilir. Açık uçlu maddeler bir öğrencinin fikirlerini organize etme, mantıklı bir argüman geliştirme, belirli konumların veya verilerin değerlendirmelerini tartışma, düşünce ve duyguları iletme veya orijinal düşünmeyi sergileme becerisiyle ilgili öğrenme hedeflerini değerlendirmeye uygundur (Brookhart ve Nitko, 2008, s.155). Açık uçlu maddeler sınıf içi

kullanımlarda etkilidir. Çoktan seçmeli maddelerin aksine öğrencinin kendi yanıtını oluşturmasını sağlayan bu görevler günlük yaşamdaki zorluk ve problemlere olan bakış açısını da çeşitlendirir. Bir öğrencinin performansını ölçebilmek için öğretmenler sadece sonucu değil süreci de dikkate almalıdırlar. Çoktan seçmeli maddelerde çözüm göz ardı edilip sadece sonuca odaklanılır. Öğrencinin bilgiyi ne ölçüde öğrendiği görülmez. Öğrenci doğru yanıtı ulaşabilmek için seçenekleri eleyerek gidebilir. Şans faktörü ile doğru yanıtı bulabilir. Ancak açık uçlu madde formatında öğrenci bildiğini yazıya döküp göstermek için üst düzey düşünme becerisini kullanır. Öğrenciler açık uçlu maddeleri yanıtlarken çoktan seçmeli maddelere göre başarı düzeylerini daha net gösterirler. (Badger ve Thomas, 1991). Açık uçlu maddeler karmaşık ve derin düşünmeyi gerektirir. Öğrencileri ezberden kurtaran açık uçlu maddeler, üst düzey düşünme, eleştirel gözle görebilme, çok yönlü düşünebilme becerisine sahip olmayı gerektirir. Öğrenciyi düşünmeye, muhakeme etmeye yönlendirir. Öğrencilere yanıt özgürlüğü sağlar. Tahmin yolu ile soruyu çözmeyi en aza indirir. (Reiner vd., 2002). Öğrenciler bir açık uçlu soruyu yanıtlarken şu aşamalardan geçerler. İçerik bilgisi, içerik becerisi ve düşünme becerisi (Brookhart, 2014, s.14). Yani öğrenci soruyu okuduğunda konu ile ilgili bilgisini gözden geçirir. Daha sonra uygulamasını yapar ve en son neden bu şekilde çözdüğünü/cevapladığını açıklar.

Açık uçlu maddeler, yanıtı sınırlandırılmış ve yanıtı sınırlandırılmamış maddeler olarak ikiye ayrılır (Berberoğlu, 2006, Reiner vd., 2002). Yanıtı sınırlandırılmış açık uçlu madde çeşidinde maddenin birden çok çözümü vardır. Bu çözüm sonsuz sayıda olmasa da çok çeşitlidir. Bu tür maddelere yorumlama da katılırsa madde geliştirilmiş olur. Sınıf içi başarı testi uygulamaları için kullanışlıdır (Brookhart ve Nitko, 2008; Karakaya, 2002). Yanıtı sınırlandırılmamış açık uçlu maddeler ise her bireye özgü cevabı olabilen maddelerdir. Yanıtı sınırsızdır. Öğrenciler yanıtlarında özgürdür. Daha çok kompozisyon tipi sınavlarda kullanımı yaygındır.

Açık uçlu maddelerin özgür düşünmeye teşvik etmesi, yeni çözüm yolları üretme becerisini geliştirmesi, şans başarısını düşürmesi gibi birçok faydasının olmasına karşın bazı

dezavantajları da vardır. Bu tür maddelerin içerik yelpazesi daha dardır. Maddelerin yanıtlanma süreleri çoktan seçmeli sorulardan daha fazla olduğu için aynı süre içerisinde öğrenciye daha az açık uçlu madde sunulabilir. En önemli dezavantajlarından birisi bu maddelerin değerlendirilmesindeki puanlayıcı güvenilirliğidir. Bu maddeleri tarafsızca puanlamak zordur. Ayrıca puanlama yapmak da zaman alıcıdır. Çoktan seçmeli, doğru yanlış gibi testlerde puanlayıcı kaynaklı hatalar neredeyse yok denecek kadar azdır. Hatta büyük ölçekli sınavlarda optik formların makineler tarafından okunması puanlama hatasını neredeyse sıfıra indirir. Ancak açık uçlu maddeleri puanlarken puanlayıcının alanında uzman olması, puanlayıcı eğitimi almış olması geçerli ve güvenilir bir ölçme için gereklidir. Bu maddeleri puanlarken puanlayıcıya rehber olması açısından DPA oluşturulmalıdır. Puanlama anahtarının oluşturulması puanlamanın güvenilirliğini ve geçerliğini artıracaktır (Brookhart ve Nitko, 2008). Aynı zamanda puanlama anahtarı zamanın etkin kullanılmasını sağlar. DPA oluşturulurken zayıf, orta ve iyi düzeydeki öğrencilerin maddeye verebilecekleri tüm yanıtlar öngörülüp hazırlanır. Bu da alanında uzmanlık gerektirir. Bu anahtarlar ne kadar kapsamlı oluşturulursa puanlayıcı o kadar objektif olur.

Dereceli puanlama anahtarlarını oluştururken şu adımlara dikkat edilmelidir (Berberoğlu, 2006).

- Maddenin kaç kategoride inceleneceğine karar verilmesi,
- Maddenin tam ve doğru cevabında hangi açıklamaların yer alması gerektiğinin belirlenmesi,
- Maddede yakın doğru yanıtlar, kısmen doğru yanıtlar söz konusu ise bu yanıtların tam doğru yanıttan ne gibi eksiklerinin olduğunun net bir şekilde açıklanması, gerekirse bu durumların puanlama anahtarında örneklerle açıklanması,
- Tam doğru, yakın doğru ve kısmen doğru yanıtların puanlarının belirlenmesi,
- Eğer puanlama anahtarı oluşturulurken gözden kaçan yanıtlar olabileceği düşünülüyorsa öğrenci cevaplarına bakılarak puanlama anahtarının düzenlenmesi.

Puanlama anahtarının yukarıdaki adımlara göre oluşturulduğunda ölçmeye karışan hatalar aza indirilir. Ayrıca açık uçlu sorular hazırlık aşamasındaki emek ve puanlama durumlarından dolayı çoktan seçmeli sorulara göre daha maliyetlidir (Lane vd., 2016).

Akran Değerlendirme

Akran değerlendirme biçimlendirici (formative) değerlendirme sürecinde yer alır. Biçimlendirici değerlendirme, öğrenciye geri bildirimler vererek yanlış veya eksik öğrenilen kazanımlar için düzeltmeler yaparak öğrencilerin gelişiminin her aşamada değerlendirilmesi olarak tanımlanmıştır (McMillan, 2010). Biçimlendirici değerlendirme öğretmenlere, öğrencilerin kimlikleri hakkında bilgi verir. Aktif öğrenmeyi sağlar, öğrenci bu sayede öğrenme hedeflerine odaklanır (Brookhart ve Nitko, 2008, s.118). Akran değerlendirme öğrencilerin akranlarının çalışmaları/çözümleri hakkında değerlendirme yapmalarıdır (Topping, 2010). Akran değerlendirme yapılırken değerlendiren öğrenci kendini öğretmen yerine koyar. Diğer öğrencilerin yanıtlarından eksik bilgilerini öğrenebilir. Böylece değerlendiren ve değerlendirilen olarak her iki taraf için de kazanç sağlanmış olur (Topping, 1998). Literatüre bakıldığında öğrenciler akran değerlendirmesi yapmanın öğrenmeyi kolaylaştırdığını ifade etmişlerdir (Ballantyne, Hughes ve Mylonas, 2002). Öğrenciler kendi arkadaşlarından geri bildirim aldıklarında bu bildirimleri daha çabuk kabullenebilirler, daha samimi bulabilirler. Akran değerlendirmede öğrenciler sosyal ve iletişim becerilerini, müzakere ve ekip çalışması becerilerini geliştirebilirler (Topping, 2010).

Akran değerlendirmenin olumlu görüşlerinin yanı sıra Fachikov (1995), kaliteli akran değerlendirmesi oluşturmanın, organizasyon, eğitim ve izleme için zaman gerektirmesinden dolayı kısa ve orta vadede tasarruflu olmayabileceğini söylemiştir. Ancak sistemli ve düzenli yapıldığında, üstbilişsel faydalar olması muhtemeldir. Ayrıca alan yazında sosyal süreçlerin akran değerlendirmesinin güvenilirlik ve geçerliğini olumsuz yönde etkileyebileceğinden bahsedilmektedir (Topping, 2010). Öğrenciler sürecin başında akranlarını

değerlendirdiklerinde aralarındaki ilişkinin bozulacağından şüphe edebilirler. Ancak öğretmen tarafından olumlu geribildirim vermek, değerlendirilen kişinin kaygısını azaltacak ve olumsuz geri bildirim kabulünü gerçekleştirecektir. Sonuç olarak akran değerlendirmesi olumsuz bazı görüşlerin olmasına rağmen akranlar arasındaki ilişkiyi, iş birliğini, bağ kurmayı, empati yapmayı, kabul etmeyi, eleştiri yapmayı geliştiren bir değerlendirme türüdür.

Akran değerlendirme ile ilgili literatüre bakıldığında bu tür değerlendirmelerin öğrencilerin başarılarını, motivasyonlarını, üst biliş becerilerini, öğrenme düzeylerini artırdığı görülmektedir (Brown ve Harris, 2013; Black ve William, 1998b; Sadler, 1989; Topping, 2010). Akran değerlendirme ile öğrenciler kendi performanslarını ve akranlarının performanslarını değerlendirip öğretmenlerine yardımcı olurlar. Sadece öğretmenin değerlendirmesinin önüne geçilmiş olur. Performans/madde için ilgili DPA'nın güçlü bir şekilde oluşturulması gerekir. Böylece bu DPA'ya göre bir soruya öğrencinin hem kendisinin hem akranının tutarlı puanlar vermesi, değerlendirmeler yapması puanlayıcı güvenilirliğini de artıracaktır. Öğrencilerin kendilerine olan güvenleri, sorumluluk alma bilinçleri gelişecektir.

Bu değerlendirmenin bazı sınırlılıkları da vardır. Daha önce hiç bu şekilde değerlendirilmeye tabi tutulmayan, akran değerlendirme ile ilgili bilgi sahibi olmayan öğrenciler için öncelikle bu terim anlatılmalıdır. Bu değerlendirmenin öğrencilere ne gibi faydalar sağlayacağı öğretilmelidir. Bu zamana kadar sadece öğretmenin değerlendirici konumda olduğu, sadece öğretmenden geri bildirim alan öğrenciler için bu bilgilendirme süreci oldukça önemlidir (McMillan, 2014). Öğrencilere bir öğretmenin soruları nasıl değerlendirdiği, değerlendirirken hangi süreçleri gözettiği adım adım anlatılmalıdır.

Bu çalışmada da öğrencilere akranlarının çözümlerini nasıl değerlendirecekleri, değerlendirme aşamasında hangi puanlayıcı hatalarının sürece karışabileceği ve bu hataların önüne geçmek için nelere dikkat edilmesi gerektiği anlatılmıştır. Araştırmacı tarafından

öğrencilere soruları puanlarken puanlamaya karışabilecek hatalar hakkında puanlayıcı eğitimi verilmiştir.

Puanlayıcı Davranışları ve Puanlayıcı Eğitimi

Bir performansı değerlendiren puanlayıcılar ister istemez bazı subjektif davranışlar gösterebilirler. Değerlendiriciler arasında farklılık olması bir ölçme işleminde beklenen bir durumdur (Royal ve Hecker, 2016). Bazı puanlayıcılar diğerlerine göre daha katı ya da daha cömert olabilirler ya da zaman içinde bir önceki değerlendirmeleriyle tutarsız değerlendirmeler yapabilirler. Bu tarzdaki davranışların ölçümlerin geçerliği ve güvenirliği üzerinde önemli etkileri olabilir. Linacre (1993), bir kişinin performansını değerlendirmede puanlayıcı davranışlarının kontrol edilmesinin zor olduğunu, bu davranışları tamamen ortadan kaldırmanın mümkün olmayacağını söylemiştir. Puanlayıcılardan kaynaklanan tutarsızlıklar, değerlendirilen kişinin notu üzerinde önemli bir değişiklik getirecektir. Bu nedenle geçerli bir değerlendirme süreci için puanlayıcı hatası kaynaklarının kontrol edilmesine titizlikle uğraşılmalıdır.

Alan yazında puanlayıcı performanslarını inceleyen araştırmacılar “değerlendirici etkisi”, “puanlayıcı etkisi”, “puanlayıcı yanlılığı”, “değerlendirici hataları” gibi ifadeler kullanmışlardır (Myford ve Hecker, 2003). Puanlayıcılar kişilerin performanslarını değerlendirirken bazı davranışları gösterme eğiliminde olabilirler. Royal ve Hecker (2016) bu davranışları bir tablo halinde aşağıdaki gibi sunmuşlardır.

Tablo 1

Yaygın Puanlayıcı Hatalarının Listesi

Hata türü	Puanlayıcının eğilimi
Asimilasyon etkisi	“Aşırı” görünmekten kaçınmak için diğer puanlayıcılara benzer puanlar verme
Aktarma etkisi	Önceki öğrencinin performansının bir sonraki öğrencinin puanını etkilemesine olumlu ya da olumsuz yönde izin verme
Merkeze eğilim	Derecelendirme ölçeğinin orta kategorisini aşırı kullanma
Amigo etkisi	Öğrencilere destekçi görünmek için performansı ne olursa olsun herkese yüksek puan verme
Çatışan standartlar	Puanlayıcının kişisel derecelendirme standartları, ölçüm aracındaki standartlarla çeliştiği zaman daha düşük puan verme
Çakışan değerler	Performansta ifade edilen değerler puanlayıcınıninkilerle çeliştiği için daha düşük puan verme
Korkaklık	Öğrencinin sınıfa meydan okuması korkusuyla daha yüksek puan verme
Sürüklenme	Zamanla daha sert ya da daha yumuşak puanlamalar yapma
Aşırılık	Derecelendirme ölçeğinin yalnızca uç noktalarını kullanarak puan vermem
Tükenmişlik	Puanların yorgunluğa bağlı olarak etkilenmesine izin verme
Taraf tutma	Sevilen öğrencinin bazı kusurlarını gözden kaçırma
İlk izlenim etkisi	Sonraki bilgileri göz ardı ederek ilk izlenime göre karar verme
Halo etkisi	Öğrenciye yüksek puan verileceği şüphesi uyandırma
Gestalt olgusu	Puanlayıcının öğrenciyi ayırt edemediğinde genel izlenime dayanarak puan vermesi
Bağılılık etkisi	Öğrencinin başka biriyle ilişkisine dayalı puanlama yapma
Güvensizlik	Puanlayıcının kendini donanımlı hissetmemesi durumunda normalden daha yüksek puan vermesi
Uzunluk	Uzun yazılmış performanslara daha yüksek puan verme
Cömertlik	Sürekli olarak şişirilmiş puanlar verme
Kişilik çatışması	Puanlayıcının bir bireyin bazı özelliklerini beğenmediği için puanını düşürmesi
Puanlayıcı yetkinliği	Puanlayıcının puan veremediği durumlarda “hayır”, “gözlenmiyor”, “karar veremiyor” gibi derecelendirmenin sağlanması
Tekrar faktörü	Puanlayıcının aynı/benzer performansı gördüğünde verdiği puanı düşürmesi
Ölçek yorumu	Puanlayıcının değerlendirme ölçeğini yanlış/farklı anladığı durumlarda uygun olmayan derecelendirmeler yapması
Puan ranj kısıtlaması	Ölçülen gizli değişkende kişiler arasında ayırım yapamama
Öz	Performansından daha fazla puan verme
değerlendirme/puanlama	
Bana benziyor etkisi	Puanlayıcının kendine benzer algıladığı öğrenciye daha yüksek puan vermesi
Gözden geçirme	Bir performansın sadece sınırlı bir bölümünü puanlama
Ani kayıp	Puanlayıcının bir performansın olumsuz yönü nedeniyle düşük puan vermesi
Sempatı puanı	Puanlayıcının duygularına hitap eden bir performansa yüksek puan vermesi
Katılık	Sürekli olarak düşük puan verme
Kişisel özellik	Bireyin tek bir özelliğine (güncellik, iletişim, görünüm vb.) odaklanarak fazla puan verme

Royal, K.D. ve Hecker, K.G (2016). Rater Errors in Clinical Performance Assessments. *Journal of Medical Education*, 43(1), 5-8. <https://doi.org/10.3138/jvme.0715-112R>.

Yukarıdaki Tablo 1’de 30 tane puanlayıcı davranışı verilmiştir. Bu davranışlardan cömertlik/katılık, halo etkisi, merkeze eğilim ve farklılaşan puanlayıcı katılığı/cömertliği daha fazla karşılaşılan davranışlardır (Myford ve Wolfe, 2003).

Bir puanlayıcının puan verirken istenen özelliğe bakmaksızın sürekli yüksek puanlama yapması puanlayıcı cömertliği olarak tanımlanırken, sürekli düşük puanlama yapması ise puanlayıcı katılığı olarak tanımlanır (Myford ve Wolfe, 2003). Diğer bir deyişle, derecelendirmenin herhangi bir ölçeğin üst ucunda birikme eğilimi cömertlik; alt ucunda birikme eğilimi katılıktır. Puanlayıcı cömertliği ve puanlayıcı katılığı hatalarının ortaya çıkmasını engellemek için Myford ve Wolfe çalışmalarında bazı önerilerde bulunmuşlardır. İlk olarak puanlayıcılara derecelendirme ölçeklerindeki kategorilerin ne anlama geldiği konusunda çok detaylı açıklama yapılmasını önermişlerdir. Ayrıca puanlayıcıların aşırı cömert davranma eğilimini engellemek için olumlu ve olumsuz tarafta birkaç ölçek kategorisi planlanabileceğini söylemişlerdir. Yani derecelendirme ölçeklerinde olumlu kategoriler olduğu kadar olumsuzların da yer almaları gerektiğini belirtmişlerdir. Puanlayıcıları cömertlik ve katılık davranışlarına karşı uyarmanın bu davranışın çok sık ortaya çıkabileceğini söylemenin önemli olduğunu vurgulamışlardır. Böylece puanlayıcılar bu davranışlara karşı dikkatli olabileceklerdir. Ayrıca puanlayıcılardan her bir derecelendirme kategorisine öğrencileri dağıtmaları istenebilir. Puanlayıcılar değerlendirdikleri öğrencileri tüm kategorilere yerleştirmek durumunda kalırlar ve sadece düşük puan ya da sadece yüksek puan verme eğiliminden kaçınmış olurlar.

Halo etkisi alan yazında en çok çalışılan puanlayıcı hatalarından bir diğeridir. Bir puanlayıcının, bir öğrencinin herhangi bir özelliğini değerlendirip diğer özelliklerine de aynı puanı vermesi olarak tanımlanabilir (MyFord ve Wolfe, 2003). Başka bir deyişle, bir kişiyi genel olarak iyi ya da genel olarak kötü düşünme olarak ifade edilebilir. Bu davranışın önüne geçebilmek için puanlayıcıların böyle bir hata türü olduğunun farkında olması ve bu hatayı yapmasını engellemek için eğitilmeleri gerekir. Myford ve Wolfe (2003)'ye göre bir puanlayıcı birden fazla özelliği/soruyu puanlayacaksa bütün öğrencilerin ilk sorusunu puanladıktan sonra ikinci soruya geçmelidir. Bir öğrencinin tüm sorularını art arda puanlamak yerine tüm öğrencilerin ilk soruları puanlanıp sonra ikinci soruya geçilip bu şekilde devam edilmelidir. Ayrıca puanlayıcılar birden fazla özelliği derecelendiriyorlarsa

bu ölçeklerin yüksek ve düşük uçlarının sıralaması ara ara değiştirilip tersine çevrilebilir (MyFord ve Wolfe, 2003). Ayrıca puanlayıcının puanladığı kişiyi tanıması da bu hatanın ortaya çıkmasına neden olabilir (Jacobs ve Kozlowski, 1985). Bu nedenle öğrenci ismine bakmadan ya da her bir öğrenciye önceden numara vererek değerlendirme işe yarayabilir. Merkeze eğilim etkisi, bir puanlayıcının aşırı uç kategorilerden kaçınıp derecelendirme ölçeğinin hep orta kategorisini kullanması olarak tanımlanır (MyFord ve Wolfe, 2003). Landy ve Farr (1983), merkezi eğilim etkisini, aşırı olumlu veya olumsuz derecelendirmelerden kaçınma veya ölçeğin orta noktasında veya yakınında derecelendirme olarak tanımlamıştır. Guiford (1954)'a göre bazı puanlayıcılar aşırı yargılarda bulunmaktan çekinirler ve kişileri grubun ortalaması yönünde puanlarlar. MyFord ve Wolfe (2003), bu davranışı engellemek için diğer davranışlarda olduğu gibi puanlayıcıların eğitilmesi ve bu davranış hakkında bilgi sahibi olması istenir. Bir farkındalık yaratılır. Katılık ve cömertlikte olduğu gibi puanlayıcı kişilerin her bir değerlendirme kategorisine belli sayıda öğrenci yerleştirmeleri istenebilir. Böylece derecelendirme ölçeğinde orta puanlara yığılmanın önüne geçilebilir (MyFord ve Wolfe, 2003).

Farklılaşan puanlayıcı katılığı ve cömertliği ise öğrencilerin bazılarına daha cömert bazılarına daha katı davranma davranışdır (Knoch, Read ve Randow, 2007). Örneğin, bir puanlayıcı değerlendirdiği kişilerin cinsiyetine, yaşına, popüleritesine göre farklı puanlar verebilir.

Bu hataları belirlemek ne kadarının ölçme sonuçlarına karıştığını görmek için korelasyonel, grafiksel, regresyon modelleri, gizli sınıf modelleri, çok düzeyli modeller gibi istatistiksel teknik kullanılmaktadır. Ancak son zamanlarda Genellenebilirlik kuramı ve Madde tepki kuramı daha yaygın hale gelmiştir (Royal ve Hecker, 2016). Madde tepki kuramında Çok Yüzeyle Rasch Ölçme Modeli, puanlayıcı tutarsızlıklarına dayalı puanları belirlemek için daha çok tercih edilmektedir.

Puanlayıcı hatalarını azaltmak için puanlayıcıların eğitimi önemlidir. Puanlayıcı eğitimi genellikle madde değerlendirme sürecinden önce yapılır. Bu eğitimde, derecelendirme

sürecini standartlaştırma, puanlayıcıları puanlama prosedürlerine alıştırmaya ve onları derecelendirme sırasında beklenmedik durumlarla başa çıkmaya hazırlama amaçlanır (Taylor, 2004, s.2). Alan yazında puanlayıcı eğitiminin, puanlayıcılar arasındaki farklılıkları ortadan kaldırması, puanlayıcıların öz tutarlılığını artırması ve puanlayıcılar tarafından derecelendirme durumunun yönlerine yönelik sergilendiği bireysel önyargılarını azaltması yönünden kanıtlar vardır (Knoch, Read ve Randow, 2007). Eğitim sırasında puanlayıcılar normal süreçle hangi puanlama ölçeğini kullanacaklarsa onu kullanmalıdırlar (McNamara, 1996, s.125-127). Eğitimde puanlayıcılara bir soruyu nasıl değerlendirecekleri aşama aşama anlatılmalıdır. Oluşabilecek puanlayıcı davranışları açıklanmalıdır. Royal ve Hecker (2016) puanlayıcı eğitiminin iki aşaması olması gerektiğini söylemiştir. Birinci aşamada puanlayıcıların meydana gelecek puanlayıcı hatası türlerini bilmesini sağlamaktır. İkinci aşamada ise puanlayıcılara performans düzeyleri için ortak bir kavramsallaştırma oluşturmaya yönelik rehber olmaktır. Puanlayıcılara tüm seviyelerde örnekler sunulmalıdır. Bu örnekler sınıfta beraber değerlendirilmeli, ortak kararda buluşmaya çalışılmalıdır. Alderson vd. (1995), puanlayıcılar çok deneyimli olsalar bile yine de her uygulamadan önce derecelendirmenin güvenilirliği için verilen eğitimlerin kısa bir hatırlatmasının yapılmasının öneminden bahsetmişlerdir. Eğitim verildikten sonra gerçek bir puanlama uygulamasından önce tüm sınıf beraber birkaç soru üzerinden uygulama yaparak ortak paydada karar verilmesine çalışılmalıdır. Bu süreçte anında verilen geribildirimler puanlayıcıların davranışındaki değişiklikleri azaltmada yardımcı olacaktır (Karakaya, 2022; Şata, 2019).

Puanlayıcı Güvenirliği

Ölçmede güvenirlilik, bir ölçme aracından elde edilen puanların tutarlılığı olarak tanımlanabilir (Tekin, 1991; Atılğan, 2005). Puanlama işine insanın dahil olması hataları da beraberinde getirir. Puanlamayı yapan kişinin duygusal durumu, dikkati gibi özellikleri ölçüm sonucunu etkileyebilir. Bazı puanlayıcılar yorgunluk durumuna göre bazı öğrencilere

daha yüksek verirken, bazılarına daha düşük puanlama yapabilirler. Bu hatayı azaltmak için puanlama işini farklı kişilere yaptırmak güvenilirliği artırıcı bir önlem olacaktır. Puanlayıcı etkilerini en aza indirmek güvenilirlik düzeyini istenen seviyede tutmayı sağlayacaktır. Bunun için birden fazla puanlayıcı kullanmak, puanlama yaparken dereceli puanlama anahtarları kullanmak (Kutlu, Doğan ve Karakaya, 2010) gibi uygulamalar yapılabilir. Ayrıca puanlayıcı eğitimleri vermek de puanlayıcı etkilerini azaltmayı sağlayacaktır (İlhan ve Çetin, 2014). Ölçme aracının uzunluğu, kısalığı, yazı stili, kazanımlara uygun olmayışı gibi etkenler de ölçme aracından kaynaklanan hatalar doğurup güvenilirliği etkileyecektir. Ayrıca öğrencinin fiziksel durumu, hasta, yorgun oluşu, motivasyon eksikliği gibi nedenler de öğrenciden kaynaklı hatalar olarak ölçmeye karışabilir. Ölçmeye karışan hatalar, ölçülen özellikten, ölçme aracından, ölçmeyi yapan kişiden, ölçme yönteminden kaynaklanabilir (Tekin, 1991 s.55; Crocker ve Algina, 2008, s.105).

Açık uçlu soruların değerlendirilmesinde de tüm bu hata kaynakları ölçmenin güvenilirliğini etkileyecektir. Bu süreçte puanlayıcıdan gelen hatalar önemli bir varyans oluşturmaktadır (Şata, 2019). Bir öğrencinin sorusuna iki veya daha fazla puanlayıcı tarafından verilen puanların tutarlılığı puanlayıcılar arası güvenilirlik olarak adlandırılır. Bir öğrenciye bir puanlayıcının birden fazla puanlama yaptığı ölçmelerin sonuçları arasındaki tutarlılık ise puanlayıcı içi güvenilirlik olarak adlandırılır (Allharby, 2006, Güler, 2008).

Ölçme ve değerlendirmede klasik test kuramı (KTK), genellenebilirlik kuramı (GK) ve madde tepki kuramı (MTK) olarak üç temel kuram bulunmaktadır. Bu kuramlarda puanlayıcı güvenilirliğini sağlamak için çeşitli güvenilirlik hesaplama yöntemleri vardır. Bu kuramların hepsi de ölçmedeki hatalardan kaçınmak isterler. Ölçülen sonucun gerçek ölçüme yakın olmasını sağlamaya çalışırlar. Ölçme alanındaki ilk kuram olan klasik test kuramında (KTK) $X=T+E$ şeklindeki eşitlikte, “X”, gözlenen test puanını, “T” öğrencinin gerçek puanını, “E” ise ölçmeye karışan hatayı temsil eder (Crocker ve Algina, 1986). Hata ne kadar düşük olursa bireyin gerçek puanı ile gözlenen puanı birbirine o kadar yakın olacaktır. Klasik test kuramı tek bir değişkenlik kaynağından gelen puanı hesaba katar.

Ancak birçok durumda potansiyel hata kaynakları birden çoktur. Açık uçlu soruların değerlendirilmesinde puanlayıcılar en önemli hata kaynağı sayılabilir. Bu nedenle tüm değişkenlik kaynaklarından gelen hataları birlikte değerlendirme gerekir. Madde tepki kuramının çok yüzeyli rasch ölçme modelinde yukardaki açıklanan şekilde güvenilirlik hesaplanabilir.

Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM)

Çok yüzeyli Rasch ölçme modeli, test veya değerlendirme sonuçları üzerinde etkisi olan birden fazla değişkenin ayrıntılı bir analizini sağlamayı amaçlar (Eckes, 2009). Olasılıkların örtük bir modeli olan ÇYRÖM, temel rasch modelinin (Rasch, 1960) bir uzantısıdır. Yabancı alan yazında many-faceted Rasch measurement (Eckes, 2015; Engelhard, 1994; McNamara, 1996) veya many-faceted conjoint measurement (Linacre ve Myford, 1994) olarak yer almıştır. Linacre, Rasch'ın geliştirdiği modelin üzerine puanlayıcıların rolünü de eklemiştir. Bu model, yalnızca maddelerin zorluklarının ve öğrencilerinin yeteneklerinin değil, aynı zamanda puanlayıcıların ciddiyetinin de karşılaştırılmasına olanak sağlar (Esfandiriari, Farrokhi, Dalili, 2011). Eckes (2005), Puanlayıcıların, bir maddeyi puanlarken aynı ciddiyette olmadıklarını görmüştür. Rasch modelinde, belli bir yeteneğe sahip kişi, bu zorluktaki maddeyle karşılaştığında bu kişinin bu maddeyi doğru anlama olasılığı nedir? sorusuna yanıt aranır (Bond ve Fox, 2015). Bir soruyu doğru yanıtlama olasılığı, madde zorluğu ve kişi yeteneğine göre değişecektir. Rasch modeli doğrusal bir derecelendirmedir. Ölçümdeki yüzeylerin (örneğin, madde, öğrenci, puanlayıcılar vb.) her birini ortak bir doğrusal ölçeğe yerleştirir (Wolfe ve Chiu, 1997). Bu ölçeğe logit cetvel adı verilir. Logit cetvelde pozitif değerler, bireyler için yeteneğin yüksek olmasını, puanlayıcılar için cömert puanlamayı, maddeler için ise kolaylığı işaret eder. Benzer şekilde negatif değerler bireyler için yeteneğin düşük olmasını, puanlayıcılar için katı puanlamayı, maddeler içinse güçlüğü (zorluğu) gösterir (Güler, 2008). Ardışık kategori olasılıklarının lojistik dönüşümü bağımlı

değişken olarak, öğrenci yeterliği, madde zorluğu, puanlayıcı davranışı gibi durumlar bağımsız değişken olarak işlev görür (Eckes, 2005). ÇYRÖM’de bir bireyi etkilediği düşünülen tüm değişkenlik kaynakları dikkate alınır. ÇYRÖM’deki yüzeyler (facet) bir öğrencinin/bireyin değerlendirilmesini etkileyen faktörlerdir (Koons, 2008). ÇYRÖM analizleri FACETS paket programı ile yapılabilir. Bu program tüm yüzeylere ait sonuçları örneklemden bağımsız olarak analiz etmektedir. Bireysel değerlendirmeler ve grup değerlendirmeleri yapabilmektedir. Tüm yüzeylerin (öğrenci, puanlayıcı, madde) ve bunların kendi içinde etkileşimlerini hesaplayabilmektedir. Ölçüm sonuçlarına karışan puanlayıcı hatalarını (puanlayıcı katılığı/cömertliği, halo etkisi, merkeze eğilim, farklılaşan puanlayıcı katılığı/cömertliği) belirleyebilmektedir.

Çok yüzeyli Rasch ölçme modelinin bazı sayıltıları vardır. Bu sayıltılardan birincisi model-veri uyumudur. Yani daha yetenekli bir öğrencinin bir soruyu daha az yetenekli bir öğrenciye göre yanıtlama olasılığının daha yüksek olmasıdır (Esfandiriari, Farrokhi, Dalili, 2011). Yetenek arttıkça alınan puanlar da artar. İkinci gereksinimi/sayıltısı yerel bağımsızlıktır. Yerel bağımsızlık, bir öğrencinin sonraki maddelerdeki performansının önceki maddelere bağlı olmamasını ifade eder (Linacre, 1998). Üçüncü özelliği ise tek boyutluluktur. Tek boyutluluk, her seferinde bir gizli(örtük) veya altta yatan yeteneğin ölçülmesi anlamına gelir (Linacre, 1998). Bir başka deyişle, bir testteki tüm maddelerin aynı temel değişkeni, yapıyı ölçmesi demektir.

İki yüzeyli Rasch modelinde değişkenlik kaynağı olarak bireyler (öğrenciler) ve maddeler bulunur.

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

P_{nik} : n bireyinin i maddesinde (kategorisinde) k puan alması olasılığı

P_{nik-1} : n bireyinin i maddesinde (kategorisinde) k-1 puan alması olasılığı

B_n : n bireyinin yetenek düzeyi

D_i : i maddesinin güçlük düzeyi

C_k : k-1 puanından k puanına geçme güçlüğü

Yukarıdaki formül ölçme sürecine puanlayıcıların da katıldığı durumda yeterli olmayabilir. Çok yüzeyli Rasch ölçme modelinde genel olarak üç yüzey bulunduğu aşağıdaki formül kullanılır (Linacre, 1989).

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

P_{nijk} : n bireyinin j puanlayıcısından i kategori gözleminde k puan alma olasılığı

P_{nijk-1} : n bireyinin j puanlayıcısından i kategori gözleminde k-1 puan alma olasılığı

B_n : n bireyinin yetenek düzeyi

D_i : i maddesinin güçlük düzeyi

C_j : j puanlayıcısının katılık düzeyi

F_k : k-1 puanından k puanına geçişin güçlüğü

Bu formül ile derecelendirmeler bir logit ölçeğine dönüştürülür. Tüm yüzeyler aynı anda fakat bağımsız olarak analiz edilir. Yüzeylerin ortak kalibrasyonu, puanlayıcı katılığının, puanlanan birey ve özelliğin zorluğu ile aynı ölçekte ölçülmesini sağlar. (Myford ve Wolfe, 2003). Çok yüzeyli Rasch modelde grup düzeyinde ve bireysel düzeyde istatistiksel göstergeler bulunur. Grup düzeyindeki istatistiksel göstergeler, sabit etkili ki-kare testi ve ayırma istatistikleridir (ayırma oranı, ayırma indeksi ve ayırma indeksinin güvenilirliği). Araştırmadaki tüm yüzeyler için grup düzeyindeki istatistikler hesaplanır. Sabit etkili X^2 testinde puanlayıcıların katılık düzeylerinin anlamlı olarak farklı olmadığı kabul edilir. Yani her puanlayıcının aynı düzeyde katı olduğu kabul edilir. O halde sabit etkili ki-kare değerinin istatistiksel olarak manidar çıkması puanlayıcıların birbirlerinden farklı puanlamalar yaptıklarını gösterir (Myford ve Wolfe, 2004). Grup düzeyindeki istatistiklerden diğeri ayırma istatistikleridir. Bunlardan ilki ayırma oranı (separation ratio), puanlayıcıların katılıklarının yayılımı anlamına gelir. Bu değer 0'a ne kadar yakında

puanlayıcılar birbirlerine o kadar benzerdir denilebilir (Eckes, 2009). Ayırma oranı şu şekilde hesaplanır.

$$G_J = \frac{SD_{t(J)}}{RMSE_J} \quad (3)$$

Burada $SD_{t(J)}$ gerçek standart sapma, $RMSE_J$ ise ortalama ölçüm hatasıdır. G_J değeri ne kadar yüksek olursa puanlayıcıların katılık durumları o kadar yayılmış demektir. Örneğin, ayırma oranı 7,50 olması bir yüzeyde puanlayıcı katılık farklılıklarının ölçüm hatasından 7 kat fazla olduğu anlamına gelir (Eckes, 2009). Ayırma oranı değeri kullanılarak bulunan bir istatistik de ayırma indeksi (seperation index)'tir. Ayırma indeksi de ayırma oranı ile benzer şekilde yorumlanabilir. Örneğin ayırma indeksi 3,0 bulunduğunda puanlayıcıların katılık davranışının farklı üç gruba ayrılabilceği şeklinde yorum yapılabilir (Eckes, 2009). Ayırma indeksi aşağıdaki formül ile hesaplanır.

$$H_J = \frac{4G_J + 1}{3} \quad (4)$$

Grup düzeyinde bir diğer istatistik ise ayırma indeksinin güvenilirliği (seperation index reliability)'dir. Bu istatistik, puanlayıcıların birbirlerinden nasıl güvenilir bir şekilde ayrıldığı konusunda bilgi verir. Madde ve öğrenci yüzeyi için ayırma indeksinin güvenilirliğinin yüksek olması istenirken, puanlayıcı yüzeyi için bu değer in sıfıra yakın olması istenir. Çünkü sıfıra yakın olduğunda puanlayıcı davranışları birbirine benzerdir denebilir (Myford ve Wolfe, 2003). Ayırma istatistiği hesaplanırken aşağıdaki formül kullanılır.

$$R_J = \frac{G^2}{1 + G^2} \quad (5)$$

Rasch-cohen's kappa istatistiği grup düzeyindeki istatistiklerden bir diğeridir. Çok yüzeyli Rasch modelinde gözlenen ve beklenen uyum yüzdeleri puanlayıcı yüzeyine ait puanlayıcı güvenilirliği için hesaplanmaktadır. Gözlenen ve beklenen uyum yüzdeleri aynı olduğunda

kappa değeri 0,00 çıkar ki bu da puanlayıcının çok iyi puanlama yaptığını göstermektedir (Eckes, 2009). Pozitif çıkan değerler puanlayıcıların uyumunun iyi olduğunu belirtirken, negatif çıkan değerler puanlayıcıların düşük uyuma sahip olduklarını gösterir. Rasch-cohen kappa değeri aşağıdaki formül ile hesaplanır.

$$\text{Rasch- Cohen's Kappa} = \frac{\text{Gözlenen uyum yüzdesi} - \text{beklenen uyum yüzdesi}}{100 - \text{beklenen uyum yüzdesi}} \quad (6)$$

Çok yüzeyli Rasch ölçme modelinde bireysel düzeyde hesaplanan istatistiklerden ilki her bir yüzeyin elemanları için hesaplanan logit değerleridir. Bu değerler maddeler (sorular) için zorluğu veya kolaylığı, öğrenciler için başarılı veya başarısız olmayı, puanlayıcılar için katılığı veya cömertliği gösterir. Bu logit değerleri ilk olarak Rasch analiz çıktısında değişken haritası (kalibrasyon haritası) üzerinde gösterilir. Araştırmada kaç yüzey varsa tüm yüzeyler aynı kalibrasyon haritasında çıktı olarak sunulur. Logit cetvelinde pozitif değerler cömert puanlayıcı, kolay madde ve başarılı öğrenci olarak yorumlanırken; negatif değerler katı puanlayıcı, zor madde ve başarısı düşük öğrenci olarak yorumlanır (Eckes, 2009; İlhan, 2015). Puanlayıcılar hakkında katı/cömert olarak yorum yapabilmek için bakılan logit değerlerinde karşılaştırma için kritik bir değer olmaması nedeniyle kesin bir karar vermek için logit değerleri genellikle t-değerlerine dönüştürülür. Her bir puanlayıcı için hesaplanan t-değeri, araştırma verileri doğrultusunda t-dağılım tablosundan bakılan t-değeri ile karşılaştırılır. Hesaplanan t-değerinin, t-dağılım tablosundaki değerden mutlak değerce büyük ya da küçük olmasına göre katı ya da cömert puanlayıcı olarak yorumlanır (Myford ve Wolfe, 2003). Logit ölçülerini t-değerlerine çevirmek için aşağıdaki formül kullanılır.

$$t\text{-değeri} = \frac{\text{puanlayıcının lojit ölçümü} - \text{logit değerleri ortalaması}}{\text{logit ölçümünün standart hatası}} \quad (7)$$

Bireysel düzeydeki bir diğer istatistik iç uyum (infit) ve dış uyum (outfit) uygunluk istatistikleridir. Bir bireyin gözlenen puanı ile beklenen puanı arasındaki farkı temel alan uygunluk istatistikleri düşük yetenekli bir bireyin zor bir soruyu yapabilmesi veya yüksek yetenekli bir bireyin kolay bir soruyu yapamaması olasılığını kestirir (Atılgan, 2004). Her bir yüzeye ait iç uyum (infit) ve dış uyum (outfit) değerleri Rasch çıktısında yer alır. Bu istatistiklerin beklenen değeri “1” dir. Ancak Linacre (2002), iç uyum ve dış uyum değerleri için 0,50 ile 1,50 arasını kabul edilebilir aralık olarak önermiştir. Uyum istatistiklerinin 0,50’den küçük, 1,50’den büyük olması beklenen değer ve gözlenen değer arasındaki farkın (değişimin) fazla olduğunu gösterir (Myford ve Wolfe, 2003).

Araştırmalarda puanlayıcı yanlılığı için yanlılık istatistikleri de Rasch çıktısında yer alır (bias analysis). “*puanlayıcı x birey*” yanlılık analizi için Rasch formülü aşağıdaki şekile dönüşür.

$$\log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = B_n - D_i - C_j - F_k - I_{nj} \quad (8)$$

Burada denkleme eklenen I_{nj} *puanlayıcı x birey* yanlılığına ilişkin etkileşimi gösterir. Araştırmada iki yüzey ya da ikiden fazla yüzeyin etkileşimine bakıldığında da aynı şekilde hesap yapılır.

Rasch analizinde puanlama ölçeğindeki derecelendirmenin uygunluğu için “kategori istatistikleri” de hesaplanır. Hem bireysel düzeyde hem de grup düzeyinde çıktı olarak sunulur. Puanlayıcı davranışlarından merkeze yönelim davranışı araştırıldığında kategori istatistikleri incelenir (Myford ve Wolfe, 2003). Burada beklenen ve gözlenen değer, dış uyum istatistiği yer alır ve uyum dışı değer 0,5 ile 1,5 arasında değer alıp almadığı incelenir.

Puanlayıcı Eğitimi ve Akran Değerlendirme ile İlgili Araştırmalar

Bu bölümde puanlayıcı eğitimi ve akran değerlendirmesi ile ilgili araştırmaya katkıda bulunabilecek ilgili araştırmalara kronolojik sıra ile yer verilmiştir.

Weigle (1998)' in yaptığı çalışmada deneyimli ve deneyimsiz puanlayıcıların puanlama katılımı incelenmiştir. Sekiz deneyimsiz ve sekiz deneyimli olmak üzere toplam 16 puanlayıcı 60 tane öğrencinin kompozisyonlarını değerlendirmişlerdir. Veriler Çok Yüzeyle Rasch Ölçme Modeli ile analiz edilmiştir. Araştırmada üç yüzey (dil, retorik kontrol ve içerik) vardır. Araştırma sonuçlarına göre puanlayıcı eğitimi verilmeden önce deneyimsiz puanlayıcıların deneyimli puanlayıcılardan daha katı oldukları sonucuna varılmıştır. Puanlayıcı eğitimi verildikten sonra deneyimli ve deneyimsiz puanlayıcılar arasındaki katılım farkında anlamlı bir azalma olduğu görülmüştür. Puanlayıcı eğitiminin daha objektif puan verme konusunda etkililiği ortaya konmuştur.

Nakamura (2002) yaptığı çalışmada İngilizce sunum becerilerinin değerlendirilmesinde akran ve öğretmen puanlarını ÇYRÖM ile incelemiştir. Üç tane yüzeyin (öğrenciler, maddeler ve puanlayıcılar) kullanıldığı araştırmada, 12 üniversite öğrencisinin sunumu bir öğretmen ve dört akranı tarafından puanlanmıştır. Araştırma sonuçlarına göre akran değerlendirme yapmanın öğrenci performanslarını olumlu yönde etkilediği görülmüştür.

Matsuno (2006), öz ve akran değerlendirmesi ile ilgili bir çalışma yapmıştır. Çalışmaya 91 öğrenci ve 4 öğretmen katılmıştır. Öğrencilerin öz ve akran değerlendirmeleri öğretmen değerlendirmeleriyle karşılaştırılmıştır. Araştırmada Çok Yüzeyle Rasch Ölçme Modeli kullanılmıştır. Başarılı öğrencilerin öz değerlendirme yaparken kendi kompozisyonlarına düşük puan verdikleri saptanmıştır. Akran değerlendiriciler en hoşgörülü değerlendiriciler olmuştur. Yüksek başarılı öğrencilerin puanlamada ciddi, düşük başarılı öğrencilerin puanlamada daha hafif davrandığı sonucuna ulaşılmıştır. Akran değerlendirmesinde tutarlı puanlar görülmüştür. Öz değerlendirmenin kişiye özgü olduğu, bu nedenle resmi değerlendirmenin bir parçası olarak sınırlı katkı sağlayabileceği düşünülmüştür. Akran değerlendirmenin yazılı olarak değerlendirmede fayda sağlayacağı söylenmiştir.

Öğretmenlerin ÇYRÖM kullanarak akran değerlendirmedeki puanlayıcıların yanlış davranışları hakkında onları bilgilendirerek kendilerini geliştirmelerine yardımcı olabilecekleri sonucuna varılmıştır.

Baştürk (2010) araştırmasında öğrencilerin bilimsel araştırma ödevlerini çok yüzeyli Rasch ölçme modeli ile incelemiştir. Araştırma Pamukkale Üniversitesi'nde Rehberlik ve Psikolojik Danışmanlık bölümünde öğrenim gören 20 öğrenci ile yapılmıştır. Rasch ölçümü sonuçlarında iç uyum ve dış uyum değerlerinin alan yazına uygun aralıkta kaldığı gözlenmiştir. Puanlayıcıların katılık ve cömertlik davranışları birbirinden farklılık göstermemiştir. Yükseköğretimde çok yüzeyli Rasch ölçme modelinin kullanımının uygun olacağı yönünde fikrini belirtmiştir.

Farrokhi ve Esfandiari (2011) tarafından yapılan çalışmada ikinci dil yazma becerilerinin değerlendirilmesinde halo etkisini saptanmaya çalışılmıştır. Çalışmada İran'daki iki üniversitede 188 lisans öğrencisi tarafından yazılan 188 makaleyi öz değerlendirme, akran değerlendirme ve öğretmen değerlendirmesi ile değerlendiren 194 değerlendirici bulunmaktadır. Toplanan veriler analitik DPA ile puanlanmıştır. Çalışma sonuçlarına göre grup düzeyinde puanlayıcıların herhangi bir halo etkisi göstermediği, bireysel düzeyde tüm puanlayıcı türlerinde puanlayıcıların halo etkisi gösterdiği saptanmıştır. Öz değerlendirmede diğer iki değerlendirmeye göre daha fazla halo etkisi görüldüğü ortaya koyulmuştur.

Fahim ve Bijani (2011) ikinci dil yazma becerilerinin değerlendirilmesinde puanlayıcı eğitimi vermenin puanlayıcı davranışına etkisini incelemiştir. Araştırmada 12 puanlayıcı yer almaktadır. 60 öğrencinin kompozisyonları bu puanlayıcılar tarafından değerlendirilmiştir. IELTS tarafından geliştirilen puanlama anahtarının kullanıldığı bu çalışmada puanlayıcı eğitimi öncesinde 60 kompozisyonun 45 tanesi puanlanmış, daha sonra puanlayıcı eğitimi verilmiş ve ardından 15 tanesi puanlanmıştır. Sonuca göre puanlayıcı eğitiminin puanlayıcı katılığını düşürdüğü ve puanlayıcı tutarlılığını artırdığı saptanmıştır.

Güler (2014)'ün bir çalışmasında açık uçlu istatistik sorularının analizi çok yüzeyli rasch ölçme modeli ile yapılmıştır. Psikolojik Danışmanlık ve Rehberlik bölümü yüksek lisans

İstatistik I dersinde 10 açık uçlu madde kullanılmış ve bu maddelere verilen yanıtlar ÇYRÖM ile analiz edilmiştir. Araştırmaya 55 öğrenci ve 3 öğretim üyesi puanlayıcı katılmıştır. Araştırmada üç değişkenlik kaynağı olarak, öğrenciler, maddeler ve puanlayıcılar bulunmaktadır. ÇYRÖM sonuçlarına göre açık uçlu madde kullanımının öğrencilerin gelecekte girebilecekleri yaşamlarını etkileyecek kararların alındığı sınavlarda kullanılabilir olarak bulunmuştur.

Karakaya (2015), portfolyo değerlendirmesinde öz akran ve öğretmen değerlendirmelerini çok yüzeyli rasch ölçme modeli ile analiz etmiştir. Çalışmasında 74 öğretmen adayı yer almaktadır. Çalışmanın rasch modeli için dört yüzey bulunmaktadır. Bu yüzeyler, öğretmen adaylarının cinsiyeti, öğretmenlerin başarıları, öz, akran ve öğretmen puanlayıcıların katılığı/cömertliği ve portföy değerlendirilmesinde kullanılan kriterlerdir. Araştırma için araştırmacı tarafından yedi ölçütlü bir DPA oluşturulmuştur. Araştırma sonucunda başarı düzeyi yüksek olan öğrencilerin portfolyo performanslarının daha iyi olduğu, kız öğrencilerin erkek öğrencilere göre portfolyo performanslarının daha iyi olduğu sonucuna varılmıştır. Puanlayıcı türleri karşılaştırıldığında öz değerlendirmenin en yumuşak, akran değerlendirmenin ise en şiddetli olduğu sonucu ortaya çıkmıştır.

Yüzüak, Yüzüak ve Kaptan (2015), Genetik ve Biyoteknoloji konusunda verilen performans görevinin poster sunumunun değerlendirmesini akran ve öğretmen değerlendirmesi kullanarak çok yüzeyli rasch ölçme modeli ile analiz etmişlerdir. Bu araştırmada 7.sınıfta okuyan 50 öğrenci ve bir öğretmen yer almıştır. Öğrenciler beşer kişilik 10 gruba ayrılıp grup olarak materyal hazırlamışlardır. Akran ve öğretmen değerlendirmeleri sonucunda fen eğitiminde performans değerlendirmesi yapılırken akran değerlendirmesinin etkili bir şekilde kullanılabileceği sonucuna varılmıştır.

İlhan (2015), sunduğu doktora tezinde standart ve solo taksonomisine dayalı DPAlar kullanılarak puanlanan açık uçlu matematik sorularında puanlayıcı etkilerini çok yüzeyli rasch ölçme modeline göre incelemiştir. Yapılan çalışmada, 8.sınıf okuyan 104 öğrenci ve yedi matematik öğretmeni vardır. Çalışma için araştırmacı sekiz tane açık uçlu matematik

sorusu geliřtirmiş ve standart DPA, SOLO taksonomisine uygun DPA kullanmıştır. Verilen yanıtları ÇYRÖM ile analiz etmiştir. Sonuç olarak standart DPA ve SOLO taksonomisine dayalı DPAların kullanıldığı puanlamalarda birbirinden yetenekleri farklı olan öğrencilerin ayırt ediciliğini yüksek bir güvenilirlik ile hesaplamıştır. Ayrıca madde güçlük düzeylerinin de anlamlı olarak farklılaştığı belirtilmiştir. Standart puanlama DPAlarının kullanıldığı puanlamada, tutarsızlık, halo etkisinin olmadığı belirtilmiştir. Merkeze yönelme, puanlayıcı katılımı/cömertliği yönüyle anlamlı farklılık bulunmuştur. SOLO taksonomisi ile yapılan ölçmede ise puanlayıcı katılım/cömertliği, halo etkisi, merkeze yönelme gibi hatalar bulunmamıştır.

Salmaner (2015), yüksek lisans tezinde ortaokul beşinci sınıf öğrencilerinin yazma davranışlarının değerlendirilmesinde öz, akran ve öğretmen değerlendirmesini incelemiştir. Araştırmada çok yüzeyli Rasch ölçme modeli kullanılmıştır. Yapılan analiz sonucunda öz değerlendirmelerin en cömert, akran ve öğretmen değerlendirmelerinin ise neredeyse birbirlerine yakın ve katı olduğunu bulmuştur.

Şata (2019), sunduğu doktora tezinde performans değerlendirmesine karışan puanlayıcı hatalarını belirlemeye çalışmıştır. Araştırmada deney ve kontrol grublu ön test, son test kontrol grublu yarı deneysel desenle çalışmıştır. Kontrol grubunda 23 deney grubunda 22 olmak üzere toplam 45 puanlayıcıya ön test uygulamış ve Uluslararası İngilizce Dil Test Sistemi tarafından hazırlanmış olan yazma görevini puanlamalarını istemiştir. Ön test sonucunda gerekli analizleri yapıp grupların benzer olduklarını bulduktan sonra deney grubuna puanlayıcı eğitimi vermiştir. Bu eğitimin içeriğini puanlayıcı hatası eğitimi ve referans çerçevesi eğitimi desenlerinin birleştirilmesi ile oluşturduğu programla belirlemiştir. Verilen eğitimden sonra son testi uygulamıştır. Araştırmada puanlayıcıların katılımı/cömertliği, merkeze yönelim davranışları, halo etkisi davranışları ve puanlayıcı katılımı/cömertliği yani yanlılık davranışları incelenmiştir. Bu araştırmayı çok yüzeyli Rasch ölçme modelini kullanarak FACETS programında yapmıştır. Araştırma sonucunda puanlayıcı eğitimi öncesinde deney ve kontrol gruplarında, puanlayıcı katılımı/cömertliği,

yanlılık davranışlarının hep grup düzeyinde hem de bireysel düzeyde ortaya çıktığını bulmuştur. Halo ve merkeze eğilim davranışlarının ise bireysel düzeyde ortaya çıktığını görmüştür. Puanlayıcı eğitimi sonrasında tekrar yapılan Rasch analizi sonucunda deney grubunda yine de puanlayıcı hatalarının ortaya çıktığını ancak önceye göre azaldığını ortaya koymuştur. Verilen eğitimin puanlayıcı katılığı/cömertliğini giderme açısından %15’lik bir değişimle açıklamıştır.

Şahin, Taşdelen Teker ve Güler (2016) araştırmalarında akran değerlendirmede çok yüzeyli Rasch ölçme modelini kullanmışlardır. Araştırmaya Bilimsel Araştırma dersi alan 91 lisans öğrencisi ve dersi veren öğretim üyesi katılmıştır. Dersi alan 85 öğrencinin hazırladığı projeleri 6 akranı ve öğretim elemanı değerlendirmiş bütüncül dereceli puanlama anahtarı ile değerlendirmiştir. Araştırma sonucunda katılık cömertlik davranışlarının farklılaştığı görülmüş, başarısı yüksek öğrencilerin, başarısı orta ve düşük olan öğrencilere göre daha katı davrandıkları bulunmuştur. Öğrencilerin akran değerlendirmeye karşı görüşlerinin olumlu olduğu görülmüştür.

Öztürk, Şahin ve İlhan (2019), araştırmalarında akran değerlendirme sürecinde analitik DPA kullanmayı ve genel izlenim puanlaması yapmayı karşılaştırmışlardır. Araştırmada 66 lisans öğrencisi vardır. Bu öğrencilerden 6 tanesi akran puanlayıcısı olarak seçilmişlerdir. Sunum becerilerinin değerlendirildiği bu çalışmada öğrenciler sunum yaparken öğretmen ve 6 akran değerlendiricisi hem analitik DPA ile hem de genel izlenimle puanlama yapmışlardır. Elde edilen veriler çok yüzeyli Rasch ölçme modeline göre analiz edilmiş ve her iki yöntemde de yüksek güvenirlik ile öğrencilerin birbirlerinden ayırt edilebildiği bulunmuştur. Ancak analitik DPA kullanımının daha ayrıntılı bir şekilde yetenek düzeylerini ayırabildiği görülmüştür. Akran değerlendirme sürecinde her iki yöntemin yetenek kestirimleri arasında pozitif yönlü güçlü bir korelasyon bulunmuştur.

Tobaş (2020), yüksek lisans tezinde performans değerlendirmesinde puanlayıcı davranışlarını incelemiştir. Araştırmasında çok yüzeyli Rasch ölçme modelini kullanmıştır. Araştırmada 57 tane ortaokul 7.sınıf öğrencilerinin yazma performansının belirlenmesinde

puanlayıcı davranışları araştırılmış ve puanlayıcı güvenilirliği hesaplanmıştır. Puanlayıcı eğitimi alan 11 öğretmenin öğrencilerin yazma becerilerini puanlaması üzerine kurulu bir araştırmadır. Araştırmada puanlayıcı davranışları olarak alan yazında en çok ortaya çıkan puanlayıcı katılığı/cömertliği, halo etkisi, merkeze yönelim ve farklılaşan puanlayıcı katılığı/cömertliği (yanlılık) araştırılmıştır. Yapılan Rasch ölçme sonucunda öğrencilerin performansını değerlendirirken puanlayıcıların katı veya cömert davrandıklarını ancak merkeze eğilim ve halo davranışı göstermediklerini bulmuştur. Yanlılık davranışı incelendiğinde puanlayıcıların grup düzeyinde yanlı davranmadıkları fakat bireysel düzeyde yanlı davrandıkları ortaya çıkmıştır.

BÖLÜM III

YÖNTEM

Bu bölümde araştırmanın modeli, çalışma grubu, veri toplama araçları ve puanlayıcı eğitimi yer almaktadır.

Araştırma Deseni

Araştırmada öğrencilere puanlayıcı eğitimi verilip akran değerlendirme sürecinde puanlayıcı davranışları incelenmiştir. Çalışmada ön test son test kontrol gruplu yarı deneysel desen kullanılmıştır. Yarı deneysel desende katılımcılar deneysel işlemde önce ve sonra aynı değişken ile ölçülmüşler (Büyüköztürk, 2011).

Çalışma Grubu

Araştırmanın çalışma grubunu Ankara ili Mamak ilçesinde bulunan bir devlet lisesinde öğrenim gören 18 i erkek 47 si kız olmak üzere 65 tane 10.sınıf öğrencisi oluşturmaktadır. Bu öğrencilerden 32 tanesi 10A şubesinde, 33 tanesi 10B şubesinde öğrenim görmektedir. Öğrencilerin okulda uygulanan matematik sınavları not ortalamalarının benzer olması (10A için 55,81; 10B için 55,36) ve öğrencilerin daha önce herhangi bir puanlayıcı eğitimi

almamış olmaları gerekçesiyle rastgele iki gruba ayrılmışlardır. 10A deney grubu, 10B ise kontrol grubu olarak seçilmiştir. Araştırmaya katılım gönüllük esasına göre yapılmıştır.

Veri Toplama Araçları

Veri toplama aracı olarak açık uçlu matematik soruları ve dereceli puanlama anahtarı kullanılmıştır.

Açık Uçlu Maddelerin Hazırlanması

Lise 10.sınıf öğrencilerinin puanlayıcı davranışlarını belirlemek amacıyla matematik alanında 10 tane açık uçlu soru yazılmıştır. Bu maddeler 9. ve 10. sınıf kazanımlarını içermektedir. Maddelerde ayrıntılı görsellere, tablolara yer verilip anlatılmak istenen içerik gereksiz kelime kullanılmadan özgün bir şekilde anlatılmaya çalışılmıştır. Sorular öğrencilerin dikkatini çekecek, gerçek yaşam durumlarına uyacak öykülerle anlatılmıştır. Maddeler yazıldıktan sonra 1'i ölçme ve değerlendirme alanında yüksek lisans yapmış 3 matematik öğretmeni, 3 ölçme ve değerlendirme uzmanı ve 1 Türk Dili ve Edebiyatı öğretmeni olmak üzere 7 uzman görüşü alınmış ve sorularda gerekli düzeltmeler yapılmıştır. Gazi Üniversitesi Etik Komisyonu'ndan ve Milli Eğitim Bakanlığı Araştırma, Yarışma ve Sosyal Etkinlik biriminden ilgili izinler alınmıştır. Daha sonra bu maddeler Ankara ilinde üç farklı devlet lisesinde 103 öğrenciye pilot uygulama olarak uygulanmıştır. Her sorunun madde güçlük indeksleri ve madde ayırt edicilikleri hesaplanmıştır. Madde ayırt edicilik indeksi aşağıdaki formül ile hesaplanmıştır (Karakaya ve Şata, 2022).

$$D = \frac{\text{Üst grubun maddedeki ortalaması} - \text{alt grubun maddedeki ortalaması}}{\text{maddeye verilebilecek en yüksek puan} - \text{maddeye verilebilecek en düşük puan}} \quad (9)$$

Buradaki madde ayırt edicilik indeksi D değeri,

- 0,30 ve üstü değerde ise madde teste kabul edilir
- 0,20 ile 0,29 arasında ise madde düzeltilir
- 0,19 değerinin altında ise madde testten çıkarılır (Crocker ve Algina, 1986).

Maddeler için güçlük indeksi ise aşağıdaki formül yardımı ile hesaplanmıştır (Karakaya & Şata, 2022).

$$p = \frac{\text{Maddeden alınan ortalama puan} - \text{maddeye verilen en düşük puan}}{\text{maddeye verilen en yüksek puan} - \text{maddeye verilen en düşük puan}} \quad (10)$$

Madde güçlük indeksi 0' a yaklaştıkça madde zor, 1'e yaklaştıkça madde kolay demektir. Aslında burada madde güçlüğü'nün ideal değeri ölçmenin amacına göre değişebilir. Örneğin çok eleyici bir sınav için güçlük indeksi 0'a yakın maddeler kullanılabilir. Geçerli ve güvenilir bir ölçme için 0,40 ile 0,60 arasındaki madde güçlüğü ideal sayılabilir (Karakaya ve Şata, 2022).

Esas uygulama için madde güçlük ve madde ayırt edicilik indeksleri yardımıyla en uygun değerlere sahip 4 tane soru seçilmiştir. Bu sorular için araştırmacı tarafından DPA oluşturulmuştur. Her iki gruba da deneysel işlem olan puanlayıcı eğitiminden önce kısa bir bilgilendirme yapılmıştır ve 3 akranının 4 sorusunu değerlendirmeleri sağlanmıştır. Daha sonra deney grubuna puanlayıcı eğitimi verilmiş ve son test olarak her iki gruptaki öğrenciler ön testte puanladıkları aynı 3 akranının kağıtlarını değerlendirmişlerdir.

Dereceli Puanlama Anahtarının Oluşturulması

Araştırmacılar ve uygulayıcılar, hayatın çeşitli yerlerinde derecelendirme ölçekleri kullanır. Örneğin, bir işveren işe alma kararına yardımcı olmak veya mevcut çalışanlarının performansını değerlendirme için derecelendirme ölçekleri kullanır. Okullarda öğretmenler öğrencilerin performansları için ya da öğrenciler kendi ürünlerini veya akranının ürününü

değerlendirmek için bu ölçekleri kullanırken, müdür ise öğretmenleri değerlendirmek için kullanır (Myford ve Wolfe, 2003, sy.5). Puanlayıcılar bir ürünü/soruyu değerlendirirken bu değerlendirmeleri gözleme, yorumlamaya, kişisel muhakemeye dayandırabilirler. Guilford (1954), derecelendirme ölçekleriyle derecelendirme yapan uygulayıcıların bir dereceye kadar nesnel olabildiklerini söylemiştir. Ayrıca, nicel yargılar için insanların kararlarına güvenilse de insanların değerlendirmedeki zayıflıklarına karşı tetikte olunması gerektiğini vurgulamıştır. Popham (1990), Guilford'un bu görüşünü desteklemiştir ve derecelendirme ölçekleri, derecelendirme işlemi ve değerlendiricilerin muhtemel hata kaynakları olabileceğini belirtmiştir. Puanlama anahtarlarının tasarlanması ve oluşturulması aşamasında birtakım sorunlar olabilir. Mesela puanlayıcılar için tam olarak neye puan vermeleri istendiği açık ve net bir şekilde yazılmamış olabilir veya maddenin cevabı açıkça tanımlanmamış olabilir. Kısmi doğru yanıtların olduğu maddeler için bu yanıtlar aşamalı bir şekilde örnekler içererek verilmemiş olabilir ya da sınırlar belirtilmemiş olabilir. Ölçme araçlarının bu tip sorunları derecelendirme işlemi olumsuz etkiler ve böylece puanlayıcıların araçları tutarsız ve gelişigüzel uygulamasına sebep olabilir. Bu tür hataların oluşmaması için derecelendirme anahtarları titizlikle ve aşamalı bir şekilde hazırlanmalıdır. Gerekirse pilot uygulama sonrası öğrencilerin verdikleri yanıtlardan da cevap anahtarı hazırlanması konusunda yardım alınabilir (Myford ve Wolfe, 2003).

Bu çalışmada açık uçlu maddelerin puanlanmasında holistik (bütünsel) puanlama anahtarı kullanılmıştır. Bütünsel puanlama anahtarında her bir madde için derecelendirilmiş puanlamalar bulunur. Her madde için ayrı puanlama anahtarı hazırlanmamış ve öğrencilerin puanlama yaparken eksik öğrenmelerini görebilmeleri ve kavram yanılgılarını giderebilmeleri sağlanmıştır. Açık uçlu sorular için bütünsel puanlama anahtarı oluşturulurken aşağıdaki adımlar izlenmiştir (Karakaya, 2022).

1. Amaca uygun olarak hangi puanlama anahtarının oluşturulacağına karar verilmesi

2. Puanlama anahtarında performans düzeyleri için ölçüt sayısına karar verilmesi. Yanıtların, *tam doğru yanıtlar, uzak doğru yanıtlar, kısmen doğru yanıtlar, yanlış yanıtlar* gibi kategorize edilmesi.
3. Tam doğru, kısmen doğru ve yanlış doğru yanıtlar için tüm olası yanıtlar yazılmaya çalışılmalıdır. Bunun için maddeyi cevaplandıran öğrenciler gibi düşünmek fayda sağlayabilir. Uzak doğru yanıtların, tam doğru yanıtlardan hangi eksikliklerinin olduğu açıkça belirtilmelidir.
4. Yanıtları yazarken açık, anlaşılır ve sade bir dil kullanılmasına özen gösterilmelidir.
5. Yanıtların binişik olmamasına dikkat edilmesi gerekmektedir.

Araştırma için hazırlanan DPA için uzman görüşü alınmıştır. Onların dönütleri göz önünde bulundurulmuş, yapılan pilot uygulama sonrasında olası yanıtlar kontrol edilmiştir. Bu uygulama puanlama anahtarının geçerliğini ve güvenilirliğini artırmaya yönelik yapılmıştır. 3 ölçme ve değerlendirme uzmanı, 3 matematik öğretmeni ve 1 Türk dili ve edebiyatı öğretmeni olmak üzere toplam 7 uzmandan görüş alınmıştır. Bu görüşler doğrultusunda maddenin Türkçe dil kurallarına uygunluğu, matematikteki ilgili konu alanını temsil edip etmediği hakkında bilgi sahibi olunmuştur. Açık uçlu maddelerin değerlendirilmesindeki kriterler için (1) gerekli/önemli, (2) önemli değil, (3) gereksiz anlamına gelecek şekilde bir form kullanılmış ve uzmanlardan maddeleri bu kriterlere göre değerlendirmeleri istenmiştir. Düzenlenen puanlama anahtarı için kapsam geçerlik indeksi (KGO) hesaplanmıştır. Lawshe (1975) tarafından geliştirilen bu indeks aşağıdaki şekilde hesaplanmaktadır. Lawshe tekniğinde uzmanlar arasında bir uyuşma olup olmadığına bakılır.

$$KGO = \frac{N_U}{N/2} - 1 \quad (11)$$

Burada,

N_U = Maddeye “gerekli” diyen uzman sayısı

N = Görüş bildiren toplam uzman sayısı

Kapsam geçerlik indeksi -1 ile +1 arasında bir değer alır. -1 mutlak red, +1 mutlak kabul anlamına gelir. Mutlak kabul 0,99 olarak da kabul edilmektedir. Yani tüm uzmanlar maddeyi uygun olarak değerlendirirse maddenin KGO değeri +1 çıkacaktır. Tam tersi şekilde uzmanların hepsi uygun olmadığını söylerse maddenin KGO değeri -1 çıkar. Uzmanların %50'sinden daha fazlasının o madde için "gerekli" demesi gerekir. Aksi halde KGO değeri 0 çıkacaktır. KGO değeri 0 veya negatif değer aldığında, bu maddelerin kapsam geçerliğinden söz edilemez ve maddelerin çıkarılması gerekir (Lawshe, 1975). Lawshe tarafından 1975' te geliştirilen KGO için madde seçimine yönelik kritik değerler, 2012 yılında Wilson, Pan ve Schumsky tarafından güncellenip revize edilmiştir.

Tablo 2
Kriter Seçimi için Geçerlik Oranları

	Tek kuyruklu hipotezler için anlamlılık düzeyi					
	.10	.05	.025	.01	.005	.001
N (uzman sayısı)	Çift kuyruklu hipotezler için anlamlılık düzeyi					
	.20	.10	.05	.02	.01	.002
5	.573	.736	.877	.99	.99	.99
6	.523	.672	.800	.950	.99	.99
7	.485	.622	.741	.879	.974	.99
8	.453	.582	.693	.822	.911	.99
9	.427	.548	.653	.775	.859	.99
10	.405	.520	.620	.736	.815	.977
15	.311	.425	.506	.601	.665	.798
20	.287	.368	.438	.520	.576	.691
25	.256	.329	.392	.465	.515	.618
30	.234	.300	.358	.425	.470	.564
40	.203	.260	.310	.368	.407	.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> adresinden erişilmiştir.

Tablo 2 incelendiğinde, örneğin 5 uzmanın görüşüne başvurulduysa KGO değeri 0,05 anlamlılık düzeyinde .877 değerine eşit ya da bu değerden büyük olmalıdır. Uzman sayısına göre hesaplanan KGO değeri kriterin kapsam geçerliği hakkında bilgi verir. Araştırmada 7 uzmandan görüş alındığı için 0,05 anlamlılık düzeyi için eşik değer 0,622'dir. Bu nedenle KGO değeri 0,622 değerinden büyük eşit olan kriterler esas uygulama formuna dahil edilmiştir. Aşağıda tablo 3'te her madde için hesaplanan KGO değerleri verilmiştir.

Tablo 3

Ölçme Aracının Maddelerine Yönelik Hesaplanan KGO Değerleri

Maddeler	Gerekli/Önemli	Gerekli değil	Yetersiz	KGO
1.Geçme notu	6	1	0	0,714
2.Yemek kartı	6	1	0	0,714
3.Yakıt zammı	7	0	0	1,000
4.Çark çevirme	6	0	1	0,714
Kapsam Geçerlik İndeksi (KGİ)				0,785

Tablo 3'te maddeler ile ilgili hesaplanan KGO değerlerinin eşik değer olan 0,622'den büyük olduğu görülmüştür.

DPA'da her madde için bulunan tam doğru, yakın doğru, kısmen doğru ve yanlış yanıtlar için hesaplanan KGO değerleri Tablo 4'te verilmiştir.

Tablo 4

Dereceli Puanlama Anahtarının Kriterlerine Yönelik Hesaplanan KGO Değerleri

Maddeler	DPA kriterleri	Gerekli/Önemli	Gerekli değil	Yetersiz	KGO
1.Geçme notu	Tam doğru	5	1	1	0,428
	Yakın doğru	6	1	0	0,714
	Kısmen doğru	7	0	0	1,000
	Yanlış yanıtlar	7	0	0	1,000
2.Yemek kartı	Tam doğru	6	1	0	0,714
	Yakın doğru	7	0	0	1,000
	Kısmen doğru	5	2	0	0,428
	Yanlış yanıtlar	7	0	0	1,000
3.Yakıt zammı	Tam doğru	7	0	0	1,000
	Yakın doğru	6	0	1	0,714
	Kısmen doğru	7	0	0	1,000
	Yanlış yanıtlar	7	0	0	1,000
4.Çark çevirme	Tam doğru	7	0	0	1,000
	Yakın doğru	7	0	0	1,000
	Kısmen doğru	7	0	0	1,000
	Yanlış yanıtlar	7	0	0	1,000
Kapsam Geçerlik İndeksi (KGİ)					0,874

Tablo 4'te görüldüğü gibi DPA'daki 1. maddenin tam doğru yanıtları ve 2. maddenin kısmen doğru yanıtlarının KGO değerleri 0,428 olarak bulunmuştur. Bu değer eşik değer olan 0,622 değerinden küçük olduğu için DPA'nın bu kısımları tekrar gözden geçirilip uzman görüşleri doğrultusunda revize edilmiştir. Maddeler ve DPA'ların kriterlerini KGİ değeri KGO değerlerinin ortalaması alınarak hesaplanmış sırasıyla 0,785 ve 0,874 olarak bulunmuştur. Bu değerlerin ikisi de 0,622 değerinden büyüktür. Ölçme aracının kapsam geçerliğinin anlamlı olduğu söylenebilir.

Verilerin Toplanması

Araştırma yapılacak sınıflarda öğrencilere araştırmadan bahsedilmiştir. Bu araştırmaya katılarak bilime sağlayacakları katkıdan bahsedilmiş ve gönüllülük esasına göre istemezlerse araştırmaya katılmak zorunda olmadıkları söylenmiştir. Öğrencilerden gerekli olumlu dönütler alındıktan sonra öğrenci velileri de araştırma ile ilgili bilgilendirilmiş ve velilerden yazılı olarak izin belgesi ile onay alınmıştır. Uygulamanın herhangi bir notla değerlendirilmeyeceği sadece bilimsel bir araştırma için kullanılacağı tekrar hatırlatılmıştır. Veri toplama aşamasında deney ve kontrol grubuna açık uçlu maddenin tanımı, açık uçlu maddenin bireye kazandırdığı kazanımlar ile ilgili bilgi verilmiştir. Çoktan seçmeli maddeler ve açık uçlu maddeler arasındaki farklar anlatılmış ve açık uçlu maddelerin üst düzey becerileri geliştirme avantajlarından bahsedilmiştir. Ayrıca akran değerlendirmeden bahsedilmiş ve öğrencilerin eğer eğitilirse öğretmenler gibi puanlama yapabilecekleri belirtilmiş, öğrencilerin dikkati ve ilgisi çekilmeye çalışılmıştır. Yine her iki gruba da araştırma öncesinde DPA tanıtılmış ve bir soruyu puanlarken DPA'nın nasıl kullanılacağı kısaca gösterilmiştir. Her iki grubun benzer ön deneyime sahip olması sağlanmaya çalışılmıştır. Tüm bu bilgilendirmeler iki ders saatinde yapılmıştır. Öğrencilerin süreç hakkındaki soruları şeffaf bir şekilde yanıtlanmıştır. Sonrasında her iki gruba da esas uygulamada kullanılacak olan 4 soru uygulanmıştır. Daha sonra 63 öğrencinin kağıtları okunup ön test sonuçlarının benzer olduğu görülmüştür (Tablo 17). Benzer sonuçlardan dolayı 10A sınıfı deney grubu, 10B sınıfı kontrol grubu olarak seçilmiştir. Öğrencilere deney ya da kontrol grubu olduklarından bahsedilmemiştir. Ön test sonuçları arasından yüksek performans gösteren, orta düzeyde performans gösteren ve düşük performans gösteren 3 tane kağıt seçilmiştir. Bunun yapılmasındaki amaç öğrencilerin hep iyi ya da hep kötü yanıt içeren kağıtları okumasının önüne geçmektir. Daha sonra bu seçilen 3 öğrencinin kağıdı her iki gruptaki öğrenci sayısı kadar çoğaltılmış ve her iki gruptaki öğrenciler DPA'yı kullanarak 3 akranının kağıtlarını puanlamışlardır. Böylece ön test işlemi tamamlanmıştır.

Son test yapılmadan önce deney grubuna her hafta bir saat olmak üzere yedi saat boyunca puanlayıcı eğitimi verilmiştir. Puanlayıcı eğitiminin genel hatlarından oluşan çizelge EK 3’te sunulmuştur. Eğitimin ilk saatinde açık uçlu madde kavramı detaylıca anlatılmış ve çeşitli örnekler gösterilmiştir. Açık uçlu maddelerde birden çok doğru cevabın olabileceği öğrencilere açıklanmıştır. Ayrıca akran değerlendirmenin öneminden ve kişiye kattığı olumlu kazanımlardan bahsedilmiştir. Eğitimin ikinci haftasında öğrencilere DPA ile ilgili bilgilendirmeler yapılmıştır. Puanlamada objektifliği sağlayabilmek için puanlama anahtarlarının öneminden bahsedilmiş ve örnek puanlama anahtarları gösterilmiştir. Eğitimin üçüncü haftasında ilk iki haftasından anlatılanlar kısaca tekrar edilmiştir ve puanlayıcıların en sık yaptıkları hatalardan bahsedilmiştir. Öğrencilere katılık, cömertlik ve yanlılık davranışları açıklanmıştır. Her bir puanlayıcı davranışı için somut örnekler verilmiş ve bu davranışların tanımlarının iyice pekiştirilmesi sağlanmıştır. Eğitimin dördüncü haftasında örnek uygulamalara geçilmiştir. Örnek uygulamalarda PISA soruları kullanılmıştır. Tahtadan örnek soru açılmış ve öğrencilere bu soruları yanıtlamaları için yeterli süre verilmiştir. Yanıtlama işlemi bittikten sonra sıra arkadaşları ile kağıtlarını değiştirmeleri söylenmiş ve yine tahtaya yansıtılan DPA ile öğrencilerin akranlarını puanlamaları istenmiştir. Daha sonra öğrencilerin akranlarına neden o puanları verdiklerini açıklamaları istenmiş ve sınıfta verilen puanın doğru olup olmadığı tartışılmıştır. Eğitimin beşinci, altıncı, yedinci haftalarında da örnek uygulamalar yaptırılmıştır. Eğitimde örnek uygulamalara daha fazla vakit ayrılmasının sebebi yanıtlara verilen puanlarda ortak paydada buluşmaya çalışılmasıdır. Deney grubundaki öğrencilerin puanlama konusunda hazır olduğu düşünüldüğünde son test uygulamasına geçilmiştir. Son test uygulamasında deney ve kontrol grubundaki öğrenciler ön testte yanıtlarını değerlendirdikleri üç akranının kağıtlarını DPA’yı kullanarak tekrar değerlendirmişlerdir. Bu işlem yaklaşık iki ders saati sürmüştür.

Verilerin Analizi

Araştırma kapsamında veriler, çok yüzeyli Rasch ölçme modeli ve genellenebilirlik kuramı ile analiz edilmiştir. Araştırmada öğrenci, puanlayıcı, cinsiyet, başarı ve madde olmak üzere beş yüzey bulunmaktadır. Çok yüzeyli Rasch analizi için FACETS programı, Genellenebilirlik kuramı için EduG programı kullanılmıştır. Yapılan G kuramı analizleri sonucunda yüzeylerin kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri bulunmuştur. ÇYRÖM analizine başlamadan önce bu analizin gerekli varsayımları karşılayıp karşılamadığına bakılmıştır. Bu varsayımlar yerel bağımsızlık, tek boyutluluk ve model veri uyumu varsayımlarıdır. Tek boyutluluk varsayımı için geliştirilen analitik DPA'nın ölçütlerinin tek bir yapıyı ölçmesi gerekmektedir. Yapılan açımlayıcı faktör analizi (AFA) sonucunda ölçme aracının tek bir yapıyı ölçtüğü bulunmuştur (açıklanan varyans 25,42). Yerel bağımsızlık varsayımı için tek boyutluluk varsayımının sağlanması ön şarttır. Yerel bağımsızlık, ölçülmek istenen yetenek sabit tutulduğunda, kişilerin o maddeye verdikleri cevapların bağımsız olması durumudur (Erkuş vd, 2017). Hambleton ve Swaminathan (2013), tek boyutluluk varsayımının sağlanmasının yerel bağımsızlık varsayımının sağlanması için yeterli olabileceğini söylemişlerdir. Bu nedenle dereceli puanlama anahtarının yerel bağımsızlık varsayımını karşıladığı söylenebilir. Üçüncü varsayım olan model veri uyumu için artık değerler incelenmiştir. Bunun için ± 2 aralığının dışındaki standartlaştırılmış artık değerlerin (z puanı) toplam veri sayısının %5'inden az olması ve ± 3 aralığının dışındaki artık verilerin toplam veri sayısının %1'inden az olması gerekmektedir (Linacre, 2017). Analizde kullanılan toplam veri sayısı $65 \times 3 \times 4 \times 2 \times 5 = 7800$ 'dür. Ön test için ± 2 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı 39 (%0,5) iken ± 3 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı 11 (%0,14) olarak bulunmuştur. Son test uygulamasında ise ± 2 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı 32 (%0,41) iken ± 3 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı 12 (%0,15) olarak bulunmuştur. Ön test ve son test

verilerinin model-veri uyumunu sağladığı görülmüştür. Tüm varsayımlar karşılanıp analiz kısmına geçilmiştir.



BÖLÜM IV

BULGULAR VE YORUMLAR

Bu bölümde araştırma soruları ve problemlerine ilişkin bulgulara ve yorumlara yer verilmiştir.

Birinci Araştırma Sorusuna İlişkin Bulgular ve Yorumlar

Ön test sonucunda deney ve kontrol grubunun yüzeylere ait varyans bileşenleri ve toplam varyans yüzdeleri ile ilgili veriler Tablo 5’te sunulmuştur.

Tablo 5

Ön Test Sonucunda Deney Ve Kontrol Grubuna Ait Varyans Bileşenleri Ve Toplam Varyans Yüzdeleri

Varyans Kaynağı	Deney Grubu		Kontrol Grubu	
	Kestirilen Varyans Bileşeni	Varyans Yüzdesi	Kestirilen Varyans Bileşeni	Varyans Yüzdesi
O	0,414	32,4	0,449	30,9
P	0,024	1,9	0,054	3,8
M	0,266	20,9	0,344	23,7
OxP	0,060	4,8	0,002	0,2
OxM	0,135	10,6	0,201	13,9
PxM	0,013	1,1	0,017	1,2
OxPxM	0,360	28,3	0,381	26,3
Toplam		100		100
G Katsayısı	0.91			
Phi Katsayısı	0.78			

Not: O: Öğrenci, P: Puanlayıcı, M: Madde, OxP: Öğrenci ve puanlayıcı etkileşimi, OxM: Öğrenci ve madde etkileşimi, PxM: Puanlayıcı ve madde etkileşimi, OxPxM: Öğrenci, puanlayıcı ve madde etkileşimini göstermektedir.

Tablo 5 incelendiğinde deney grubunda öğrenci yüzeyi için varyans yüzdesinin 32,4, kontrol grubunda ise 30,9 olduğu görülmektedir. Bu varyans yüzdelерinin yüksek çıkması öğrencilerin ölçüldükleri özelliklerine göre ayrılabilеdikleri anlamına gelir ki bu da istenen bir durumdur. Maddeler için varyans yüzdeleri deney grubu için 20,9, kontrol grubu için 23,7 olarak bulunmuştur. Bu yüzdeleri maddelerin güçlük düzeyleri bakımından farklılık gösterdiğini belirtmektedir. Puanlayıcılara ait varyans yüzdesi deney grubunda 1,9, kontrol grubunda ise 3,8 olarak bulunmuştur. Deney ve kontrol grubunda puanlayıcıların katılık/cömertlik olarak farklı davranış sergiledikleri söylenebilir. Puanlayıcı x madde için kestirilen varyans değeri deney grubunda 1,1, kontrol grubunda ise 1,2'dir. Puanlayıcıların maddeler üzerinde az da olsa farklı puanlamalar yaptıkları söylenebilir. Öğrenci x madde etkileşiminin varyans yüzdesi deney grubunda 10,6 iken kontrol grubunda 13,9 olarak bulunmuştur. Öğrencilerin maddelere göre bağıl konumlarının değışkenlik gösterdiğі söylenebilir. Öğrenci x puanlayıcı ortak etkisine bakıldığında varyans yüzdeleri deney grubunda 4,8, kontrol grubunda ise 0,2 olarak bulunmuştur. Deney grubunda puanlayıcıların öğrencileri puanlarken daha katı/cömert davrandıkları söylenebilir. Kontrol grubunda ise katı/cömert davranan öğrencilerin daha az olduğu yorumu yapılabilir.

Öğrencilerin puanlayıcılar ile madde ortak etkileşimine bakıldığında varyans yüzdelерinin deney ve kontrol gruplarında sırasıyla 28,3 ve 26,3 olduğu görülmektedir. Bu durum ölçülemeyen hata kaynaklarının varlığını gösterebilir.

Puanlayıcı eğitimi sonrasında yapılan son testte deney ve kontrol grubunun yüzeylere ait varyans bileşenleri ve toplam varyans yüzdeleri ile ilgili veriler Tablo 6'da sunulmuştur.

Tablo 6

Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubuna Ait Varyans Bileşenleri ve Toplam Varyans Yüzdeleri

Varyans Kaynağı	Deney Grubu		Kontrol Grubu	
	Kestirilen Varyans Bileşeni	Varyans Yüzdesi	Kestirilen Varyans Bileşeni	Varyans Yüzdesi
O	0,172	12,7	0,163	12,5
P	0,053	3,9	0,067	5,2
M	0,098	7,2	0,177	13,6
OxP	0,019	1,4	0,034	2,7
OxM	0,560	41,2	0,303	23,3
PxM	0,103	7,6	0,195	15,0
OxPxM	0,354	26,0	0,359	27,6
Toplam		100		100
G Katsayısı	0.61			
Phi Katsayısı	0.54			

Not: O:Öğrenci, P:Puanlayıcı, M:Madde, OxP: Öğrenci ve puanlayıcı etkileşimi, OxM: Öğrenci ve madde etkileşimi, PxM: Puanlayıcı ve madde etkileşimi, OxPxM: Öğrenci, puanlayıcı ve madde etkileşimini göstermektedir.

Tablo 6’da deney grubu için varyans yüzdeleri incelendiğinde öğrenci, madde yüzeylerinin ve öğrenci x madde ve öğrenci x puanlayıcı x madde etkileşimlerine ait yüzdelerin azaldığı diğer yüzeylere ait etkileşimlerin arttığı görülmüştür. Puanlayıcı eğitimi sonrasında deney grubunda puanlayıcılara ait varyans yüzdesinin 1,9’dan 3,9’a çıkması puanlayıcı davranışlarının katılık/cömertliği açısından fazlaştığı şeklinde yorumlanabilir. Kontrol grubunda ise öğrenci ve madde yüzeyinde açıklanan varyans yüzdelерinde bir azalma söz konusuyken, diğer yüzey ve etkileşimlerde artma mevcuttur.

İkinci Araştırma Sorusuna İlişkin Bulgular ve Yorumlar

“Öğrenci, puanlayıcı, cinsiyet, başarı ve madde yüzeylerinin ön test ve son test ölçüm raporlarının değişken tablosu, uygunluk değerleri ve model ölçümleri nasıldır?” araştırma sorusuna ilişkin tablolar aşağıda verilmiştir. Öncelikle ön test ve son teste ilişkin değişken haritası (variable map), daha sonra da yüzeylerin ön test ve son testlerinden elde edilen ölçüm raporları gösterilmiştir. Ön teste ait değişken haritası Şekil 1’de gösterilmiştir.

öğrenci (-0.33) olduğu görülmektedir. Üçüncü sütunda puanlayıcılar yer almaktadır. Bu sütunun en üstünde yer alan 42 numaralı puanlayıcının (2.12 logit) en cömert, 45 numaralı puanlayıcının (-1.55 logit) ise en katı olduğu söylenebilir. Bu sütunda yukarıdan aşağıya inildikçe puanlayıcılar daha katı, yukarı doğru çıkıldıkça puanlayıcılar daha cömert olmaktadır. Logit değerlerinin eksi olması puanlayıcıların katı, artı olması cömert olduğu anlamına gelir. Dördüncü sütunda cinsiyet değişkeni yer almaktadır. Puanlayıcılar arasında katılık/cömertlik durumlarının farklılık gösterdiği görülmektedir. Bunun için ölçüm raporları incelenmelidir. 5.sütunda puanlayıcıların matematik dersinde 2022-2023 eğitim-öğretim yılı içerisinde yapılan yazılı notları yer almaktadır. Başarı durumları farklı olan puanlayıcıların genelde benzer puanlamalar yaptıkları görülmüştür. Logit değerleri en düşük -0.07 iken en yüksek 0.08 olarak bulunmuştur. Altıncı sütunda ise maddeler yer almaktadır. Logit değeri artı olan maddeler, logit değeri eksi olan maddelere göre daha kolay maddelerdir. Buna göre en kolay madde 4.madde (1.81 logit), en zor madde ise 1. Madde (-1.52 logit) olarak bulunmuştur.

Son teste ait değişken haritası aşağıda Şekil 2’de gösterilmiştir.

Measr	Ogrenci	Puanlayici	Cinsiyet	Basari	Maddeler	Scale
2	+	+	+	+	+	(4)
	60					
		56 6				
					4	---
1	+	30	+	+	+	+
		29 53				
		43 58 63				
		11 40 42 48 9				
		31				3
		18 5				
		13 2 35 36 41 52				
		37			3	
		12 17 21 22 23 32 49 57 60				
*	0	* 26	* 19 20 39 47 51 59 61 8	* Erkek Kiz	* Iyi Gecer Orta Pekiyi	* --- *
					Gecmez	
		14 38 44 62				
		15 24 25 26 28 45				
		50 7				
		55				
		16 4				2
		27 3 46			2	
		1 30 64			1	
-1	+	+	+	+	+	+
		34				
		33 54				---
		10				
		65				
-2	+	+	+	+	+	(1)
Measr	Ogrenci	Puanlayici	Cinsiyet	Basari	Maddeler	Scale

Şekil 2. Son teste ait değişken haritası

Şekil 2 incelendiğinde en başarılı öğrencinin 60 numaralı (1.50 logit) en başarısız öğrencinin 26 numaralı (-0.02 logit) olduğu görülmektedir. Yine ön testte olduğu gibi erkek ve kız puanlayıcıların puanlama davranışlarında bir farklılık görülmemektedir. Başarı notları incelendiğinde iyi not alan öğrencilerin daha cömert, geçmez not alan öğrencilerin daha katı olduğu söylenebilir. Puanlayıcılar sütununa bakıldığında son testte en cömert puanlayıcıların 56 (1.31 logit) ve 6 (1.25 logit) numaralı puanlayıcılar oldukları görülmektedir. Bu puanlayıcıların matematik dersindeki başarı durumları “iyi” kategorisindedir. En katı puanlayıcılar 65 (-1.35 logit) ve 10 (-1.30) numaralı puanlayıcıdır. Bu puanlayıcıların başarı durumları sırasıyla “orta” ve “geçer” kategorisinde yer almaktadır. Ön test ve son testin

değişken haritaları biçimsel olarak karşılaştırıldığında son testteki dağılımın normal dağılıma daha yakın olduğu görülmektedir.

Aşağıda araştırmada kullanılan yüzeylerin her birine ait ölçüm raporları incelenmiştir.

Sırasıyla önce ön teste ilişkin raporlar daha sonra son teste ilişkin raporlar verilmiştir.

Tablo 7’de ön testte öğrenci yüzeyine ilişkin ölçüm değerleri yer almaktadır.

Tablo 7

Öğrenci Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu

Öğrenci no:	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	İç Uyum	Dış Uyum
60	3,68	3,85	2,88	0,12	1,31	1,04
30	3,31	3,56	1,75	0,10	0,74	1,03
26	2,33	2,22	-0,33	0,09	1,10	1,09
Ortalama	3,11	3,21	1,43	0,10	1,05	1,05
S(Evren)	0,57	0,71	1,33	0,01	0,23	0,03
S(Örneklem)	0,70	0,87	1,63	0,02	0,29	0,03
Model, Evren	RMSE= 0,10 Düzeltilmiş S= 1,33 Ayırma Oranı= 12,70 Ayırma İndeksi= 17,26 Güvenirlilik =0,99					
Model Örneklem	RMSE= 0,10 Düzeltilmiş S= 1,63 Ayırma Oranı = 15,57 Ayırma İndeksi =21,09 Güvenirlilik= 1,00					
Model Ki-Kare (Sabit Etkili):	511,0 sd=2 p=,00 Model Ki-Kare (Normal): 2,0 sd=1 p= 0,16					

Not: S: standart sapma: sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü

Tablo 7’de cevap kağıtları puanlanan 3 öğrencinin ölçüm değerleri verilmiştir. 2,88-(-0,33)=4,21 logit gibi bir aralıkta öğrencilerin yeteneklerinin değiştiği görülmüştür. Öğrencilerin yetenek düzeylerinin ortalaması 1,43 olarak hesaplanmıştır. Örneklem standart sapması 1,63’tür. İç uyum ve dış uyum istatistiklerinin 0,74 ile 1,31 arasında değiştiği ve ortalamasının 1,05 olduğu bulunmuştur. İç uyum ve dış uyum değerlerinin 1 olması model ve veri arasındaki uyumun çok iyi olduğunu gösterir. 0,50 ile 1,50 arasında olması kabul edilebilir bir durumdur (Alharby, 2006). Öğrenci yüzeyi için model veri uyumunun iyi olduğu söylenebilir. Ayırma istatistiklerine bakıldığında ise ayırma oranının 15,57, ayırma indeksinin 21,09 olduğu görülür. Ayırma indeksinin güvenirliliği 1,00’dir. Bu değerlere göre öğrencilerin yetenek düzeylerinin birbirinden farklı olduğu ve yeteneklerine göre iyi bir şekilde ayırt edildiği söylenebilir. Ki-kare değeri öğrencilerin yetenek düzeylerinin birbirinden farklılaşıp farklılaşmadığını gösteren bir istatistiktir. Bu değer istatistiksel olarak anlamlı olduğu belirlenmiştir ($X^2 = 511,0$, $sd=2$, $p=0,00<0,05$).

Tablo 8’de puanlayıcı yüzeyinin ön teste göre ölçüm raporu verilmiştir.

Tablo 8

Puanlayıcı Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu

Puanlayıcı No	Gözlenen Uyum	Beklenen Uyum	Logit Değeri	Standart Hata	İç Uyum	Dış Uyum	Rasch-Kappa	t değeri
D1	1,92	2,14	-0,52	0,44	2,54	8,97	-0,18	-1,18
D2	2,33	2,65	0,56	0,50	1,06	0,69	0,17	1,12
D3	2,00	2,26	-0,32	0,45	0,68	0,54	0,27	-0,71
D4	1,75	1,88	-0,90	0,43	0,83	0,61	0,05	-2,09
D5	2,25	2,57	0,32	0,48	0,43	1,57	-0,08	0,67
D6	2,50	2,79	1,10	0,55	0,90	0,63	0,15	2,00
D7	1,92	2,14	-0,52	0,44	3,75	4,21	-0,39	-1,18
D8	1,75	1,88	-0,90	0,43	0,58	0,45	0,15	-2,09
D9	2,58	2,85	1,42	0,58	0,79	0,67	0,05	2,45
D10	1,58	1,61	-1,26	0,43	1,67	1,86	0,07	-2,93
D11	1,83	2,01	-0,71	0,43	1,94	1,65	-0,04	-1,65
D12	2,25	2,57	0,32	0,48	0,76	0,51	0,22	0,67
D13	2,00	2,26	-0,32	0,45	0,57	0,50	0,08	-0,71
D14	2,42	2,73	0,82	0,52	1,35	0,78	0,10	1,58
D15	2,17	2,48	0,09	0,47	0,73	0,63	0,27	0,19
D16	2,08	2,37	-0,12	0,46	0,85	0,59	0,32	-0,26
D17	2,08	2,37	-0,12	0,46	0,51	0,40	0,19	-0,26
D18	2,33	2,65	0,56	0,50	0,23	0,20	0,41	1,12
D19	2,17	2,48	0,09	0,47	0,49	0,39	0,37	0,19
D20	1,83	2,01	-0,71	0,43	0,38	0,63	0,14	-1,65
D21	2,42	2,73	0,82	0,52	0,44	2,47	0,05	1,58
D22	2,17	2,48	0,09	0,47	0,48	0,38	0,45	0,19
D23	2,17	2,48	0,09	0,47	0,33	0,27	0,24	0,19
D24	2,00	2,26	-0,32	0,45	1,90	2,07	-0,08	-0,71
D25	2,25	2,57	0,32	0,48	0,83	0,60	0,03	0,67
D26	1,83	2,01	-0,71	0,43	1,22	0,86	0,20	-1,65
D27	2,00	2,26	-0,32	0,45	0,53	1,02	-0,12	-0,71
D28	2,42	2,73	0,82	0,52	1,07	0,69	0,13	1,58
D29	2,25	2,57	0,32	0,48	0,91	1,82	-0,01	0,67
D30	2,50	2,79	1,10	0,55	1,30	1,45	0,09	2,00
D31	2,50	2,79	1,10	0,55	1,30	1,45	0,05	2,00
D32	2,33	2,65	0,56	0,50	0,50	0,34	0,20	1,12
K1	1,67	1,75	-1,08	0,43	0,81	0,58	0,21	-2,51
K2	2,08	2,37	-0,12	0,46	0,94	0,77	0,03	-0,26
K3	1,92	2,14	-0,52	0,44	0,73	0,66	-0,05	-1,18
K4	2,42	2,73	0,82	0,52	0,51	0,38	0,02	1,58
K5	2,00	2,26	-0,32	0,45	0,86	0,70	0,21	-0,71
K6	2,00	2,26	-0,32	0,45	0,76	0,58	0,25	-0,71
K7	2,33	2,65	0,56	0,50	0,50	0,37	0,29	1,12
K8	2,25	2,57	0,32	0,48	0,87	0,56	0,23	0,67
K9	2,50	2,79	1,10	0,55	0,84	0,79	0,03	2,00
K10	2,75	2,93	2,19	0,68	0,31	0,13	0,09	3,22
K11	2,42	2,73	0,82	0,52	0,78	0,49	0,07	1,58
K12	1,75	1,88	-0,90	0,43	2,35	1,79	0,07	-2,09
K13	1,50	1,47	-1,55	0,43	1,03	-0,79	0,15	-3,60
K14	1,58	1,61	-1,26	0,43	1,84	2,35	-0,03	-2,93
K15	2,08	2,37	-0,12	0,46	0,65	0,54	0,28	-0,26
K16	2,42	2,73	0,82	0,52	0,50	0,31	0,26	1,58
K17	1,67	1,75	-1,08	0,43	1,99	1,37	0,20	-2,51
K18	1,67	1,75	-1,08	0,43	0,80	0,77	0,08	-2,51
K19	2,25	2,57	0,32	0,48	1,30	0,96	0,08	0,67
K20	1,92	2,14	-0,52	0,44	1,20	2,87	-0,01	-1,18
K21	1,92	2,14	-0,52	0,44	0,59	0,47	0,16	-1,18
K22	2,17	2,48	0,09	0,47	0,73	0,63	0,25	0,19
K23	1,83	2,01	-0,71	0,43	0,94	0,67	0,20	-1,65
K24	2,42	2,73	0,82	0,52	1,48	0,85	0,16	1,58
K25	2,08	2,37	-0,12	0,46	0,72	1,28	0,45	-0,26

Puanlayıcı Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu (devamı)

K26	2,17	2,48	0,09	0,47	1,26	1,08	0,08	0,19
K27	2,08	2,37	-0,12	0,46	0,55	0,46	0,16	-0,26
K28	2,17	2,48	0,09	0,47	0,77	0,54	0,10	0,19
K29	2,17	2,48	0,09	0,47	1,54	1,17	0,27	0,19
K30	2,33	2,65	0,56	0,50	1,15	0,82	0,18	1,12
K31	1,75	1,88	-0,90	0,43	1,60	1,31	0,08	-2,09
K32	1,75	1,88	-0,90	0,43	0,91	0,75	0,11	-2,09
K33	2,33	2,65	0,56	0,50	0,57	0,41	0,22	1,12
Ortalama	2,11	2,36	0,00	0,47	0,99	1,05		
S(Evren)	0,28	0,35	0,74	0,05	0,61	0,21		
S(Örneklem)	0,28	0,36	0,75	0,05	0,61	1,22		

Model, Evren RMSE= 0,48 Düzeltilmiş S= 0,57 Ayırma Oranı = 1,20
Ayırma İndeksi= 1,93 Güvenirlik =0,59

Model, Örneklem RMSE= 0,48 Düzeltilmiş S=0,58 Ayırma Oranı= 1,21
Ayırma İndeksi=1,95 Güvenirlik=0,60

Model Ki-Kare (Sabit Etkili): 147,1 sd=64 p= ,00 Model Ki-Kare (Normal) : 47,0 sd =63 p=,93

Not: S: standart sapma sd: serbestlik derecesi, RMSE:hata kareleri ortalamasının karekökü D:deney grubu
K:kontrol grubu

Tablo 8’de deney grubunda 32, kontrol grubunda 33 olmak üzere toplam 65 puanlayıcının logit ölçüleri, ölçmenin standart hatası, uyum değerleri, rasch-kappa ve t değerleri verilmiştir. Puanlayıcılar arası güvenilirlik ile ilgili bir istatistik olan Kappa değeri Rasch analiz çıktısında yer almamaktadır. Bu nedenle ilgili sütundaki kappa değerleri aşağıdaki formül yardımıyla her puanlayıcı için hesaplanmıştır.

$$\text{Kappa değeri} = \frac{\text{Gözlenen Uyum Yüzdesi} - \text{Beklenen Uyum Yüzdesi}}{100 - \text{Beklenen Uyum Yüzdesi}} \quad (12)$$

Ayrıca yine Rasch analizi çıktısında yer almayan t-değeri aşağıdaki formül ile hesaplanmıştır. Puanlayıcıların katılık ve cömertlik davranışları gösterip göstermediğinin belirlenmesinde t değerinin hesaplanması gerekir (Myford ve Wolfe, 2003).

$$t \text{ değeri} = \frac{\text{puanlayıcının logit ölçüsü} - \text{logit değerleri ortalaması}}{\text{logit ölçüsü standart hatası}} \quad (13)$$

Puanlayıcıların logit ölçülerinin -1,45 ile 2,19 arasında değiştiği ve puanlayıcıların katılık/cömertliklerine ilişkin aralığın ise 2,19-(-1,45) =3,64 logit olduğu görülmektedir. Bu değer yüksek olması puanlayıcıların performanslarının geniş bir aralığa yayıldığını gösterir. Tabloda logit değerlerinin eksi olması puanlayıcıların katılık davranışı

sergilediğini, artı olması ise cömert davrandıklarını göstermektedir. Uyum istatistiklerine bakıldığında uyum içi ve uyum dışı değer ortalamalarının sırasıyla 0,99 ve 1,05 olduğu belirlenmiştir. Model veri uyumuna uygun oldukları söylenebilir. Ayırma istatistiklerine bakıldığında ise ayırma oranı 1,21, ayırma indeksi 1,95 ve ayırma indeksinin güvenilirliği 0,60 olarak bulunmuştur. Bu değerlerin düşük çıkması puanlayıcıların aynı sorulara aynı puanları veremediğini gösterir. Bu zaten beklenen bir durumdur. Puanlayıcılar maddeleri puanlarken birbirlerinden farklı davranmışlardır. Puanlayıcı yüzeyi için grup düzeyinde istatistiksel bir farklılık olup olmadığına bakmak için sabit etkili ki-kare değeri incelenmiştir. $X^2=147,1$, $sd=64$, $p=0,00<0,05$ olarak bulunmuştur. Buna göre ki-kare değerinin anlamlı olduğu görülür. Deney ve kontrol grubundaki puanlayıcıların grup düzeyinde puanlayıcı davranışlarının farklılaştığı söylenebilir.

Tablo 9’da cinsiyet yüzeyine ilişkin ön test sonuçları verilmiştir.

Tablo 9

Cinsiyet Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu

Cinsiyet	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	İç Uyum	Dış Uyum
Erkek	3,06	3,40	-0,04	0,11	1,13	1,57
Kız	3,12	3,44	0,04	0,07	0,98	0,85
Ortalama	3,09	3,42	0,00	0,09	1,05	1,21
S(Evren)	0,03	0,02	0,04	0,02	0,07	0,36
S(Örnekleme)	0,05	0,03	0,06	0,03	0,11	0,51
Model, Evren	RMSE= 0,09 Düzeltilmiş S= 0,00 Ayırma Oranı = 0,00 Ayırma İndeksi= 0,33 Güvenirlik =0,00					
Model, Örnekleme	RMSE= 0,09 Düzeltilmiş S=0,00 Ayırma Oranı= 0,00 Ayırma İndeksi=0,33 Güvenirlik=0,00					
Model Ki-Kare (Sabit Etkili):	0,50 sd=1 p= 0,49					

Not: S: standart sapma: sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü

Tablo 9 doğrultusunda erkek ve kız puanlayıcıların iç ve dış uyum değerleri normal aralıklardadır. Logit değerlerine bakıldığında erkek puanlayıcıların kız puanlayıcılara göre biraz daha katı davrandıkları söylenebilir. (Erkekler için logit değeri -0,04, kızlar için logit değeri 0,04). Erkek ve kızlar için grup düzeyindeki farklılık olup olmadığına bakmak için sabit etkili ki kare değerine bakılır. $X^2=0,50$, $sd=1$ ve $p=0,05<0,49$ olarak bulunmuştur. Erkek ve kızların puanlayıcı davranışlarının istatistiksel olarak farklılaştığı söylenemez. Benzer davranışlar sergilenmişlerdir.

Tablo 10’da başarı yüzeyine ilişkin ön test sonuçları verilmiştir.

Tablo 10

Başarı Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu

Başarı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	İç Uyum	Dış Uyum
Pekiyi	3,08	3,38	-0,05	0,27	0,85	1,01
İyi	3,08	3,41	-0,02	0,14	0,87	0,92
Orta	3,17	3,46	0,08	0,13	0,84	0,87
Geçer	3,05	3,39	-0,07	0,10	1,01	0,89
Geçmez	3,13	3,45	0,05	0,11	1,29	1,44
Ortalama	3,11	3,42	0,00	0,15	0,97	1,03
S(Evren)	0,04	0,03	0,06	0,06	0,17	0,21
S(Örnekleme)	0,05	0,03	0,07	0,07	0,19	0,24
Model, Evren	RMSE= 0,16 Düzeltilmiş S= 0,00 Ayırma Oranı = 0,00 Ayırma İndeksi= 0,33 Güvenirlilik =0,00					
Model, Örnekleme	RMSE= 0,16 Düzeltilmiş S=0,00 Ayırma Oranı= 0,00 Ayırma İndeksi=0,33 Güvenirlilik=0,00					
Model Ki-Kare (Sabit Etkili): 1,1 sd=4 p= 0,89 Model Ki-Kare (Normal) : 0,9 sd 3 p=0,83						

Not: S: standart sapma: sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü

Tablo 10’deki 65 puanlayıcının 2022-2023 eğitim öğretim yılında matematik dersinin 1.yazılısından aldıkları notlara göre pekiyi (85-100 arası), iyi (70-84 arası), orta (60- 69 arası), geçer (50-59 arası) ve geçmez (0-49 arası) olarak sınıflandırılmasıyla oluşan ölçümler görülmektedir. Geçmez ve orta not alanların daha cömert, geçer ve pekiyi not alanların daha katı puanlayıcı oldukları söylenebilir. İç uyum ve dış uyum değerleri uygun aralıktadır. Ayırma oranı ve ayırma indeksi değerlerine bakıldığında puanlayıcıların okulda aldıkları matematik notlarının yüksek ve düşüklüğüne göre puanlayıcı davranışlarının birbirinden farklılaşmadığı görülmektedir. Deney ve kontrol grubu için grup düzeyinde farklılaşma olup olmadığına bakmak için X^2 değeri incelendiğinde grupların birbirlerinden farklılaşmadığı görülmektedir ($X^2= 1,1$, sd=4 , p=0,89 > 0,00).

Son olarak Tablo 11’de madde yüzeyine ilişkin ön test sonuçlarına yer verilmiştir.

Tablo 11

Madde Yüzeyinin Ön Test Ölçümlerine İlişkin Ölçüm Raporu

Madde(Soru)	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	İç Uyum	Dış Uyum
1.Madde	2,48	2,40	-1,52	0,10	1,19	1,29
2.Madde	2,72	2,75	-1,05	0,10	0,76	0,77
3.Madde	3,48	3,71	0,77	0,13	0,84	0,53
4.Madde	3,75	3,89	1,81	0,16	1,57	1,62
Ortalama	3,11	3,19	0,00	0,12	1,08	1,05
S(Evren)	0,52	0,63	1,35	0,02	0,31	0,43
S(Örnekleme)	0,60	0,73	1,56	0,03	0,36	0,49
Model, Evren	RMSE= 0,12 Düzeltilmiş S= 1,34 Ayırma Oranı = 10,79 Ayırma İndeksi= 14,72 Güvenirlilik =0,99					
Model, Örnekleme	RMSE= 0,12 Düzeltilmiş S= 1,55 Ayırma Oranı= 12,47 Ayırma İndeksi=16,96 Güvenirlilik=0,99					
Model Ki-Kare (Sabit Etkili):	441,0 sd=3 p= 0,00 Model Ki-Kare (Normal) : 3,0 sd =2 p=0,22					

Not: S: standart sapma: sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü

Tablo 11 incelendiğinde başarı testindeki maddelerin güçlük düzeylerinin -1,52 ile 1,81 arasında değiştiği görülmektedir. Buradaki logit değerinin artı olması maddenin güçlük düzeyinin kolay, eksi olması maddenin zor olduğunu gösterir. Buna göre en kolay maddenin 4. madde, en zor maddenin 1.madde olduğu söylenebilir. Maddelerin logit değerleri ortalaması 0,00 ve standart sapması 1,55 olarak bulunmuştur. İç uyum değerlerinin 0,76 ile 1,57 arasında, dış uyum değerlerinin 0,53 ile 1,62 değerleri arasında değiştiği görülmektedir. İç ve dış uyum değerlerinin ortalamaları sırasıyla 1,08 ve 1,05 olarak bulunduğu için maddeler kabul edilebilir uygunluk değerlerine sahiptir. Ayrıca maddeler yüzeyine ait ayırma oranının 12,47, ayırma indeksinin 16,96 ve ayırma indeksi güvenirliliğinin 0,99 çıkması maddelerin birbirinden güçlüklerine göre ayırt edilebildiğini göstermektedir. Ayrıca maddelerin güçlük düzeylerinin birbirinden ayırt edilip edilmediğine bakmak için X^2 değeri incelenmiştir. $X^2=441,0$, $sd=3$, $p=0,00 < 0,05$ olarak bulunmuştur. Bu değerler istatistiksel olarak anlamlıdır.

Araştırmaya ilişkin değişken tablosu ve öğrenci puanlayıcı, cinsiyet, başarı ve madde yüzeylerine ilişkin ölçüm raporları incelendikten sonra son test ölçümlerinin incelenmesine geçilmiştir. Aşağıda yüzeyler için son test ölçüm raporları yer almaktadır.

İlk olarak Tablo 12’de öğrenci yüzeyine ilişkin son test ölçüm raporu sunulmuştur.

Tablo 12

Öğrenci Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu

Öğrenci no:	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	İç Uyum	Dış Uyum
26	2,44	2,50	-0,02	0,07	1,10	1,10
30	3,16	3,39	0,97	0,08	0,90	1,26
60	3,46	3,64	1,50	0,09	1,05	0,84
Ortalama	3,02	3,18	0,82	0,08	1,02	1,06
S(Evren)	0,43	0,49	0,63	0,01	0,09	0,17
S(Örnekleme)	0,52	0,60	0,78	0,01	0,11	0,21
Model, Evren	RMSE= 0,08 Düzeltilmiş S= 0,63 Ayırma Oranı = 7,86 Ayırma İndeksi= 10,81 Güvenirlilik =0,98					
Model, Örnekleme	RMSE= 0,08 Düzeltilmiş S= 0,77 Ayırma Oranı= 9,65 Ayırma İndeksi=13,20 Güvenirlilik=0,98					
Model Ki-Kare (Sabit Etkili):	193,3 sd=2 p= 0,00 Model Ki-Kare (Normal) : 2,0 sd=1 p=0,16					

Not: S: standart sapma: sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü

Tablo 12'ye bakıldığında iç uyum ve dış uyum değerlerinin bire yakın değerler olduğu ve model veri uyumunun iyi olduğu söylenebilir. Örnekleme için hesaplanan ayırma oranı, ayırma indeksi ve ayırma indeksinin güvenilirliğinin yüksek olduğu ve öğrenci performanslarını birbirinden iyi bir şekilde ayırt edilebildiği görülmektedir. İstatistiksel olarak öğrenci performanslarının ayırt edilip edilmediğine bakmak için ki-kare değerine bakıldığında bu değer anlamlı olduğu görülmüştür ($X^2=193,3$, $sd=2$, $p=0,00<0,05$). Logit değerleri incelendiğinde değerlerin -0,02 ile 1,50 arasında değiştiği görülür. Logit değerleri arasındaki fark ise 1,52'dir. Bir logit değerinin eksi olması o öğrencinin performansının düşük, artı olması performansının yüksek olduğunu gösterir. 60 numaralı öğrencinin performansı en yüksek, 26 numaralı öğrencinin performansı en düşüktür.

Puanlayıcı yüzeyinin son test ölçüm değerleri aşağıdaki Tablo 13'te verilmiştir.

Tablo 13

Puanlayıcı Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu

Puanlayıcı No	Gözlenen Uyum	Beklenen Uyum	Logit Değeri	Standart Hata	İç Uyum	Dış Uyum	Rasch-Kappa	t-değeri
D1	2,50	2,57	-0,77	0,34	1,26	1,45	0,15	-2,26
D2	3,25	3,49	0,33	0,38	1,01	0,71	0,25	0,87
D3	2,58	2,79	-0,57	0,34	1,10	0,98	0,21	-1,68
D4	2,75	2,90	-0,45	0,34	0,86	0,73	0,10	-1,32
D5	3,33	3,54	0,44	0,39	0,66	0,47	0,31	1,13
D6	3,67	3,80	1,25	0,51	1,10	0,81	0,11	2,45
D7	2,83	3,03	-0,32	0,34	0,62	0,65	-0,03	-0,94
D8	3,08	3,30	0,02	0,36	1,28	0,96	0,34	0,06
D9	3,42	3,59	0,55	0,41	0,52	0,41	0,40	1,34
D10	2,08	2,03	-1,35	0,35	0,91	1,05	0,07	-3,86
D11	3,33	3,60	0,57	0,39	0,97	1,18	0,09	1,46
D12	3,17	3,35	0,10	0,37	0,84	0,69	0,33	0,27

Puanlayıcı Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu (devamı)

D13	3,25	3,49	0,33	0,38	0,59	0,45	0,19	0,87
D14	2,92	3,21	-0,10	0,35	0,73	0,62	0,22	-0,29
D15	2,83	3,11	-0,22	0,34	0,40	0,52	0,11	-0,65
D16	2,75	2,84	-0,51	0,34	1,78	2,26	0,03	-1,50
D17	3,08	3,36	0,11	0,36	1,90	1,62	0,06	0,31
D18	3,33	3,52	0,39	0,39	0,66	0,47	0,46	1,00
D19	3,08	3,26	-0,03	0,36	1,28	0,96	0,34	-0,08
D20	3,00	3,27	-0,02	0,35	1,22	1,03	0,07	-0,06
D21	3,08	3,32	0,06	0,36	1,20	1,36	-0,03	0,17
D22	3,17	3,35	0,10	0,37	0,88	0,65	0,38	0,27
D23	3,17	3,35	0,10	0,37	0,45	0,44	0,30	0,27
D24	2,92	3,14	-0,19	0,35	0,86	0,75	0,31	-0,54
D25	2,92	3,12	-0,22	0,35	1,67	4,03	0,07	-0,63
D26	2,83	3,11	-0,22	0,34	0,98	1,05	0,10	-0,65
D27	2,58	2,75	-0,60	0,34	0,40	0,66	0,03	-1,76
D28	2,92	3,14	-0,19	0,35	0,86	0,75	0,31	-0,54
D29	3,50	3,68	0,79	0,43	1,26	1,49	0,15	1,84
D30	2,42	2,55	-0,79	0,34	0,37	0,35	0,04	-2,32
D31	3,33	3,58	0,53	0,39	1,61	1,05	0,14	1,36
D32	3,08	3,32	0,06	0,36	1,68	4,43	0,12	0,17
K1	2,25	2,27	-1,06	0,34	1,03	0,86	0,14	-3,12
K2	2,25	2,31	-1,02	0,34	0,54	0,70	-0,13	-3,00
K3	3,25	3,48	0,32	0,38	0,57	0,46	0,30	0,84
K4	3,25	3,48	0,32	0,38	0,52	0,51	0,15	0,84
K5	3,17	3,44	0,24	0,37	0,63	0,50	0,17	0,65
K6	3,00	3,23	-0,07	0,35	0,55	0,51	0,22	-0,20
K7	3,08	3,30	0,02	0,36	0,67	0,82	0,11	0,06
K8	3,42	3,61	0,60	0,41	0,99	0,68	0,25	1,46
K9	3,25	3,48	0,32	0,38	0,99	0,85	0,15	0,84
K10	3,33	3,60	0,57	0,39	0,56	0,57	0,13	1,46
K11	3,42	3,66	0,73	0,41	1,39	1,52	0,06	1,78
K12	2,92	3,18	-0,14	0,35	1,34	1,98	-0,07	-0,40
K13	2,92	3,14	-0,19	0,35	2,05	3,90	-0,02	-0,54
K14	2,75	2,75	-0,61	0,34	1,41	1,34	0,22	-1,79
K15	3,00	3,30	0,02	0,35	1,36	1,13	0,08	0,06
K16	3,42	3,61	0,60	0,41	2,21	2,10	0,01	1,46
K17	3,08	3,36	0,11	0,36	0,98	0,77	0,15	0,31
K18	2,75	3,01	-0,34	0,34	1,97	2,50	-0,05	-1,00
K19	3,00	3,30	0,02	0,35	0,49	0,50	0,12	0,06
K20	3,33	3,47	0,30	0,39	0,50	0,38	0,07	0,77
K21	3,58	3,69	0,84	0,46	1,46	1,10	0,23	1,83
K22	2,25	2,27	-1,06	0,34	0,51	0,66	-0,16	-3,12
K23	2,83	3,00	-0,35	0,34	1,22	1,07	0,17	-1,03
K24	3,67	3,81	1,31	0,51	1,57	0,70	0,10	2,57
K25	3,17	3,35	0,10	0,37	1,30	1,28	0,39	0,27
K26	3,50	3,66	0,73	0,43	0,65	0,55	0,32	1,70
K27	3,00	3,27	-0,02	0,35	0,88	0,75	0,19	-0,06
K28	3,08	3,32	0,06	0,36	0,63	0,50	0,31	0,17
K29	3,08	3,30	0,02	0,36	0,88	0,74	0,26	0,06
K30	2,92	3,18	-0,14	0,35	0,76	0,69	0,20	-0,40
K31	3,42	3,66	0,73	0,41	1,84	1,13	0,08	1,78
K32	2,58	2,51	-0,84	0,34	0,93	0,84	0,21	-2,47
K33	2,08	1,98	-1,35	0,35	0,91	1,05	-0,13	-3,86
Ortalama	3,2	3,21	0,00	0,37	1,03	1,06		
S(Evren)	0,36	0,42	0,56	0,04	0,46	0,81		
S(Örnekleme)	0,37	0,42	0,56	0,04	0,46	0,82		

Model, Evren RMSE= 0,37 Düzeltilmiş S= 0,41 Ayırma Oranı = 1,11

Ayırma İndeksi= 1,82 Güvenirlik =0,55

Model, Örnekleme RMSE= 0,37 Düzeltilmiş S= 0,42 Ayırma Oranı= 1,13

Ayırma İndeksi=1,84 Güvenirlik=0,56

Model Ki-Kare (Sabit Etkili): 137,5 sd=64 p= 0,00 Model Ki-Kare (Normal) : 45,0 sd=63 p=0,96

Not: S: standart sapma: sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü D: deney grubu, K:kontrol grubu

Tablo 13'e bakıldığında puanlayıcıların logit değerlerinin -1,35 ile 1,31 arasında değiştiği görülmektedir. Logit değerlerinin aralığı $1,31 - (-1,35) = 2,66$ olarak bulunur. Bu değer puanlayıcıların katılık/cömertlik davranışlarının aralığıdır. Logit değeri artı olan puanlayıcıların cömertlik davranışı gösterdiği, logit değeri eksi değer olan puanlayıcıların katılık davranışı sergilediği söylenebilir. İç uyum ve dış uyum değerlerinin ortalamaları sırasıyla 1,03 ve 1,06 olarak bulunmuştur. Ayırma istatistiklerine bakıldığında puanlayıcıların birbirlerinden çok farklı puanlar verdikleri söylenemez. Bu nedenle güvenilirlik düşük çıkmış denilebilir. Puanlayıcı yüzeyinde analiz sonucundaki güvenilirlik aslında puanlayıcılar arasındaki güvenilirliği değil, farklılığın güvenilirliğini verir (Haiyang,2010). Grup düzeyinde farklılaşma olup olmadığına bakmak için X^2 değerine bakılır. $X^2 = 137,5$ sd= 64 $p = 0,00 < 0,05$ olarak bulunmuştur. Deney ve kontrol grubundaki puanlayıcıların grup düzeyinde farklılaştığı söylenebilir.

Tablo 14'te cinsiyet yüzeyinin son test ölçümlerine yer verilmiştir.

Tablo 14

Cinsiyet Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu

Cinsiyet	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	İç Uyum	Dış Uyum
Erkek	3,11	3,32	0,05	0,09	1,01	0,91
Kız	2,98	3,25	-0,05	0,05	1,02	1,12
Ortalama	3,05	3,28	0,00	0,07	1,02	1,02
S(evren)	0,06	0,03	0,05	0,02	0,01	0,11
S(örneklem)	0,09	0,05	0,07	0,02	0,01	0,15
Model, Evren	RMSE= 0,07 Düzeltilmiş S= 0,00 Ayırma Oranı = 0,00 Ayırma İndeksi= 0,33 Güvenirlik =0,00					
Model, Örneklem	RMSE= 0,07 Düzeltilmiş S= 0,00 Ayırma Oranı= 0,00 Ayırma İndeksi=0,33 Güvenirlik=0,00					
Model Ki-Kare (Sabit Etkili): 0,90 sd=1 p= 0,35						

Not: S: standart sapma: sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü

Tablo 14'te son teste ait cinsiyet yüzeyinin ölçüm değerleri incelendiğinde uyum değerlerin uygun aralıklarda olduğu görülür. Logit ölçülerine bakıldığında -0,05 logit ölçüsüne sahip erkeklerin 0,05 logit ölçüsüne sahip kızlardan daha cömert oldukları söylenebilir. Ancak ayırma istatistiklerinin düşük çıkması ve istatistiksel anlamlılık için bakılan ki-kare değerinin de düşük çıkmasında dolayı erkek ve kızların puanlayıcı davranışlarının benzer olduğu görülür ($X^2 = 0,90$, sd= 1, $p = 0,35 > 0,05$).

Tablo 15’te başarı yüzeyinin son test ölçüm raporu verilmiştir.

Tablo15

Başarı Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu

Başarı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	İç Uyum	Dış Uyum
Pekiyi	3,06	3,30	0,03	0,21	1,26	2,08
İyi	3,15	3,35	0,09	0,11	0,98	0,86
Orta	3,04	3,29	0,01	0,10	1,05	1,32
Geçer	2,98	3,25	-0,05	0,08	1,09	0,99
Geçmez	2,95	3,22	-0,08	0,08	0,92	0,93
Ortalama	3,04	3,28	0,00	0,12	1,06	1,23
S(Evren)	0,07	0,04	0,06	0,05	0,12	0,45
S(Örnekleme)	0,08	0,05	0,07	0,05	0,13	0,51
Model, Evren	RMSE= 0,13 Düzeltilmiş S= 0,00 Ayırma Oranı = 0,00 Ayırma İndeksi= 0,33 Güvenirlik =0,00					
Model, Örnekleme	RMSE= 0,13 Düzeltilmiş S=0,00 Ayırma Oranı= 0,00 Ayırma İndeksi=0,33 Güvenirlik=0,00					
Model Ki-Kare (Sabit Etkili): 1,8 sd=4 p= 0,78 Model Ki-Kare (Normal) : 1,2 sd 3 p=0,74						

Not: S: standart sapma: sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü

Tablo 15’te deney ve kontrol grubu öğrencilerinin 2022-2023 eğitim öğretim yılı ikinci yarısında matematik dersinden aldıkları notlara göre gruplandırılmış durumlarına göre son test ölçümleri yer almaktadır. Tabloya göre uyum değerlerinin normal aralıkta olduğu söylenebilir. Logit değerlerinin -0,08 ile 0,09 arasında değiştiği görülmektedir. Ayırma istatistikleri incelendiğinde puanlayıcıların davranışlarının matematik dersindeki başarılarına göre ayırt edilmediği görülmektedir. Sabit etkili ki-kare değerinin $X^2=1,8$, $sd=4$, $p=0,078>0,05$ olarak bulunması istatistiksel olarak başarı düzeylerine göre puanlayıcıların farklı davranışlar sergilemediklerini göstermiştir. Alan yazın incelendiğinde Güler, Taşdelen Teker ve Şahin (2016)’ın yaptığı akran değerlendirme araştırmasında puanlayıcıların başarı düzeylerine göre puanlayıcı davranışları incelenmiş başarı düzeyleri düşük olan öğrencilerin daha cömert, başarı düzeyi yüksek olan öğrencilerin daha katı oldukları bulunmuştu. Bu çalışmada herhangi bir fark bulunamamıştır.

Son olarak Tablo 16’da madde yüzeyine ait son test ölçümleri gösterilmiştir.

Tablo 16

Madde Yüzeyinin Son Test Ölçümlerine İlişkin Ölçüm Raporu

Madde(Soru)	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	İç Uyum	Dış Uyum
1.Madde	2,54	2,59	-0,75	0,08	1,23	1,32
2.Madde	2,63	2,71	-0,64	0,08	0,80	0,83
3.Madde	3,20	3,38	0,15	0,09	0,96	0,81
4.Madde	3,70	3,80	1,25	0,13	1,17	1,30
Ortalama	3,02	3,12	0,00	0,10	1,04	1,06
S(Evren)	0,47	0,49	0,80	0,02	0,17	0,25
S(Örnekleme)	0,54	0,57	0,93	0,02	0,20	0,28
Model, Evren	RMSE= 0,10 Düzeltilmiş S= 0,80 Ayırma Oranı = 8,07 Ayırma İndeksi= 11,10 Güvenirlik =0,98					
Model, Örnekleme	RMSE= 0,10 Düzeltilmiş S= 0,92 Ayırma Oranı= 9,34 Ayırma İndeksi=12,79 Güvenirlik=0,99					
Model Ki-Kare (Sabit Etkili):	210,8 sd=3 p= 0,00 Model Ki-Kare (Normal) : 3,0 sd =2 p=0,23					

Not: S: standart sapma: sd: serbestlik derecesi, RMSE: hata kareleri ortalamasının karekökü

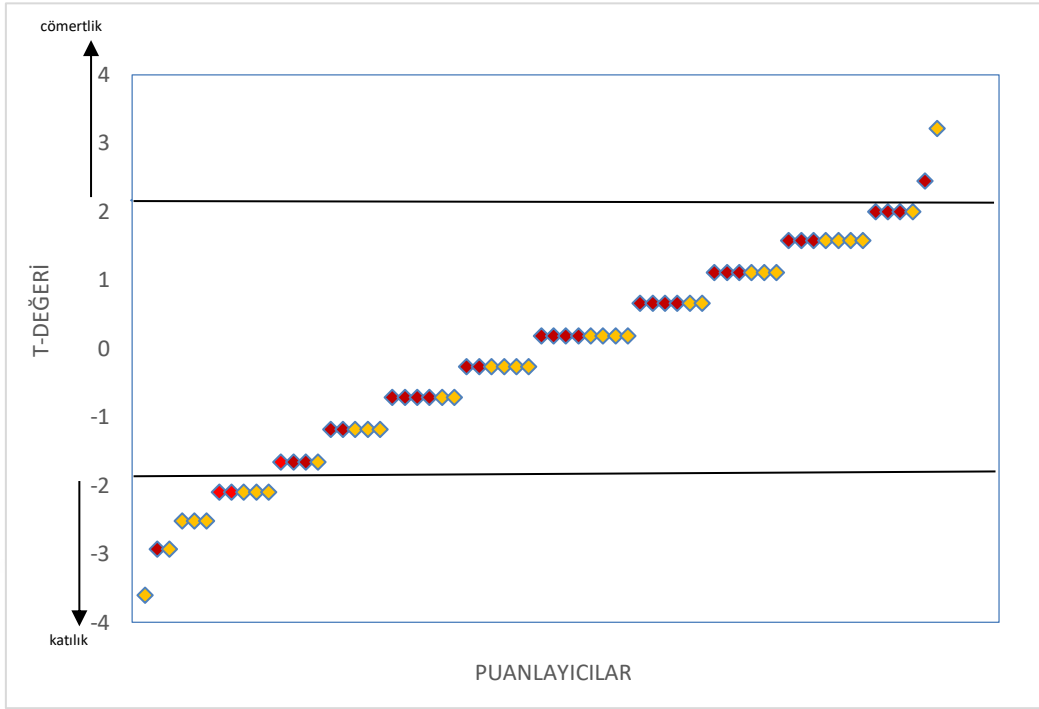
Tablo 16 incelendiğinde başarı testindeki maddelerin güçlük düzeylerinin -0,75 ile 1,25 arasında değiştiği görülmektedir. En kolay maddenin 4. madde, en zor maddenin 1.madde olduğu söylenebilir. Maddelerin logit değerleri ortalaması 0,00 ve standart sapması 0,93 olarak bulunmuştur. İç uyum değerlerinin 0,80 ile 1,23 arasında, dış uyum değerlerinin 0,81 ile 1,32 değerleri arasında değiştiği görülmektedir. İç ve dış uyum değerlerinin ortalamaları sırasıyla 1,04 ve 1,06 olarak bulunduğu için maddeler kabul edilebilir uygunluk değerlerine sahiptir. Ayrıca maddeler yüzeyine ait ayırma oranının 9,34, ayırma indeksinin 12,79 ve ayırma indeksi güvenirliliğinin 0,99 çıkması maddelerin birbirinden güçlüklerine göre ayırt edilebildiğini göstermektedir. Maddelerin güçlük düzeylerinin birbirinden ayırt edilip edilmediğine bakmak için X^2 değeri incelenmiştir. $X^2=218,8$, $sd=3$, $p=0,00 < 0,05$ olarak bulunmuştur. Bu değerler istatistiksel olarak anlamlıdır.

Üçüncü Araştırma Sorusuna İlişkin Bulgular ve Yorumlar

Üçüncü araştırma sorusu olan “Puanlayıcı eğitimi öncesinde deney ve kontrol grubundaki puanlayıcıların puanlamalarındaki katılık/cömertlik ve farklılaşan puanlayıcı katılığı cömertliği durumları nasıldır?” sorusuna yönelik bulgular ve yorumlar aşağıda verilmiştir. Burada puanlayıcı davranışları teker teker alt başlıklar halinde sunulmuştur.

Puanlayıcı Eğitimi Öncesi Deney ve Kontrol Grubundaki Puanlayıcıların Katılık/Cömertlik Durumları

Puanlayıcı eğitimi öncesinde puanlayıcıların katılık/cömertlik davranışı gösterip göstermediğini belirlemek için grup düzeyinde ve bireysel düzeyde istatistiksel göstergelere bakılmıştır. Puanlayıcı katılığı/cömertliği için grup düzeyinde bakılması gereken hususlar, ayırma oranı, ayırma indeksi ve ayırma indeksi güvenilirliğinin yüksek olması ve sabit etkili ki kare değerinin istatistiksel olarak anlamlı olmasıdır. Ayırma indeksi, ayırma oranı ve ayırma indeksinin güvenilirliği yüksek bulunmuştur (tablo 8). Ayrıca puanlayıcı yüzeyinin sabit etkili ki-kare istatistiği de istatistiksel olarak anlamlı bulunmuştur. Grup düzeyinde deney ve kontrol grubunda puanlayıcı davranışlarının farklılaştığı görülmüştür. Bireysel düzeydeki istatistiksel göstergeler için puanlayıcıların değişken haritasındaki konumlarına ve her bir puanlayıcı için logit değeri ve lojit değeri ortalaması kullanılarak bulunan t değerinin istatistiksel anlamlılığına bakılır. Şekil 1’de verilen ön teste ait değişken haritası ve Tablo 8’de her bir puanlayıcı için hesaplanan t değerleri incelenip kritik t-değeri ile karşılaştırılmıştır. Araştırmada 32 deney grubu ve 33 kontrol grubu öğrencileri için kritik t değeri $sd=31$ ve anlamlılık düzeyi ,05 ile $t_{kritik}=2,042$ olarak bulunmuştur. Şekil 1’deki değişken haritasının puanlayıcı sütununa bakıldığında puanlayıcıların farklı yerlerde konumlandıkları görülmektedir. 42 numaralı öğrencinin daha cömert, 45 numaralı puanlayıcının daha katı olduğu söylenebilir. 15 ve 60 numaralı öğrencilerin logit cetvelinin sıfır noktasında bulunmaktadır. Deney ve kontrol grubu öğrencilerinin katılık cömertlik durumları Şekil 3’te grafiksel olarak gösterilmiştir.



Şekil 3: Deney ve kontrol grubunun ön test t-değerleri (kırmızı simgeler deney grubunu, sarı simgeler kontrol grubunu göstermektedir.)

Şekil 3'te kırmızı simgeler deney grubundaki puanlayıcıları, sarı simgeler kontrol grubundaki puanlayıcıları göstermektedir. Grafik incelendiğinde deney grubunda üç puanlayıcının kontrol grubunda ise sekiz puanlayıcının katılık davranışını gösterdiği, deney grubunda ve kontrol grubunda bir puanlayıcının cömertlik davranışı gösterdiği görülmektedir. $\pm 2,042$ aralığındaki puanlayıcıların katılık ve cömertlik davranışını göstermedikleri söylenebilir. Ön test sonuçlarına göre deney ve kontrol grubundaki puanlayıcıların katılık cömertlik durumlarının birbirlerinden farklılaşıp farklılaşmadığını anlamak için bağımsız örneklem t-testi yapılmıştır. Bu testin sonuçları aşağıda Tablo 17'de verilmiştir.

Tablo 17

Deney ve Kontrol Grubunun Ön Test Puanlamalarının Bağımsız Örneklem t-testi

	Grup	N	$\bar{X}$	S	t	sd	p
Ön test	Deney	32	0,10	1,38	1,15	63	0,252
	Kontrol	33	-0,33	1,65			

Tablo 17'ye bakıldığında deney ve kontrol grubunun ön test puanlama sonuçlarının benzer olduğu bulunmuştur ($t_{63}=1,15$, $p=0,252>0,05$). Puanlayıcı eğitimi öncesinde her iki grubun katılık cömertlik davranışları gösterme düzeylerinin benzer olduğu söylenebilir. Alan yazın incelendiğinde Şata (2019) 'nın doktora tezinde deney ve kontrol grubuna ayrılmış grupların performans değerlendirme sürecinde katılık ve cömertlik davranışının değerlendirme sürecine karıştığı bulunmuştur. Yine Abu Kassım (2007)'nin çalışmasında performans değerlendirme sürecine katılık/cömertlik davranışının karıştığı bulunmuştur.

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

Puanlayıcıların maddeleri puanlama sürecinde bazı öğrencilere daha katı bazı öğrencilere daha cömert davranması sonucu ortaya çıkan yanlılık davranışı incelenirken modeldeki yüzeyler arasındaki etkileşim değerlerine bakılır. *Bu çalışmada öğrenci (o), puanlayıcı (p), cinsiyet (c), başarı (b), madde (m)* olmak üzere beş yüzey bulunduğu için incelenen etkileşimler, *puanlayıcı x öğrenci (pxo)*, *öğrenci x başarı (oxb)*, *puanlayıcı x madde (pxm)*, *puanlayıcı x öğrenci x madde (pxoxm)* şeklindedir. Hangi puanlayıcıların öğrencilerden bazılarına daha katı ya da cömert davrandığını belirlemek için *puanlayıcı x öğrenci* etkileşimine bakılmıştır. Puanlayıcıların, öğrencilerin bazı maddelere verdikleri yanıtları daha yüksek ya da daha düşük puanlayıp puanlamadıklarını belirlemek için *puanlayıcı x öğrenci x madde* etkileşimi incelenmiştir. Puanlayıcıların başarı testindeki maddelerin hepsini aynı katılıkta, cömertlikte puanlayıp puanlamadığını anlamak için *puanlayıcı x*

madde etkileşimine bakılmıştır. Puanlayıcıların okul başarısı ile öğrencileri puanlama davranışının değişip değişmediğini incelemek için *öğrenci x başarı* etkileşimine bakılmıştır. Grup düzeyinde puanlayıcıların yanlılık davranışları için ki kare istatistiğinin anlamlılığı, bireysel düzeyde yanlılık için ise t değerlerinin ± 2 aralığının dışında kalıp kalmadığı incelenir (Bond & Fox, 2015). t değeri ± 2 aralığının dışında kalan puanlayıcıların yanlı davrandıkları söylenir.

Tablo 18’de *puanlayıcı x öğrenci* etkileşimine yönelik veriler yer almaktadır.

Tablo 18

Deney ve Kontrol Grubuna Uygulanan Ön Testin Puanlayıcı x Öğrenci Etkileşim Analizinde Yanlı Olan Etkileşimleri

Puanlayıcı	Öğrenci	Gözlenen Puan	Beklenen Puan	Yanlılık Büyüklüğü	Standart Hata	t-değeri
D1	60	11	14,30	-1,85	0,71	-2,62
D7	30	8	12,42	-2,23	0,72	-3,08
D24	60	12	14,55	-1,54	0,72	-2,15
D30	60	14	15,60	-1,84	0,82	-2,24
D31	60	14	15,60	-1,84	0,82	-2,24
K17	60	10	13,45	-1,79	0,70	-2,54
Sabit etkili ki-kare: 124,4 sd: 195 p=1,00						

Çalışmada 32 deney ve 33 kontrol grubunda bulunan 65 puanlayıcı 3 öğrencinin açık uçlu matematik sorularını puanlamışlar ve bu puanlamaya yönelik “puanlayıcı x öğrenci” etkileşiminde 195 olası etkileşim oluşmuştur. Bu tabloda sadece yanlı davranışlar gösteren puanlayıcılara yer verilmiştir. Tablo 18’de verilen sabit etkili ki-kare değerinin istatistiksel olarak anlamlı olmadığı görülmüştür ($X^2=124,4$, $sd=195$, $p=1,00 >0,05$). Puanlayıcıların grup düzeyinde farklılaşan puanlayıcı katılığı/cömertliği (yanlılık) davranışı göstermedikleri söylenebilir. Bireysel düzeyde puanlayıcıların yanlılık düzeyleri incelendiğinde deney grubunda beş öğrencinin, kontrol grubunda bir öğrencinin yanlılık davranışı gösterdikleri tespit edilmiştir (t değerleri ± 2 aralığının dışında). Bu öğrencilerin hepsinin t değerlerinin negatif olduğu ve gözlenen puanlarının beklenen puanlardan daha düşük olduğu için katılık davranışı gösterdikleri söylenebilir.

“Puanlayıcı x madde” etkileşimine yönelik sonuçlar Tablo 19’da gösterilmiştir.

Tablo 19

Deney ve Kontrol Grubuna Uygulanan Ön Testin Puanlayıcı x Madde Etkileşim Tablosu

Puanlayıcı	Madde	Gözlenen Puan	Beklenen Puan	Yanlılık Büyüklüğü	Standart Hata	t-değeri
K26	1	10	7,62	1,71	0,93	1,84
K24	1	11	8,72	2,00	1,12	1,78
D10	3	11	9,02	1,79	1,12	1,59
K18	3	11	9,27	1,60	1,12	1,43
...	...	...	...	...	...	...
D5	4	10	11,51	-1,84	0,93	-1,99
D11	4	8	10,84	-2,29	0,80	-2,85
D10	4	7	10,29	-2,36	0,79	-3,01
D7	4	8	11,00	-2,48	0,80	-3,09
Sabit etkili ki-kare: 171,3 sd:260 p=1,00						

Tablo 19’da deney ve kontrol grubundaki puanlayıcıların puanlama yaparken hangi maddelerde daha cömert hangi maddelerde daha katı davrandıkları gösterilmiştir. Çalışmada 65 puanlayıcı ve 4 madde olduğu için 260 etkileşim meydana gelmiştir. 260 etkileşimden t-değeri en yüksekten en düşüğe olacak şekilde bazıları tabloda gösterilmiştir. Tablo 19’a göre grup düzeyinde istatistiksel bir gösterge olan sabit etkili ki-kare değeri istatistiksel olarak anlamlı çıkmamıştır ($X^2=171,3$, sd: 260, $p=1,00$). Deney ve kontrol grubundaki puanlayıcıların maddeleri puanlarken farklılaşan puanlayıcı katılığı/cömertliği (yanlılık) davranışını göstermediği yorumu yapılabilir. Bireysel düzeyde istatistiksel göstergeler incelendiğinde deney ve kontrol grubundaki bazı puanlayıcıların yanlı etkileşimlerinin anlamlı olduğu görülmüştür. t-değeri ± 2 aralığının dışında kalan puanlayıcılar maddeleri puanlarken yanlı davranmışlar diyebiliriz. Deney grubundan üç öğrencinin t-değerleri belirlenen aralığın dışındadır. Bu öğrencilerin gözlenen puanları beklenen puanlarından düşük olduğu için ilgili maddeye (4.madde) gerekenden daha az puan verdikleri yani katı davranış gösterdikleri belirlenmiştir.

“Öğrenci x başarı” etkileşimine yönelik sonuçlar Tablo 20’de verilmiştir.

Tablo 20

Deney ve Kontrol Grubuna Uygulanan Ön Testin Öğrenci x Başarı Etkileşim Tablosu

Puanlayıcının Mat Başarısı	Öğrenci	Gözlenen Puan	Beklenen Puan	Yanlılık Büyükliği	Standart Hata	t-değeri
5(Pekiyi)	60	46	44,09	0,79	0,75	1,06
4(İyi)	60	165	161,42	0,35	0,33	1,05
2(Geçer)	30	270	260,27	0,30	0,18	1,68
1(Geçmez)	26	175	171,13	0,11	0,17	0,66
5(Pekiyi)	30	40	39,45	0,11	0,45	0,24
3(Orta)	26	128	125,38	0,10	0,20	0,52
4(İyi)	26	103	101,09	0,09	0,22	0,41
3(Orta)	60	194	193,76	0,02	0,30	0,07
1(Geçmez)	30	238	239,93	-0,07	0,19	-0,36
1(Geçmez)	60	264	265,94	-0,11	0,23	-0,46
3(Orta)	30	173	175,86	-0,14	0,22	-0,64
2(Geçer)	26	175	180,94	-0,15	0,16	-0,96
2(Geçer)	60	288	291,78	-0,17	0,21	-0,81
4(İyi)	30	139	144,48	-0,29	0,23	-1,28
5(Pekiyi)	26	25	27,46	-0,42	0,42	-1,01
Sabit etkili ki-kare değeri: 11,0 df: 15 p=0,76						

Tablo 20’de deney ve kontrol grubundaki puanlayıcıların matematik başarıları, kağıtları değerlendirilen öğrenciler, beklenen ve gözlenen puanlar, t-değerleri gibi ölçümler verilmiştir. Grup düzeyinde etkileşime bakıldığında sabit etkili ki-kare değerinin istatistiksel olarak manidar olmadığı görülmüştür ($X^2=11,0$ sd: 15 $p=0,76>0,05$) . Bireysel düzeyde istatistikler incelendiğinde t-değerlerinin hepsinin beklenen ± 2 aralığında oldukları görülmektedir. Puanlayıcıların öğrencilerin matematik testlerini puanlarken, matematik başarı düzeylerinin puanlama davranışlarına etki etmediği görülmüştür.

“Puanlayıcı x Öğrenci x Madde” yüzeylerinin etkileşimine yönelik bulgular Tablo 21’de verilmiştir.

Tablo 21

Deney ve Kontrol Grubuna Uygulanan Ön Testin Puanlayıcı x Öğrenci x Madde Etkileşim Analizinde Yanlı Olan Etkileşimleri

Puanlayıcı	Öğrenci	Madde	Gözlenen Puan	Beklenen Puan	Yanlılık Büyüklüğü	Standart Hata	t-değeri
K14	26	2	3	1,22	3,35	1,20	2,78
D7	26	1	3	1,27	3,08	1,20	2,56
D7	26	2	3	1,41	2,61	1,20	2,16
K26	26	1	3	1,46	2,47	1,20	2,05
K29	26	1	3	1,46	2,47	1,20	2,05
D6	30	4	4	3,97	-3,08	1,51	-2,03
K9	30	4	4	3,97	-3,08	1,51	-2,03
D30	30	4	4	3,97	-3,08	1,51	-2,03
D31	30	4	4	3,97	-3,08	1,51	-2,03
K10	30	3	4	3,98	-3,13	1,51	-2,07
D6	60	3	4	3,98	-3,17	1,51	-2,09
D30	60	3	4	3,98	-3,17	1,51	-2,09
D31	60	3	4	3,98	-3,17	1,51	-2,09
K9	60	3	4	3,98	-3,17	1,51	-2,09
D24	30	4	3	3,90	-2,53	1,20	-2,10
D27	30	4	3	3,90	-2,53	1,20	-2,10
D15	60	4	4	3,98	-3,20	1,51	-2,11
D19	60	4	4	3,98	-3,20	1,51	-2,11
D22	60	4	4	3,98	-3,20	1,51	-2,11
D23	60	4	4	3,98	-3,20	1,51	-2,11
K22	60	4	4	3,98	-3,20	1,51	-2,11
K26	60	4	4	3,98	-3,20	1,51	-2,11
K28	60	4	4	3,98	-3,20	1,51	-2,11
K29	60	4	4	3,98	-3,20	1,51	-2,11
D7	30	3	2	3,66	-2,70	1,25	-2,15
D9	30	4	4	3,98	-3,40	1,51	-2,24
D5	60	4	4	3,98	-3,42	1,51	-2,26
D12	60	4	4	3,98	-3,42	1,51	-2,26
D29	60	4	4	3,98	-3,42	1,51	-2,26
D25	60	4	4	3,98	-3,42	1,51	-2,26
K8	60	4	4	3,98	-3,42	1,51	-2,26
K19	60	4	4	3,98	-3,42	1,51	-2,26
K25	30	4	3	3,91	-2,74	1,20	-2,27
D9	60	3	4	3,98	-3,49	1,51	-2,30
D10	30	4	2	3,74	-2,99	1,02	-2,38
D2	60	4	4	3,99	-3,67	1,51	-2,42
D18	60	4	4	3,99	-3,67	1,51	-2,42
K7	60	4	4	3,99	-3,67	1,51	-2,42
D32	60	4	4	3,99	-3,67	1,51	-2,42
K30	60	4	4	3,99	-3,67	1,51	-2,42
K33	60	4	4	3,99	-3,67	1,51	-2,42
D30	60	1	2	3,78	-3,15	1,25	-2,51
D31	60	1	2	3,78	-3,15	1,25	-2,51
D14	60	4	4	3,99	-3,93	1,51	-2,59
D21	60	4	4	3,99	-3,93	1,51	-2,59
D28	60	4	4	3,99	-3,93	1,51	-2,59
K4	60	4	4	3,99	-3,93	1,51	-2,59
K11	60	4	4	3,99	-3,93	1,51	-2,59
K16	60	4	4	3,99	-3,93	1,51	-2,59
K24	60	4	4	3,99	-3,93	1,51	-2,59
D29	30	4	3	3,94	-3,17	1,20	-2,63
D5	30	4	3	3,94	-3,17	1,20	-2,63
K10	30	4	4	3,99	-4,17	1,51	-2,75
D6	60	4	4	3,99	-4,21	1,51	-2,78
D30	60	4	4	3,99	-4,21	1,51	-2,78
D31	60	4	4	3,99	-4,21	1,51	-2,78
K9	60	4	4	3,99	-4,21	1,51	-2,78
K10	60	3	4	3,99	-1,26	1,51	-2,81
K20	30	4	2	3,87	-3,73	1,25	-2,98
D9	60	4	4	3,99	-4,53	1,51	-2,99
D21	30	4	3	3,97	-3,68	1,20	-3,05

Deney ve Kontrol Grubuna Uygulanan Ön Testin Puanlayıcı x Öğrenci x Madde Etkileşim Analizinde Yanlı Olan Etkileşimleri (devamı)

K10	60	4	4	4	-5,30	1,51	-3,50
D1	60	4	2	3,96	-4,86	1,25	-3,88
Sabit etkili ki-kare değeri: 964,8 sd: 780 p=0,00							

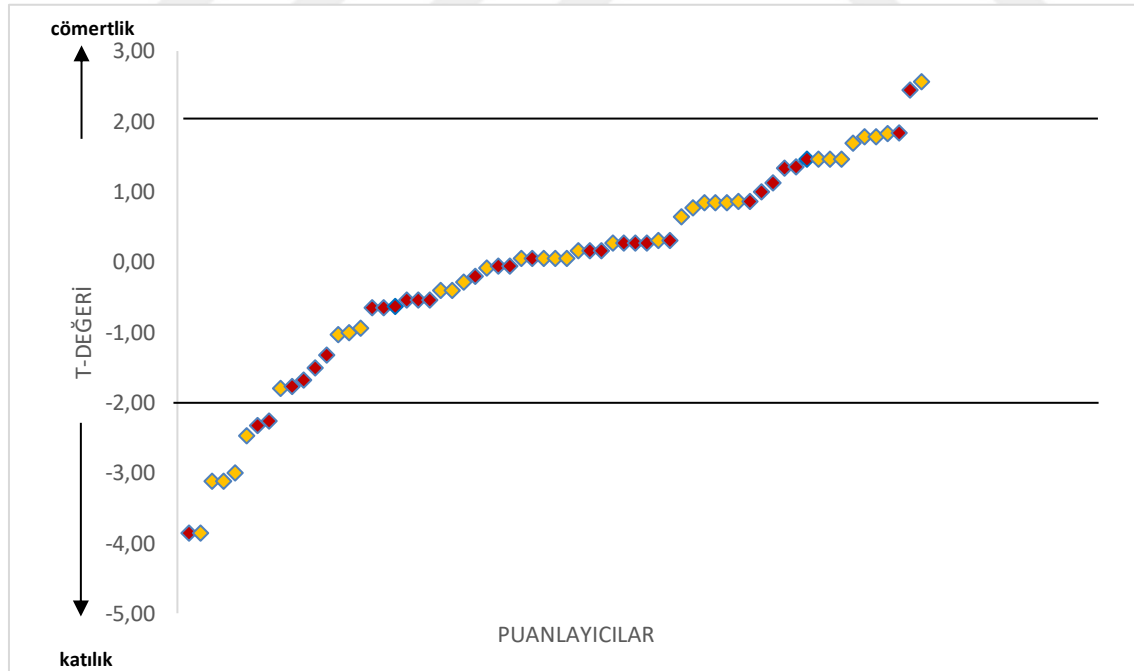
Tablo 21’de puanlayıcıların bazı öğrencilere hangi maddelere göre olması gerekenden düşük ya da yüksek puanlar verip vermedikleri incelenmiştir. 64 puanlayıcı 3 öğrenci ve 4 maddenin 780 etkileşimi olmuştur. Tabloda oluşan tüm etkileşimler yerine t-değerine göre anlamlı olan etkileşimler gösterilmiştir. İlk olarak grup düzeyinde sabit etkili ki-kare değeri incelenmiş ve bu değerin istatistiksel olarak manidar olduğu görülmüştür ($X^2=964,8$ sd:780 $p=0,00<0,05$). Yani puanlayıcılar grup düzeyinde farklılaşan puanlayıcı katılığı/cömertliği (yanlılık) davranışını göstermişlerdir. Yanlılık davranışının bireysel düzeyde ortaya çıkıp çıkmadığı için t-değerleri incelendiğinde deney ve kontrol grubundaki bazı öğrencilerin yanlı etkileşimlerinin anlamlı olduğu görülmüştür. “puanlayıcı x öğrenci x madde” etkileşiminde oluşan toplam 780 etkileşimde 63 tanesinin istatistiksel olarak manidar olduğu bulunmuştur. Deney grubunda iki etkileşimde farklılaşan puanlayıcı cömertliği, otuz altı etkileşimde farklılaşan puanlayıcı katılığı bulunurken, kontrol grubunda üç etkileşimde farklılaşan puanlayıcı cömertliği, yirmi iki etkileşimde farklılaşan puanlayıcı katılığı davranışı görülmüştür. Araştırma kapsamındaki sonuçlar ile alan yazında benzer sonuçlar görmek mümkündür. Yan (2014) tarafından yapılan İngilizce iletişim becerilerinin ölçüldüğü bir araştırmada öğrencilerin puanlanmasında puanlayıcı yanlılığı davranışı görülmüştür.

Dördüncü Araştırma Sorusuna İlişkin Bulgular ve Yorumlar

Deney ve kontrol gruplarının ön testteki puanlayıcı davranışları belirlendikten sonra deney grubuna puanlayıcı eğitimi verilmiştir. Puanlayıcı eğitiminden sonra iki gruba son test uygulanmış ve puanlayıcıların puanlama davranışlarına tekrar bakılmıştır.

Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubundaki Puanlayıcıların Katılık/Cömertlik Durumları

Deney grubuna puanlayıcı eğitimi verildikten sonra her iki gruba da uygulanan son test sonucunda puanlayıcıların grup düzeyinde ve bireysel düzeyde katılık/cömertlik durumları incelenmiştir. Grup düzeyinde katılık/cömertlik davranışı için ayırma oranı, ayırma indeksi ve ayırma indeksinin güvenilirliği değerlerine bakılmıştır. Tablo 8’de puanlayıcı yüzeyine ait son test ölçümleri verilmiştir. Tablo 8’e bakıldığında sabit etkili ki-kare değerinin istatistiksel olarak anlamlı olduğu görülmüştür ($X^2=137,5$ sd:64 $p=0,00<0,05$). Grup düzeyinde farklı davranışlar sergilenmiştir. Bireysel düzeyde puanlayıcı katılığı/cömertliği için puanlayıcıların logit ölçümleri ve t-değerleri incelenmiştir (Tablo 11, t-değeri sütunu). Her bir puanlayıcı için hesaplanan t-değerleri incelenip kritik t-değeri ile karşılaştırılmıştır. Araştırmanın son testinde 32 deney, 33 kontrol grubu öğrencisi için kritik t değeri sd=31 ve anlamlılık düzeyi ,05 ile $t_{kritik} = 2,042$ olarak bulunmuştur. Gruplara ait t-değerleriyle elde edilmiş grafik Şekil 4’te verilmiştir.



Şekil 4:Deney ve kontrol grubunun son test t-değerleri (kırmızı simgeler deney grubunu, sarı simgeler kontrol grubunu göstermektedir.)

Şekil 4 incelendiğinde kontrol grubunda beş puanlayıcının katılık davranışı gösterdiği, bir öğrencinin ise cömertlik davranışı gösterdiği bulunmuştur. Deney grubunda ise üç öğrencinin puanlama yaparken katı, bir öğrencinin cömert davrandığı görülmüştür. Geri kalan diğer puanlayıcıların t-değerlerim -2,042 ile +2,042 değerleri arasında kalmış, puanlayıcılar herhangi bir katılık/cömertlik davranışı göstermemişlerdir. Deney ve kontrol gruplarının son testinden elde edilen verilerle (logit değerleri) bu grupların birbirlerinden farklı olup olmadığını anlamak için bağımsız örneklem t-testi yapılmıştır.

Tablo 22

Deney ve Kontrol Grubunun Son Test Puanlamalarının Bağımsız Örneklem t-testi (Logit Ölçüleri Farkı Alınmış)

	Grup	N	$\bar{X}$	S	t	sd	p
Logit Ölçüleri Farkı	Deney	32	-0,51	0,70	-5,02	63	0,00
	Kontrol	33	0,50	0,90			

Tablo 22’de deney ve kontrol gruplarındaki puanlayıcıların son testlerinden elde edilen logit ölçüleri ile ön testlerinden elde edilen logit ölçüleri farkları alınıp bağımsız örneklem t-testi yapılan sonuçları yer almaktadır. Bu sonuca göre deney ve kontrol gruplarının son testlerinin iki grubu ayırmada istatistiksel olarak manidar olduğu söylenebilir ($t_{63} = -0,502$, $p=0,00 < 0,05$). Puanlayıcı eğitimi sonrası puanlayıcıların katılık/cömertlik davranışlarına bakıldığında hem deney hem de kontrol grubunda bu davranışların ortaya çıktığı görülmektedir. Eğitim öncesinde deney grubunda katılık/cömertlik davranışı gösterme oranı %12,5 iken, eğitim sonrasında yine %12,5 olarak bulunmuştur. Kontrol grubunda ise ön testte %33 olan oran son testte %18’e düşmüştür. Alan yazında da benzer çalışmalar görmek mümkündür. Polat (2018)’in araştırmasında puanlayıcılara eğitim verilmesine rağmen katılık/cömertlik davranışı gösterdikleri sonucu ortaya çıkmıştır. Tobaş (2020)’ın yaptığı çalışmada öğretmen puanlayıcılara araştırma öncesi puanlayıcı eğitimi verilmiş olmasına rağmen puanlamalarda katılık/cömertlik davranışının ortaya çıktığı görülmüştür. Weitz vd. (2014) tarafından yapılan çalışmada puanlayıcı eğitimi sonrasında puanlayıcıların daha katı

puanlamalar yaptıkları tespit edilmiştir. Ayrıca Kondo (2010)'nın çalışmasında ise puanlayıcı eğitimi ve sonrasında puanlayıcıların katılık/cömertlik davranışlarının benzer olduğu sonucuna varılmıştır. Mevcut çalışmada deney grubunun puanlayıcı eğitimi almış olmasına rağmen katılık/cömertlik davranışlarının benzer olması alan yazındaki bazı çalışmalarla benzerlik göstermektedir.

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

Yanlılık durumları için ilk olarak grup düzeyindeki istatistikler incelenmiştir. Bunun için ki-kare istatistiği kullanılır. Analiz sonucunda *puanlayıcı x öğrenci* etkileşiminin ki-kare değeri istatistiksel olarak anlamlı çıkmamıştır ($X^2 = 87,1$, $sd = 195$, $p = 1,00 > 0,05$). Grup düzeyinde yanlılık davranışının puanlamaya karışmadığı söylenebilir. Daha sonra bireysel düzeyde yanlılık davranışı ise gerekli istatistikler incelenmiştir. Deney ve kontrol grubunda toplamda 65 puanlayıcı 3 öğrencinin yanıtlarını değerlendirdiği için toplamda 195 etkileşim oluşmuştur. Bireysel düzeyde yanlılık için t değerinin ± 2 aralığının dışında kalıp kalmadığına bakılır (Bond & Fox, 2015). t değeri bu aralık dışında kalan puanlayıcıların yanlı davrandıkları söylenebilir. Bu etkileşimlerden istatistiksel olarak anlamlı olanlar Tablo 23'te verilmiştir.

Tablo 23

Deney ve Kontrol Grubuna Uygulanan Son Testin Puanlayıcı x Öğrenci Etkileşim Analizinde Yanlı Olan Etkileşimler

Puanlayıcı	Öğrenci	Gözlenen Puan	Beklenen Puan	Yanlılık Büyüklüğü	Standart Hata	t-değeri
K13	26	13	9,08	1,26	0,61	2,06
Sabit etkili ki-kare: 87,1 sd:195 p= 1,00						

Tablo 23 incelendiğinde deney grubundaki öğrencilerden yanlı davranış gösteren puanlayıcının olmadığı görülmektedir. Ön test sonucunda deney grubundan 1, 7, 24, 30 ve 31 numaralı dört puanlayıcının yanlı davrandığı bulunmuştu (%12,5). Puanlayıcı eğitimi

sonrası yanlış davranışın azaldığı söylenebilir. Kontrol grubunda ise ön testte bir puanlayıcının yanlış davrandığı bulunmuştu. Son test sonrasında da bir puanlayıcının yanlış davrandığı görülmektedir. Sonuç olarak deney grubunda puanlayıcı yanlışlığının azaldığı, kontrol grubunda ise değişmediği görülmektedir. Şata (2019) tarafından yapılan araştırmada da puanlayıcı eğitiminin yanlışlık davranışını azalttığı bulunmuştur. Ayrıca Fahim ve Bijani (2011) ve May (2008) tarafından yapılan araştırmalarda da benzer olarak puanlayıcı eğitiminin yanlışlığı azalttığı bulunmuştur.

Puanlayıcı eğitimi sonrası “*puanlayıcı x madde*” etkileşimine yönelik sonuçlar Tablo 24’te gösterilmiştir.

Tablo 24

Deney ve Kontrol Grubuna Uygulanan Son Testin Puanlayıcı x Madde Etkileşim Tablosu

Puanlayıcı	Madde	Gözlenen Puan	Beklenen Puan	Yanlışlık Büyüklüğü	Standart Hata	t-değeri
K18	2	11	6,82	2,07	1,00	2,08
K12	2	11	7,44	1,84	1,00	1,85
K33	1	7	4,54	1,10	0,62	1,78
D17	1	12	7,80	2,38	1,40	1,70
...	...	...	...	...	...	...
K13	4	9	11,06	-1,26	0,66	-1,90
D21	3	7	9,86	-1,21	0,62	-1,96
K21	3	9	11,16	-1,37	0,66	-2,07
K16	2	6	9,44	-1,39	0,64	-2,18
D32	4	9	11,27	-1,51	0,66	-2,28
D25	4	7	11,06	-2,06	0,62	-3,33
Sabit etkili ki-kare: 178,2 sd: 260 p= 1,00						

Tablo 24’te deney ve kontrol grubundaki puanlayıcıların maddeleri puanlarken yanlış davranıp davranmadıklarına ilişkin ölçüm değerleri verilmiştir. Araştırmada 65 puanlayıcı ve 4 madde olduğu için 260 etkileşim meydana gelmiştir. 260 etkileşimden t-değeri en yüksekten en düşüğe olacak şekilde bazıları tabloda gösterilmiştir. Grup düzeyinde istatistiksel gösterge olan sabit etkili ki-kare değeri anlamlı çıkmamıştır ($X^2=178,2$, sd: 260, $p=1,00$). Bireysel düzeyde istatistiksel göstergeler incelendiğinde deney ve kontrol grubundaki bazı puanlayıcıların yanlış etkileşimlerinin anlamlı olduğu görülmüştür. t-değeri ± 2 aralığının dışında kalan puanlayıcılar maddeleri puanlarken yanlış davranmışlar diyebiliriz. Deney grubundan iki öğrencinin t-değerleri belirlenen aralığın dışındadır

(%6,25). Bu öğrencilerin gözlenen puanları beklenen puanlarından düşük olduğu için ilgili maddeye gerekenden daha az puan verdikleri yani katı davranış gösterdikleri belirlenmiştir. Ön test sonucuna göre deney grubunda üç öğrencinin t değerleri ± 2 aralığının dışında idi (%9,37). Puanlayıcı eğitimi sonra deney grubundaki puanlayıcıların maddeleri puanlarken yanlış davranışlarında az da olsa bir azalma meydana gelmiştir. Puanlayıcı eğitimi öncesi deney grubundaki D7, D10 ve D11 numaralı puanlayıcılar yanlış davranış (katılık) göstermişlerdi. Son testte bu kişilerin maddeleri puanlamada yanlış davranmadıkları görülmüştür. Kontrol grubunda ise ön testte t değerleri istenen aralıkta idi. Son testte kontrol grubundaki üç öğrencinin maddeleri puanlarken yanlış davranış gösterdiği söylenebilir.

Puanlayıcı eğitimi sonrası “Öğrenci x başarı” etkileşimine yönelik sonuçlar Tablo 25’de verilmiştir.

Tablo 25

Deney ve Kontrol Grubuna Uygulanan Son Testin Öğrenci x Başarı Etkileşim Tablosu

Puanlayıcının Mat Başarısı	Öğrenci	Gözlenen Puan	Beklenen Puan	Yanlılık Büyüklüğü	Standart Hata	t-değeri
1(Geçmez)	26	175	169,83	0,10	0,14	0,71
3(Orta)	60	184	181,17	0,12	0,21	0,58
5(Pekiyi)	60	43	41,92	0,20	0,45	0,45
2(Geçer)	30	252	249,30	0,05	0,14	0,38
3(Orta)	30	167	165,85	0,03	0,17	0,20
4(iyi)	26	115	114,36	0,02	0,17	0,11
5(Pekiyi)	26	30	29,77	0,03	0,34	0,08
4(iyi)	30	145	144,95	0,00	0,20	0,01
2(Geçer)	60	273	273,65	-0,02	0,16	-0,10
4(İyi)	60	156	156,69	-0,04	0,23	-0,16
2(Geçer)	26	190	192,06	-0,04	0,13	-0,27
1(Geçmez)	30	220	222,59	-0,05	0,14	-0,37
1(Geçmez)	60	243	245,58	-0,07	0,16	-0,42
5(Pekiyi)	30	37	38,32	-0,17	0,35	-0,47
3(orta)	26	124	127,98	-0,10	0,16	-0,64
Sabit etkili ki-kare değeri: 2,3 df: 15 p=1,00						

Tablo 25’te deney ve kontrol grubundaki puanlayıcıların matematik başarıları, kağıtları değerlendirilen öğrenciler, beklenen ve gözlenen puanlar, t-değerleri gibi ölçümler verilmiştir. Grup düzeyinde etkileşime bakıldığında sabit etkili ki-kare değerinin istatistiksel olarak manidar olmadığı görülmüştür ($X^2=2,3$ sd: 15 $p=1,00>0,05$) . Bireysel düzeyde istatistikler incelendiğinde t-değerlerinin hepsinin beklenen ± 2 aralığında oldukları

görülmektedir. Puanlayıcıların öğrencilerin matematik testlerini puanlarken, matematik başarı düzeylerinin puanlama davranışlarına etki etmediği görülmüştür. Ön testten farklı olarak tüm puanlayıcıların t değerleri 0'a oldukça yaklaşmıştır.

Puanlayıcı eğitimi sonrası “Puanlayıcı x Öğrenci x Madde” yüzeylerinin etkileşimine yönelik bulgular Tablo 26’da verilmiştir.

Tablo 26

Deney ve Kontrol Grubuna Uygulanan Son Testin Puanlayıcı x Öğrenci x Madde Etkileşim Analizinde Yanlı Olan Etkileşimleri

Puanlayıcı	Öğrenci	Madde	Gözlenen Puan	Beklenen Puan	Yanlılık Büyüklüğü	Standart Hata	t-değeri
D32	30	4	1	3,83	-3,41	1,35	-2,53
K12	30	4	2	3,78	-2,50	1,03	-2,43
D29	30	4	3	3,92	-2,53	1,02	-2,42
D25	30	4	1	3,78	-3,16	1,35	-2,35
K13	30	4	1	3,83	-3,16	1,35	-2,35
K18	30	4	2	3,72	-2,26	1,03	-2,20
D21	60	3	2	3,70	-2,18	1,03	-2,12
D11	30	4	3	3,89	-2,19	1,05	-2,09

Sabit etkili ki-kare değeri: 464,4 sd: 780 p=1,00

Tablo 26’da puanlayıcıların bazı öğrencilere hangi maddelere göre olması gerekenden düşük ya da yüksek puanlar verip vermedikleri incelenmiştir. 64 puanlayıcı 3 öğrenci ve 4 maddenin 780 etkileşimi olmuştur. Tabloda oluşan tüm etkileşimler yerine t-değerine göre anlamlı olan etkileşimler gösterilmiştir. İlk olarak grup düzeyinde sabit etkili ki-kare değeri incelenmiş ve bu değerin istatistiksel olarak manidar olmadığı görülmüştür ($X^2=464,4$ sd:780 p=1,00 > 0,05). Yani puanlayıcılar grup düzeyinde farklılaşan puanlayıcı katılığı/cömertliği (yanlılık) davranışını göstermemişlerdir. Yanlılık davranışının bireysel düzeyde ortaya çıkıp çıkmadığı için t-değerleri incelendiğinde deney ve kontrol grubundaki bazı öğrencilerin yanlı etkileşimlerinin anlamlı olduğu görülmüştür. “puanlayıcı x öğrenci x madde” etkileşiminde oluşan toplam 780 etkileşimde 8 tanesinin istatistiksel olarak manidar olduğu bulunmuştur. Deney grubunda beş etkileşimde, kontrol grubunda ise üç etkileşimde farklılaşan puanlayıcı katılığı davranışının ortaya çıktığı görülmüştür. Ön testte deney grubunda 38 puanlayıcıda yanlılık davranışı görülmüşken puanlayıcı eğitimi sonrası yanlılık davranışını gösteren puanlayıcı sayısı 5’e inmiştir. Kontrol grubunda da ön testte 25 yanlı

davranış bulunmuşken, son testte bu sayı 3'e inmiştir. Alan yazın incelendiğinde Şata ve Karakaya (2020), tarafından yapılan çalışmada performans değerlendirilmesinde puanlayıcıların yanlış davranış gösterdikleri sonucuna ulaşılmıştır. Goodwin (2016)'nın araştırmasında grup ve bireysel düzeyde yanlış davranış tespit etmiştir. Bu durumun puanlayıcıların eğitilmesiyle çözülebileceğini ifade etmiştir.



BÖLÜM V

SONUÇ VE ÖNERİLER

Bu araştırmada açık uçlu soruların puanlanması sürecinde ortaya çıkan puanlayıcı davranışları belirlenmiş ve bu davranışları ortadan kaldırmaya yönelik uygulamalardan biri olan puanlayıcı eğitimi verilmiştir. Bu bölümde araştırmanın bulguları doğrultusunda sonuçlara ve önerilere yer verilmiştir.

SONUÇLAR

Araştırmada puanlayıcıların katılık/ cömertlik ve farklılaşan puanlayıcı davranışları yani yanlılıkları ÇYRÖM ile incelenmiştir. Analizde öğrenci, puanlayıcı, madde, başarı ve cinsiyet olmak üzere 5 yüzey bulunmaktadır. Ayrıca üç yüzey öğrenci, puanlayıcı ve madde olacak şekilde genellenebilirlik kuramına göre varyans bileşenleri ve toplam varyans yüzdeleri hesaplanmıştır. Araştırma sorularına yönelik sonuçlar aşağıda sırasıyla verilmiştir.

Birinci Araştırma Sorusuna Yönelik Sonuçlar

Genellenebilirlik kuramı için EduG programı ile yapılan analizde deney ve kontrol grubunun öğrenci, puanlayıcı ve madde yüzeylerinin ve bu yüzeylerin birbirleri ile etkileşimlerinin varyans yüzdelerine bakılmıştır. Buna göre öğrenci yüzeyinin varyans yüzdesi her iki grupta

da yüksek çıkmıştır. Bu durum öğrencilerin ölçüldükleri özelliklere göre ayrılabilindiklerini göstermektedir. Madde yüzeyi için varyans yüzdeleri her iki grup için de yüksek çıkmıştır. Bu durum ise maddelerin güçlük düzeyine göre ayrılabilindiklerini göstermiştir. Deney ve kontrol grubunun katılık/cömertlik açısından incelenmesinde ise puanlayıcı yüzeyi varyans yüzdesine bakılmıştır. Puanlayıcı eğitimi öncesinde her iki grupta da katılık/cömertlik davranışının ortaya çıktığı görülmüştür. Puanlayıcı eğitimi sonrasında yapılan son testin sonuçlarına bakıldığında deney ve kontrol grubunda da puanlayıcılara ait varyans yüzdesinin arttığı görülmüştür. Puanlayıcı eğitimi sonucunda deney grubunda katılık/cömertlik davranışı azalmamıştır denebilir.

İkinci Araştırma Sorusuna Yönelik Sonuçlar

Lise öğrencilerinin açık uçlu soruları puanlarken ortaya çıkan puanlayıcı davranışlarının puanlayıcı eğitimi ile farklılaşıp farklılaşmadığının araştırıldığı bu çalışmada beş yüzey üzerinden çalışılmıştır. Veriler ÇYRÖM ile FACETS programı yardımıyla analiz edilmiştir. Öğrenci, puanlayıcı, cinsiyet, başarı ve madde yüzeylerinin ön test ve son test ölçüm raporları ve değişken haritaları incelendiğinde şu sonuçlar ortaya çıkmıştır.

Ön testin ve son testin değişken haritaları incelendiğinde cinsiyet açısından kızlar ve erkeklerin benzer puanlamalar yaptıkları bulunmuştur. Matematik başarılarına göre bakıldığında yine ön test ve son testte puanlayıcılar benzer puanlamalar yapmışlardır. Puanlayıcıların haritada dağınık olması ve uç noktalarda puanlayıcıların olması ön test ve son testte katılık/cömertlik davranışının ortaya çıktığını göstermiştir. Öğrenci yüzeyinin ölçüm raporları incelendiğinde ön test ve son testte de öğrencilerin yeteneklerine göre iyi bir şekilde ayırt edildikleri bulunmuştur. Puanlayıcı yüzeyinin ön test ve son test sonuçlarına bakıldığında her iki testte de katılık cömertlik davranışlarının olduğu görülmüştür. Ön testte katılık cömertliğe ilişkin logit aralığı 3,64 iken son testte 2,66 bulunmuştur. Puanlayıcı eğitimi sonrasında da her iki grupta katılık/cömertlik davranışının olduğu görülmüştür.

Cinsiyet ve başarı yüzeylerinin ön test ve son test ölçüm raporları incelendiğinde sabit etkili ki kare değerinin anlamlı olmadığı görülmüştür. Puanlayıcıların cinsiyetlerinin ve başarılarının farklı olması puanlama davranışlarına etki etmemiştir denebilir. Madde yüzeyi ön test ve son test sonuçlarına bakıldığında ise maddeler her iki testte de güçlük düzeyleri bakımından iyi bir şekilde ayırt edilmiştir.

Üçüncü ve Dördüncü Araştırma Sorularına Yönelik Sonuçlar

Puanlayıcı eğitimi öncesi ve sonrası puanlayıcıların katılık/cömertlik ve yanlılık davranışlarına ilişkin sonuçlar şu şekildedir.

Eğitim öncesi deney ve kontrol grubuna yapılan bağımsız örneklem t-testi sonucunda her iki grubun ön test sonuçları benzer çıkmıştır. Puanlayıcı eğitimi öncesi deney grubunda üç puanlayıcının, kontrol grubunda ise sekiz puanlayıcının katılık davranışı gösterdiği, her iki grupta da birer puanlayıcının cömert davranış gösterdiği bulunmuştur. Ön test sonrasında deney grubu öğrencilerine 7 ders saati boyunca puanlayıcı eğitimi verilmiştir. Bunun sonucunda yapılan son testte deney grubunda üç puanlayıcının katı, bir öğrencinin ise cömert davrandığı görülmüştür. Kontrol grubunda ise beş öğrencinin katı, bir öğrencinin cömert davrandığı görülmüştür. Eğitim öncesinde deney grubunda katılık/cömertlik davranışı yüzdesi %12,5 iken eğitim sonrasında %12,5 ile yine aynı kalmıştır. Kontrol grubunda ise bu oran ön testte %33 iken son testte %18'e gerilemiştir. Yani öğrencilerin puanlayıcı eğitimi almış olmaları bu davranışlarını azaltmamıştır. Deney grubundaki puanlayıcılar bu davranışları göstermeye devam etmişlerdir.

Eğitim öncesi ve sonrası yanlılık durumu için puanlayıcı x öğrenci, puanlayıcı x madde, öğrenci x başarı, puanlayıcı x öğrenci x madde etkileşimlerine bakılmış ve aşağıdaki sonuçlara ulaşılmıştır. Ön testte puanlayıcı x öğrenci etkileşiminde puanlayıcıların grup düzeyinde yanlılık davranışı göstermedikleri bulunmuştur. Bireysel düzeyde ise deney grubunda beş, kontrol grubunda bir öğrencinin yanlılık davranışı gösterdiği tespit edilmiştir.

Son testte puanlayıcı eğitimi sonrası deney grubunda yanlılık davranışında azalma olduğu görülmüş, kontrol grubunda ise herhangi bir değişiklik olmamıştır. Puanlayıcı x madde etkileşiminde grup düzeyinde farklılık görülmemiştir. Bireysel düzeyde ise maddelere deney grubunda üç öğrencinin 4.maddeye daha katı puanlama yaptıkları görülmüştür. Eğitim sonrasında deney grubunda iki öğrencinin 4.maddeye daha katı puanlama yaptıkları bulunmuştur. Kontrol grubunda ise 2.maddeye bir öğrencinin cömert davrandığı tespit edilmiştir. Öğrenci x başarı etkileşiminde ön testte grup düzeyinde ve bireysel düzeydeki istatistikler manidar çıkmamıştır. Puanlayıcıların okuldaki matematik başarı düzeyleri puanlama davranışlarını etkilememiştir denebilir. Son testte de benzer sonuçlar bulunmuştur. t değerlerinin hepsi ± 2 aralığındadır. Ön testten farklı olarak tüm puanlayıcıların t değerleri 0'a oldukça yaklaşmıştır. Puanlayıcı x öğrenci x madde etkileşiminde ön testte puanlayıcıların grup düzeyinde yanlı davrandıkları görülürken, son testte grup düzeyinde sabit etkili ki-kare değeri anlamlı çıkmamıştır yani grup düzeyinde farklılaşma bulunmamıştır. Bireysel düzeyde ön testteki t değerleri incelendiğinde 780 etkileşimin 63 tanesinin istatistiksel olarak manidar olduğu görülmüştür. Deney grubunda 2 etkileşimde farklılaşan puanlayıcı cömertliği, 36 etkileşimde ise farklılaşan puanlayıcı katılığı bulunmuştur. Kontrol grubunda ise 3 etkileşimde farklılaşan puanlayıcı cömertliği, 22 etkileşimde farklılaşan puanlayıcı katılığı görülmüştür. Son testte ise 780 etkileşimin 8 tanesinin manidar olduğu görülmüştür. Deney grubunda 5 etkileşimde, kontrol grubunda ise 3 etkileşimde farklılaşan puanlayıcı katılığı olduğu görülmüştür. Ön test ile karşılaştırıldığında deney grubunda 38 yanlı davranış puanlayıcı eğitimi sonrasında 5'e düşmüştür. Kontrol grubunda ise ön testte 25 yanlı davranış varken son testte 3 yanlı davranış görülmüştür.

Sonuç olarak birey performansının değerlendirilme sürecinde açık uçlu maddelerdeki puanlayıcı davranışlarını belirlerken çok yüzeyli Rasch ölçme modeli kullanılabilir ve bulunan sonuçlar genellenebilirlik kuramı ile desteklenebilir. Rasch modelinin beş yüzeyli çalışmalarda daha kullanışlı olduğu aşikardır.

ÖNERİLER

- Bu araştırmada lise öğrencilerine puanlayıcı eğitimi verilmiş akranlarını değerlendirirken ortaya çıkan katılık cömertlik ve farklılaşan puanlayıcı davranışları incelenmiştir. Alan yazında lise öğrencilerine eğitim verilmesiyle alakalı herhangi bir çalışmaya rastlanmamıştır.
- Sonuçlara bakıldığında öğrencilere verilen puanlayıcı eğitiminin çok etkili olmadığı ortaya çıkmıştır. Bundan sonra lise öğrencileri ile çalışma yapmak isteyen araştırmacıların bu kişilerin yaşları itibarıyla bu çalışmayı çok fazla ciddiye almamış olabilecekleri göz önünde bulundurulmalıdır.
- Verilen puanlayıcı eğitimi 7 ders saati sürmüştür. Eğitimin süresi daha uzun tutulabilir.
- Eğitimin içeriği zenginleştirilebilir. Sınıfta örnek olarak puanlandırılan soru örnekleri çoğaltılabilir.
- Araştırmada puanlayıcı eğitimi kalabalık bir gruba verilmiştir. 32 kişiye eğitim verilmesi yerine daha küçük gruplarla benzer çalışmalar yapılabilir. Kişi sayısını düşürmek daha farklı sonuçlar doğurabilir.
- Mevcut çalışmada lise öğrencilerinin sadece katılık cömertlik ve yanlışlık davranışları incelenmiştir. Bu çalışma genişletilip akran değerlendirmeye ek olarak öz değerlendirme de yaptırılabilir. Ön test son test çalışmasında öğrencinin öz değerlendirme yaparken hangi puanlayıcı davranışları gösterdiği araştırılabilir. Yine benzer şekilde puanlayıcı davranışlarına ek olarak halo etkisi, merkeze yayılım etkisi ve ranj etkisine de bakılabilir.
- Yapılan çalışmada her ne kadar puanlayıcı eğitiminin etkisi olumlu yönde olmamış olsa da öğrenciler sürece gerçekten ilgili ve heyecanlı bir şekilde katılmışlardır. Puanlama nasıl yapılmalı konusunu öğrenirken yani öğretmenin yerine kendilerini koyduklarında eksik bilgilerini daha iyi öğrendiklerini söylemişlerdir. Son yıllarda açık uçlu soruların önem kazanması puanlayıcı eğitimlerinin de ne kadar önemli bir

yere sahip olacađını göstermektedir. Bu nedenle farklı kademelerde ilkokul veya ortaokul düzeyinde benzer çalışmalar yapılabilir.



KAYNAKLAR

- 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.
- Alharby, E. R. (2006). *A comporison 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 assesment*, Doktora tezi. The Pennsylvania State Üniversitesi, Eğitim Bilimleri Enstitüsü, Amerika.
- Alderson, J. C., & Banerjee, J. (2001). Language testing and assessment (Part 1). *Language Teaching*, 34(4), 213-236.
- Andrade, H, Du, Y, & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment inEducation: Principles, Policy & Practice*, (17)2, 199-214.
- Andrade, H. & Cizek, G. (Ed.). (2010). *Handbook of Formative Assessment*. New York: Routledge.
- Anıl, D. & Büyükkıdık, S. (2012). Genellenebilirlik kuramında dört facetli karışık desen kullanımı için örnek bir uygulama. *Journal of Measurement and Evaluation in Education and Psychology*, 3(2), 291-296.

- Atılğan, H. (2004). *Genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch modelinin karşılaştırılmasına ilişkin bir araştırma*. (Doktora tezi). <https://tez.yok.gov.tr> sayfasından erişilmiştir.
- Badger, E. & Thomas, B. (1991). Open-Ended Questions in Reading. *Practical Assessment, Research and Evaluations*. 3(4). <https://doi.org/10.7275/fryf-z044>.
- Baştürk, R. (2010). Bilimsel Araştırma Ödevlerinin Çok Yüzeyle Rasch Ölçme Modeli ile Değerlendirilmesi, *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 1(1), 51-57.
- Berberoğlu, G. (2006). *Sınıf içi ölçme ve değerlendirme teknikleri*, İstanbul : Morpa Kültür.
- Brookhart, S. M. & Nitko, A. J. (2008). *Assessment and grading in classrooms*. New Jersey: Pearson.
- Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*, Alexandria, Virginia: ASCD.
- Brookhart, S. M. (2014). *How to design questions and tasks to assess student thinking*. Amerika Birleşik Devletleri: ASCD.
- Brown, J. D. & Hudson, T. (1998). The alternatives in language assessment, *TESOL Quarterly*, 32(4), 653-675.
- Bond, T. G. & Fox, C. M. (2015). *Applying the rasch model: Fundamental measurement in the human sciences*. New York: Routledge.
- Borman, W. C. (1997). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance*, 20, 238-252.
- Boztunç Öztürk, N., Şahin, M. G., & İlhan, M. (2019). An analysis of scoring via analytic rubric and general impression in peer assessment. *Turkish Journal of Education*, 8(4), 258- 273.

- Bukta, K. (2013). *Rating EFL written performance*, Great Britain: Versita.
- Büyüköztürk, Ş., Çakmak, K. E., Akgün, E. Ö., Karadeniz, Ş. & Demirel, F. (2014). *Bilimsel Araştırma Yöntemleri*, Ankara: Pegem.
- Büyüköztürk, Ş. (2011). *Deneyisel desenler- öntest- son test kontrol grubu desen ve veri analizi*. Ankara: Pegem
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.
- Çetin, B. & İlhan, M. (2017). Standart ve SOLO taksonomisine dayalı rubrikler ile puanların 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/5082/2558> sayfasından erişilmiştir.
- Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, 5, 194-197.
- Doğan, N., & Bıkmaz Bilgen, Ö. (2017). Puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması, *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 8(1), 63-78.
- Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, 2(3). 197-221.
- Eckes, T. (2009). Many-facet Rasch measurement. Amerika: Routledge.
- 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>.
- Erkuş, A., Sünbül, Ö., Ömür Sünbül, S., Yorulmaz, S. & Aşiret, S. (2017). *Psikolojide ölçme ve ölçek geliştirme-2: Ölçme araçlarının psikometrik nitelikleri ve ölçme*. Ankara: Pegem.

- 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.
- Falchikov, N. (1995). Peer feedback marking: developing peer assessment, *Innovations in Education and Training International*, 32, 175-187.
- Farrokhi, F., & Esfandiari, R. (2011). *Theory and Practice in Language Studies*, 1(11), 1531-1540.
- Farrokhi, F., Esfandiari, R., & Dalili, M. V. (2011). Applying the many-facet rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment, *World Applied Sciences Journal, (Innovation and Pedagogy for Lifelong Learning)*, 15, 70-77.
- 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, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.
- 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>
- Güler, N. & Taşdelen Teker, G. (2015). Açık uçlu maddelerde farklı yaklaşımlarla elde edilen puanlayıcılar arası güvenilirliğin değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 6(1), 12-24.

- Güneş, S. (2020). *Açık uçlu matematik sorularının değerlendirilmesinde puanlayıcı güvenirliğinin genellenebilirlik kuramına göre incelenmesi*, Yüksek Lisans Tezi, Mersin Üniversitesi, Eğitim Bilimleri Enstitüsü, Mersin.
- Gronlund, N.E. (1998). *Assessment of students achievement*. Boston: Allyn and Bacon.
- Haiyang, S. (2010). An application of classical test theory and many facet rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics*, 33(2), 87-102,
- Haladyna, T. M. (1997). *Writing Test Items to Evaluate Higher Order Thinking*. Boston: Allyn & Bacon.
- Hambleton, R. K. & Jones, R. W. (1993). Comparison between classical test theory and item response theory and their applications to the test development. *Educational Measurement*, 12, 38-47.
- Heidi L. Andrade, Ying Du & Kristina Mycek (2010). Rubric-referenced self assessment and middle school students' writing. *Assessments in education: Principles, Policy & Practice*. 17:2, 199-214. doi: 10.1080/09695941003696172.
- 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.
- İlhan, M. (2015). *Standart ve solo taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli rasch ölçme modeli ile incelenmesi*. Doktora tezi, Gaziantep Üniversitesi, Eğitim Bilimleri Enstitüsü, Gaziantep.
- İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerdeki klasik test kuramı ve çok yüzeyli rasch ölçme modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması, *Hacettepe Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 31(2), 346-368.

- Jacobs, R. & Kozlowski, S. W. (1985). A closer look at halo error in performance ratings. *Academy of Management Journal*, 28(1), 201-212. https://www.jstor.org/stable/256068?read-now=1&seq=1#page_scan_tab_contents sayfasından 17.02.2023 tarihinde erişilmiştir.
- Karakaya, İ. (2015). Comparison of self, peer and instruction assessments in the portfolio assessments by using many facet rasch model. *Journal of Education and Human Development*, 4(2), 182-192.
- Karakaya, İ. (Ed.). (2022). *Açık Uçlu Soruların Hazırlanması, Uygulanması ve Değerlendirilmesi*. Ankara: Pegem.
- Karasar, N. (1994). *Bilimsel araştırma yöntemi*. Ankara: Nobel.
- Knoch, U., Read, J. & Randow, J. (2007). Re-training writing raters online: How does compare with face-to-face training?. *Assessing Writing*, 12(2). 26-43. <https://doi.org/10.1177/0265532210384252>.
- Kocaoğlu, S. (2023). *Açık Uçlu ve Çoktan Seçmeli Maddelerden Oluşan Madde Takımlarının Güvenirliliğe Etkisinin Genellenebilirlik Kuramı ile İncelenmesi*, Yüksek Lisans Tezi, Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.
- Koons, H. H. (2008). *The reading-writing connection: An investigation of the relationship between reading ability and writing quality across multiple grades and three writing discourse modes*. Doctoral Dissertation. University of North Carolina, ABD.
- Kutlu, Ö., Doğan, C. D. & Karakaya, İ. (2014). *Eğitimde ölçme ve değerlendirme-performansa ve portfolyoya dayalı durum belirleme*. Ankara: Pegem.
- Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and application*. San Diego, CA: Academic Press.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28, 563-575.

- Lee, M., Peterson, J. J., & Dixon, A. (2010). Rasch calibration of physical activity self-efficacy and social support scale for persons with intellectual disabilities. *Research in Developmental Disabilities*, 31(4), 903-913.
- Linacre, J. M. (1989). *Many- facet Rasch measurement*. Chicago: Mesa.
- Linacre, J. M. (1993). Generalizability theory and many facet rasch measurement. *Annual Meeting Of The American Educational Research Association'da sunulmuş bildiri*. (ED 364 573). Atlanta Georgia. 26.09.2022 tarihinde <https://files.eric.ed.gov/fulltext/ED364573.pdf> sayfasından erişilmiştir.
- Linacre, J. M. (2017). *A user's guide to FACETS: Rasch-model computer programs*. Chicago: MESA Press.
- Matsuno, S. (2009). *Self-,peer-, and teacher- assessments in Japanese university EFL writing classrooms*. Doktora Tezi. Temple University
- 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>
- McMillan, J. H. (2010). The Practical Implications of Educational Aims and Contexts for Formative Assessment. In H. L. Andrade & G. J. Cizek (Eds.). *Handbook of Formative Assessment* (pp. 41-58). New York: Routledge.
- 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.
- Nakamura, Y. (2002). Teacher assessment and peer assessment in practice. *Educational Studies*, 44, 45-85.
- Ölçme, Seçme ve Yerleştirme Merkezi (2015). Yazılı Sınav (*Açık Uçlu Sorularla Sınav*) 17.09.2022 tarihinde <https://www.osym.gov.tr/TR,721/yazili-sinav-acik-uclu-sorularla-sinav-04022015.html> internet sitesinden erişilmiştir.

- Özbaşı, D. & Arcagök, S. (2021). Öğrenci Projelerinin Genellenebilirlik Kuramı ile İncelenmesi. *Eğitimde Kuram ve Uygulama*, 17(2), 69-78. Doi: 10.17244/edu.1024532
- Popham, W. J. (1990). *Modern educational measurement: A practitioner's perspective*. Englewood Cliffs. NJ: Prentice Hall.
- Reiner, C. M., Bothell, T. W., Sudweeks, R. R., & Wood, B. (2002). *Preparing effective essay questions*. Oklahoma: New Forums Pres.
- 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>
- Shalvelson, R. J. & Webb, N. M. (1991). *Generalizability theory: a primer*. USA: Sage.
- Şahin, M. G., Güler, N., & Taşdelen Teker, G. (2016). An Analysis of Peer Assessment through Many Facet Rasch Model, *Journal of Education and Practice*, 7(32), 172-181.
- Şata, M. (2019). *Performans değerlendirme sürecinde puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisinin incelenmesi*. Doktora Tezi. Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara, Türkiye.
- Şata, M. & Karakaya, İ. (2020). *Investigation of the use of electronic portfolios in the determination of student achievement in higher education using many-facet Rasch measurement model*. *Educational Policy Analysis and Strategic Research*, 15(7-21).
- Tan, Ş., & Erdoğan, A. (2004). *Öğretimi planlama ve değerlendirme*. Ankara: Pegem.
- Taştan, Z. (2021). Genellenebilirlik kuramında çok yüzeyli desenlerin incelenmesi. *Türkiye Sosyal Araştırmalar Dergisi*, 25(3), 743-756.
- Tekin, H. (1991). *Eğitimde ölçme ve değerlendirme*. Ankara: Yargı.

- Temizkan, M. (2009). Akran deęerlendirmesinin konuřma becerisinin geliřtirilmesi zerindeki etkisi. *Mustafa Kemal niversitesi Sosyal Bilimler Enstits Dergisi*, 6(12), 90-112.
- Tobař, C. (2020). *Performansın deęerlendirilmesinde farklılařan puanlayıcı davranıřlarının ok yzeyli rasch lme modeli ile incelenmesi*. Yksek Lisans Tezi, Gazi niversitesi, Eęitim Bilimleri Enstits, Ankara, Trkiye.
- Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, 68, 249-276.
- Topping, K. J. (2010). Peers as a Source of Formative Assessment. In H. L. Andrade & G. J. Cizek (Eds.). *Handbook of Formative Assessment* (pp. 61-74). New York: Routledge.
- Uysal, K. (2008). *ęrencilerin lme ve deęerlendirme srecine katılması: akran deęerlendirme ve z deęerlendirme*. Yksek Lisans Tezi, Abant İzzet Baysal niversitesi, Sosyal Bilimler Enstits, Bolu.
- 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>.
- Wagemaker, H. (Ed). (2020). *Reliability and validity of international large-scale assessment*. Springer: Switzerland.
- Wan, S. (2017). *An application of many-facet Rasch model to the analysis of the implicit associate test: Disentangling implicit gender-science stereotype*. Doctoral dissertation, University of Georgia.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263-287. <https://doi.org/10.1177/026553229801500205>

- Wolfe, E. W. & Chiu, C. W. (1997). Detecting rater effects with a multi-faceted rating scale model. *The Annual Meeting of the National Council on Measurement in Education Chicago, IL, March 25-27*. <https://eric.ed.gov/?id=ED408324> sitesinden 24.02.2023 tarihinde 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>
- Yurdugöl, H. (2005). *Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması*. XIV. Ulusal Eğitim Bilimleri Kongresi'nde sunulmuş bildiri, Pamukkale Üniversitesi Eğitim Bilimleri Fakültesi, Denizli.
- Yüzüak, A. V., Yüzüak, B. & Kaptan, F. (2015). Performans görevinin akran gruplar ve öğretmen yaklaşımları doğrultusunda çok-yüzeyli rasch ölçme modeli ile analizi. *Eğitim ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 6(1), 1-11.

EKLER



EK 1. Gazi Üniversitesi Etik Komisyonu İzin Belgesi

Evrak Tarih ve Sayısı: 12.01.2023-E.558560



T.C.
GAZİ ÜNİVERSİTESİ REKTÖRLÜĞÜ
Etik Komisyonu

Sayı : E-77082166-302.08.01-558560
Konu : Bilimsel ve Eğitim Amaçlı

12.01.2023

Dağıtım Yerlerine

Üniversitemiz Eğitim Bilimleri Enstitüsü Eğitim Bilimleri Ana Bilim Dalı Yüksek Lisans Öğrencisi Burçin GÖRGÜLÜ ÖZTÜRK'ün, Prof.Dr.İsmail KARAKAYA'nın danışmanlığında yürüttüğü *"Açık Uçlu Matematik Maddelerinin Puanlanmasında Öz ve Akran Değerlendirmesinin Puanlayıcı Davranışına Etkisinin Çok Yüzeyle Rasch Ölçme Modeli ile İncelenmesi"* adlı tez çalışması ile ilgili konu Komisyonumuzun 10.01.2023 tarih ve 01 sayılı toplantısında görüşülmüş olup,

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

Bilgilerinizi rica ederim.

Araştırma Kod No: 2023 - 42

Prof. Dr. İsmail KARAKAYA
Komisyon Başkanı

Ek:1 Liste

DAĞITIM

Gereği:

Sayın Prof. Dr. İsmail KARAKAYA

Bilgi:

Eğitim Bilimleri Enstitüsü Müdürlüğüne

Evrak Tarih ve Sayısı: 24.11.2023-E.807705



T.C.
GAZİ ÜNİVERSİTESİ REKTÖRLÜĞÜ
Etik Komisyonu



Sayı : E-77082166-302.08.01-807705
Konu : Bilimsel ve Eğitim Amaçlı

24.11.2023

Dağıtım Yerlerine

Daha önce 12.01.2023 tarih v eE.558560 sayılı yazımız ile onay alan 2023-42 kod numaralı, Üniversitemiz Eğitim Bilimleri Enstitüsü Eğitim Bilimleri Ana Bilim Dalı Yüksek Lisans Öğrencisi Burçin GÖRGÜLÜ ÖZTÜRK'ün, Prof.Dr.İsmail KARAKAYA'nın danışmanlığında yürüttüğü "Açık Uçlu Matematik Maddelerinin Puanlanmasında Öz ve Akran Değerlendirmesinin Puanlayıcı Davranışına Etkisinin Çok Yüzeyle Rasch Ölçme Modeli ile İncelenmesi" başlıklı tez çalışması hakkında ilgililerden alınan 17.11.2023 tarihli dilekçe Komisyonumuzun 21.11.2023 tarih ve 20 sayılı toplantısında görüşülmüş olup,

Dilekçede; tez savunması aşamasında jüri tarafından tez başlığının değiştirilmesi nedeniyle, çalışmanın başlığının "Açık Uçlu Maddelerin Puanlanmasında Öğrencilere Verilen Puanlayıcı Eğitiminin Puanlayıcı Davranışlarına Etkisinin Çok Yüzeyle Rasch Ölçme Modeli ile İncelenmesi" olarak değiştirilmesi ile ilgili talebin uygun olduğuna oybirliği ile karar verilmiş ve karara ilişkin imza listesi ekte gönderilmiştir.

Bilgilerinizi rica ederim.

Prof. Dr. İsmail KARAKAYA
Komisyon Başkanı

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

Bilgi:
Eğitim Bilimleri Enstitüsü Müdürlüğüne

EK 2. MEB Araştırma İzni

Evrak Tarih ve Sayısı: 21.02.2023-E.593206



T.C.
ANKARA VALİLİĞİ
Milli Eğitim Müdürlüğü

Sayı : E-14588481-605.99-70661175
Konu : Araştırma İzni

20.02.2023

GAZİ ÜNİVERSİTESİ REKTÖRLÜĞÜNE

- İlgi: a) 09.02.2023 tarihli ve 584238 sayılı yazınız.
b) MEB Yenilik ve Eğitim Teknolojileri Genel Müdürlüğünün 2020/2 nolu Genelgesi.

Üniversiteniz Eğitim Bilimleri Enstitüsü Yüksek Lisans Öğrencisi Burçin GÖRGÜLÜ ÖZTÜRK'ün "Açık Uçlu Matematik Maddelerinin Puanlanmasında Öz ve Akran Değerlendirmesinin Puanlayıcı Davranışına Etkisinin Çok Yüzeyle Rasch Ölçme Modeli İle İncelenmesi" konulu tezi kapsamında okul ve kurumlarda uygulanacak olan veri toplama araçları ilgi (b) Genelge çerçevesinde incelenmiştir.

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

Bilgilerinizi ve gereğini rica ederim.

Harun FATSA
Vali a.
Milli Eğitim Müdürü

Ek:
Uygulama araçları (9 sayfa)
Dağıtım:
Gereği:
Gazi Üniversitesi
Bilgi:
25 İlçe MEM

EK 3. Puanlayıcı Eğitimi Planı

Süre (dk)	EĞİTİMİN İÇERİĞİ	HEDEFLenen ÖĞRENCİ KAZANIMLARI	KULLANILAN YÖNTEM
40 dk	Açık uçlu madde kavramı hakkında bilgilendirme Biçimlendirici ölçme araçlarından akran değerlendirme kavramının açıklanması ve önemi	Öğrenciler açık uçlu madde hakkında bilgi sahibi olur. Açık uçlu maddelerin özelliklerini söyler. Akran değerlendirmenin tanımını yapar. Akran değerlendirmenin avantajlarını açıklar.	Anlatım Soru Cevap
40 dk	Dereceli puanlama anahtarları hakkında bilgilendirme	Öğrenciler DPA'nın tanımını yapar. Öğrenci açık uçlu maddeleri DPA yardımıyla değerlendirebileceğini açıklar.	Anlatım Soru Cevap
40 dk	Puanlayıcı eğitimi hakkında bilgilendirme En sık yapılan puanlayıcı hataları (cömertlik/katılık , halo etkisi ve farklılaşan puanlayıcı cömertliği ve katılığı)	Öğrenciler puanlayıcı eğitiminin tanımını söyler. Öğrenciler en sık yapılan puanlayıcı hatalarını bilir. Bu hataların değerlendirme sürecindeki etkilerini söyler.	Anlatım Tartışma Soru cevap
80 dk	Örnek soru uygulamaları, DPA'ya göre akran değerlendirilmesinin yapılması	Öğrenciler, öğretmen eşliğinde, örnek açık uçlu soruları nasıl puanlayacağını söyler. Puanlama yaparken puanlayıcı hatalarından kaçınmaya çalışır.	Bireysel puanlama Akran puanlaması Sınıf içi tartışma
80 dk	Örnek soru uygulamaları, geri dönüt verme, puanlama örnekleri yapma	Öğrenciler soruları nasıl puanlayacaklarına ilişkin fikir birliği yapar. Öğrenciler soruları puanlama açısından ortak bir anlayış içinde olurlar.	Bireysel puanlama Akran puanlaması Sınıf içi tartışma

EK 5. Dereceli Puanlama Anahtarları

Yemek Kartı

Kodlar	Yanıtlar	Puan
20	Tam Doğru Yanıtlar	3
	$f(x) = 100x + 2000$ $f(x) = x + 2000$ $f(x) = x^2 + 2000$ $f(x) = 50x + 2000$ vb.. şeklinde yukarıdaki bağıntılara benzeyen <i>cevabı en az 2000 Türk lirası</i> olan yanıtlar. Bu bağıntıda çocuğu olmayan çalışan ve 3 çocuğu olan çalışan için doğru hesaplama yapan yanıtlar: Örneğin; $f(x) = 100x + 2000$ yazar ve çocuğu olmayan çalışan için: $f(0) = 100 \cdot 0 + 2000 = 2000$ TL yardım alır. 3 çocuklu bir çalışan için: $f(3) = 100 \cdot 3 + 2000 = 2300$ TL yardım alır. Şeklinde açıklama yapan yanıtlar.	
	Yakın Doğru Yanıtlar	
15	Yanıt olarak bir fonksiyon kuralı yazmayan ancak çocuğu olmayana 2000, 1 çocuğu olana 2100, 2 çocuğu olana 2200, 3 çocuğu olana 2300 şeklinde doğru bir artış yazan yanıtlar	2
14	Sorudaki kurala uygun bağıntının yazıldığı ve sadece çocuksuz ya da 3 çocuklunun bir tanesinin hesaplandığı yanıtlar	2
	Kısmen Doğru Yanıtlar	
12	Sorudaki kurala uygun doğru bir bağıntının yazıldığı ancak çocuksuz ve 3 çocukluya göre hesaplamanın <u>yapılmadığı</u> veya bunların yanlış hesaplandığı yanıtlar	1
10	Soruya uygun bağıntının yazılmadığı sadece çocuksuz ya da 3 çocuklu çalışanın alabileceği miktarı yazan yanıtlar	1
	Yanlış Yanıtlar	
7	Bağıntıyı kurala uygun kurmayan ve yanlış hesaplamalar yapılan yanıtlar	0
6	Boş bırakılan yanıtlar	0
	Diğer Yanıtlar	0
4	Soru ile ilgili alakasız yanıtlar	0

Yakıt Zammı

Kodlar	Yanıtlar	Puan
	Tam Doğru Yanıtlar	
20	Öğrenci Aylin hanımın görüşüne katılmadığını yazar. Benzin için; motorin için; 21,50 TL de 4 TL zam 20 TL de 4 TL zam 100 TL de x 100 TL de x $x = 400/21,50 = 18,60$ (%18,60 zam) $x = 400/20 = 20$ bulunur. (%20 zam) şeklinde işlemlerle doğru cevabı gösterip “ benzine yapılan zam oranı yüzde olarak 18,60 iken motorine yapılan zam oranı %20 dir” şeklinde bir açıklama yapan yanıtlar.	3
	Yakın Doğru Yanıtlar	
15	Aylin hanıma görüşüne katılmadığını belirten ancak benzindeki artışın motorinden daha az olduğunu işlem yapmadan sözel biçimde mantıklı bir açıklama ile belirten yanıtlar.	2
13	Aylin hanımın görüşüne katılmadığını belirtip küçük işlem hataları olsa da motorine yapılan zam yüzdesini daha çok bulan yanıtlar	2
	Kısmen Doğru Yanıtlar	
12	Herhangi bir işlem yapılmayan sadece “motorine gelen zam, benzine gelen zamdan daha fazladır” şekline açıklama yapılan yanıtlar.	1
10	Benzin ya da motorinden herhangi birinin yüzdesini hesaplayıp diğerini hesaplamayan yanıtlar.	1
	Yanlış Yanıtlar	
7	Aylin hanıma katıldığını ve Aylin hanımın açıklamasının doğru olduğunu açıklayan yanıtlar	0
6	Benzin ve motorine yapılan zam yüzdesini eşit bulan yanıtlar	
5	Boş bırakılan yanıtlar	0
	Diğer Yanıtlar	0
4	Soru ile ilgili alakasız yanıtlar	0

Kitap Satışı

Kodlar	Yanıtlar	Puan
	Tam Doğru Yanıtlar	
20	Gama yayınevi ile çalışırsa; $100 \cdot 8 = 800$ TL $200 \cdot 10 = 2000$ TL ve toplamda $800 + 2000 = 2800$ TL kazanır. Omega yayınevi ile çalışırsa; $300 \cdot 4 + 5 \cdot 300 = 1200 + 1500 = 2700$ TL kazanır. “Ahmet’in yerinde olsaydım Gama yayınevi ile anlaşırdım çünkü 100 TL daha fazla kazanırım” “Ahmet’in yerinde olsam Gama yayınevi ile anlaşırdım çünkü daha fazla kazanırım” şeklindeki yanıtlar	3
	Yakın Doğru Yanıtlar	

15	İşlemleri doğru bir şekilde gösterip hangi yayınevini seçtiğini belirtmeyen yanıtlar	2
13	Gerekli işlemleri yaparken küçük bir işlem hatası yapıp yine de Gama yayınevini seçerdim çünkü daha fazla kazanırdım şeklinde açıklama yapılan yanıtlar	2
	Kısmen Doğru Yanıtlar	
12	Sadece Gama yayınevinden alacağı ya da sadece Omega yayınevinden alacağı ücreti bulup gerekçenin açıklanmadığı yanıtlar	1
10	İşlemler yapılmadan sadece “Gama yayınevini seçerdim” şeklindeki yanıtlar.	1
	Yanlış Yanıtlar	
7	Yanlış hesaplamalar yapılan ve Omega yayınevinin seçildiği yanıtlar	0
6	Boş bırakılan yanıtlar	0
	Diğer Yanıtlar	
4	Soruyla alakasız çözümler yapanlar.	0

Çark Çevirme

Kodlar	Yanıtlar	Puan
	Tam Doğru Yanıtlar	
20	“Kemal’in birinci Naz’ın sonuncu gelmesi için sarı rengin en çok, kırmızı rengin ise en az olması gerekmektedir. Böylece Kemal sarı rengi daha çok bölmeye boyarsa çarkı çevirdiğinde sarının gelme olasılığı yükselecektir.” Şeklinde bir açıklama yapan öğrenciler. Bunun için; Sarı 7 bölme, pembe 2 bölme, kırmızı 1 bölme; Sarı 6 bölme, pembe 3 bölme, kırmızı 1 bölme; Sarı 5 bölme, pembe 4 bölme, kırmızı 1 bölme; Sarı 5 bölme, pembe 3 bölme, kırmızı 2 bölme; vb. Şeklinde bölmeleri paylaştıran yanıtlar.	3
	Yakın Doğru Yanıtlar	
15	Kemal’in birinci ve Naz’ın sonuncu olması için sarının en çok kırmızısını en az olmasını yazan, ancak yaptığı dağıtımda bu kurala uyarınca sayıların toplamını 10 yapamaması	2
13	Açıklama yapmayıp sadece renkleri doğru sayılarda dağıtabilen yanıtlar	2
	Kısmen Doğru Yanıtlar	
12	Sadece sarının en çok kırmızısının en az olması gerektiğini yazan yanıtlar	1
10	Herhangi iki renk için kaç bölme boyayacağını yazıp diğer renk için yorum yapılmamış yanıtlar	1
	Yanlış Yanıtlar	
7	Açıklama yapmayıp sayıları yanlış dağıtım yapan yanıtlar	0
6	Boş bırakılan yanıtlar	0
	Diğer Yanıtlar	0

4	Soru ile ilgili alakasız yanıtlar	0
---	-----------------------------------	---





GAZİLİ OLMAK AYRICALIKTIR..