



**FARKLI EĞİTİM DESENLERİNE GÖRE VERİLEN ÇEVİRİM İÇİ PUANLAYICI
EĞİTİMİNİN PUANLAMA DAVRANIŞLARI ÜZERİNE ETKİSİNİN
İNCELENMESİ**

Nazira Tursynbayeva

**DOKTORA TEZİ
EĞİTİM BİLİMLERİ ANA BİLİM DALI**

**GAZİ ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ**

MART, 2025

TELİF HAKKI VE TEZ FOTOKOPİ İZİN FORMU

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

YAZARIN

Adı : Nazira
Soyadı : Tursynbayeva
Bölümü : Eğitimde Ölçme ve Değerlendirme
İmza :
Teslim tarihi :

TEZİN

Türkçe Adı : Farklı Eğitim Desenlerine Göre Verilen Çevrim İçi Puanlayıcı Eğitimlerinin Puanlama Davranışları Üzerine Etkisinin İncelenmesi
İngilizce Adı : An Examination of the Effect of Online Rater Training Based on Different Educational Designs on Rater Behaviors

ETİK İLKELERE UYGUNLUK BEYANI

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

Yazarın Adı Soyadı : Nazira Tursynbayeva

İmza :

JÜRİ ONAY SAYFASI

Nazira Tursynbayeva tarafından hazırlanan “Farklı Eğitim Desenlerine Göre Verilen Çevrim İçi Puanlayıcı Eğitiminin Puanlama Davranışları Üzerine Etkisinin İncelenmesi” adlı tez çalışması aşağıdaki jüri tarafından oy birliği ile Gazi Üniversitesi Eğitimde Ölçme ve Değerlendirme Anabilim Dalı’nda Doktora tezi olarak kabul edilmiştir.

Danışman: Prof. Dr. İsmail Karakaya
(Eğitimde Ölçme ve Değerlendirme, Gazi Üniversitesi)

Üye: Prof. Dr. Dilara Bakan Kalaycıoğlu
(Eğitimde Ölçme ve Değerlendirme, Gazi Üniversitesi)

Üye: Prof. Dr. Melek Gülşah Şahin
(Eğitimde Ölçme ve Değerlendirme, Gazi Üniversitesi)

Üye: Prof. Dr. Celal Deha Doğan
(Eğitimde Ölçme ve Değerlendirme, Ankara Üniversitesi)

Üye: Doç. Dr. Mehmet Şata
(Eğitimde Ölçme ve Değerlendirme, Van Yüzüncü Yıl Üniversitesi)

Tez Savunma Tarihi: 16/01/2025

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

Prof. Dr. Şaban Çetin

Eğitim Bilimleri Enstitüsü Müdürü



*“İki Dünya Eşiği Er Türk’ün
Beşiği Türkistan’a ve Aileme”*

*“Екі дүние есігі, Ер Түріктің
бесігі- киелі Түркістанға
және отбасыма”*

TEŞEKKÜR

Bu doktora tez çalışmasının başlangıcından bitimine kadarki süreçte yanımda olan, ayrıca bilimsel hazırlık ve doktora eğitimimden itibaren akademi ve eğitimci olarak gelişimimde önemli katkılar sunan, desteğini ve inancını her zaman hissettiğim, alçakgönüllülüğü, sabrı ve işine olan sevgisiyle eğitimcilik yönünü her zaman örnek aldığım saygıdeğer danışman hocam Prof. Dr. İsmail Karakaya'ya sonsuz teşekkürlerimi sunarım. Doktora programına başladığım günden itibaren, özellikle tezimle ilgili çalışmaları benim için bir rehber niteliği taşıyan ve akademik yolculuğumda değerli katkılarda bulunan kıymetli tez izleme komite hocalarım Doç. Dr. Mehmet Şata ve Prof. Dr. Melek Gülşah Şahin'e tez çalışmamın planlanmasından bitimine kadarki olan süreçteki vermiş oldukları katkı ve paylaşımlarından dolayı çok teşekkür ederim. Doktora sürecinde ders döneminde derslerini alma şansını bulduğum ve bilimsel ufkumu genişleten değerli tez savunma jüri üyesi hocam Prof. Dr. Dilara Bakan Kalaycıoğlu ve doktora yeterlik sınavı jüri üyem ve tez savunma jüri üyem olarak sağladığı yapıcı eleştiri, yol gösterici tavsiyeleri ve değerli görüşlerini sunan değerli hocam Prof. Dr. Celal Deha Doğan'a teşekkürlerimi sunarım.

Tezimin veri toplama sürecinde katkıda bulunan değerli Müzik Anabilim Dalı Başkanı Prof. Dr. Aytekin Albuz ve anabilim dalındaki hocalarıma sürece gösterdiği çaba ve katkıları nedeniyle saygıdeğer Doç. Dr. Çağla Serin Özparlak'a en içten teşekkürlerimi sunarım. Gerek ders döneminde gerekse tez yazım sürecinde kendisinden kazandığım bilgi ve beceriler, paylaştığı deneyimi ve fikirlerinden dolayı tezimin daha iyi hale gelebilmesinde katkı sunan değerli arkadaşım Dr. Umur Öç'e de teşekkür ederim.

Türkiye'de doktora yapabilmeme imkân sağlayan Ahmet Yesevi Üniversitesi Mütevelli Heyet başkanlığına kurumun tüm çalışanlarına ve bu süreçte bilgi ve deneyimlerini cömertçe paylaşarak beni daha iyi bir akademisyen olmaya teşvik ettikleri için Gazi Eğitimde Ölçme ve Değerlendirme ABD'deki tüm hocalarıma teşekkür ederim.

Bilim insanı olma yolculuğumda bana rehberlik eden, akademik ve insani yönleriyle örnek aldığım çok değerli hocalarım Prof. Dr. Mehmet Kutalmış'a ve Prof. Dr. Dosay Kenzhetay'a sonsuz teşekkür ederim. Verdikleri destek benim için çok kıymetlidir. Her zaman yanımda olan, sevgileri ve sabırlarıyla beni güçlendiren aileme, arkadaşlarıma ve ölçmenin neferlerine minnettarım. Onların desteği olmadan bu yolculuk eksik kalırdı.

**FARKLI EĞİTİM DESENLERİNE GÖRE VERİLEN ÇEVİRİM İÇİ
PUANLAYICI EĞİTİMİNİN PUANLAMA DAVRANIŞLARI
ÜZERİNE ETKİSİNİN İNCELENMESİ**
(Doktora Tezi)

Nazira Tursynbayeva
GAZİ ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
Mart 2025

ÖZ

Bu araştırmada, öğrencilerin müzik performansının puanlanmasında puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisinin incelenmesi amaçlanmaktadır. Bu amaç kapsamında araştırmada eşit olmayan kontrol gruplu desen kullanılmıştır. Araştırmanın iki deney ve bir kontrol grubu mevcuttur. Deney A grubunda 10 puanlayıcı, Deney B grubunda 7 puanlayıcı ve Kontrol grubunda 8 puanlayıcı olmak üzere toplamda 25 müzik alanındaki öğretim elemanlarından oluşmuştur. Yapılan bu araştırmada her gruptaki puanlayıcıların puanlama davranışları olarak şunlar incelenmiştir: katılık, cömertlik, halo etkisi, farklılaşan puanlayıcı katılığı ve puanlayıcı cömertliği. Araştırmaya katılan puanlayıcı veya değerlendiricilerin davranışları hem grup düzeyinde hem de bireysel düzeyde incelenmiştir. Araştırmada veri toplama aracı olarak müzik bölümünde birinci sınıfta eğitim gören on öğrencinin başlangıç düzey piyano performansına ilişkin iki tane eser alınmıştır. Araştırmacılar tarafından geliştirilen analitik dereceli puanlama anahtarı ile puanlayıcıların değerlendirme formu ölçme aracı olarak kullanılmıştır. Buna göre veriler çevrim içi olarak toplanarak puanlayıcıların her bir öğrenci performansını değerlendirirken gösterdikleri davranışlar incelenmiştir. Puanlayıcılardan ön uygulama alındıktan sonra Deney A grubuna Referans Çerçevesi ile Puanlayıcı Hatası; Deney B grubuna ise Referans Çerçevesi ile Performans Boyutu Eğitimi çevrim içi olarak verilmiştir. Çok Yüzeyle Rasch ölçme modeli kullanılan bu araştırmada puanlayıcıların yapmış olduğu analizlerde tümüyle çaprazlanmış desen tercih edilmiştir. Puanlayıcı eğitimi sadece iki deney grubunda uygulanarak dört haftada toplamda on iki saatlik puanlayıcı eğitimi gerçekleştirilmiştir. Araştırmada öğrenci performansı, ölçütler ve puanlayıcılar olmak üzere üç yüzey bulunduğu için her bir yüzeyin ölçüm sonuçları ve uyum istatistikleri incelenmiştir. Deney A, Deney B ve Kontrol grubu ön ve son uygulama sonuçları incelendiğinde tüm yüzeylerin model veri uyumunun iyi olduğu belirlenmiştir. Puanlayıcı eğitimi öncesinde puanlayıcıların katılık, cömertlik, halo etkisi ve yanlılık davranışlarının hem grup düzeyinde hem de bireysel düzeyde ortaya çıkma durumları incelendiğinde; Deney A, Deney B ve Kontrol ön uygulama gruplarında grup düzeyinde katılık, cömertlik ve yanlılık davranışlarının olduğu belirlenirken gruplarda grup düzeyinde

halo etkisi davranışının olmadığı belirlenmiştir. Her bir grubun bireysel düzeyde puanlayıcı davranışları incelendiğinde Deney A ön uygulama grubunda iki cömert, üç katı, iki halo etkisi ve üç tane de yanlışlık davranışlarının olduğu belirlenmiştir. Deney B grubu ön uygulama grubunda iki cömert, üç katı, bir halo etkisi ve iki tane de yanlışlık davranışlarının olduğu belirlenmiştir. Kontrol grubu ön uygulama grubunda üç cömert, üç katı, iki halo etkisi ve bir tane de yanlışlık davranışı belirlenmiştir. Puanlayıcı eğitimi sonrasında puanlayıcıların katılık, cömertlik, halo etkisi ve yanlışlık davranışlarının hem grup düzeyinde hem de bireysel düzeyde ortaya çıkma durumları incelendiğinde; Deney A ve Kontrol son uygulama gruplarında grup düzeyinde katılık, cömertlik ve yanlışlık davranışlarının olduğu belirlenirken gruplarda grup düzeyinde halo etkisi davranışının olmadığı belirlenmiştir. Deney B grubu son kontrol grubunda ise grup düzeyinde katılık ve cömertlik davranışları yer alırken halo etkisi ve yanlışlık davranışlarının yer almadığı belirlenmiştir. Bu durum Deney B grubuna verilen puanlayıcı eğitiminin halo ve yanlışlık davranışları üzerinde daha etkili olduğu ile açıklanabilir. Deney A grubuna verilen eğitimde eğitim süresi, eğitimin zamanı, eğitmen kaynaklı sınırlılıklar gibi durumlar eğitimin etkisini sınırlandırmış olabilir. Her bir grubun bireysel düzeyde puanlayıcı davranışları incelendiğinde Deney A son uygulama grubunda üç cömert dört katı, iki halo etkisi ve iki tane de yanlışlık davranışlarının olduğu belirlenmiştir. Deney B grubu son uygulama grubunda iki cömert, bir katı, iki halo etkisi ve bir tane de yanlışlık davranışlarının olduğu belirlenmiştir. Kontrol grubu son uygulama grubunda üç cömert, üç katı, üç halo etkisi ve iki tane de yanlışlık davranışı belirlenmiştir. Tüm bireysel sonuçlar grup içinde değerlendirildiğinde; Deney A grubuna verilen puanlayıcı eğitiminin puanlayıcıların grup düzeyindeki davranışlarını değiştirmemiştir. Deney A grubunda ön uygulamada katılık davranışı gösteren iki puanlayıcının puanlayıcı eğitiminden sonra davranışlarının ortadan kalktığı belirlenmiştir. Ayrıca Deney A grubunda halo etkisi davranışı gösteren iki puanlayıcının ve yanlışlık davranışı gösteren üç puanlayıcının da puanlayıcı eğitiminden sonra bu davranışları ortadan kalmıştır. Bu durum Deney A grubuna verilen eğitimin puanlayıcıların katılık, halo etkisi ve yanlışlık davranışları üzerinde etkili olduğunu göstermektedir. Deney B grubuna verilen puanlayıcı eğitiminin puanlayıcıların yanlışlık davranışlarını ortadan kaldırdığı belirlenmiştir. Ayrıca Deney B grubunda ön uygulamada iki katılık, üç cömertlik, bir halo etkisi ve bir de yanlışlık davranışlarının ortadan kalktığı belirlenmiştir. Bu durum Deney B grubuna verilen eğitimin puanlayıcıların katılık, halo etkisi ve yanlışlık davranışları üzerinde etkili olduğunu göstermektedir. Ayrıca Deney B grubuna verilen puanlayıcı eğitiminin Deney A grubuna göre daha etkili olduğu da söylenebilir. Çünkü Deney B grubuna verilen puanlayıcı eğitimi sonrasında ön uygulamada farklı puanlayıcı davranışlarına sahip olan tüm puanlayıcılar benzer davranışlara sahip olmuşlardır. Kontrol grubunda ise her iki uygulamada da puanlayıcı davranışları genel olarak benzer kalmıştır. Bu durum beklenen bir durumdur.

Anahtar Kelimeler :Puanlayıcı Eğitimi Desenleri, Puanlayıcı Davranışları, Performans Değerlendirme, Çok Yüzeyli Rasch Ölçme Modeli, Müzik Performansları
Sayfa Adedi : xvii +145
Danışman : Prof. Dr. İsmail Karakaya

AN EXAMINATION OF THE EFFECT OF ONLINE RATER TRAINING BASED ON DIFFERENT EDUCATIONAL DESIGNS ON RATING BEHAVIORS

(Ph. D Thesis)

Nazira Tursynbayeva

GAZI UNIVERSITY

GRADUATE SCHOOL OF EDUCATIONAL SCIENCES

March 2025

ABSTRACT

The purpose of this research is to investigate how rater training affects rater behaviors in the student musical performance scoring. The research used an uneven control group structure for this aim. The research included one control group and two experimental ones. There are 25 music faculty members and members overall— Experimental A group has 10 raters; Experimental B group has 7 raters—and the Control group has 8 raters. The scoring behaviors of the raters in each group were investigated in this work under the following categories: (severity, generosity, halo effect, differential rater severity and generosity). Examined both group and individual levels, the actions of the raters—or evaluators— participated in the research. Two compositions on the beginning level piano performance of ten first-year students enrolled in the music department were chosen as the data collecting instruments for the research. Measurement instruments were the analytical rubric created by the raters' assessment form and the researchers. Consequently, the data were gathered online and the raters' conduct under evaluation of every student's performance was under analysis. Rater error was investigated using the reference frame in Experimental A Group after the first application collected from the raters; the reference frame and performance dimension training modules were sent online to Experimental Group B. In this work using a many-facet Rasch measurement model, the evaluations carried out by the raters favored a totally crossed design. Only two experimental groups had rater training, and in four weeks a total of twelve hours of rater instruction were dedicated. Examined were the measurement findings and fit statistics of each of the three surfaces in the study—student performance, criteria and raters. Examining pre- and post-application findings of Experiment A, Experiment B, and the Control group revealed that the model data fit of all surfaces was excellent. It was found that there were severity, leniency and bias behaviors at the group level in Experiment A,

Experiment B and the Control pre-application groups, while it was found that there was no halo effect behavior at the group level in the groups. This was examined both at the group level and at the individual level before the rater training. Examining the rater behaviors of every group at the individual level revealed in the Experiment A pre-application group two generosity, two halo effects, and three bias behaviors. The pre-application group of Experiment B revealed two generous, one halo effect, and two biased behaviors. Pre-application the Control group revealed three generous, three rigids, two halo effects, and one bias behavior. It was found that there were severity, generosity, and bias behaviors at the group level in the Experiment A and Control post-application groups, while there was no halo effect behavior at the group level in the groups when the severity, generosity, halo effect, and bias behaviors of the raters were investigated at both the group and individual levels following rater training. While there were no halo effects and bias behavior in the Experiment B post-Control group, it was found that there were severity and generosity behaviors at the group level there. The fact that the rater training offered the Experiment B group was more successful on bias behaviors helps to explain this circumstance. Limitations resulting from the teacher, the length of the training, and the timing of the training might have reduced the impact of the instruction in the training delivered to the Experiment A group. Examining the rater behaviors of every group at the individual level revealed three generous, four strict, two halo effect, two bias behaviors in the Experiment A post-application group. The Experiment B post-application group revealed two generous, one rigid, two halo effect, one biased conduct. In the Control group post-application, three generous, three rigorous, three halo effect, two bias behaviors were found. When all individual data were aggregated within the group; the rater training provided to the Experiment A group had no effect on the group level behavior of the raters. It was found that after rater training, the rigorous conduct of the raters in Experiment A vanished. Furthermore, disappearing following the rater training were the raters displaying halo effect behavior and those displaying bias behavior in the Experiment A group. These results imply that the training provided to the Experiment A group was efficient on the raters' strictness, halo effect and bias behaviors. Rater training assigned to the group Experiment B.

Key Words :Rater Training Designs, Rater Behaviors, Performansce Assessment,
Many Facet Rasch Model, Music Performances,
Page Number : xvii +145
Supervisor : Prof. Dr. 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
TABLO LİSTESİ	xiii
ŞEKİL LİSTESİ	xvi
SİMGELER VE KISALTMALAR.....	xvii
BÖLÜM I.....	1
GİRİŞ.....	1
Problem Durumu	1
Problem İfadesi	7
Araştırma Soruları.....	7
Araştırmanın Amacı	8
Araştırmanın Önemi	8
Sınırlılıklar	10
Tanımlamalar	10
BÖLÜM II.....	11
KAVRAMSAL ÇERÇEVE	11
Performans Değerlendirme.....	11
Performans Değerlendirmeye İlişkin Kavramsal Çerçeve	11
Müzik Performanslarının Ölçülmesi	13
Problemlili Puanlayıcı Davranışları.....	14

Puanlayıcı Katılığı ve Cömertliği.....	16
Merkeze Yönelim/ Eğilim	16
Halo Etkisi	17
Farklılaşan Puanlayıcı Katılığı ve Cömertliği (Puanlayıcı Yanlılığı)	18
Puanlayıcı Davranışların Azaltılmaya Yönelik Uygulamalar	18
Birden Fazla Puanlayıcının Kullanılması.....	18
Dereceli Puanlama Anahtarı.....	19
Bütünsel Puanlama Anahtarı	20
Analitik Dereceli Puanlama Anahtarı	20
Puanlayıcı Güvenirliği	24
Klasik Test Kuramına Göre Puanlayıcı Güvenirliği Belirleme Teknikleri	25
Genellenebilirlik Kuramına Göre Puanlayıcı Güvenirliği.....	25
Madde Tepki Kuramına Dayalı Puanlayıcı Güvenirliği.....	26
Çok Yüzeyle Rasch Ölçme Modeli	26
İlgili Araştırmalar	31
Sınıf İçi Eğitimi İle İlgili Performansın Değerlendirmesinde Puanlayıcı Davranışlarının İncelenmesi	31
Müzik Performansının Ölçülmesi İle İlgili Çok Yüzeyle Rasch Ölçme Modeli Kullanarak Yapılan Çalışmalar	35
BÖLÜM III	39
YÖNTEM	39
Araştırma Deseni.....	39
Deney Grubu	40
Veri Toplama Araçları.....	42
Analitik Dereceli Puanlama Anahtarının Geliştirilmesi.....	43
Verilerin Toplanması	46
Puanlayıcı Eğitimi	47
Verilerin Analizi	50
BÖLÜM IV	52
BULGULAR VE YORUMLAR.....	52
BÖLÜM V	108
SONUÇLAR VE ÖNERİLER.....	108
KAYNAKÇA	112
EKLER.....	127
EK 1. Gazi Üniversitesi Etik Komisyon İzin Belgesi.....	128

EK 2. Başlangıç Düzey Piyano Performansının Ölçülmesine İlişkin Analitik Dereceli Puanlama Anahtarı	129
EK 3. Puanlayıcı Değerlendirme Formu	132
EK 4. Kapsam Geçerlik Oranına Göre Kriterin Belirlenmesinde Kullanılan Eşik Değerler	133
EK 5. Deney A Ön Uygulama Puanlayıcı Grubu <i>Puanlayıcı X Öğrenci Performansı</i> Etkileşimleri	134
EK 6. Deney A Son Uygulama Puanlayıcı Grubu <i>Puanlayıcı X Öğrenci Performansı</i> Etkileşimleri	136
EK 7. Deney B Ön Uygulama Puanlayıcı Grubu <i>Puanlayıcı X Öğrenci Performansı</i> Etkileşimleri	138
EK 8. Deney B Son Uygulama Puanlayıcı Grubu <i>Puanlayıcı X Öğrenci Performansı</i> Etkileşimleri	140
EK 9. Kontrol Ön Uygulama Puanlayıcı Grubu <i>Puanlayıcı X Öğrenci Performansı</i> Etkileşimleri	142
EK 10. Kontrol Son Uygulama Puanlayıcı Grubu <i>Puanlayıcı X Öğrenci Performansı</i> Etkileşimleri	144

TABLO LİSTESİ

Tablo 1. <i>Performans Değerlendirmede Karşılaşılan Yaygın Hata Türleri</i>	15
Tablo 2. <i>Puanlayıcı Eğitimi Desenleri Amaç ve Odak Noktası</i>	24
Tablo 3. <i>KTK Temelli Puanlayıcılar Arası Güvenirliğin Belirlenmesi*</i>	25
Tablo 4. <i>Yarı Deneysel Süreç</i>	40
Tablo 5. <i>Çalışma Grubundaki Katılımcıların Demografik Bilgileri</i>	41
Tablo 6. <i>Ölçme Aracının Taslak Formunda Bulunan Kriterlere İlişkin Hesaplanan KGO Değerleri</i>	45
Tablo 7. <i>Deney A Grubu “Referans Çerçevesi ve Puanlayıcı Hatası Eğitim” Eğitimi</i>	48
Tablo 8. <i>Deney B Grubu “Referans Çerçevesi ve Performans Boyutu” Eğitimi</i>	49
Tablo 9. <i>Deney A Grubu Öğrenci Performansı Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	53
Tablo 10. <i>Deney A Grubu Puanlayıcı Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	55
Tablo 11. <i>Deney A Grubu Ölçüt Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	56
Tablo 12. <i>Deney A Grubu Öğrenci Performansı Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	57
Tablo 13. <i>Deney A Grubu Puanlayıcı Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	59
Tablo 14. <i>Deney A Grubu Ölçüt Yüzeylerinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	60
Tablo 15. <i>Deney B Grubu Öğrenci Performansı Yüzeyinin Ön Test Ölçümlerine Ait Ölçüm Raporu</i>	61

Tablo 16. <i>Deney B Grubu Puanlayıcı Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	63
Tablo 17. <i>Deney B Grubu Ölçüt Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	64
Tablo 18. <i>Deney B Grubu Öğrenci Performansı Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	65
Tablo 19. <i>Deney B Grubu Puanlayıcı Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	66
Tablo 20. <i>Deney B Grubu Ölçüt Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	67
Tablo 21. <i>Kontrol Grubu Öğrenci Performansı Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	68
Tablo 22. <i>Kontrol Grubu Puanlayıcı Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	70
Tablo 23. <i>Kontrol Grubu Ölçüt Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	71
Tablo 24. <i>Kontrol Grubu Öğrenci Performansı Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	72
Tablo 25. <i>Kontrol Grubu Puanlayıcı Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	73
Tablo 26. <i>Kontrol Grubu Ölçüt Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu</i>	74
Tablo 27. <i>Madde Güçlükleri Eşitlendikten Sonra Uyum İstatistikleri</i>	83
Tablo 28. <i>Deney A Ön Uygulama Puanlayıcı x Öğrenci Performansı Etkileşimleri</i>	85
Tablo 29. <i>Deney B Ön Uygulama Puanlayıcı x Öğrenci Performansı Etkileşimleri</i>	88
Tablo 30. <i>Kontrol Ön Uygulama Puanlayıcı x Öğrenci Performansı Etkileşimleri</i>	90
Tablo 31. <i>Madde Güçlükleri Eşitlendikten Sonra Uygun İstatistikleri</i>	98
Tablo 32. <i>Deney A Son Uygulama Puanlayıcı x Öğrenci Performansı Etkileşimleri</i>	100
Tablo 33. <i>Deney B Son Uygulama Puanlayıcı x Öğrenci Performansı Etkileşimleri</i>	103

Tablo 34. *Kontrol Son Uygulama Puanlayıcı x Öğrenci Performansı Etkileşimleri*..... 105



ŞEKİL LİSTESİ

Şekil 1. Performans değerlendirmeye ilişkin kavramsal çerçeve.....	12
Şekil 2. Deney A grubu ön uygulama t değerleri	79
Şekil 3. Deney B grubu ön uygulama t değerleri.....	80
Şekil 4. Kontrol grubu ön uygulama t değerleri.....	81
Şekil 5. Deney A grubu son uygulama t değerleri	94
Şekil 6. Deney B grubu son uygulama t değerleri.....	95
Şekil 7. Kontrol grubu son uygulama t değerleri	96

SİMGELER VE KISALTMALAR

ABRSM	The Associated Board of the Royal School of Music
ADPA	Analitik Dereceli Puanlama Anahtarı
ÇYRÖM	Çok Yüzeyli Rasch Ölçme Modeli
DPA	Dereceli Puanlama Anahtarı
G Kuramı	Genellenebilirlik Kuramı
KGİ	Kapsam Geçerlik İndeksi
KGO	Kapsam Geçerlik Oranı
KTK	Klasik Test Kuramı
MTK	Madde Tepki Kuramı
NAfME	National Association for Music Education
RMSE	Root Mean Square Error
χ^2	Ki-kare İstatistiği

BÖLÜM I

GİRİŞ

Bu bölümde, araştırmanın problem durumu, amacı ve alt problemleri, önemi ortaya konarak, sınırlılıkları ile tanımlara yer verilmiştir.

Problem Durumu

Eğitimin; girdi, süreç, çıktı ve denetim veya kontrol unsurlarını bütün olarak kapsayan bir sistem şeklinde incelendiğinde, bu sürecin sonunda bireylerdeki istenilen davranışların ortaya çıkıp çıkmadığı, çıktı ise hangi düzeyde başarıldığını değerlendirmek, süreçteki zorlukların belirlenmesi, buna göre eğitim programları ile modellerindeki var olan yöntem ve tekniklerin etkisini irdelemek bu doğrultuda öğrencileri yönlendirmek için ölçme ve değerlendirmenin rolü büyüktür (Atılgan, Kan & Doğan, 2007).

Günlük yaşamda sayısız ölçme ve değerlendirme işlemleriyle iç içe yaşamaktayız. Bir durum veya nesne hakkında karar vermek, puanlayabilmek, bir yargıya varabilmek için gözlem yapmak gerekmektedir. İnsan doğasının çok boyutlu yapısından dolayı sosyal bilimciler diğer bilim dallarına göre daha çok insan davranışlarının, özelliklerinin ölçülmesi ve değerlendirilmesi üzerine odaklanmaktadır. Günümüzde bilgidен ziyade, becerinin yaygın olarak kullanıldığı ortamlarda insanlar çeşitli nedenlerle performansları, yetenekleri hakkında birçok değerlendirme süreçlerine tabi tutulurlar. Bunlar performanslar içeriğine göre birey performanslarının değerlendirildiği; eğitim, güzel sanatlar (resim ve müzik), spor, tıp, insan kaynaklarında yapılan mülakatlar ile işe alım gibi çeşitli seçme işlemi gerektiren alanlarda birden fazla puanlayıcının nesnel puanlama davranışını sergilemesi ile teknik ve uzmanlık beceriye sahip olarak bu işi yürütmeleri beklenir.

Ölçme ve değerlendirme işlemi ile ilgili yaşamdan sayısızca örnekler verilebilir. Örneğin bir yazma dersinde eğitimci tarafından öğrencilere görev olarak vermiş olduğu kompozisyonlarının kontrol edilmesinde; bir ressamın resim yapan öğrencilerinin resimlerini değerlendirmesinde; bir tıp doktorun öğrencisinin hasta bakarken sergilemiş olduğu tıbbi yeterliliklerini ortaya konulmasının gözlemlenmesinde; müzik eğitimi alanındaki bir öğrencinin piyanoyu kullanarak başlangıç düzeyine ait eserleri tüm tekniklere uygun şekilde icra etmesinde; bir sporcunun yarışma sırasında gösterdiği yeterliliğini veya işe alım sürecinde, bir işçinin becerisini sergilendiği durumların ölçülerek değerlendirildiğindedir. Bireylerin göstermiş olduğu performansların istenilen nitelikte uygun bir şekilde değerlendirilebilmesi için öncelikle geçerliği ve güvenilirliği yüksek bir puanlama aracına sonra da bu aracı nesnel bir şekilde kullanabilecek nitelikli bir puanlayıcıya gereksinim duyulmaktadır. İyi bir ölçme aracı geliştirilmiş olsa bile nitelikli, hatalardan arınık bir şekilde puanlama yapan puanlayıcı olmadığında da ölçme sonuçlarına ciddi hatalar karışmaktadır.

Eğitim ve psikolojide yapılan ölçümlere her zaman birtakım hatalar karışır. Bu hatalar ise farklı kaynaklardan ortaya çıkmaktadır. Ortaya çıkan hataların kaynağı: ölçme aracından, ölçme yapılan bireylerden, değerlendirmeyi yapan kişiler vb. olabilmektedir. Bu farklı kaynaklardan gelen hataların kontrol altında alınması geçerli ve güvenilir değerlendirmeler için önem arz etmektedir. Dolayısıyla ölçme ve değerlendirme araştırmalarının amacı, bu hatayı kontrol ederek en aza indirmeye çalışmaktır. Tüm geliştirilen kuramların temelinde güvenilir, geçerli, doğru ve tutarlı ölçümler yaparak hatasız sonuçlar elde etme çabası vardır (Hambleton, 2004).

Performans değerlendirme sürecinde de geçerli ve güvenilir ölçümler yaparak amaç doğrultusunda farklı hata kaynaklarından gelebilecek hataları kontrol ederek doğru karar vermek söz konusudur. Ancak performans değerlendirme süreci, çoktan seçmeli maddelerinin puanlanmasından daha karmaşıktır (Ames & Luecht, 2018). Örneğin seçimli cevap gerektiren maddelerin puanlanmasında değerlendirciden kaynaklı hatalar ya hiç olmamakta ya da olabildiğince sınırlı olmaktadır. Diğer taraftan performans değerlendirme sürecinde ise hata kaynakları arasında değerlendirciler/puanlayıcılar da yer almaktadır. Bu nedenle performans değerlendirme sürecinde, değerlendircilerden kaynaklı hataların dikkatli bir şekilde incelenmesi ve değerlendircilerin ne derece geçerli ve güvenilir bir puanlama yaptıklarının izlenmesi gerekmektedir. Bunun için de değerlendircilerin

puanlayıcı davranışları incelenmeli ve hatalı puanlama davranışları var ise bunlar eğitim yoluyla giderilmelidir.

Ölçme aracıda yer alan değerlendirme ölçütlerinin tutarlı ve doğru bir şekilde uygulanması sınava girenlerin yeteneklerini tam olarak temsil eden derecelendirmeler ile sağlanabilir (Downing, 2006). Bununla birlikte geleneksel değerlendirme ölçütleri puan tabanlı değerlendirmeler için değerlendirme ölçütünü yorumlamaları gerektiğinden, değerlendiriciler üzerinde önemli bir bilişsel yük ve sorumluluk oluşturabilir (Wolfe, 1997). Değerlendiriciler ilgili performansı ölçmek için puanlama anahtarındaki ölçütleri analiz etmeli ve değerlendirilmesi gereken ürünün kalitesini bu ölçütlere göre değerlendirmelidir. Dolayısıyla ölçme aracıdaki ölçütler asıl uygulamaya geçilmeden önce ön uygulamalara tabi tutularak uzman görüşleri doğrultusunda revize edilmelidir. Yorumlayıcı yargılar nedeniyle, değerlendiricilerin değerlendirme ölçütünü tutarlı bir şekilde yorumlayıp yorumlamadıklarını belirlemek de çok önemlidir. Değerlendiriciler arası güvenirliliğin yetersizliği (yani değerlendiriciler arasındaki yetersiz fikir birliği) ve değerlendirici içi güvenirliliğin yetersizliği (yani değerlendiricinin kendi değerlendirmeleri arasında tutarsızlık), performans değerlendirme puanlanmasında ölçüt puanlama anahtarının kullanımında zorluk oluşturmaya devam etmektedir (Ezike, & Ames, 2021).

Bir müzik sınav (performans) değerlendirmesi; sınava giren öğrenci, sınav esnasında kullanılan yönergenin, görevin, seçilen eserin kendisi, puanlayıcı ve puanlama anahtarını da içeren kapsamlı bir süreci içermektedir (Hamp-Lyons, 1990; Kenyon, 1992; McNamara, 1996). Öğrenci performansının değerlendirilmesinde bireylerden istenen görevin yönergesinin açık ve net olması, puanlama anahtarının amaca göre seçilmesi önemlidir (Kutlu, Doğan & Karakaya, 2017). Amaca tam olarak uymayan ölçme aracının seçilmesi, acemi puanlayıcıların puanlama yapması gibi birden fazla faktör bireyler hakkında yanlış karar verme eğilimine götürebilir. Dolayısıyla performansın ölçüldüğü noktada mümkün olduğu kadar hatasız puanlama yapabilecek puanlayıcıya ve iyi hazırlanmış puanlama aracına ihtiyaç duyulur. Buradaki asıl amaç performans değerlendirmesinde olabildiğince nesnel değerlendirme-yapabilmektir.

Performansı değerlendirme süreci oldukça kapsamlıdır. Bir performans değerlendirme tasarımının geliştirilmesi, bir değerlendirmenin oluşturulması, performans değerlendirmesinin yönetilmesi, dereceli puanlama anahtarının (rubric) geliştirilmesi, puanlayıcıların eğitilmesi, operasyonel puanlamanın yapılması ve madde düzeyi ile test düzeyinde analizlerin tamamlanmasını içeren bir süreçtir (Johnson, 2009).

Performans değerlendirme ile ilgili yapılan birçok araştırmanın sonucunda değerlendirenden kaynaklı puanlamadaki tutarsızlığa yol açan farklı değerlendirici davranışları bulunmaktadır (Bech, 1993; Barrett, 2001; Fox, 2010; Idris & Zakaria, 2016; May 2008; Tajeddin & Alemi, 2014; Şata, 2019; van der Linden ve Hambleton, 1997). Puanlayıcıların farklı davranış sergilemelerinin nedenleri olarak önyargılı davranma, puanlamaya ilişkin eğitim almama, ayrıntılı hazırlanmamış dereceli puanlama anahtarı kullanımı, deneyim eksikliği gibi bazı nedenler söylenebilir. Bu farklı davranışların nedenleri sınıflandırıldığında merkezi eğilim, halo etkisi, aralık (ranj) daralması, cömertlik ve katılık şeklinde örnekler verilebilir (Engelhard, 1994; Myford & Wolfe, 2003; Saal, Downey & Lahey, 1980). Bir diğer çalışma araştırmacılar Johnson, Penny ve Gordon (2009) ve Linn, Baker ve Dunbar (1991) derecelendiriciler tarafından puanlanan puanlama anahtarının nesnel değil de öznel olarak puanlanmasından kaynaklı olduğuna ilişkin sonuçlara ulaşmışlardır. Popham (1997) herhangi bir performans değerlendirildiğinde ölçme aracı olan dereceli puanlama anahtarının performans değerlendirme sürecinin değerlendirilmesinde önyargıya neden olabileceğini iddia etmektedir. Dereceli puanlama anahtarında birden fazla boyutun ya da niteliğin dikkate alınması ve puanlama sürecinin uzun sürmesi durumlarında puanlayıcılarda yorgunluk, bıkkınlık gibi durumların ortaya çıkabileceği vurgulanmaktadır (Myford & Wolf, 2003).

Performans değerlendirme sürecinde; puanlayıcılar arası tutarlılığının yüksek olması verilen kararların doğruluğu açısından önemlidir. Aradaki tutarlılığı yüksek tutmak için puanlayıcıların sık sık kaliteli puanlayıcı eğitimlerinden geçmesi gerekmektedir. Puanlama işlemine başlamadan önce puanlayıcıların ölçme aracına yani dereceli puanlama anahtarına ilişkin uzlaşmaya varmaları, bireylerin performansları değerlendirildiğinde puanlayıcılar arası tutarlılığın yüksek çıkması ve bu tutarlılığın sürdürülmesi (Johnson, Penny & Gordon, 2009) eğitilmiş puanlayıcılara bağlı olduğu vurgulanmaktadır. Dolayısıyla puanlayıcıları eğitmek performans değerlendirme sürecinin önemli bir parçası ve bileşeni olarak görülmektedir.

Performans değerlendirme sürecinde insan puanlayıcıları kullanmada bazı sınırlılıkları olsada, insan değerlendiricileri sayısızca avantajları vardır. Zhang (2013), bireylerin değerlendirilmesinde kullanılan insan değerlendiricilerin görev almasının üç avantajını belirtmiştir: a) sınava girenlerin sunduğu bilgi ve becerileri değerlendirmek için bilişsel süreçlerin uygulanması, b) karar vermeyi bilgilendirmek amaçlı önceki bilgilerin bütünleştirilmesi, c) öğrencinin veya bireyin işinin kalitesini içeriğin anlaşılmasına göre değerlendirmesini sağlaması.

Puanlayıcı eğitimlerinin genel olarak ölçümlere karışan puanlayıcı kaynaklı varyansı azaltarak, daha geçerli ve güvenilir ölçümler yapmak ve süreci olumsuz yönde etkileyecek hususları ortadan kaldırmak veya en aza indirmek için yapıldığı söylenebilir. Bu amaç doğrultusunda puanlayıcı eğitim programları uygulanmaktadır. Bu programlar; eğitim, eğitim içeriği ve uygulamanın değerlendirilmesine dair dönüt olmak üzere üç öge içermektedir (Hauenstein & McCusker, 2017). Alanyazın incelendiğinde puanlayıcı eğitime yönelik uluslararası alanda birçok çalışmanın olduğu Türkiye’de ise sınırlı düzeyde çalışmanın olduğu söylenebilir (Görgülü, 2023; Onkun, 2024; Öç, 2024; Şata, 2019). Ayrıca puanlayıcı eğitimlerinin farklı eğitim desenleri kullanılarak verildiği görülmektedir. Bunlar arasında yoğun olarak kullanılan puanlayıcı eğitimi desenleri arasında; performans boyutları eğitimi, puanlayıcı hatası eğitimi ve referans çerçevesi eğitimi, davranış gözlem eğitimi, puanlayıcı değişkenliği eğitimi vb. olduğu görülmektedir (Aguinis, Mazurkiewicz & Heggstad, 2009; Antonioni ve Woehr, 2001; Bernardin & Pence, 1980; Borman, 1975; Cooper, 1981; Feldman vd., 2012; Landy & Farr, 1980; Neck, 1996; Neck, Stewart & Manz, 1995; Noonan & Sulsky, 2001; Roch, Woehr, Mishra & Kieszynska, 2012; Stamoulis & Hauenstein, 1993; Woehr & Huffcutt, 1994; Zedeck & Cascio, 1982). Çalışmanın asıl konusu olan müzik performansının değerlendirilmesine yönelik alan yazın incelendiğinde herhangi bir puanlayıcı eğitimi verilerek puanlayıcı davranışlarının incelenmediği söylenebilir. Bu nedenle bu alanda hem birden farklı desenlerin bir arada eğitimlerde kullanılarak değerlendirici davranışlarının irdelenmesi önem kazanmaktadır. Bazı performansların puanlanmasına yönelik uygulanan puanlayıcı eğitimi desenleri; becerilerin doğası gereği birleştirilerek yürütüldüğü de görülmektedir (Hauenstein & McCusker 2017; Rosales Sánchez, 2019).

Performans değerlendirme sürecinde puanlayıcıdan kaynaklı hataları en aza indirmek için verilen puanlayıcı eğitimlerinin genel olarak yüz yüze verildiği söylenebilir. Örneğin Türkiye’de bu alanda oldukça sınırlı çalışma bulunmakta olup, bu çalışmaların tamamının yüz yüze verildiği anlaşılmaktadır (Görgülü, 2023; Onkun, 2024; Öç, 2024; Şata, 2019). Puanlayıcı eğitimlerinde yer alacak bireylerin veya uzmanların bazen bir araya gelmeleri mümkün olamayabilir. Örneğin konu alan uzmanı kişilerin sayıca çok az olması, farklı şehirlerde yaşıyor olmaları, ekonomik nedenler, zaman gibi faktörler nedeniyle çevrimiçi eğitim yoluyla da eğitimler yürütülebilmektedir. Teknolojinin gelişmiş olması, pandemi sürecinde birçok eğitim faaliyetlerinin çevrimiçi olarak yürütülmesi, kişilerde belirli bir deneyimi kazandırdığı çevrimiçi eğitimlerin çeşitli alanlarda kullanımının yaygınlaştırdığı

söylenbilir. Hem öğrenme hem de öğretme ile ilgili eğitim amaçları gün geçtikçe daha da yaygın olmaya başlamıştır (Cornelius, 2014; Cornelius & Gordon, 2013; Gegenfurtner & Ebner, 2019; Gegenfurtner, Schwab & Ebner, 2018; McKinney, 2017; Wang & Hsu, 2008).

Çevrim içi eğitimlerin son zamanlarda yaygınlaşmasının başında: geniş çapta gerçek zamana dayalı katılımcılar arasındaki coğrafi uzaklıkları ortadan kaldırmaya yardımcı olması gelmektedir (Amhag, 2015; Gegenfurtner vd, 2017; Johnson & Schumacher, 2016; Stout vd, 2012; Zomenou vd, 2015). Web tabanlı seminerlerin veya eğitimdeki dijitalleşme sadece öğrencileri değil yetişkinlerin eğitimlerinin tasarlanarak sunulması ile paylaşılması gelişmekte ve değişmekte olan dünya çağında bir eğilim olarak karşımıza çıkmaktadır (Bell vd., 2017; Gegenfurtner vd, 2014; Goe vd., 2018; Martin vd., 2014; McKinney, 2017; Schmid vd., 2014).

Başta Türkiye ve Kazakistan olmak üzere birçok ülkede alanyazın incelendiğinde çevrimiçi olarak puanlayıcı eğitimlerine yönelik herhangi bir çalışmaya rastlanılmamıştır. Ayrıca müzik performansının değerlendirilmesinde değerlendirici eğitimleri yaparak bireylerin puanlama davranışlarının da irdelenmediği görülmektedir.

Bu çalışmada; üniversitede öğrenim gören öğrencilerin müzik performansının değerlendirilmesine ilişkin puanlayıcı davranışlarının birden farklı puanlayıcı eğitim desenleri birleştirilerek iki farklı eğitimin uygulama öncesi ve uygulama sonrası incelenmiştir. Ayrıca bu eğitimin çevrimiçi olarak sunulması ise çalışmanın bir diğer özgün yönünü göstermektedir. Bu amaç doğrultusunda piyano çalgısına yönelik öğrencilerin başlangıç düzeyindeki eserlerden iki tanesine ait performansları alınmıştır. Bu performansların değerlendirilmesinde başta piyano alanındaki kişilerin sayıca sınırlı, farklı kurum ve şehirlerde olması, kişilerin müzik öğretmenliği ile konservatuvarlarda öğretim elemanı olarak çalışmaları ve dolayısıyla günlük yüklerinin fazla olması ve yüz yüze bir araya gelmelerinin çok sınırlı olması nedeniyle çevrimiçi olarak puanlayıcı eğitimi yürütülmüştür.

Bu çalışmanın problemini; müzik eğitiminde piyano çalgısı başlangıç düzeyine ait öğrenci performanslarının puanlanması sürecinde; verilen puanlayıcı eğitimlerinin etkisinin incelenmesi oluşturmaktadır.

Problem İfadesi

“Referans Çerçevesi ile Puanlayıcı Hatası” ve “Referans Çerçevesi ile Performans Boyutu” puanlayıcı eğitimleri kullanılarak verilen puanlayıcı eğitiminin; puanlayıcı davranışları üzerine etkisi nasıldır?

Araştırma Soruları

Çalışmanın araştırma problem ifadesine ilişkin aşağıdaki sorulara cevap aranmıştır;

1. Puanlayıcı eğitimi öncesinde piyano başlangıç düzeyine ilişkin eserin puanlanması sürecinde;
 - a. Deney A grubundaki öğretim elemanlarından oluşan puanlayıcı grubunun ön uygulamada ve son uygulamada elde edilen *öğrenci performansı*, *ölçüt ve puanlayıcı* yüzeylerine ilişkin toplanan verilerin uygunluk değerleri ile model ölçümleri nasıldır?
 - b. Deney B grubundaki öğretim elemanlarından oluşan puanlayıcı grubunun ön uygulamada ve son uygulamada elde edilen *öğrenci performansı*, *ölçüt ve puanlayıcı* yüzeylerine ilişkin toplanan verilerin uygunluk değerleri ile model ölçümleri nasıldır?
 - c. Kontrol grubundaki öğretim elemanlarından oluşan puanlayıcı grubunun ön uygulamada ve son uygulamada elde edilen *öğrenci performansı*, *ölçüt ve puanlayıcı* yüzeylerine ilişkin toplanan verilerin uygunluk değerleri ile model ölçümleri nasıldır?
2. Puanlayıcı eğitimi uygulama öncesi piyano başlangıç düzeye ilişkin eserin puanlanması sürecinde;
 - a. Deney A, Deney B ve Kontrol grubu ön uygulamada yer alan puanlayıcıların katılık ve cömertlik davranışları nedir?
 - b. Deney A, Deney B ve Kontrol grubu ön uygulamada yer alan puanlayıcıların halo davranışları nedir?
 - c. Deney A, Deney B ve Kontrol grubu ön uygulamada yer alan puanlayıcıların farklılaşan puanlayıcı katılığı ve cömertliği (yanlılık) davranışları nasıldır?

3. Puanlayıcı eğitimi uygulama sonrasında piyano başlangıç düzeye ilişkin eserin puanlanması sürecinde;
- a) Deney A, Deney B ve Kontrol grubu son uygulamada yer alan puanlayıcıların katılık ve cömertlik davranışları nedir?
 - b) Deney A, Deney B ve Kontrol grubu son uygulamada yer alan puanlayıcıların halo davranışları nedir?
 - c) Deney A, Deney B ve Kontrol grubu son uygulamada yer alan puanlayıcıların farklılaşan puanlayıcı katılığı ve cömertliği (yanlılık) davranışları nedir?

Araştırmanın Amacı

Bu araştırmanın amacı; başlangıç düzey piyano öğrenci performanslarının ölçülmesine ilişkin iki farklı puanlayıcı eğitiminin çevrimiçi uygulanmasıyla; puanlayıcı davranışlarındaki değişimin incelenmesidir. Puanlayıcı eğitimi kapsamında Deney A Grubuna “Referans Çerçevesi ile Puanlayıcı Hatası”, Deney B Grubuna “Referans Çerçevesi ile Performans Boyutu” eğitim desenleri birleştirilerek puanlama davranışları üzerine etkileri karşılaştırılmıştır. Ayrıca çalışmada kontrol grubu bulunmaktadır. Bu grup üzerinde herhangi bir puanlayıcı eğitimi yürütülmemiştir. Araştırmanın amacına göre literatürde performans değerlendirmede tercih edilen puanlayıcı eğitimlerinden her iki grup için referans çerçevesi eğitimi ortak olarak belirlenmiş olup her iki deney grubunda uygulanmıştır.

Araştırmanın bu genel amacına ulaşması için öncelikle alanda bulunmaması nedeniyle öğrencilerin piyano başlangıç düzeyine ait performanslarının ölçülmesinde kullanılmak üzere analitik (analytic rubric) dereceli puanlama anahtarının geliştirilmesi hedeflenmiştir.

Araştırmanın Önemi

Türkiye’deki performansın ölçümünde ulaşılabilen kaynakların incelenmesinde ortaya çıkan puanlayıcı davranışlarını konu alan çalışmalar gün geçtikçe sınırlı olsa artış göstermektedir. Bu çalışmalar gerek eğitim gerekse farklı disiplinlerdeki (spor, güzel sanatlar vs. gibi) performans değerlendirme sürecinde kullanılan puanlayıcı davranışlarının incelenmesi ile ilgilidir. Bu çalışmalar (Akın & Baştürk, 2012; Atılğan, 2005; Baştürk, 2010; Erman Aslanoğlu ve Şata, 2023; Görgülü Öztürk, 2023; Karakaya, 2015; Onkun Özgür,

2024; Öztürk Gübeş, 2021; Şahan, 2018; Şahin, Taşdelen Teker ve Güler, 2016; Tobaş, 2020; Tursynbayeva, Öç & Karakaya, 2024; Wesolowski, 2015; Wesolowski, 2019) yer almaktadır.

Performans değerlendirmede sürecinde puanlayıcı eğitim desenlerinin kullanımı yapılan değerlendirmenin tutarlılığını ve nesnelliğini arttırmayı hedefleyerek geliştirilmiş sistematik yaklaşımlar sunmaktadır. Konuyla ilgili alan yazında altı eğitim deseni mevcuttur. Bunlardan ilki öz liderlik eğitimi, performans değerlendirmede puanlayıcıların öz farkındalıklarını geliştirerek kendi performanslarını yönetmelerini sağlarken, ikinci eğitim olan davranış gözlem eğitimi performansın detaylı ve objektif şekilde kaydedilmesini öngörmektedir (Eckes, 2015). Eğitim desenlerinin üçüncüsü olan performans boyutu eğitiminde ise değerlendiricilerin belirli performans bileşenlerine odaklanmasını sağlayarak puanlama sürecinde sistematik bir yapı oluşturmayı hedeflemektedir (Weigle, 2002). Bu eğitimler, değerlendirme sürecindeki subjektif önyargıları en aza indirerek puanlayıcıların daha güvenilir ve adil kararlar vermelerine katkıda bulunur (Barkaoui, 2010).

Bununla birlikte dördüncü eğitim deseni olan puanlayıcı hatası eğitimi, değerlendiricilerin sistematik hata kaynaklarını fark etmelerine ve bu hataları minimize etmelerine yardımcı olmaktadır (Myford ve Wolfe, 2003). Referans çerçevesi olan beşinci eğitiminde ise puanlayıcıların ortak standartlar doğrultusunda değerlendirme yapmalarını sağlayarak farklı puanlayıcılar arasındaki tutarlılığı artırmaktadır (Lumley, 2005). Son eğitim deseni olan puanlayıcı değişkenliği eğitimi, puanlama sürecinde ortaya çıkan varyasyonların nedenlerini anlamaya ve bu faktörleri kontrol altına almayı amaçlamaktadır. Bu eğitim desenleri, performans değerlendirme sürecinde nesnelliği arttırarak ölçme araçlarının geçerlik ve güvenilirliğini yükseltmekte ve eğitim süreçlerinde daha adil sonuçlar elde edilmesini sağlamaktadır (Weigle, 2002).

Alanyazında Türkiye’de puanlayıcıların çeşitli eğitimler kullanılarak puanlayıcı davranışlarına etkisine ilişkin de sınırlı düzeyde çalışmaya rastlanıldığı söylenebilir (Çetin & İlhan, 2017; Şata, 2019; Öç, 2024). Özellikle birden fazla eğitim kullanılması, her iki deney grubunda ikişer eğitim desenlerinin birleştirilerek sunulması ve puanlayıcı davranışlarındaki değişimin incelenmesi, puanlayıcı kaynaklı hataların azaltılmasına yönelik yeni bilgiler sunacak olması açısından önem taşımaktadır. Bu uygulamanın da müzik eğitiminde ilk defa yapılacak olması ve bu eğitimin de çevrimiçi olarak yürütülmesi çalışmanın özgün yönünü göstermektedir. Dolayısıyla çalışma hem müzik eğitiminde ilk defa uygulanmış olması hem iki farklı eğitimin etkisinin incelenmesi hemde bütün bu

uygulamalarında çevrim içi olarak yürütülmesi bağlamında çalışmanın önemini ortaya çıkarmaktadır.

Sınırlılıklar

Araştırmanın sınırlılıkları aşağıdaki gibi sıralanmıştır;

- Araştırmada puanlanan performanslar; Gazi Üniversitesindeki Güzel Sanatlar Eğitim Bölümü Müzik Anabilim dalının bahar döneminde eğitim gören birinci sınıf 10 öğrencinin piyano başlangıç düzeyi ilişkin performansları ile sınırlıdır.
- Öğrenci performanslarının puanlanması için geliştirilen Analitik Dereceli Puanlama Anahtarının geçerlik çalışması; uygulanacak birey sayısının ve puanlayıcı sayısının sınırlı olması nedeniyle Açıklayıcı Faktör Analizi uygulanmadan sadece uzman görüşleri doğrultusunda Kapsam Geçerlik İndeksleri kullanılarak belirlenmiştir. Bu durum ölçme aracı için geçerlik kanıtı sunma açısından bir sınırlılık oluşturmuş olabilir. Dolayısıyla geçerlik kanıtı uzman görüşleri ile sınırlıdır.
- Öğrencilerin piyano çalgısı başlangıç düzeyi performansları “czerny 12. etüt ve Alman Dansı” isimli eserlere aittir. Dolayısıyla başlangıç düzeyi performanslar bu eserlerle sınırlıdır.
- Araştırmada en sık görülen puanlayıcı davranışlarından; puanlayıcı katılığı, puanlayıcı cömertliği, halo etkisi ve farklılaşan puanlayıcı katılığı ve cömertliği (yanlılık) olmak üzere beş tanesi araştırılmıştır.
- Bu çalışmada puanlayıcı eğitimleri “Referans Çerçevesi Eğitimi, Puanlayıcı Hatası Eğitimi ve Performans Boyutu” eğitim ile sınırlıdır.

Tanımlamalar

Performans Değerlendirme: Gerçek hayata yönelik görevler doğrultusunda ortaya konan eylemin belirli ölçütlere göre puanlayıcılar tarafından değerlendirilmesidir (Johnson, 2009).

Çok Yüzeyli Rasch Modeli: Madde tepki kuramına bağlı bir model olarak ölçüt, puanlayıcı ve öğrenci/aday gibi farklı yüzeylerin tek bir ölçek üzerinde sergilendiği modeldir (Linacre, 2017).

BÖLÜM II

KAVRAMSAL ÇERÇEVE

İlgili bölümde araştırmanın problemine ve araştırmaya ait kuramsal temellerle beraber, performans değerlendirme sürecinde ortaya çıkan puanlayıcılardan kaynaklı çeşitli puanlayıcı davranışlarının belirlenmesi ve puanlayıcı eğitimlerine ilişkin ilgili araştırmalara yer verilmiştir.

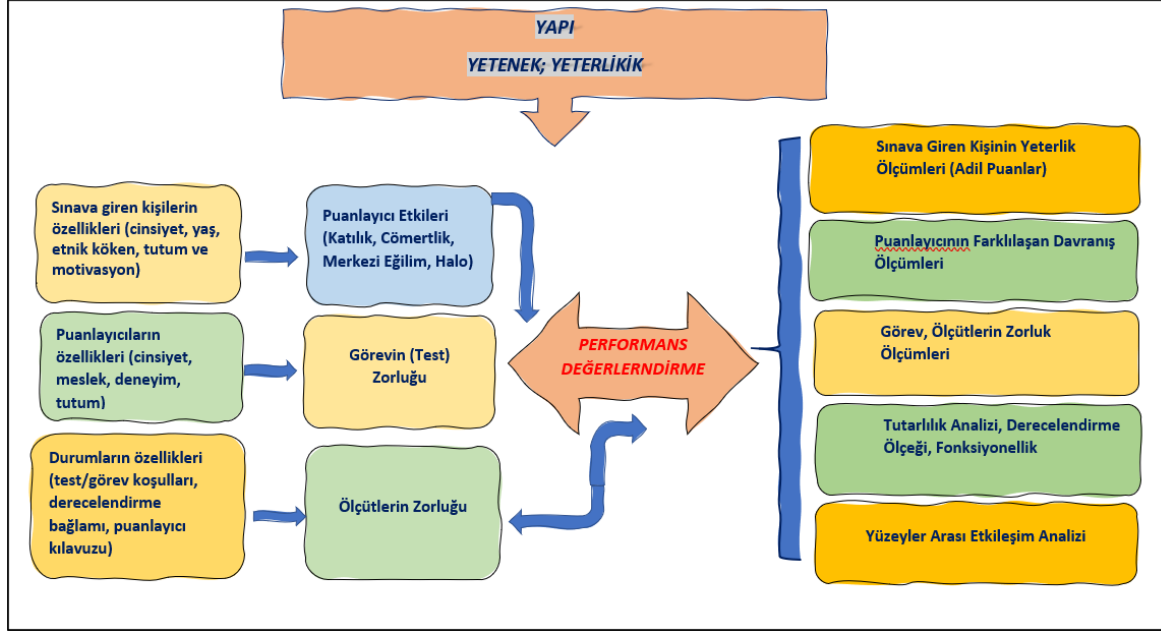
Performans Değerlendirme

Performans değerlendirme genel olarak iki ana bölümü kapsar. Bunlar öğrenciye verilen görev ve puanlama ölçütlerinden oluşan puanlama anahtarıdır. Öğrencinin kendisine verilen yanıtlardan birini seçmesi istendiği durumda performans değerlendirme, çoktan seçmeli veya doğru yanlışın seçilmesini gerektiren testlerden farklı olarak öğrencinin kendi yanıtını oluşturması beklenir. Örneğin yazılı bir performans değerlendirmesinde öğrenciden dil bilgisine dayalı veya noktalama işaretleri hakkında çoktan seçmeli soruları yanıtlamak yerine kendisinden istenilen görevle ilgili yazması beklenir. Performans değerlendirmede hazırlanan değerlendirme görevleri bir ürün performansı ya da genişletilmiş yazılı yanıt gerektiren bir görev olabilir (Kutlu, Doğan & Karakaya, 2017; Perlman, 2003). Asıl olan öğrenciye verilecek görevin veya performansın kendisinde eleştirel düşünme, yaşamla ilişkilendirme, yaratıcı düşünme, diğer disiplinlerle bağlantı kurma gibi becerileri kullanmasını teşvik edecek şekilde hazırlanmasıdır. Performans değerlendirme görevleri tüm alanı kapsadığından dolayı her alana özgü çeşitli örnekler verilebilir. Bunlar: sözlü sunumlar, deneme, sanat eserleri, bilim fuarı projeleri, araştırma projeleri, müzik performansları, doktora yeterlik sınavları, ehliyet sınavları ve açık uçlu matematik problemleridir.

Performans Değerlendirmeye İlişkin Kavramsal Çerçeve

Çok Yüzeyle Rasch ölçme modeli, örneğin bir testte veya bir performans görevinde yer alan; testi cevaplayan ile ilgili testteki maddelerin birer değişken olduğu düşünüldüğünde bu değişkenlere ek

olarak birden fazla değişkeni veya yüzeyleri ekleyerek basit Rasch modelini (Rasch,1960/1980; Wright ve Stone, 1979) genişleten ve yüzeyler arası ayrıntılar üzerinde potansiyel bir etkiye sahip olabilecek birden fazla değişkenin ince ayrıntısına kadar analizini ortaya koyan bir ölçme modelidir (Eckes, 2005). Kavramsal çerçeve şekil 1’ de açıklanmıştır.



Şekil 1. Performans değerlendirmeye ilişkin kavramsal çerçeve

Şekil 1’de performans değerlendirmeye ilişkin yapı ve yapıyla uzaktan veya yakından ilişkili olan faktörler yer almaktadır. Buna göre sınava giren adayın veya bireyin performans değerlendirmesi sürecini etkileyebilecek Eckes (2015) tarafından geliştirilen kavramsal çerçeve geliştirilmiştir.

Puanlayıcı veya Değerlendirici: Performans değerlendirmedeki puanlayıcılardan oluşan istenmeden de olsa meydana gelen yanlış davranışları (katılık, cömertlik, yanlılık) ÇYRÖM ile bu değerlendirici etkileri tespit edilebilir.

Performans Görevleri: Performans değerlendirme sürecinde kullanılan sorular veya testlerdir. Verilen görevin zorluk düzeyindeki değişkenliğin de ÇYRÖM ile etkileri tespit edilebilir.

Test Denekleri veya Katılımcılar: Performansı değerlendirilen veya eylemleri inceleme altında olan bireyler ya da öğrencilerdir. ÇYRÖM bireylerin beceri setlerine ilişkin tahminde bulunur.

Değerlendirme Sistemi: Performans değerlendirmede kullanılan derecelendirme sistemi veya ölçütlerin yapısının ve alt boyutlarının bilinmesi gerekir.

Yüzeyler Arasındaki Etkileşim: Şekil 1’de, sınava giren kişiyi ödev veya göreve bağlayan bu tür etkileşim, sınava giren kişinin çıktısı üzerinde bireysel bir etkiye sahip olabilir. Alternatif olarak,

zor bir görev seçmenin, daha az zor bir görev seçmekten daha düşük bir puanla sonuçlanması muhtemeldir çünkü sınava giren kişiler zorluk bakımından farklılık gösteren bir dizi görev arasından seçim yapabilirler. Genel etki, sınava giren kişinin puanlarının değişkenliğinde bir artış olacaktır; bu da sınava giren kişiyle ilgisi olmayan bir değişkenlikte bir artışın sonucunu gösterir.

Müzik Performanslarının Ölçülmesi

Değerlendirme hem eğitim hem de müzik performansının ölçülmesinde önemli yönlerindendir. Değerlendirme süreci şu şekilde özetlenebilir: “eğitimle ilgili kararlar almak ve sürekli iyileştirme sağlamak amacıyla öğrenci performansı veya program etkinliği hakkında bilgi toplama, analiz etme, yorumlama ve uygulama yapmak” olarak tanımlamışlar. Dolayısıyla gerek eğitim gerek müzik gibi performansın ölçüldüğü alanlarda etkili değerlendirmelerin planlanması ve yapılması son derece önemlidir. Çünkü etkili değerlendirme öğrencilerin belirli bir öğrenme seviyesi veya sonucu için beklenen bilgi ve becerilere göre ne kadar iyi performans sergiledikleri ile ilgili bir durum paylaşımı ile sınırlı kalmayıp, paydaş olan öğretmen ve öğrencilerin var olan yöntemlerini tekrardan revize yapmaya olanak tanımakla beraber aynı zamanda müfredat uygulamalarının uygulanabileceği alanları da ortaya çıkarır.

Çeşitli performansların değerlendirildiği (eğitim, güzel sanatlar, spor, tıp gibi) disiplinlerde performans değerlendirmelerin sonuçları; topluluklar, eğitimciler, yöneticiler ve öğrenciler için oldukça önemlidir. Dolayısıyla müzik gibi alanlarda yapılan sınavlarda puanlayıcıların değerlendirmeleri her öğrencinin gerçek performans yeteneğini adil bir şekilde yansıtabilmesi için oldukça önemlidir. Puanlayıcıların bireyleri puanlarken puanlama araçlarını kullanarak elde edilen sonuçların genellikle puanlayıcıların puanlama davranışları ile örtüşmesi veya üzerinde değerlendirme yapılan performansların kendisiyle pek bağlantılı olmamasından kaynaklıdır (Engelhard, 2002).

Müzik performansının değerlendirilmesi, farklı görüşlere yol açabilecek kişisel ifade ve yorumlamayı içeren içsel öznelliğinden dolayı zordur. Ulusal Müzik Eğitim Derneği (NAfME) ve Kraliyet Müzik Okulları Birliği (ABRSM) gibi müzik kurumları, müzik performanslarının değerlendirmelerini belli bir standartlara uygun veya karşılaştırma yapılabilmesi için bütünsel yaklaşımla değerlendirme yöntemlerini kullanmanın müziğin doğasına aykırı olduğunu belirtmişlerdir. Bu nedenle, müzik performans değerlendirmeleri için doğru, geçerli ve güvenilir ölçme araçları ile ölçülmeye ihtiyaç duyulmaktadır (Colwell, 2007; Wesolowski, 2014).

Problemli Puanlayıcı Davranışları

Performanslar değerlendirme sürecinde puanlayıcılar istemeden de olsa bireyleri veya öğrencileri puanlarken kişisel deneyim, kültür, sosyalleşme, çevresel faktörler ve çeşitli önyargılardan dolayı farklı puanlama davranışlarını sergileme eğiliminde olabilirler. İstenmeden yapılan davranış olmasına rağmen bu durum elde edilen ölçme sonucunu zedeleyerek sonuçta geçerliği ve güvenirliği düşük ölçme yapılmasına sebebiyet verebilir. Ölçmenin geçerlik ve güvenirliğini zedeleyen bu tür eylemler alan yazında; puanlayıcı davranışları, problemli veya sorunlu puanlayıcı davranışlar, puanlayıcı etkisi, puanlayıcı hatası ve puanlayıcı yanlılığı olarak ifade edilmiştir (Farrokhi vd., 2011; Haladyna, 1997; Myford & Wolfe, 2004; Royal & Hecker, 2016).

Alanyazında puanlayıcı davranışlarının birden fazla türü mevcuttur. Bunlar halo etkisi, merkeze eğilim, puanlayıcılardan kaynaklı katılık veya cömertlik, ranj/ aralık sınırlaması, tutarsızlık gibi farklı adlarla belirtilmektedir (Myford ve Wolfe, 2004). Tablo 1’de performans değerlendirmede sıklıkla puanlayıcıların karşı karşıya kaldıkları hata türleri verilmiştir (Johnson, Penny & Gordon, 2009; Royal & Hecker, 2016; Wesolowski, Wind & Engelhard, 2015):

Tablo 1

Performans Değerlendirmede Karşılaşılan Yaygın Hata Türleri

Hatalar	Puanlayıcı Davranışları
Asimilasyon etkisi	Diğer puanlayıcılara uyum sağlamak amacıyla bilinçli olarak benzer puanlar verme
Aktarma etkisi	Öğrencinin genel akademik başarısının, bireysel performans değerlendirmesine yansıtılması (+/-)
Merkeze eğilim	Puanlama yaparken çoğunlukla ölçeğin orta kategorilerini tercih etme
Amigo etkisi	Öğrencileri desteklemek veya motive etmek amacıyla performansı ne olursa olsun yüksek puan verme
Uyuşmayan standartlar	Ölçme aracındaki standartlarla kişisel değerlendirme kriterleri arasındaki farklılık nedeniyle tüm öğrencilere düşük puan verme
Uyuşmayan değerler	Puanlayıcının öznel değerleri ile değerlendirilen performans arasında uyumsuzluk olması nedeniyle düşük puan verme
Çekimserlik	Öğrencinin itiraz etme ya da karşı gelme ihtimali nedeniyle yüksek puan verme
Kayma etkisi	Zaman zaman artan şekilde daha katı veya daha cömert puanlar verme eğilimi gösterme
Aşırılık	Ölçme aracının sadece uç değerini kullanarak ara değerleri göz ardı etme
Yorgunluk	Fiziksel veya zihinsel yorgunluk nedeni ile değerlendirme sürecinde puanların etkilenmesi
Taraf tutma	Tanıyan bir öğrencinin değerlendirilmesinde bazı hatalarını görmezden gelinmesi
İlk izlenim	Öğrencin sonraki performansını göz ardı ederek bıraktığı ilk izlenime göre değerlendirme
Halo etkisi	Öğrencide şüphe veya beklenti uyandırarak gerçek performansından daha yüksek puan verme
Gestalt olgusu	Performansın farklı bileşenleri ayırt edilemediğinden öğrenciye genel bir izlenime dayalı puan verme
Bağlantı etkisi	Öğrencinin başka bireylerle olan bağlantısının puanı etkilemesi
Güvensizlik	Puanlayıcının düşük bir puanı açıklamada yetersiz kalmasından dolayı yüksek puan vermesi
Uzunluk	Daha uzun sürede gerçekleştirilen performans daha yüksek puanlama
Cömertlik	Genel olarak sürekli olarak yüksek puan vermesi
Kişilik çatışması	Öğrenciye yönelik sahip olunan olumsuz yargılardan dolayı düşük puan verme
Puanlayıcı yeterliği	Doğrudan gözlemlenmeyen veya değerlendirilemeyen durumlarda yine de puan/not verilmesi
Tekrar etkisi	Aynı performansın tekrar edilmesi nedeni ile düşük puan verme
Ölçme Aracını	Ölçme aracının yanlış anlaşılması nedeniyle hatalı puanlar verme
Yorumlama	
Puan-aralığı sınırlaması	Ölçülmek istenilen gizli değişkenleri performansta doğru ayırt edememe
Öz-puanlama	Performansı gerektiğinden fazla inceleme ve değerlendirme
Benzerlik etkisi	Puanlayıcıya yakın veya benzer performans gösteren öğrencilere yüksek puan verme
Yüzeysel puanlama	Performansın çok az bir bölümünü dikkate alarak veya inceleyerek puanlama yapma
Ani kayıp	Puanlayıcıda olumsuz bir tepki veya izlenim oluşturan performans daha düşük puanlama
Sempati puanı	Puanlayıcın duygularına yönelik veya ödüllendirme amacı ile ortaya konan bir performans değerlendirme
Katılık	Genel olarak sürekli olarak daha düşük puanlar verme
Kişisel özellik etkisi	Öğrencinin bireysel özelliklerini (zaman yönetimi, iletişim veya fiziksel görünüm) dikkate alarak performansını puanlama

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: designing, scoring, and validating performance tasks*. New York: Guilford.

Royal, K. D., & Hecker, K. G. (2016). Rater errors in clinical performance assessments. *Journal of Veterinary Medical Education*, 43(1), 5-8.

Wesolowski, B.C., Wind, S.A., Engelhard G. (2015). Rater fairness in music performance assessment: evaluating model data fit and differential rater functioning. *Music Science*, 19(2), 147–70.

Tablo 1’deki kullanılan yaygın hatalar performans değerlendirmesi için en önemli geçerlik tehditlerindendir. Bu davranışlar hem puanlayıcı kişiliğinde hem de ölçme aracından kaynaklı olabilir (Royal & Hecker, 2016). Tablo 1’deki puanlayıcı davranışlarından en sık meydana gelen hata türleri şunlardır; puanlayıcıdan kaynaklı katılık ve cömertlik, farklılaşan katılık ve cömertlik, tutarsızlık söylenilebilir (Myford & Wolfe, 2003). Araştırmada kullanılan puanlayıcı davranışları sırası ile aşağıda açıklanmıştır.

Puanlayıcı Katılığı ve Cömertliği

Performans değerlendirmedeki varyansı etkileyen ranj daralması ve merkezi eğilim hatalarının aksine puanlayıcı katılığı ve cömertliği dağılımın yerini ve ortalamasını etkileyen değerlendirici hata biçimidir. Tanım olarak puanlayıcı cömertliği şu şekilde açıklanmıştır: bir puanlayıcının derecelendirme ölçeğinin orta noktasının üstünde puanlama yapmasıdır. Puanlayıcı katılığı ise; cömertlik davranışın aksine derecelendirme ölçeğinin orta noktasının altında puan verme eğilimidir (Saal, Downey & Lahey, 1980). Kişiyi, ölçütü ve zamanın geçişini etkileyen pek çok faktör puanlayıcı cömertliği ile katılığını da etkileyebilir ve bu faktörler kişilik nitelikleri ile değildir (Guilford, 1954; Lynch & McNamara, 1998). McNamara (1996)'ya göre kesme noktasındaki bireyler, puanlayıcının katı mı yoksa cömert mi olduğuna bağlı olarak bir avantaja veya dezavantaja ya da sınavı geçmesi ile kalması gibi duruma sahip olabilir. Puanlayıcının davranışından dolayı birey sıralamasında ciddi bir kayma söz konusu olabilir. Puanlayıcıların cömertlik ve katılık davranışlarının azaltmak veya önüne geçilmesi için önerilen birkaç teknik mevcuttur (Myford & Wolfe, 2003). Bu yaklaşımlar puanlayıcılar ile ölçme aracıda etkilidir.

Ölçme aracı ile ilgili yapılacak adımlar: üzerinde ölçme işlemi yapılacak olan özellik ayrıntılı açıklanmalı ve ölçekteki her bir kategoriye özgü ortak açıklamalar ile tanımlamalar verilmelidir. Böylece puanlayıcı ölçme aracıdaki kategoriler ayrımını yaparak onların ne tür anlam içerdiğinden haberdar olur. Ölçme aracıdaki bulunan kategoriler eşit olmak şartı ile (yarısı olumlu ve yarısı olumsuz) geliştirilmelidir. Puanlayıcı ile ilgili yapılacak adımlar: değerlendirme esnasında puanlayıcılar bireylere katı veya cömert davranmalarının sonucunda birey hakkında ciddi bir yanlış veya doğru karar alınabileceğinden ve bu yanlışın önüne geçilmesi için de eğitim almalılar.

Merkeze Yönelim/ Eğilim

Aralık kısıtlaması veya merkezi eğilim puanlayıcıların hedef çalışanların performans seviyeleri arasında yeterli ayrımı yapmayan derecelendirmeler atadığı bir dağılım hatasıdır. Puanlayıcının değerlendirmiş olduğu performans farklılıklarına bakılmaksızın her insanı veya bireyi ortalama olarak değerlendirme eğilimidir (Myford & Wolfe, 2003; Saal vd., 1980). Puanlayıcıların ölçme aracıdaki dereceli puanlama anahtarını kullanma şeklini, bireyi puanlarken göstermiş olduğu davranışlarını merkezleme biçiminde öğrenilebilir (Engelhard, 1994). Wind (2019)'a göre bu davranışı sergileyen puanlayıcılar genellikle

yüksek yetenek düzeyine sahip olan öğrencilere düşük puanlar ve yetenek düzeyleri düşük olan öğrencilere yüksek puan verme eğilimindedir.

Myford ve Wolf (2003) tarafından puanlayıcının birey performansını değerlendirmesinde merkeze eğilim davranışlarını azaltmaya yönelik stratejiler aşağıdaki gibi önerilmiştir:

Dereceli puanlama anahtarında yer alan her bir ölçüt ve düzeylerinin tanımlanması gerekmektedir. Performans düzeyleri olarak da ifade edilen düzeylerin tanımlanması hem görevi yerine getirecek bireyler için hem de puanlamayı yapacak öğretmen ve öğrenciler için yönerge niteliği taşımaktadır. Bu durum görevi yerine getirecekler için açıklayıcı bilgi sunması nedeniyle yapılacak ölçümlerin daha geçerli olmasına, hem puanlama yapacak kişiler için daha tutarlı ölçümlerin oluşmasına katkı sunması da güvenilirliğe ilişkin katkılar olarak değerlendirilebilir. Böylece puanlayıcı ölçme aracındaki her bir kategoriye ilişkin açıklığı ve seviyeler arasındaki ayrımı fark ettiğinde daha iyi değerlendirme yapabilir. Ölçme aracını geliştirirken abartılı şekilde kategorilere bölmeden kaçınılmalıdır. Çünkü puanlayıcıların çok azı öğrencileri ölçme aracındaki ilgili kategorilere atamaktadır.

Halo Etkisi

Gözlemcinin bir denek hakkında, ilk görüşmeden sonra yaptığı yargıların o deneklerin ne kadar iyi performans gösterdiğini etkilemesine izin verdiğinde oluşan ilk izlenim hatasıdır (Latham, Wexley ve Pursell, 1975). Bu nedenle bir konunun, o kişinin gerçek davranışlarına dayanmayan hatalı şekilde yüksek veya düşük derecelendirmeleri meydana getirebilir. Ayrıca bir performans ölçümünün yapı geçerliğini de zedeleyebilir. Halo davranışını sergileyen puanlayıcı genellikle ölçeğin kriterleri arasında gerekenden tam olarak veya benzer şekilde daha yüksek puan alır (Kim, 2020). Bir performans puanlama sürecinde halo etkisini bulmak için dört farklı yaklaşım önerilmiştir. Bu teknikler Saal vd. (1980) ile örtüşmektedir;

Üzerinde ölçme işlemi yapılan yapıya ait ölçütler ile puanlamalar içi korelasyon değerlerini inceleyerek inceleme sonucunda çıkan korelasyonun yüksek bulunması ölçülen özellikleri iyi ayırım yapamadığı ve halo etkisinin meydana geldiğine işaret eder.

Araştırmaya dahil edilen yüzeylerden puanlayıcı ile birey arasındaki etkileşim varyans analizi ile incelenebilir. İnceleme sonucundaki etkileşim istatistiksel olarak anlamlı ve etkileşim puanlama sonucundaki varyansın büyük bir kısmını açıklarsa bu durumda halo

etkisinin varlığı gösterir. Halo etkisinin tespit edilmesinde sabit etkili ki-kare, ayırma indeksi ve ayırma indeksinin güvenilirliği kullanılır.

Farklılaşan Puanlayıcı Katılığı ve Cömertliği (Puanlayıcı Yanlılığı)

Değerlendiricilerin bazı öğrencilere karşı daha katı ve hoşgörülü olma olasılığı daha yüksektir (Knoch, Read & Randow, 2007). Bu davranışa farklılaşan puanlayıcı katılığı veya cömertliği denir. Puanlayıcı önyargısının bir puanlayıcının bazı kişilere cinsiyet veya yaş gibi özellikleri dikkate alarak daha yüksek veya daha düşük puan verme eğiliminde olmasıdır. Bu gibi durumlarda yapılan ölçme işleminden elde edilen verilerin geçerliği ile güvenilirliği puanlayıcıların davranışlarından dolayı zarar görecektir (Kumar, 2005).

Puanlayıcı Davranışların Azaltılmaya Yönelik Uygulamalar

Performans değerlendirme sürecinde bireylerin üst düzey zihinsel becerilerinin nesnel şekilde ölçerek değerlendirme gibi avantajı vardır, fakat yapılan bu ölçme ve değerlendirmenin nesnelliğinin sağlanamaması dezavantaj getiren bir durumdur. Bunun için performansların ölçülmesi ve değerlendirilmesi kapsamında puanlayıcıdan meydana gelen etkilerin en aza indirmek veya benzer hale getirmeye yönelik birçok öneri öne sürülmüştür. Bu öneriler birden fazla puanlayıcı performans görevinde yer alacaksa doğrudan puanlayıcı eğitiminde geçmeleri gerektiği görüşünü savunanlardır (Ebel, 1951; Haladyna, 1997; İlhan & Çetin, 2014; Kubiszyn & Borich, 2013; Woehr, 1994) ve verilen eğitimin yanında da puanlama anahtarının kullanımına (Kutlu vd., 2014) ilişkin görüşler yer almaktadır.

Birden Fazla Puanlayıcının Kullanılması

Performans değerlendirmede birden fazla puanlayıcının veya değerlendiricinin kullanılması sürecin sağlıklı yapılabilmesi için zorunludur. Benzre şekilde müzik performanslarının da değerlendirilmesinde birden fazla puanlayıcının puanlama sürecine katılım göstermesi önemlidir. Nesnelliğin sağlanabilmesi için puanlayıcıların davranışlarını en aza indirmek veya benzer hale getirmek için birden fazla puanlayıcının kullanımıdır (Gronlund, 1977). Örneğin doktora tez savunmasında birden fazla jüri üyesinin yer alması, yine benzer şekilde

doktora yeterlik sınavlarındaki hem yazma hem de sözlü kısımlarında birden fazla puanlayıcının olmasıdır.

Dereceli Puanlama Anahtarı

Dereceli puanlama (rubric) kelimesinin Türkçe anlamı incelendiğinde sözlükte birden fazla anlam içerdiği görülmüştür. Bunlardan en yaygın olarak kullanılan Merriam-Webster'in çevrim içi sözlüğündeki ilk anlamı “yetkili bir kural” şeklinde tanımlanmıştır. Sözlükteki tanımlar arasında en manidar olanı ise “akademik makaleleri, projeleri veya testleri derecelendirmek, puanlamak için belli başlı ölçütler listesini kapsayan bir rehber, kılavuz” şeklinde açıklanan dördüncü tanımıdır (Karakaya ve Fişne, 2023). Benzer bir açıklamaya göre: ADPA, bir performansın en önemli bileşenlerini puanlama için boyutlara bölen bir puanlama kılavuzudur. Her boyuta ayrı bir puan atanır ve bu puanlar daha sonra toplanarak nihai puan hesaplanır (Johnson, Penny & Gordon, 2009).

Performans değerlendirmesinde sık kullanılan dereceli puanlama anahtarı hem puanlama işini yürüten puanlayıcı hem de performansı değerlendirilen öğrenci veya birey için avantajlıdır. Kullanımın basit olmasının yanı sıra, puanlayıcının performans değerlendirme sürecinde nesnellik ile sistematikliği korumasına yardımcı olur. Bireyler veya öğrencilerin avantajlı olmasının sebebi; göstermiş olduğu performansına ilişkin puanlanırken ne tür ölçütlere göre puanlandığını daha önceden bildiği için daha iyi performans göstermesine olanak sağlayarak öğrencinin kendisindeki gelişimi görmesine de katkıda bulur (Johnson, Penny & Gordon, 2009; Mertler, 2001).

DPA, birkaç aşamayı içeren yöntemsel bir yaklaşım kullanarak geliştirilmelidir. DPA keyfi veya düzensiz bir sırayla geliştirilirse, yapılan ölçümlerin psikometrik özellikleri zedelenmiş olabilir. DPA geliştirilirken, bir ölçme aracında bulunması gereken özelliklerin son derece dikkatlice geliştirilmesine öncülük verilmelidir: Goodrich (1997), Haladyna (1997), Nitko ve Brookhart, (2007), Johnson, Penny ve Gordon, (2009) ile Kutlu vd. (2023)'nin önerileri ışığında DPA geliştirme aşamalarındaki belli başlı özellikler aşağıdaki gibi özetlenebilir;

1. İlgili yapının ölçülesine ilişkin (bilgi, beceri veya davranışın) ölçülme amacının belirlenmesi,
2. İlgili yapının ölçülmesi amacıyla yapıyı temsil eden uygun ölçütlerin belirlenerek taslağının oluşturulması,

3. Ölçütlerin değerlendirilmesinde uygun puanlama anahtarının seçimine ilişkin kararın verilmesi,
4. Değerlendirme işlemi öncesinde öğrencilere hangi ölçütler kapsamında değerlendirilecekleri ile ilgili ölçütlerin öğrencilerle veya bireylere paylaşılması,
5. Yönergeye göre ön uygulamanın yapılması (uygun puanlama anahtarına göre puanlanması),
6. Ölçme aracının uzmanlar tarafından (alan, dil, ölçme) anlam, içerik, yazım kuralları vb. gibi incelemelerden geçirilmesi,
7. Ölçme aracının nihai formunun oluşturulmasıdır.

Bütünsel Puanlama Anahtarı

Bütünsel değerlendirme ölçütü, değerlendirilecek performansın bölümlere ayrılmadan tek bir puanla derecelendirir (Güler, 2019, s. 82; Kutlu vd., 2017, s. 56). Dolayısıyla bu ölçme aracını kullanarak puanlayıcı birey performansını veya çalışmasını bütün olarak değerlendirir ve puanlar (Korkmaz, 2004, s. 234). Nihai sonuç veya ürün önemli olduğunda, performans değerlendirme için zaman kısıtlı olduğu durumlarda ve değerlendirilmekte olan performans bileşenlerini alt boyutlara ayırmanın mümkün olmadığı durumlarda tercih edilir (Mertler, 2001).

Analitik Dereceli Puanlama Anahtarı

Analitik dereceli puanlama anahtarı, ölçülecek performansın her aşamasını ayrıntılı olarak açıklayan bir puanlama aracıdır (Kutlu vd., 2017). Her bir performansçının değeri, performans bileşen parçasına göre değerlendirilir. ADPA'nın kullanımı bir bireyin kendisindeki güçlü ve zayıf yönlerini ortaya koyarak performans hakkında daha kapsamlı ve ayrıntılı bilgi verir (Gronlund, 1998; Kutlu vd., 2017, s. 60). Mertler (2001) (Kutlu vd., 2017) ADPA'nın dikkate alınması gereken birden fazla önemli hususun olduğunu vurgulamışlar ve performans çok boyutlu olduğunda, öğrencilerin yapmış olduğu ürünlerine veya performanslarına dair güçlü ve zayıf yönlerinin ayrıntılı belirlenmesini gerektiren durumlar için yararlı bir araç olduğunu belirtmişlerdir.

Puanlayıcı Eğitimi

Bireylerin performanslarının değerlendirildiği süreçte yapılan ölçümlere karışan hataları veya davranışları benzer hale getirmek için başvurulacak yöntemlerden biri puanlayıcı eğitimidir. Puanlama eğitiminin amacı olarak şunlar söylenilebilir: puanlama sürecinin standart hale getirilmesi, puanlayıcıların puanlama davranışlarının benzer hale getirilmesi, puanlayıcılarla ortak bir anlayış geliştirmesi puanlama esnasında meydana gelebilecek öngörülemez olaylara hazırlıklı olunmasını amaçlar (Robertson, Park, Gillman & Vergis, 2020; Şata, 2022). Puanlayıcı eğitiminde puanlayıcılar arasında ortak bir anlayışa varmak çok önemlidir. Buna göre herhangi bir puanlayıcı eğitiminin başlangıcında ilk adım olarak şunların yapılması önemlidir (İlhan, 2016). Öncelikle puanlayıcılar üzerinde değerlendirme yapacakları test veya performans görevleri ile performans değerlendirmesinde kullanılacak olan ölçme aracı hakkında ayrıntılı bilgilendirilmelidir. Diğer bir ifade ile, a) ölçülen yapı ve boyutları (bilişsel, duyuşsal), b) ölçülmesi amaçlanan performans düzeyleri (başlangıç, orta ve ileri), c) her bir performans düzeyindeki (başlangıç, orta ve ileri), yapıyla ilgili ölçütlerin bilinmesi, d) puanlama ölçeğinde bulunan derecelemelerin ve e) performans sergileyen bireylerin göstermiş olduğu performansın zorluk veya kolaylık düzeylerine ilişkin puanlayıcılar arasında ortak bir anlaşma veya uzlaşmaya gitmek puanlayıcı eğitiminin ilk adımlarındandır. Puanlama eğitimlerinin ikinci adımı üzerinde ölçme işlemi yapılan yapı ile ilgili performans temsil eden örnek uygulamaların birden çok kez yaptırılmasıdır. Bu sayede puanlayıcı asıl performans değerlendirmesine geçmeden önce örnek uygulamalar yaparak ölçülmek istenilen yapı ile ve o yapıyı ölçme için geliştirilen ölçme aracındaki ölçütlerin kullanımına ilişkin beceri kazanmış olur. Böylece puanlayıcılar hem ölçülen yapı hem de ölçme aracının kullanımına ilişkin fikir birliği sağlayabilirler. Puanlayıcı eğitiminin son adımında ise performans değerlendirmede görev alan puanlayıcıların yapmış olduğu puanlamalara ilişkin geri dönütler verilerek davranışların incelenmesidir (Pulakos, 1991; Şata, 2022). Alan yazın incelendiğinde performans değerlendirme sürecinde görev alan puanlayıcı veya değerlendiricilerin nesnel puanlama davranışlarını geliştirmek için en sık kullanılan altı eğitim deseni mevcuttur. Bu eğitim desenleri kısaca aşağıdaki gibi ele alınmıştır;

Birinci eğitim stratejisi olan *Öz Liderlik* bu eğitim stratejisinde, puanlayıcıların kendi değerlendirme süreçlerini anlamalarına ve iyileştirmelerine odaklanır. Bu süreçte bireyler kişisel sorumluluk olarak değerlendirme hatalarını minimize etmeyi hedefler (Neck, 1996). Öz liderlik eğitimini üç bileşen tanımlar: davranış odaklı teknikler, doğal ödül sistemleri ve

yapıcı düşünce kalıpları (Houghton & Neck, 2002). Doğal ödül sistemleri içsel motivasyon ve yapıcı düşünce kalıplarına atıfta bulunurken, davranış odaklı stratejiler kişisel hedefler öz gözlem, ödül ve ceza gibi yönleri içerir (Furtner, Tutzer & Sachse, 2018). Öz düzenleme tekniklerini kullanmak, davranışı temel alan yaklaşımlarda istenilen durum ile birey arasındaki farkı en aza indirmeyi amaçlamaktır (Furtner, Sachse & Exenberger, 2012). Değerlendiricinin bilişsel yapısının doğruluğu veya yanlışlığı puanlama davranışını etkileyeceğinden dolayı yapılacak ölçmeler bireye özgü olmalıdır (Neck vd., 1995).

İkinci eğitim stratejisi olan *Davranış Gözlem Eğitiminde* puanlayıcıların performansla ilgili davranışları gözlemleyip doğru şekilde kaydetmeleri üzerine odaklanır. Gözlem sırasında önyargıların ve hatalı izlenimlerin engellenmesi için eğitim verilir (Woehr & Huffcutt, 1994).

Performans boyutları ile ilgili davranışların gözlenmesi ve hatırlanmasını arttırmak için performansın değerlendirmesine yönelik önemli bazı noktaların not alınması, puanlama veya eğitim süreci hakkında günlüklerin tutulması, sıklık hesaplama gibi becerileri puanlayıcılara kazandırmayı amaçlar (Newman, Kinney & Farr, 2016; Woehr & Huffcutt, 1994). Gözlem kayıtlarının başarılı bir şekilde tutulması değerlendirme sürecinde meydana gelebilecek beklenmedik tehlikeyi atlatarak puanlayıcılardan kaynaklı etkileri azaltmaya yardımcı olabilir (Antonioni & Woehr, 2001).

Üçüncü eğitim stratejisi olan *Performans Boyutu Eğitiminde* ise değerlendiricilerin ölçme aracındaki belirlenen ölçütlerin daha iyi anlamaları ve bu ölçütlere dayalı olarak doğru kararlar alınmasına yöneliktir. Özellikle diğer eğitim stratejilerinden farklı olarak durum belirlenmesi istenilen performansın boyutlarına odaklanır (Feldman vd., 2012). Puanlayıcıların performans hakkında bilgi sahibi olmak için kullanılabilecek iki yaklaşım mevcuttur. İlki puanlayıcılar puanlama işlemi yapmadan önce ölçme aracındaki ölçütleri gözden geçirdiklerinde ve ölçülmesi istenilen yapı hakkında ölçme aracını geliştirmeye katkıda bulundukları zaman puanlayıcıların performans boyutuna aşina olarak puanlamaya devam edebilirler (Woehr, 1994). Eğitim esnasında eğitimciler performans boyutu eğitimini verirken ölçme aracındaki her bir ölçütün kullanılarak nasıl doğru puan vereceklerini göstermek için ölçülen yapının boyutlarında örnek vererek puanlamanın doğru yapılmasını sağlarlar. Performans örnekleri oyunlar, filimler, metin veya hikâye şeklinde verilebilir. Buradaki asıl amaç istenilen sonuçla ilgisi olmayan puanlayıcı eylemlerini azaltarak veya ortadan kaldırarak verilen puanların güvenilirliğini arttırmaktır (Feldman vd., 2012). Özetle

performans boyutu eğitiminde amaç, puanlama sırasında yapılan hata sayısının az veya çok olmasından ziyade doğruluğunu arttırmaktır (Woehr & Huffcutt, 1994).

Dördüncü eğitim stratejisi olan *Puanlayıcı Hatası Eğitimidir*, performansın puanlanması sürecinde meydana gelebilecek sistematik hataları tanımlamayı ve bu hataları nasıl azaltılabileceğine dair puanlayıcıları eğitmeyi hedefler (Myford & Wolfe, 2003, 2004). Örneğin alan yazındaki puanlayıcı davranışlarına karışan 30'dan fazla hata türleri örnek olarak verilebilir (cömertlik, katılık, merkeze eğilim vb.). Bu eğitim stratejisinin temel varsayımı, performans değerlendireme sık karşılaşılan yaygın puanlayıcı davranışlarına hâkim olmanın ve puanlayıcıları bu hata türlerinde kaçınmayı teşvik etmekle doğrudan puanlayıcı davranışlarında azalmaya veya daha etkili puanlama davranışı sergileme yoluna gidilebilir (Woehr & Huffcutt, 1994).

Beşinci eğitim stratejisi olan *Referans Çerçevesi*, puanlayıcılara performans görevinin değerlendirilmesi esnasında puanlamaya ilişkin ortak bir anlayış geliştirmesini amaç edinerek puanlayıcılara performansla ilgili referans çerçevesi sunar. Puanlayıcıların performansın değerlendirilmesine ilişkin geliştirilen ölçme aracının kullanımını öğrenmelerini sağlayan bir eğitim stratejisidir (Woehr ve Huffcutt, 1994). Puanlayıcıların ölçülen tüm performans boyutları için benzer bir zihinsel model geliştirmeleri ve puanlama sırasında bu zihinsel modeli kullanarak puanlama yapmalarını sağlamak referans çerçevesi eğitiminin temel amaçlarından (Athey & McIntyre, 1987; Woehr, 1994). Diğer bir ifade ile RÇE, puanlayıcı davranışlarını azaltarak puanlama doğruluğunun daha iyi hale gelmesi için ortak bir zihinsel model geliştirmeyi baz alırken diğer eğitim stratejileri geri bildirim, standart belirleme veya puanlayıcı motivasyonunu artırma gibi stratejiler kullanılarak puanlayıcı güvenirliliğini arttırmayı amaçlayarak bu eğitimin kullanım sıklığını diğer eğitimlere göre artırır.

Altıncı eğitim stratejisi olan *Puanlayıcı Değişkenliği Eğitimidir*, performans puanlama sırasında aynı grup içinde ya da farklı puanlayıcılar arasında tutarlılığı sağlamayı hedefler. Bu strateji değerlendiriciler arasındaki varyasyonun etkisi minimize etmek için başvuru yöntemlerdendir (Stamoulis & Hauenstein, 1993). PDE, referans çerçevesi ile puanlayıcı hatası eğitim stratejilerinden ayıran en önemli özelliği puanlayıcılara puanlama işlemi bittikten sonra puanlama davranışları ile ilgili geri dönütlerin sağlanmasıdır (Mubeena, 2010). Diğer eğitim stratejileriyle kıyasla PDE, performans değerlendirmesi ile ilgili yapılan araştırmalarda az tercih edilmiştir. Bunun nedeni puanlayıcı değişkenliği eğitiminin,

referans çerçevesi eğitim yaklaşımıyla birçok yönden benzer olmasıdır (Şata, 2022). Tüm verilen belgilerden hareketle puanlayıcı eğitimleri aşağıdaki gibi Tablo 2’de özetlenmiştir.

Tablo 2

Puanlayıcı Eğitimi Desenleri Amaç ve Odak Noktası

Eğitim Türü	Amaç	Odak Noktası
Öz Liderlik	Puanlayıcının öz farkındalığını artırarak kendi performansını yönetebilmesini sağlamak.	Öz disiplin geliştirme, hedefe uygun davranışlar sergileme.
Davranış Gözlem Eğitimi	Performansın detaylı gözlemlenmesi ve kaydedilmesini öğretmek.	Davranışların objektif şekilde kaydedilmesi.
Performans Boyutu	Puanlayıcıların, belirli performans boyutlarına odaklanarak puanlama yapmalarını sağlamak.	Performansı alt boyutlara ayırıştırma ve her birini bağımsız olarak değerlendirme.
Puanlayıcı Hatası	Puanlayıcı hatalarını fark ettirerek değerlendirme tutarlılığını artırmak.	Hata farkındalığı ve azaltılması.
Referans Çerçevesi	Puanlamada tutarlılığı artırmak için puanlayıcıların ortak bir çerçevede eğitilmesi.	Ortak standartlar, tutarlı ve adil değerlendirme yapma.
Puanlayıcı Değişkenliği	Puanlayıcıların varyasyon kaynaklarını anlamalarını ve değerlendirme sürecinde bunları kontrol etmelerini sağlamak.	Değerlendirme varyasyonlarının nedenleri, önlenmesi veya azaltılması.

Puanlayıcı Güvenirliği

Bilimde aynı sorunu çözmek için çok sayıda kuram geliştirilmiştir. Dolayısıyla bilim adamları tarafından geliştirilen bu kuramların gerçek dünyada ne kadar doğru ve etkili olduğunu test etmek gerekir. Sonuç olarak, aynı amaca hizmet eden kuramların karşılaştırılması, hangi kuramın daha fazla bilgi verdiği, etkili olup olmadığı ve gerçek dünyadaki durumlarla uyumlu olup olmadığı konusunda test edilmesi gerekir. Kuramların karşılaştırılması, işleyen veya işlemeyen yönlerini belirlemek için gerekli olduğunu vurgulanarak, benzerlikleri ve farklılıkları ortaya koymanın bilimsel çalışmaların önemli bir parçası olduğunu belirtir (Atılğan, 2004).

Sosyal bilimler araştırmaları ve eğitimdeki temel fikirlerden biri bir ölçme aracının rastgele hatalardan arınmış olma derecesidir. Araştırma veya puanlama sürecinde toplanan ölçümlerin güvenilirliği zorunluluktur (Downing & Yudkowitz, 2009, s. 11). Tekdüze puanlama söz konusu olduğunda, tüm ölçüm testleri eşit puanlama işlemlerine uygun olmayabilir. Bu nedenle, bir ölçüm sürecinde birden fazla puanlayıcıdan elde edilen ölçümler arasındaki tutarlık veya benzer puanlama davranış sergilenip sergilenmediği veya güvenilirlik düzeyinin sorgulanması yapılan ölçümün geçerli olmasından geçmektedir. Belli bir ölçüm için iki veya daha fazla puanlayıcı arasındaki uyum veya tutarlılık düzeyine puanlayıcılar arası güvenilirlik denir (Cohen & Swerdlik, 2013; s.150-151).

Klasik Test Kuramına Göre Puanlayıcı Güvenirliği Belirleme Teknikleri

Ölçmenin amacı, ölçülen özelliğin gerçek değerine yaklaşımdır. Çünkü ölçümlere çeşitli hatalar karışır. Bu nedenle gerçek değeri elde etmek pratikte sağlanamaz; gözlenen puanlardan kestirilmeye çalışılır. Klasik test kuramı gerçek puanı gözlenen puandan kestirmeye çalışan kuramlardan biridir. Klasik test kuramına gerçek puan modeli de denilmektedir. Klasik test kuramının temel denklemi şöyle ifade edilmektedir: $T=X+E$ (T: Gerçek puan, X: Gözlenen puan, E: Random hatalar). Klasik Test Teorisi, bireyin gözlenen özelliğinin içinde hem gerçek puan hem de hata puanı bulunduğunu kabul eder (Baykul, 2010; Crocker & Algina, 2006). Geleneksel klasik test kuramına göre, performans değerlendirme sürecinde puanlayıcılar arasındaki uyuma derecesini belirlemek için çok sayıda katsayı hesaplanır. Cohen ve Swerdlik'e (2013) göre, performans değerlendirmede puanlayıcılar arası tutarlılık düzeyinin belirlenmesinin en basit yollarından biri korelasyon katsayısının hesaplanmasıdır. Diğer teknikler tablo 3' de özetlenmiştir.

Tablo 3

*KTK Temelli Puanlayıcılar Arası Güvenirliğin Belirlenmesi**

Özellikler İstatistiksel Yöntemler	Toplam Puan üzerinden uyum düzeyine bakar	Uyum düzeyine her ölçüt için ayrı ayrı baka.	2 puanlayıcı arasındaki uyum düzeyine bakar	2' den fazla puanlayıcı arasındaki uyum düzeyine bakar.
Cohen's Kappa	-	+	+	-
Fleiss' Kappa	-	+	-	+
r	+	-	+	-
rs	+	-	+	-
Kendall W	+	-	-	+
Sınıf İçi	+	-	+	+
Korelasyon Katsayısı				
r: Pearson momentler çarpımı korelasyon katsayısı				
rs: Spearman Brown sıra farkları korelasyon katsayısı				

Kutlu, Ö., Doğan, C. D., & Karakaya, İ. (2017). *Ölçme ve değerlendirme performansa ve portfolyoya dayalı durum belirleme* (5. Baskı). Ankara: Pegem Akademi, s. 87) uyarlanmıştır.

Genellenebilirlik Kuramına Göre Puanlayıcı Güvenirliği

Psikoloji ve eğitim gibi büyük ölçekli kavramlar ve değerlendirmelerle ilgilenilen alanlar için genellenebilirlik kuramı sağlam temellerden biridir. Çünkü bu kuram güçlü bir istatistiksel araç olan varyans analizini kullanır. Temel çıkış noktası KTK'ya dayanır fakat

varyans analizini kullanmasından dolayı güçlü bir kuramdır. KTK'daki uygulamalarda birçok hata kaynağını aynı anda ortaya çıkarmak mümkün değilken; G kuramına göre uygulamalardaki birden fazla hata kaynağını açıklayabilir (Güler, Kaya Uyanık & Taşdelen Teker, 2012; s.3).

Atılğan (2004)' a göre G kuramı, ölçme işlemi sırasında birden fazla olası hata kaynağının ortaya çıktığı durumlarda, güvenilirlik tek bir analizde hesaplanır. Böylece tek bir güvenilirlik katsayısı veya değer hesaplanmış olur. G kuramının kullanılarak yapıldığı çalışmada hem bağıl hem de mutlak değerlendirme için yine tek güvenilirlik katsayısı bağımsız olarak hesaplanabilir. Bu farklılıklarla kalmayıp, ölçüm modelinin seçilmesi ve aracın geliştirilmesine yardımcı olmakla beraber, G çalışması yürütülerek yapılan ölçüm hatasının nedenlerini açıklamaya olanak sağlar. Karar çalışması yapılarak ölçümün, ölçme aracının geliştirilmesinde veya gelecekteki uygulamalarda nasıl kullanılması gerektiği konusunda bilgi sunarken, G çalışması bir ölçüm aracının geliştirilmesi ve gelecekteki uygulamalarda kullanımı için değişkenlik kaynaklarından ortaya çıkan hataların en aza indirmesi konusunda avantaj sunar. G kuramı hata kaynakları ve bunlar arasındaki etkileşimlerden bahseder fakat hangi bireyin veya hangi nesnenin hata ürettiği hakkında ayrıntılı bilgi sağlamaz. Bunun için ayrıntılı bilgileri Madde Tepki kuramı kullanılarak elde edilebilir.

Madde Tepki Kuramına Dayalı Puanlayıcı Güvenirliği

Rasch, MTK kuramının temelini oluşturur. İki ve üç parametrelili modellerin belli bir örneği olan Rasch modeli tek parametrelili lojistik model olarak adlandırılır. Rasch modeli, birkaç kullanım özelliklerinden dolayı araştırmacıların dikkatini çekmektedir. Bu özelliklerden biri, daha az parametreye sahip olması nedeniyle işi kolaylaştırmasıdır. Bir diğeri ise daha geniş modellerle karşılaştırıldığında parametre tahminleri ile çok daha az sorun olmasıdır (Hambleton & Swaminathan, 1991, s. 47). Nesnel bir ölçüm sürecini yürütmek için, Rasch ölçüm modeli analizindeki ham puan yerine ham puanın standart dönüşünü kullanarak hesaplama yapar (Linacre'den aktaran, Baştürk & Işıkoğlu, 2008). Linacre (1989), Rasch modeline değerlendirici etkileri gibi değişkenlik kaynaklarını dahil ederek çok Yüzeyli Rasch modelini oluşturmuştur.

Çok Yüzeyli Rasch Ölçme Modeli

Bireyler üzerinde ölçülen performansların birinci değişkenlik kaynağı bireylerin bir görevi yapabilme yeteneğidir. Ölçme işlemine tabi tutulan bireylerin yetenek düzeylerini farklı

olması beklenen bir durumdur. İkinci değişkenlik kaynak ise değerlendirmenin yapılabilmesinde kullanılan ve bireylerden istenilen görevlerdir. Üçüncü değişkenlik kaynak olarak bütün bu değerlendirme sürecinde aktif rol alan puanlayıcı veya değerlendircidir. Dördüncü değişkenlik kaynağı bu değerlendirme işinin doğru yapılabilmesinde önemli paya sahip olan ölçme aracıdır.

Performansa dayalı durum değerlendirmeler veya yazılı sınavların değerlendirilmesini içeren süreçlerde, puanlayıcıların heterojenliği, değerlendirme sonuçlarını etkileyen önemli faktörlerdendir. Bu puanlayıcılardaki farklılık (heterojenlik) puanlayıcıların hem fiziksel hem de zihinsel özellikleri bireysel tepkileri derecelendirme stratejileri ve algılama biçimlerinden kaynaklanmaktadır. Her bir puanlayıcının deneyimleri, bilişsel stilleri, dikkat düzeyleri ve değerlendirme bağlamına verdikleri anlam nedeniyle aynı performansı farklı şekilde veya bakış açısı ile yorumlayabilir. Bu durum, bireyler arasındaki farklılıklara dayalı olarak yorumlamalarda çeşitliliğe ve bazen de tutarsızlıkların meydana gelmesine sebebiyet verebilir. Örneğin, bir puanlayıcı ölçme aracındaki belirli bir ölçütü vurgularken, diğeri farklı bir ölçüte odaklanabilir veya aynı ölçütleri uygularken daha katı veya daha cömert davranış eğiliminde olabilir. Tüm bu faktörler, performans değerlendirme kapsamının öznel yönlerini artırarak elde edilen sonuçların güvenilirlik ve geçerliğini etkileyebilir (Atılğan, 2004; Lunz & Wright 1997).

Görüldüğü gibi performans değerlendirme sürecinde birden fazla potansiyel hatanın bir arada ayrıntılı olarak incelenmesine olanak sağlayan MTK'nın uzantısı olan Çok Yüzeyle Rasch ölçme modeline başvurulur. Basitçe ÇYRÖM aşağıdaki denklem ile açıklanabilir (Linacre, 1989): İki kategorili (1-0) olan ve sadece birey ve maddelerden oluşan çoktan seçmeli maddeler gibi puanlamalar için temel veya basit Rasch modeli aşağıdaki gibi yazılabilir (Linacre, 2023, s. 17):

$$\log\left(\frac{P_{ni}}{1 - P_{ni}}\right) = B_n - D_i \quad 1.1$$

Bu formüle göre;

P_{ni} : n numaralı öğrencinin i numaralı maddede/görevde başarılı olma olasılığı

B_n : n numaralı öğrencinin yetenek düzeyi

D_i : i numaralı maddenin/görevin güçlüğü olarak açıklanmaktadır.

Yukarıdaki formülde görüldüğü gibi bireylere ya da öğrencilere ve maddenin güçlüğü/zorluk derecesi yüzeylerinden oluşan iki Yüzeyli (öğrenci ve madde), bir parametrelili (madde güçlüğü, zorluğu) Rasch modeli mevcuttur. Andrich (1978), tek parametrelili modeldeki matematiksel özellikleri koruyup geliştirerek tutum ölçekleri gibi çok kategorili puanlamalarda kullanıma uygun halo getirmiştir. İki yüzeyden oluşan ve çok kategorili puanlamalara ait Rasch modeli aşağıdaki gibi yazılabilir (Linacre, 2023):

$$\log\left(\frac{P_{nik}}{P_{ni(k-1)}}\right) = B_n - D_i - F_k \quad 1.2$$

Bu formüle göre;

P_{nik} : n numaralı öğrencinin i numaralı maddede veya testte k puanı gözlenme olasılığı

$P_{ni(k-1)}$: n numaralı öğrencinin i numaralı maddede veya testte k-1 puan gözlenme olasılığı

B_n : n numaralı öğrencinin görevi yaparken gösterdiği yetenek düzeyi

D_i : i numaralı maddenin veya testin güçlüğü/ zorluğu

F_k : k-1 puanına göre k puanı almanın zorluğu

şeklinde açıklanabilir.

Yukarıdaki gösterilen formüle göre birey ve madde veya test güçlüğü/ zorluğu yüzeylerinden oluşan iki Yüzeyli (öğrenci ve madde), çok kategorili puanlamalara uygun Rasch modeli verilmiştir. Bu modelde sıklıkla çoktan seçmeli maddeler ve öğrenciler için kullanılan iki kategoriden oluşan (1-0) puanlama için kullanılan temel Rasch modelinden bireysel puanı hesaplamada madde veya test zorluk, güçlük seviyelerini dikkate alması bakımından farklılık göstermektedir.

Performansların sık sık ölçüldüğü disiplinlerde iki yüzeyli çalışmalar yeterli bilgi sağlamayabilir. Çeşitli değerlendiricilerin davranışlarının ele alındığı araştırmalarda birey, öğrenci ve madde veya test yüzeyine üçüncü bir yüzey olarak karşımıza çıkan puanlayıcı veya değerlendirici yüzeyi eklenebilir. Aşağıda üç yüzeyli çok kategorili puanlama içeren temel Rasch formülü yer almaktadır. (Linacre, 2023):

$$\log\left(\frac{P_{nij k}}{P_{nij(k-1)}}\right) = B_n - D_i - C_j - F_k \quad 1.3$$

Bu formüle göre;

P_{nijk} : j numaralı değerlendiricinin n numaralı öğrencinin i numaralı maddesine veya görevine k puan verme olasılığı

$P_{nij(k-1)}$: j numaralı değerlendiricinin n numaralı öğrencinin i numaralı maddesine veya görevine k-1 puan verme olasılığı

B_n : n numaralı öğrencinin yetenek düzeyi

D_i : i numaralı maddenin güçlüğü ya da zorluğu

C_j : j numaralı puanlayıcının katılık ya da cömertlik düzeyi

F_k : k-1 puanına göre k puanı almanın zorluğu

şeklinde açıklanabilir.

Yukarıdaki gösterilen formüle göre (birey, madde ve puanlayıcı), çok kategorili puanlamaya uygun Rasch modeli verilmiştir. Üç yüzeyli modelin yukarıda açıklanan iki modelli ölümlemlerden farkı bireyin veya öğrencinin performans görevinden elde edilen puanı belirlenirken maddelerin veya görevlerin zorluk düzeyleri ile performans değerlendirmede puanlayıcılardan kaynaklı yaygın hata türlerinde değerlendiricinin katılığı ya da cömertliği gibi davranışların istatistiğini de hesaplama katmasıdır (Linacre, 2023, s. 17). İlgili araştırma öğrencilerin başlangıç düzey piyano performanslarının değerlendirilmesinde öğrenci performansları puanlayıcılar tarafından puanlanmıştır. Dolayısıyla üç yüzeyli (öğrenci performansı x ölçüt x puanlayıcı) araştırmanın Rasch modeli şu şekilde açıklanabilir:

$$\log\left(\frac{P_{bmpx}}{P_{bmp(x-1)}}\right) = B_b - D_m - C_p - F_x \quad 1.4$$

Bu formüle göre;

P_{bmpx} : p numaralı puanlayıcının b numaralı öğrencinin m numaralı performans görevine x puan verme olasılığı

$P_{bmp(x-1)}$: p numaralı puanlayıcının b numaralı öğrencinin m numaralı performans görevine x-1 puan verme olasılığı

B_b : b numaralı öğrencinin yetenek düzeyi

D_m : m numaralı maddenin zorluğu ya da güçlüğü

C_p : p numaralı puanlayıcının katılık ya da cömertlik düzeyi

F_x : x-1 puanına göre x puanı almanın zorluğu

Yukarıdaki üç yüzeyle Rasch formülü araştırmanın modelini belirlemiştir. Dolayısıyla araştırma yukarıdaki modele göre tasarlanmıştır.

Araştırmanın Rasch modellerinden hangisiyle modelleneyeceği belirlendikten sonra (öğrenci performansı, ölçüt ve puanlayıcı) yüzeyleri tek bir doğrusal ölçek (logit cetveli) üzerinde konumlanır. Dolayısıyla belirlenen yüzeylerin tamamının gözlenme şansı olur. Performans ölçümünde seçilen maddenin veya görevin (eser) zorluk derecesi, öğrencinin verilen görevi yapabilmesine ilişkin yeteneği ve bu iki yüzeyi değerlendirecek olan puanlayıcının puanlama esnasında göstermiş olduğu puanlama davranışları ölçülmüş olacaktır. Her yüzeyin analiz sonucuna ilişkin logit tahminler, standart hata ve uygunluk istatistikleri bireysel düzeyler için hesaplanır (Linacre, 2023, s. 16; Myford & Wolfe, 2003).

Bireysel istatistiklerle ilgili incelenmesi gereken değerlerin başında logit değeri gelmektedir. Bu değer sıralı ölçek düzeylerindeki verilerden toplanan eşit aralıklı doğrusal bir ölçü birimidir. Bu birim sayesinde kişiler hakkında ayrıntılı bilgi sağlar (Linacre, 2023, ss. 309). ÇYRÖM analiz sonucunda logit değerleri logit cetvelinde yer alır. Elde edilen logit değerlerin düzeyine göre bireylerdeki yetenek ve verilen görevin veya maddenin zorlukları hakkında yorum yapılabilir.

Bireysel istatistiklerle ilgili incelenmesi gereken değerlerin ikincisi standart hata değeridir. Bu değer ile sıralı veriler eşit aralıklı verilere dönüştürmede kullanılarak meydana gelen dönüşümün anlamlı etki yaratabilmesi için gereken en küçük değere standart hata denir (Linacre, 2023, s. 320). ÇYRÖM’de puanlayıcı davranışlarının incelenmesinde logit ve standart hata değerlerinden faydalanılır. Puanlama işlemi yapan puanlayıcıların davranışları bireysel düzeyde ele alındığında katılık veya cömertlik eğilimlerinin belirlenmesi için logit değerleri t- değerlerine göre dönüştürülerek hesaplanır (Myford & Wolfe, 2003).

Logit ölçüsünün t- değerine dönüştürülmesi için kullanılan denklem:

$$t\text{-değeri} = \frac{\text{puanlayıcıya ait logit değeri} - \text{logit değerlerinin ortalaması}}{\text{logit değerinin standart hatası}} \quad 1.5$$

1.5 eşitliğindeki formüle göre hesaplanan t değerinin t dağılım tablosunda yer alan t

değerinden büyük çıkması istatistiksel olarak anlamlı olduğunu ifade eder. Küçük çıkması ise anlamsız olduğunu açıklar. Bu değer pozitif ve negatif olması ile ilgili de puanlama işlemini yapan puanlayıcıların puanlama davranışlarından söz edilebilir. Diğer bir ifade ile t değeri pozitif olan puanlayıcının birey veya öğrenci performansını değerlendirirken cömert davrandığı tersi düşünüldüğünde de katı davrandığına ilişkin çıkarım yapılabilir.

Uygunluk istatistik değerleri de ÇYRÖM’ de incelenmesi gereken üçüncü değerlerdendir.

Gözlenen puan ile beklenen puan arasındaki farka uygunluk istatistikleri denir (Linacre, 2023).

ÇYRÖM sadece bireysel düzeyler için istatistikler hesaplamayıp, aynı zamanda her bir yüzey hakkında ki-kare istatistikleri ve ayırma istatistikleri, ayırma oranı ve güvenilirlik indeksi hesaplar. Her bir yüzey için yüzeye ilgili istatistikler de hesaplanmaktadır. Hesaplanan istatistikler öncelikle madde veya göreve ilişkin kategori istatistikleri, puanlayıcı veya değerlendirici yüzeyine ait puanlayıcı güvenilirlik istatistiği ayrıca yanlılık için de yanlılık indeksi hesaplanmaktadır.

İlgili Araştırmalar

Sınıf İçi Eğitim İle İlgili Performansın Değerlendirmesinde Puanlayıcı Davranışlarının İncelenmesi

Bu kısımda; araştırmanın içeriğiyle ilişkili alanda yürütülmüş bazı araştırmalara yer verilmektedir.

Karakaya (2015), Çok Yönlü Rasch ölçme yaklaşımını kullanarak portfolyo değerlendirmesinde öğretmen, aday cinsiyeti, öğretmen başarısı ve öğretmen adayının hem akranını hem de kendilerinin değerlendirmeleri üzerine araştırma yapmıştır. 74 öğretmen adayı ile çalışılarak öğretmenin hem katılık hem de cömertlik davranışları dikkate alındığında dörtlü grup arasında Rasch modeline göre değerlendirme işlemi yapılmıştır. Araştırmacı, çalışması için yedi ölçütlü bir DPA geliştirmiştir. Araştırma, öğretmen adayları bayan grubundaki portfolyoların erkek öğretmen adayı grubuna göre daha iyi performans gösterdiği ve yüksek başarı düzeyine sahip öğrencilerin portfolyolarının genel olarak daha iyi performans gösterdiği sonucuna ulaşmıştır. Puanlayıcı türlerinin karşılaştırılması sonucunda akran değerlendirmesinin en katı, öz değerlendirmenin ise en cömert olduğu bulunmuştur.

Şahin, Taşdelen Teker ve Güler (2016) yapmış oldukları araştırmalarında lisans öğrencilerine yönelik akran değerlendirmesinde çok Yüzeyle Rasch ölçme modelini kullanmışlardır. Araştırmanın çalışma grubunda lisans 91 öğrenci ile dersin öğretim elemanı katılım göstermiştir. Araştırmanın değerlendirmesi kısmında bilimsel araştırma dersini alan 85 öğrencinin hazırladığı projeleri 6 akran ve öğretim elemanı tarafından değerlendirilmiştir. Değerlendirmede ölçme aracı olarak projelerin puanlanmasında bütünsel puanlama anahtarı kullanılmıştır. ÇYRÖM sonucunda puanlayıcıların puanlama davranışlarındaki cömertlik ve katılık düzeylerinde farklılık bulunmuş ve öğrenciler birbirlerini değerlendirirken başarılı olan öğrencilerin başarı düzeyi orta ve düşük olanlara kıyasla puanlama esnasında daha katı puanlama davranış gösterdikleri bulunmuştur. Çalışma sonucunda öğrencilerin akran değerlendirmesine ilişkin görüşlerinin olumlu yönde olduğu bulunmuştur.

Şata (2019) tarafından sunulan doktora tezinde üniversite öğrencilerinin yazılı performans olarak kompozisyon değerlendirme sürecindeki nesnelliği engelleyen sorunlu puanlama davranışlarını belirlemeyi ve bu davranışları ölçümlerde kontrol altına almak için deneysel olarak puanlayıcı eğitimlerinin etkisini belirlemeyi amaçlamıştır. Bu araştırma, deney ve kontrol gruplarını içeren ön test-son test uygulamalı yarı deneysel bir tasarım ile gerçekleştirilmiştir. Araştırmanın çalışma grubu, üniversitede lisans öğrencilerinden 23'ü kontrol grubunda ve 22'si deney grubunda toplam olmak üzere 45 puanlayıcıdan oluşmuştur. Araştırmacı performans değerlendirme sürecinde puanlayıcılardan kaynaklı problemli puanlayıcı davranışlarının en çok araştırıldığı altı davranış üzerinde durarak bağımsız örneklem t-testi, genellenebilirlik teorisi ve ÇYRÖM ile analiz edilmiştir. Analiz edilen davranışlar sırası ile şu şekildedir: Halo etkisi, merkezi eğilim, değerlendirici hoşgörüsü, değerlendirici katılığı ve farklılaştırılmış değerlendirici hoşgörüsü ve katılığıdır. İlgili araştırmada, öğrencilerin ikinci dil akademik yazma becerileri performans olarak belirlenmiştir. Bu bağlamda, İngilizce öğretmenliği birinci sınıf öğrencilerine IELTS tarafından yayımlanan yazma görevi verilmiş ve öğrencilerden bir kompozisyon yazmaları talep edilmiştir. Ardından araştırma ile ilgili kişisel bilgi formu öğrencilere dağıtılarak doldurmuştur. Öğrenci kompozisyonlarının değerlendirilmesi amacıyla, bir araştırmacı ve alan uzmanı tarafından akademik yazma becerileri için analitik dereceli puanlama anahtarı oluşturulmuştur. Geliştirilen ölçme aracının güvenilirlik ve geçerlilik analizi gerçekleştirilmiş, akademik yazma becerilerinin değerlendirilmesinde bu aracın kullanılabilir olduğu tespit edilmiştir. Araştırmaya katılan puanlayıcılara, referans çerçevesi eğitimi ve puanlayıcı hatası eğitimi desenlerinin birleştirilmesiyle oluşturulan bir puanlayıcı

eđitimi, her hafta birer saat olmak üzere toplam dört hafta boyunca dört saatlik bir puanlayıcı eđitimi verilmiştir. Araştırmanın bulguları değerlendirildiğinde, puanlayıcı eđitimi öncesinde deney ve kontrol grubunda; puanlayıcı katılımı, puanlayıcı cömertliđi, farklılaşan puanlayıcı katılımı ve farklılaşan puanlayıcı cömertliđinin hem grup düzeyinde hem de bireysel düzeyde gözlemlendiđi, halo ve merkeze yönelim davranışlarının ise yalnızca bireysel düzeyde belirlendiđi saptanmıştır. Puanlayıcı eđitimi sonrası incelenen problemli puanlayıcı davranışlarının yeniden ortaya çıktığı, ancak deney grubundaki bu davranışların ölçümlere karışma düzeylerinin istatistiksel olarak kontrol grubuna kıyasla daha az olduđu tespit edilmiştir. Puanlayıcı eđitiminin, puanlayıcı katılımı ve cömertliđi üzerindeki etkisinin önemli olduđu ve katılım ile cömertlik davranışındaki deđişimin %15'inin bu eđitimle açıklanabildiđi tespit edilmiştir. Puanlayıcı eđitiminin puanlayıcı güvenilirliđi üzerindeki etkisi incelendiğinde, puanlayıcılar arası güvenilirliğin arttığı ve bu etkinin orta düzeyde olduđu belirlenmiştir. Ayrıca, puanlayıcı güvenilirliđindeki deđişimin %10'unun puanlayıcı eđitimi ile açıklanabildiđi tespit edilmiştir. Yapılan analizler sonucunda, her iki yöntemin benzer sonuçlar ürettiđi tespit edilmiştir. Bu araştırmanın bulguları dođrultusunda, jüri üyelerinin katıldığı ve kritik kararların alındığı durumlarda puanlayıcılara yönelik eđitim verilmesi önerilmektedir.

Görgülü Öztürk (2023), sunduđu yüksek lisans tezinde Ankara ilinde bir devlet okulunun 10. sınıftaki lise öğrencisinden 32'si deney, grubu 33'ü kontrol grubunda toplam 65 öğrencinin matematik açık uçlu maddeleri puanlama davranışlarındaki puanlayıcı eđitimin etkisini araştırmıştır. Öncelikle, araştırmacı öğrencilere üç farklı başarı düzeyinde kâğıt seçerek puan verilmiş. Puanlayıcı eđitimi verilmeden önce, tüm grupların puanlayıcı davranışlarının istatistiksel olarak benzer olduđu sonucuna varılmıştır. ÇYRÖM ve genellenebilirlik kuramı ile ilgili analizler için akran deđerlendirmesi yapılmıştır. Daha sonra kontrol grubuna konuyla ilgili ön test uygulanmıştır. Deney grubundaki, puanlayıcı eđitimi öncesinde var olan katılım ve cömertlik davranışlarının puanlayıcı eđitimi sonrasında yüzde olarak deđişmediđi, ancak kontrol grubu bu davranışları azaltıldığına dair sonuca ulaşmıştır.

Erman Aslanođlu ve Şata (2023) tarafından yazılan bilimsel araştırmada zaman içinde deđerlendiricinin ciddiyetini ölçmek için kullanılan bir yöntem olan akran deđerlendirmesinde deđerlendirici kayması sorununu araştırmışlar. Çalışmaya, Fen ve Teknoloji dersinde 8 sözlü sunum yapan 28 yedinci sınıf öğrencisi katılmıştır. Araştırmada, deđerlendirici ciddiyetindeki zamana bađlı kaymayı analiz etmek için ÇYRÖM'ne

başvurulmuştur. Analiz, bireysel düzeydeki bazı değerlendiricilerin zaman içinde daha ciddi veya daha hoşgörülü olma eğiliminde olduğunu, ancak grup düzeyinde önemli bir değerlendirici kayması gözlemlenmediğini göstermiştir. Çalışma, etkili öğrenme sonuçlarını garantilemek için alternatif değerlendirme yöntemlerinde güvenilirlik ve geçerlilik sorunlarını dikkate almanın önemini vurgulamaktadır.

Tursynbayeva, Öç ve Karakaya (2024), tarafından yazılan bilimsel araştırma makalesinde okullardaki sınıf içi akran değerlendirme ile ilgili yapılan çalışmada çok Yüzeyli Rasch ölçme modeli, ortaokul öğrencilerinin akranlarını değerlendirme becerilerini geliştirmek için puanlayıcı eğitiminin çeşitli puanlayıcı davranışlarına etkisini incelemek için kullanılmıştır. Araştırma, tek gruplu ön test son test deseniyle desenlenmiştir. Araştırmada öğrenci, puanlayıcı ve ölçme aracındaki bulunan ölçüt olmak üzere üç yüzey yer almıştır. Puanlayıcıların tümü (öğrenciler), tüm öğrencileri puanladığı için çaprazlanmış (öxöxp) desen olarak tasarlanmıştır. Ölçme aracı yazma becerilerinin ölçülmesine ilişkin araştırmacılar tarafından toplamda yedi ölçüt olarak geliştirilerek psikometrik özellikleri test edilmiştir. Araştırmanın çalışma grubu 2021-2022 eğitim öğretim yılında Ankara ili Keçiören ilçesine bağlı bir devlet okulunun yedinci sınıfında öğrenim gören 23 öğrenciden oluşmuştur. Çalışma grubundaki 23 öğrenciden kompozisyon yazmaları istenmiştir. Kompozisyonlar araştırmacılar tarafından incelenmiş ve 10 tanesi bireylerin değerlendirilmesi için seçilmiştir. Deneysel işlem öncesinde öğrencilerden kompozisyonları değerlendirmeleri istenmiş ve ardından dört hafta boyunca öğrencilere sekiz saatlik bir puanlayıcı eğitimi verilmiştir. Puanlayıcı eğitimi sonrasında öğrencilerden tekrar kompozisyonları değerlendirmeleri istenmiştir. Araştırmada; puanlayıcı katılığı, puanlayıcı cömertliği, farklılaşan puanlayıcı katılığı ve farklılaşan puanlayıcı cömertliği olmak üzere dört puanlayıcı davranışı ÇYRÖM ile incelenmiştir. Araştırma sonuçları incelendiğinde puanlayıcı eğitiminin katılık, cömertlik, farklılaşan puanlayıcı katılığı ve cömertliği gibi davranışların azalmasına katkıda bulunduğu gözlenmiştir.

Öç, (2024) tarafından sunulan doktora tezinde sınıf içi ölçme ve değerlendirme kapsamında kullanılan açık uçlu soruların puanlanmasındaki puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisini ÇYRÖM ve Mann-Whitney U testi ve Wilcoxon testleri kullanılarak analiz edilmiştir. Araştırma, 2022-2023 eğitim-öğretim yılı bahar döneminde Sivas Merkez'deki beş okulun 8. sınıfında öğrenim gören 70 öğrenci ile yapılmıştır. Öğrenci kâğıtları iyi, orta ve geliştirilmeli olarak üç gruba ayrılmış ve her gruptan 10 kâğıt analiz dışında tutulmuştur. Araştırma, 14 Türkçe ve 20 matematik öğretmenliği öğrencisinden

oluşan deney ve kontrol grupları ile gerçekleştirilmiştir. Puanlayıcı davranışları olarak puanlayıcı katılığı, cömertliği, merkeze eğilim, Halo etkisi gibi altı farklı davranış incelenmiştir. Veri toplama aracı olarak, araştırmacı tarafından geliştirilen açık uçlu başarı testleri ve analitik ve bütünsel dereceli puanlama anahtarı kullanılmıştır. Deney grubuna dört hafta boyunca dört saatlik bir puanlayıcı eğitimi verilmiştir. Sonuçlar, puanlayıcı eğitim öncesi ve sonrası Türkçe ve matematik puanlayıcı gruplarında istatistiksel farklılıklar olduğunu göstermiştir. Eğitim sonrası puanlayıcı davranışlarında azalma gözlemlenmiş, Türkçe grubunda %60,2 ve matematik grubunda %57,8 değişim puanlayıcı eğitimi ile açıklanmıştır. Matematik puanlayıcılarının davranışlarındaki azalma daha belirgin bulunmuştur. Sonuç olarak, açık uçlu soruların puanlanmasında puanlayıcılara eğitim verilmesi önerilmektedir.

Sonuç olarak ulaşılabilen veya bu araştırmada verilen belli başlı eğitim alanında yürütülen sınıf içi ölçme ve değerlendirmede uygulamalarında öğrenci performansların değerlendirmesi birçok çalışmaya konu olmuştur (Karakaya, 2015; Şahin vd., 2016; Şata 2019; Görgülü Öztürk, 2023; Erman Aslanoğlu ve Şata, 2023; Tursynbayeva vd., 2024; Öç, 2024). Yapılan bu çalışmalar özetlendiğinde: sınıf içi ölçme ve değerlendirme uygulamalarında e-portfolyo kullanılması ile öğretmen adayı, cinsiyeti, öz ve akran değerlendirmesi; bilimsel araştırma dersindeki projelerin akran değerlendirmesinde; yazma kompozisyonlarının değerlendirmesinde; matematik açık uçlu soruların değerlendirilmesinde ve puanlayıcı davranışlarının belirlenmesinde; öğrencinin yeteneği, öğrenciye verilen ödevin zorluk derecesi, puanlayıcı davranışlarının Çok Yüzeyle Rasch Ölçme Modeline göre yapıldığı ve sadece öğrencinin performansı, öğrenciyi değerlendirmek için geliştirilen ölçme aracındaki ölçütlerin ne kadar iyi çalışıp çalışmadığı ve puanlayıcı davranışlarının ortaya konulması ile sınırlı olmayıp daha derinlemesine inerek puanlama yapan puanlayıcı davranışlarını ortadan kaldırmaya yönelik puanlayıcılara puanlama eğitimlerinin verildiği çalışmaların olduğu görülmüştür. Verilen puanlayıcı eğitimlerinde problemli puanlama davranışlarının azalmasına yönelik katkı sunduğu yapılan çalışmaların bulgularıdır.

Müzik Performansının Ölçülmesi İle İlgili Çok Yüzeyle Rasch Ölçme Modeli Kullanarak Yapılan Çalışmalar

Atılğan, (2005) çalışmada, 2002 ve 2003 eğitim öğretim yılında İnönü üniversitesinde yapılan müzik öğretmenliği özel yetenek seçme sınavları verileri kullanılarak ÇYRÖM'e

göre analiz edilmiştir. Birey veya adayın, yetenek ölçüleri, puanlayıcıların katılık /cömertlik, görev güçlük düzeyleri ve uygunluk istatistikleri ile puanlayıcı birey, görev birey ve puanlayıcı görev etkileşimleri arasında yanlılıklar araştırılmıştır. Deşifre, çalma ve söyleme olmak üzere üç boyutta, adayların birbirinden bağımsız olarak değerlendirme yapan müzik öğretmeni alanından dört öğretim elemanı ile özel yetenek seçme sınavından elde edilen veriler kullanılmıştır. Araştırma sonucuna göre değerlendiricilerin tüm adaylar için puanlama davranışlarının birbirine yakın fakat manidar düzeyde farklı olduğu 10 aday ve bir görevin uygun olmadığı puanlayıcı-birey, görev-birey ve puanlayıcı görev etkileşimleri incelendiğinde yanlı davranışlar tespit edilmiştir.

Öztürk Gübeş (2021) çalışmasında bir devlet üniversitesinde 2018-2020 eğitim öğretim yılında uygulanan öğrenci alımındaki müzik yetenek sınavlarının izlenmesinde kalite kontrolünün sağlanması için elde edilen verileri çok yüzeyli Rasch ölçme modeline göre incelemiştir. Bu çalışmada müziksel şarkı söyleme ve çalma sınavlarının toplam puanları kullanılmıştır. Çalışmanın katılımcıları 164 aday ve adayların performanslarını değerlendiren beş uzman grubundan ve değerlendirmede kullanılan iki görev olacak şekilde araştırmada üç Yüzeyli model kullanılmıştır. Yüzeylerin tamamı modele uyum sağlamış öğrenciler için ayırma indeksi güvenilirliği çok yüksek bulunmuş. Yapılan müzik yetenek sınavının performans açısından öğrencileri farklılaştırdığı görülmüştür. Puanlayıcıların ayırma indeksleri sonucunda ise güvenilirlik sıfır (0,00) olarak bulunmuş ve puanlayıcıların adayların müzik performanslarını çok benzer katılık ve cömertlik seviyelerde değerlendirmişler. Göreve ilişkin bulgular, müzik amaçlı şarkı söyleme görevlerinin müzik aletleri çalmayı içeren görevlerden daha zor olduğunu orta koymuştur. Yanlılık çalışma sonuçları ise, göreve göre puanlayıcılara veya öğrenci etkileşimlerine göre puanlayıcıya bağlı bir yanlılık olmadığı ancak göreve ilişkin öğrenci etkileşimlerinin bazı yanlı ölçümlerin olduğu bulunmuştur.

Akın ve Baştürk (2012) çalışmasında Güzel Sanatlar ve Spor Lisesi müzik öğretmenliği alanında eğitim gören öğretmen adaylarının keman eğitimi ile ilgili kazanmış oldukları temel becerileri puanlayıcıların katılık/cömertlik gibi davranış sergileme güçlüğüne göre ÇYRÖM ile değerlendirmiştir. Araştırmanın örneklemi 3 farklı Anadolu Güzel Sanatlar Lisesinde “Keman eğitimi” dersine devam eden 27 öğrenciden oluşmuştur. Değerlendirmede ölçme aracı olarak “Keman Çalma Becerilerini Değerlendirme Ölçeği” kullanılmıştır. Araştırma sonucunda müzik eğitiminde özellikle performansa dayalı öğrenmelerin değerlendirilmesinde ÇYRÖM etkin olarak kullanılabileceği sonucu bulunmuştur.

Wesolowski (2015) araştırmış olduğu makalesinde müzik performans değerlendirmesinde puanlayıcı adaletini model veri uyumu ve farklılaşan puanlayıcı davranışlarını Çok Yüzeyle Rasch kısmi kredi ölçüm modeli kullanarak incelemiştir. Model kapsamına uzman puanlayıcıların (n=24) dört okul düzeyindeki bireylerin (ortaokul, lise, üniversite ve profesyonel müzisyen) değerlendirmede katılık davranışlarının değişim gösterip göstermediğini araştırmıştır. Etkileşim analizlerine göre, farklı puanlayıcı davranışlarının hem puanlayıcı grubunda hem de bireysel puanlayıcılar için logit ölçeklerinde beklenen konumlara göre ortaya çıktığı bulunmuştur. Yani uzman puanlayıcılar, dört okul seviyesindeki grupları değerlendirirken her zaman tutarlı olarak katı davranış sergilemedikleri ortaya çıkmıştır. Etkileşim analiz sonucuna göre 92 etkileşimden 14'ünü (%15,2) istatistiksel açıdan anlamı olduğu bulunmuştur. Bu sonuç puanlayıcılardan 10'unun en az bir okul düzeyinde farklı katılık eyleminde bulunduğunu göstermiştir.

Wesolowski (2016) yapmış olduğu çalışmada caz büyük orkestra performansını değerlendirmek için üç başarı düzeyinde performans gösteren öğrenci (düşük n=30, orta n=49, yüksek n=23) grubunu dikkate alarak ölçek geliştirmiştir. Geliştirilen Caz Büyük Orkestrası Performansı Değerlendirme ölçeği 18 madde ve dört boyuttan oluşup üç performans başarıları düzeyinde değerlendirilerek karşılaştırma yapılmıştır. Dört boyutun açıklanan varyans oranı %63,32 ve güvenirlik değeri ise 0.84 alfa olarak hesaplanmıştır. Ayırma analizi sonucunda dört boyutlu ölçek maddelerinin düşük ve orta düzeyde performans gösteren öğrencilerin gruplara %88,5 doğrulukta ayırt etmiştir.

Wesolowski (2019) çalışmasında Çok Yüzeyle Rasch Derecelendirici Doğruluğu (MFR-RA) ölçüm modeli kullanılarak ikincil düzey solo müzik performansı değerlendirmesi bağlamında puanlama kalitesi incelenmiştir. Buna göre araştırma bağlamında iki soruya cevap aranmıştır: puanlayıcı kararlarının değerlendirme sürecindeki doğruluğu; puanlayıcı yargılarının ölçme aracındaki ölçütlerin her biri için verilen puanların doğruluğu ile ilgilidir. Elde edilen sonuçlar doğruluk puanlarının genellikle Çok Yönlü Rasch Derecelendirici Doğruluğu modelinin beklentileriyle uyumlu olduğu sonucunu göstermiştir. Bu sonuca ek olarak puanlayıcılar diğer yüzeylere göre ölçüt ve öğrenci performanslarından daha iyi konumda olduğu bulunmuştur. Diğer bir ifade ile puanlayıcılar öğrenci performanslarını puanlarken doğru puanlama davranışları sergilediklerini göstermiştir. Puanlayıcı yüzeyindeki en doğru puanlama yapan puanlayıcı için 0,54 logit (SE=0,05) ile en az doğru puanlayıcı için 0,24 logit (SE=0,04) arasında değişim göstermiş. Ölçme aracındaki genelinin kullanımına ilişkin puanlayıcı doğruluğunun zorluğu 0,91 logit (SE=0,08) ile -1,83 logit

(SE=0,17) arasında deęiřmiřtir. Puanlayıcının ölçme aracındaki boyutları kullanarak puanlama yapma doęruluęu 0,25 logit (SE=0,08) ile -0,68 logit (SE=0,17) řeklinde bulunmuřtur.

Arařtırmanın asıl odaęı olan mőzik performansının ölçőlmesi ile ilgili hem yurt içi hem de yurt dıřındaki ulařılabilen kaynaklar incelendięinde; mőzik yetenek sınavlarındaki aday yeteneęi, ölçme aracı ve puanlayıcı davranıřlarının incelenme ve mőzik alanındaki řalgılara yönelik ölçme araçlarının geliřtirmesini kapsayan řalıřmaların olduęu görőlmőřtür. Dolaysıya eęitim veya sosyal bilimlerdeki gibi ayrıntılı olarak puanlayıcı davranıřlarını azaltmaya yönelik puanlayıcı eęitimlerinin verildięi řalıřmalara rastlanmamıřtır.



BÖLÜM III

YÖNTEM

Bu bölümde; araştırmanın türü ve deseni, çalışma grubu, verilerin toplanması ve veri toplama aracı ile verilerin analizine yer verilmiştir.

Araştırma Deseni

Bu çalışmada; “Referans çerçevesi ile Puanlayıcı Hatası” ve “Referans çerçevesi ile Performans Boyutu” puanlayıcı eğitimleri kullanılarak verilen puanlayıcı eğitiminin; puanlayıcı davranışları üzerine etkisi incelenmiştir. Bir diğer ifade ile deney gruplarına iki farklı puanlama eğitiminin puanlayıcıların puanlama davranışları üzerine etkisi ele alındığından deneysel araştırma söz konusudur. Çalışmada bireyler grupları pratik ve etik nedenlerden dolayı tesadüfi olarak atama işlemi gerçekleştirilemediğinden ve öğretim görevlileri ile öğretim üyelerine verilen puanlayıcı eğitiminin öğrenci performanslarının puanlanmasındaki puanlayıcı davranışları üzerine etkisi araştırıldığı için yarı deneysel desen türünden eşdeğer olmayan grup desenlerinden; öntest- son test eşdeğer olmayan grup eseni ile tasarlanmıştır. Bu desen türüne göre, çalışmadaki katılımcılar koşullara rastgele atanmadığı zaman kullanılan denekler arası desen türüdür (İlhan & Gezer, 2023, s. 120). Bu desene göre araştırmacı araştırma öncesinde mevcut olan grupları araştırmanın içeriğine ve amacına göre iki deney ve bir kontrol gruplarında işleme tabi tutulacağını belirlemiştir. Deneysel desenin doğası gereği üzerinde uygulama işlemi yapılacak olan ön test ve son testi alan grupların puanları arasında etkisi incelenen değişkenin sonucunda anlamlı bir farkın çıkması yürütülen deneysel işlem sonucuna işaret eder (Büyüköztürk, 2021; Campbell & Stanley, 1963). Çalışmanın deneysel süreci ile ilgili bilgiler Tablo 4’te gösterilmiştir.

Tablo 4

Yarı Deneysel Süreç

Grup	Ön Uygulama	Deneysel Süreç (Çevrimiçi Eğitim)	Son Uygulama
Deney A (n=10)	X ₁	Referans Çerçevesi	Y ₁
Deney B (n=7)	X ₂	Puanlayıcı Hatası	Y ₂
Kontrol (n=8)	X ₃	Referans Çerçevesi Performans Boyutu	Y ₃
		X	

Tablo 4'te görüldüğü üzere başlangıç düzey piyano performansına ilişkin öğrencilerin performanslarını puanlamak için iki deney grubu, bir kontrol grubu olmak üzere üç grupta toplam 25 öğretim elemanları yer almaktadır.

Çalışma Grubu

Araştırmanın deney grubunda, başlangıç düzey piyano performanslarını sergileyen öğrenci grubu ile öğrenci performanslarını değerlendiren puanlayıcı grubundan oluşmaktadır.

Bu araştırmadaki deney grubundaki öğrenci grubu Gazi Üniversitesindeki Güzel Sanatlar Eğitim Bölümü Müzik Anabilim dalının 2023-2024 eğitim öğretim yılının bahar döneminde eğitim gören ve araştırmaya gönüllü olarak katılım sağlayan birinci sınıf (n=10) öğrencileridir.

Çalışmanın asıl grubunu, puanlayıcı eğitimlerinin uygulandığı deney grupları ile kontrol grubu olmak üzere üç grup oluşturmaktadır. Deney A, B ve Kontrol gruplarındaki öğrenci performanslarını değerlendiren puanlayıcılar ise (n=17); Gazi Üniversitesindeki Güzel Sanatlar Eğitim Bölümü Müzik Anabilim dalı ile öğretim görevlileri ile öğretim üyelerinden oluşmaktadır. Kontrol grubundaki puanlayıcılar (n=8) ise Türkiye'deki üç farklı üniversitedeki Güzel Sanatlar Eğitim bölümünün müzik bölümündeki öğretim görevlilerinden oluşmaktadır.

Araştırmanın öğrenci grubunda herhangi bir kayıp veya deneysel süreçten çekilme gibi durum söz konusu değildir. Puanlayıcı grubunda ise; ilk uygulamaya toplamda 30 değerlendirici katılımı ile başlanmıştır. Sonrasında ise beş değerlendirici sağlık ve iş yükünün çok olması gibi durumlardan dolayı araştırmaya katılım göstermemişlerdir. Her iki uygulamaya iki deney ve kontrol gruplarından toplamda 25 puanlayıcı (deney-A=10; deney-B=7; kontrol=8) katılım sağlamıştır. Puanlayıcıların araştırmanın deneysel sürecinde yer almasının ön şartı olarak daha önce puanlayıcı eğitimini görmemiş olmalarına ve puanlayıcı

grubunun heterojen olmasına dikkat edilmiştir. Diğer bir ifade ile her grupta hem deneyimli hem de az deneyimli olan öğretim görevlilerinin yer alması göz önünde bulundurularak çevrim içi olarak deneysel sürece geçilmiştir.

Aşağıda Tablo 5'te araştırmanın çalışma grubunda yer alan katılımcı puanlayıcıların demografik bilgileri yer almaktadır.

Tablo 5

Çalışma Grubundaki Katılımcıların Demografik Bilgileri

Puanlayıcı	Cinsiyet	Kıdem (yıl)	Unvan	Alan
P1	Erkek	3	Arş. Gör.	Piyano
P2	Erkek	20	Dr. Öğr. Üyesi	Viyola
P3	Kadın	2	YL	Piyano
P4	Kadın	14	Dr. Öğr. Üyesi	Flüt, piyano
P5	Erkek	4	Arş. Gör.	Viyolonsel
P6	Kadın	14	Arş. Gör.	Viyolonsel
P7	Erkek	16	YL	Viyolonsel
P8	Erkek	3	Arş. Gör.	Flüt
P9	Erkek	20	Dr. Öğr. Üyesi	Gitar
P10	Erkek	2	Öğr. Gör	Çello
P11	Erkek	14	Doç. Dr.	Viyola
P12	Kadın	13	Doç. Dr.	Piyano,
P13	Kadın	7	Dr Öğrencisi	Piyano, viyola
P14	Kadın	7	YL	Piyano
P15	Kadın	18	Doç. Dr.	Keman
P16	Kadın	13	Dr. Öğr. Üyesi	Viyola-Piyano
P17	Kadın	20	Doç. Dr.	Piyano
P18	Kadın	18	Doç. Dr.	Piyano
P19	Kadın	12	Arş. Gör. Dr.	Piyano
P20	Kadın	10	Doç. Dr.	Piyano
P21	Erkek	11	Dr.Öğr. Üyesi	Flüt
P22	Erkek	10	Dr. Öğr. Üyesi	Flüt
P23	Kadın	9	Dr.Öğr. Üyesi	Piyano
P24	Erkek	31	Prof. Dr.	Viyola
P25	Kadın	20	Prof. Dr.	Piyano

Tablo 5 incelendiğinde araştırmaya Deney A, Deney B ve Kontrol grupları olmak üzere 3 grupta toplamda 25 puanlayıcı katılmıştır. Puanlayıcılara öğrenci performansını değerlendirmekle beraber demografik bilgilerine ilişkin 4 sorudan oluşan anket uygulanmıştır. Buna göre puanlayıcıların cinsiyet, kıdem, unvan, alan gibi sorular sorulmuştur.

Dolayısıyla birinci değişken olan puanlayıcıların cinsiyet açısından incelendiğinde 11'i erkek, 14'ü ise kadın puanlayıcılardan oluşmuştur. İkinci değişken olan kıdem değişkenine göre puanlayıcıların kıdemi 3 -31 yıl arasında değişim göstermektedir. Üçüncü değişken unvana

göre, puanlayıcılar hem deneyimli ve deneyimsiz olmak üzere YL, doktora, araştırma görevlisi, doçent doktor ve profesörlerden oluşmaktadır. Dördüncü değişken olan puanlayıcıların çalgı alanı sorulduğunda puanlayıcıların 12'si piyano, 7'si viyola, 4'ü flüt ve diğer çalgıları çalan puanlayıcılardan oluşmuştur.

Veri Toplama Araçları

Bu araştırmanın veri toplama araçları müzik performans görevlerine ait performanslar (etüt ve eser), performansların puanlanmasında kullanılmak üzere geliştirilen analitik dereceli puanlama anahtarı ve puanlayıcıların kişisel bilgilerine yönelik formdan oluşmaktadır.

Müzik bölümü birinci sınıf öğrencilerinin başlangıç düzey piyano performanslarının ölçülmesi amacıyla bir etüt (czerny 12) ve bir eser (Alman Dansı) performans olarak seçilmiştir. Bu performansların seçilmesinde hem alan uzmanlarının görüşlerine dayanarak hem de öğrencilerin birinci sınıf olmaları ve düzeylerinin farklı olması, eğitime daha yeni başlamaları gibi durumlar göz önünde bulundurularak müzik bölümünde okuyan her öğrencinin çalabileceği başlangıç seviyesindeki performansların/ parçanın seçilmesine karar verilmiştir. Seçilmiş olan parçalar ölçme aracındaki kritik davranışları temsil eder düzeydedir. Öğrencilere piyano çalma görevinin dönemle ilgili herhangi bir not verme amaçlı değil sadece akademik araştırma amacı ile kullanılacağı gönüllülük esasının ön planda olacağı ve kendilerinin video kayıtları ile kişisel bilgilerinin hiçbir yerde paylaşılmayacağı duyurulmuştur.

Öğrencilerden başlangıç düzey piyano çalgısına yönelik belirlenen performansları sergilemeleri için veriler tek tek, yüz yüze toplanmıştır. Veri toplamadan önce araştırmacı araştırmaya gönüllü olarak katılan öğrencilere araştırmanın etik izin onayından ve çalışmanın içeriğinde bahsetmiştir. Ardından alan uzmanı ile ilgili bölümdeki piyano odasına geçilerek öğrencinin piyanoya oturup uygun pozisyona geçmesi ve eserin notaları ile ön hazırlıklar (fakültedeki piyano aletlerinin eski olmasından dolayı) yapma işlemi her öğrenci için beş dakikalık bir sürede tamamlanmıştır. Öğrenci kendisini hazır hissettiğinde öncelikle etüt czerny 12'yi, sonrasında ise ikinci performans olan Alman Dansı toplamda her iki performans için ikişer dakikada çalmıştır. Buna göre araştırmaya gönüllü olarak katılan on öğrencinin ön hazırlık ile performansları icra etmeleri öğrenci başına yedi dakika olup toplamda bütün öğrenciler için 70 dakikalık süreyi kapsamıştır. Başlangıç düzey piyano için toplamda birinci sınıf öğrencilerinden 10 öğrenci araştırmaya katılım göstermiştir.

Öğrencilerinden elde edilen video kayıtları 2023-2024 eğitim öğretim yılının bahar döneminde toplanmıştır. Toplanan veriler araştırmacı tarafından öğrencilerin yüzleri blurlaştırılmış ve her birinin video kayıtları ayrı ayrı dosyalarda kodlanarak muhafaza edilmektedir.

Analitik Dereceli Puanlama Anahtarının Geliştirilmesi

Günümüzde gerek üniversitedeki öğrenciler gerek daha alt düzeylerdeki öğrenciler aldığı derslerde ya da tabi tutulduğu sınavların değerlendirilmesinde, “Değerlendirmede kendilerinden istenilen ölçütün ne olduğu, puanlamalar nasıl yapıldığı?” gibi birtakım sorulara açıklık getirilerek yanıtlanmasını istemektedirler. Bundan dolayı, öğretmenler veya öğretim görevlileri çeşitli değerlendirme biçimleri kullanarak öğrencilerin öğrenmeye ilişkin deneyimlerinin iyileştirmek için giderek baskı altındalar. Bu baskıdan kurtulmanın yolu olarak da analitik veya bütünsel yaklaşımla (görevlerin genel veya göreve özel) hazırlanan dereceli puanlama anahtarını kullanmaktır (Cox vd., 2015). Bu araç alan yazında dereceli puanlama anahtarı (rubric) adı altında yaygın olarak kullanılmaktadır.

Bu çalışmada da performansların puanlanma sürecinde kullanılmak üzere analitik türden bir dereceli puanlama anahtarı geliştirilmiştir. Ölçme aracını geliştirmeden önce alanyazın taranmış; Dünya çapında müzik performansının değerlendirilmesinde sınavlar uygulayan ABRSM (Associated Board of the Royal Schools of Music) kurumunun derecelendirme ölçütlerine ve araçlarına bakılmıştır. Müzik alanındaki piyano çalgısı üzerinde çalışan uzman akademisyenlerle görüşmeler yapılmıştır.

Müzik bölümü birinci sınıf öğrencilerinin başlangıç düzeyi piyano çalgısına yönelik performanslarının ölçülmesi ile ilgili araştırmacı, iki alan uzmanı ve ölçme ve değerlendirme uzmanı bir kişi ile birlikte taslak olarak bir dereceli puanlama anahtarı (DPA) geliştirilmiştir. DPA'nın geliştirilme sürecinde aşağıdaki adımlar izlenmiştir (Goodrich, 1997; Karakaya & Fişne, 2023; Kutlu, Doğan & Karakaya, 2023);

- (1) İlgili yapının ölçülmesine ilişkin (bilgi, beceri veya davranışın) ölçülme amacının belirlenmesi,
- (2) İlgili yapının ölçülmesi amacıyla yapıyı temsil eden uygun ölçütlerin belirlenerek taslağın oluşturulması,

- (3) Ölçütlerin değerlendirilmesinde uygun puanlama anahtarının seçimine ilişkin kararın verilmesi,
- (4) Değerlendirme işlemi öncesinde öğrencilere hangi ölçütler kapsamında değerlendirilecekleri ile ilgili ölçütlerin öğrencilerle veya bireylere paylaşılması,
- (5) Yönergeye göre ön uygulamanın yapılması (uygun puanlama anahtarına göre puanlanması),
- (6) Ölçme aracının uzmanlar tarafından (alan, dil, ölçme) anlam, içerik, yazım kuralları vb. gibi incelemelerden geçirilmesi,
- (7) Ölçme aracının nihai formunun oluşturulmasıdır.

Bu çalışma kapsamında öğrencilerin başlangıç düzeyi piyano performansına ilişkin becerilerinin değerlendirilmesi amaçlandığından konuyla ilgili analitik dereceli puanlama anahtarı yukarıdaki ölçütler göz önüne alınarak tasarlanmıştır. Buna göre üzerinde ölçme işlemi yapılacak olan performans: öğrencilerin başlangıç düzeyi piyano çalgısına ilişkin becerilerinin değerlendirilmesidir. Performansın gösterge davranışları belirlenmeden önce performans değerlendirmede dereceli puanlama anahtarının geliştirilmesi ile ilgili (Brookhart, 2013; Kutlu vd. 2023) alan uzmanlarının çalışmaları incelenmiştir. Turgut & Dalkıran, (2021) ise piyano performansı için sekiz ölçütten oluşan DPA geliştirmiştir. Sever ve Dogan (2023) ilkökul öğretmenleri için piyano çalgısına yönelik bir dereceli puanlama anahtarı geliştirme çalışması yapmışlardır. Alanyazın ve uzman görüşleri doğrultusunda belirlenen ölçütler aşağıda yer almaktadır:

1. Duruş oturuş (postür) ve el pozisyonu
2. Notasyon doğruluğu (doğru nota ile seslendirme)
3. Ritmik hız boyutuna uygunluk
4. Hız terimine uygunluk ve sabitlik
5. Parmak numarası kullanımı (parçayı çalarken doğru parmak numaralarını kullanmak)
6. Dinamikleri sergileme (eser içindeki tekniklerin doğru sergilenmesi)
7. Ton veya Tını bütünlüğü
8. Sağ el- sol el koordinasyonu
9. Cümleleme ve müzikal kimliğin yansıtması

10. Teknik ve müzikal bütünlük

Taslak formda toplamda on ölçüt ve beş dereceyi (çok başarılı, başarılı, orta, zayıf, çok zayıf) kapsayacak şekilde geliştirilmiştir. Taslak olarak geliştirilen DPA, alan uzmanları ile ölçme uzmanlarının görüşlerine sunulmuş, her bir ölçüte ilişkin görüşler; gerekli, düzeltilmeli ve gereksiz olacak şekilde alınmıştır. Uzman görüşlerinin Kapsam Geçerlik İndeksleri ve Kapsam Geçerlik Oranları (KGO) hesaplanarak her bir ölçüte ilişkin hesaplamalar yapılmıştır. Kapsam Geçerlik Oranlarının belirlenmesinde Wilson, Pan ve Schumsky, (2012)'de önerdiği (EK 4) eşik değerler ölçüt olarak dikkate alınmıştır. Tablo 6'da ise uzman görüşleri doğrultusunda elde edilen KGO ve KGİ verilmiştir.

Tablo 6

Ölçme Aracının Taslak Formunda Bulunan Kriterlere İlişkin Hesaplanan KGO Değerleri

Ölçütler	Gerekli	Düzeltilmeli	Gereksiz	KGO
Duruş-Oturuş (Postür) ve El Pozisyonu	16	0	0	1
Notasyon Doğruluğu	16	0	0	1
Ritmik Boyut İfadesi	16	0	0	1
Hız Terimine Uygunluk ve Sabitlik	16	0	0	1
Parmak Numarası Kullanımı	16	0	0	1
Artikülasyon ve Nüansları Sergileme (Seslendirme)	13	3	0	0,63
Ton / Tını Bütünlüğü (Balans)	14	2	0	0,75
Sağ el / Sol el Koordinasyonu ve Senkronizasyon	16	0	0	1
Cümleleme ve Müzikal Kimlik Yansıtmı	13	3	0	0,63
Teknik ve Müzikal Bütünlük	15	1	0	0,88
Kapsam geçerlik indeksi (KGİ)				0.89

Not: Başlangıç düzey piyano performansına ilişkin analitik dereceli puanlama anahtarı EK 2'de verilmiştir.

Tablo 6'da görüldüğü gibi başlangıç düzey piyano performansına ilişkin ADPA'daki ölçütlerin kapsam geçerlik oranları n=16; 11 müzik alan uzmanı, 5 ölçme uzmanı ile 1 dil uzmanının görüşleri doğrultusunda hesaplanmıştır. ADPA'daki ölçütlerin kapsam geçerlik oranları hesaplandıktan sonra ölçme aracının tamamı için kapsam geçerlik indeksi (KGİ) incelenir (Lawshe, 1975). Dereceli puanlama anahtarının son haline alınacak maddelerin KGO değerlerinin ortalamaları ile KGİ hesaplanır (Yurdugül, 2005).

$$KGO = \frac{N_G}{N/2} - 1 \quad (1.6)$$

Bu değer eşitlik 1.6'daki gibi Excel programında hesaplanmıştır. Bu formülde KGO kriterin kapsam geçerlik oranını, N_G ilgili kriterin “uygun” olduğunu ifade eden uzman

sayısını, N kriter hakkında görüş bildiren uzman sayısını göstermektedir. KGO, +1 ile -1 arasında değer almaktadır.

Uzman görüşlerine göre her ölçüte KGO hesaplama formülü uygulanmış ve her bir ölçütün KGO değeri belirlenmiştir. Buna göre 1 değerinin altında olan ölçütlerin KGO'ları: Artikülasyon ve Nüansları Sergileme (Seslendirme), Cümleleme ve Müzikal Kimlik Yansıtmı, Ton, Tını Bütünlüğü (Balans), Teknik ve Müzikal Bütünlük şeklinde bulunmuştur. Piyano başlangıç düzeyini bir bütün olarak ele aldığımızda on boyutun toplamına ilişkin KGİ 0,89 yüksek değer olarak bulunmuştur. Dereceli puanlama anahtarının son hali (EK-2'de) verilmiştir.

Verilerin Toplanması

Bu çalışmadaki verilerin toplanması iki aşamadan oluşmuştur.

İlk aşamada Müzik Anabilim dalında eğitim gören 10 öğrencinin performansları (czerny 12, Etüt) olmak üzere iki tane eser video kaydına alınmıştır. Video kaydına alınmadan önce etik kurul izin belgesi öğrencilere gösterilerek gönüllülük esas alınmıştır. Gönüllü olan on öğrenciye çalışmanın içeriğı tanıtılmış ve etik ilkeler gereğı kayıt altına alınan öğrenci performansları sadece bilimsel çalışma için kullanılacağına da değinilmiştir. Her bir öğrenciden kayıt altına alınan videonun uzunluğu hem czerny 12. etüdü, hem de Alman Dansı için her biri birer dakikalık uzunlukta olup toplamda her öğrenci iki dakikalık performans sergilemiştir. Kayıt altına alınan her bir öğrencinin yüzleri blurlaştırma işleminden geçtikten sonra puanlayıcılara ön ve son uygulama veya puanlama için izletilmiştir.

İkinci aşamada ise yine aynı şekilde Müzik Anabilim dalında eğitim öğretim elemanı olarak çalışan deney A ve B gruplarındaki on yedi akademisyenden puanlayıcı olacak şekilde birinci aşamadaki öğrencilerden elde edilen video kayıtlarından bir tanesi Alman Dansı asıl uygulama başlamadan önce deneme amaçlı izletilmiş ve çevrim içi Zoom, Google Meets gibi ortamlarda izletilerek araştırmacı tarafından oluşturulan puanlayıcı formunda yer alan on ölçüt dikkate alınarak ilk uygulamaya ilişkin puanlamalarını yapmaları istenmiştir. İlk uygulama yaklaşık olarak 60 dakikada yapılmıştır. İlk puanlama işlemi çevrim içinde gerçekleştirildiğinden dolayı ve internet sıkıntısı gibi araştırmacıların kontrol edemediğı durumlarda video iki defa izletilmiştir.

Puanlayıcı Eğitimi

Alan yazında performans değerlendirmelerine yönelik en sık eleştirilen hususlardan bir de verilen görevin nesnel puanlanmaması ile bu sürecin karışık olmasıdır. Bu karışıklığı ortadan kaldırabilmenin tek yolu amaca hizmet edecek şekilde geliştirilen ölçme aracı ile bu işin puanlanmasında görev alacak puanlayıcıların nesnel puanlamaya ilişkin puanlayıcı davranışlarının objektif olması ve yanlılık davranışı gibi hatalı davranışlardan arınık olmasıdır. Bunun içinde puanlayıcıların puanlama davranışlarının ayrıntılı olarak incelenmesi ve takip edilmesi gerekmektedir.

Performans değerlendirme sürecinde görev alan puanlayıcı veya değerlendiricilerin nesnel puanlama davranışlarını geliştirmek için en sık kullanılan altı eğitim modeli mevcuttur. Bunlar öz liderlik, davranış gözlem, performans boyutu, puanlayıcı hatası, referans çerçevesi, puanlayıcı değişkenliği eğitim modelleridir.

Bu çalışmada Deney A grubu için “referans çerçevesi eğitimi ve puanlayıcı hatası eğitimi” ve Deney B grubu için “referans çerçevesi ile performans boyutu eğitimi” eğitimleri oluşturulmuştur. Kontrol grubu için herhangi bir eğitim kullanılmamıştır. Her iki grup için oluşturulan puanlayıcı eğitimleri aşağıda Tablo 7 ve Tablo 8’de verilmiştir.

Deney A ve Deney B grubundaki puanlayıcılara ayrı ayrı olmak üzere toplamda dört hafta (12 saat) eğitimler verilmiştir. Eğitimin öncesinde Deney A, Deney B ve Kontrol grubu için WhatsApp uygulaması ve e-mail ile ortak bir grup kurulmuştur. Uygulama öncesinde çalışmanın genel amaçları katılımcılarla paylaşılmış ve gönüllü olanlar çalışmaya devam etmişlerdir. Araştırmacı tarafından her hafta eğitim başlamadan önce puanlayıcılara oturum linki ile eğitim içeriği paylaşılmıştır. Eğitimler zoom programı kullanılarak yürütülmüştür. Eğitimlerde; araştırmacının kendisi, alan uzmanı bir akademisyen ve bir ölçme değerlendirme uzmanı eğitimci olarak katılmıştır. Eğitimlerin önemli bir kısmını araştırmacının kendisi yürütmüştür. Eğitim sonrası son uygulamalarda çevrimiçi olarak yapılmıştır.

Tablo 7

Deney A Grubu “Referans Çerçevesi ve Puanlayıcı Hatası Eğitim” Eğitimi

1. Gün Ön Uygulanmanın Alınması Süre 60 Dakika Puanlayıcıların Kişisel Bilgi Formunun Doldurulması (Google Forms)					
2. Gün Puanlayıcı Eğitimine Giriş Süre 80 Dakika					
Konu ve Eğitim içeriği	Objektif	Puanlamanın	Hedefler	Materyaller	Yöntem
Eğitimdeki Önemi;			Değerlendiriciler değerlendirici eğitiminin amacını, önemini ve kullanım alanlarını bilirler.	Konu ile ilgili yurt dışı ve yurt içi çalışmalardan örnekler Sunum	Anlatım Soru Cevap
1. Değerlendirici Eğitimi ile İlgili Genel Bilgiler:					
a. Amaç					
b. Önem					
c. Kullanım Alanları					
2. Geniş Ölçekli Sınavlar					
3. Müzik Alanındaki Sınavlar					
a. Özel Yetenek Sınavları ABRSM					
Uygulama-1					
Dereceli Geliştirilmesi ve Kullanımı	Puanlama	Anahtarının	Puanlama sürecinde DPA geliştirilmesi ve kullanımını bilirler.	Müzik alanındaki örnek DPA	Uygulama Soru Cevap Grup Tartışması
Uygulama -2					
Örnek Öğrenci video performanslarının değerlendirilmesi			Puanlama sürecinde DPA geliştirilmesi ve kullanımını bilirler.	Örnek öğrenci video kayıtları	Uygulama Soru Cevap Grup Tartışması
3. Gün Referans Çerçevesi x Puanlayıcı Hatası Eğitimleri Süre 130 Dakika					
Konu ve Eğitim içeriği			Hedefler	Materyaller	Yöntem
Referans Çerçevesi Eğitimi			Değerlendiriciler her bir puanlama boyutunun tanımını ve boyutlara ilişkin hangi davranışların göstergesi olduğunu bilirler.	Konu ile ilgili yurt dışı ve yurt içi çalışmalardan örnekler Sunum	Anlatım Soru Cevap
1. Puanlayıcılara her bir puanlama boyutunun tanımının yapısı					
2. Her bir boyutun hangi davranışların göstergesi olduğunu belirlemek					
1. Sınıflandırma			Değerlendiriciler ilgili performansa ilişkin basit davranışların sıralanmasını ve değerlendirilmesini bilirler.		
2. Ağırlıklandırma					
3. Değerlendirme					
Puanlayıcı Hatası Eğitimi			Değerlendiriciler yaygın olarak yapılan değerlendirici hatalarını bilir ve bu hataların puanlama/değerlendirme üzerindeki olası etkilerini açıklar	Konu ile ilgili yurt dışı ve yurt içi çalışmalardan örnekler Sunum	Anlatım Soru Cevap
1. Alan yazındaki yaygın karşılaşılan hataların açıklanması					
2. Alan yazındaki yaygın karşılaşılan hataların tanımlanması					
3. Olası hata kaynakları					
4. Tipik yanlışlıklar türleri		Puanlama türleri			
Uygulama -3					

Ön	Uygulama	Sonuçlarının	Puanlayıcılar puanlama sürecinde puanlayacağı performansa hâkim olmayı ve puanlamada karşılaşılan yaygın hata davranışlarını bilirler.	FACETS İstatistiksel Program Çıktıları	Anlatım Soru Cevap Grup Tartışması
Deney Sürecine 1 Hafta Ara Verilmesi					
4. Gün Son Uygulamanın Alınması Süre 60 Dakika					
Uygulama 4 Süre 30 Dakika					
Son	Uygulama	Sonuçlarının	Puanlayıcılar puanlama sürecinde puanlayacağı performansa hâkim olmayı ve puanlamada karşılaşılan yaygın hata davranışlarını bilirler.	FACETS İstatistiksel Program Çıktıları	Anlatım Soru Cevap
Puanlayıcılarla Paylaşılması	Paylaşılması ve Ön	Ön			
Uygulama Sonucu ile Karşılaştırılması					

Tablo 8

Deney B Grubu “Referans Çerçevesi ve Performans Boyutu” Eğitimi

1. Gün Ön Uygulanmanın Alınması Süre 60 Dakika				
Puanlayıcıların Kişisel Bilgi Formunun Doldurulması (Google Forms)				
2. Gün Puanlayıcı Eğitimine Giriş Süre 80 Dakika				
Konu ve Eğitim içeriği		Hedefler	Materyaller	Yöntem
Eğitimdeki Objektif Puanlamanın Önemi; Değerlendirici Eğitimi ile İlgili Genel Bilgiler: d. Amaç e. Önem f. Kullanım Alanları Geniş Ölçekli Sınavlar Müzik Alanındaki Sınavlar b. Özel Yetenek Sınavları ABRSM		Değerlendiriciler değerlendirici eğitiminin amacını, önemini ve kullanım alanlarını bilirler.	Konu ile ilgili yurt dışı ve yurt içi çalışmalardan örnekler Sunum	Anlatım Soru Cevap
Uygulama-1				
Dereceli Puanlama Geliştirilmesi ve Kullanımı	Anahtarının	Puanlama sürecinde DPA geliştirilmesi ve kullanımını bilirler.	Müzik alanındaki örnek DPA	Uygulama Soru Cevap
Uygulama -2				
Örnek Öğrenci video performanslarının değerlendirilmesi		Puanlama sürecinde DPA geliştirilmesi ve kullanımını bilirler.	Örnek öğrenci video kayıtları	Uygulama Soru Cevap
3. Gün Referans Çerçevesi x Performans Boyutu Eğitimleri Süre 130 Dakika				
Konu ve Eğitim içeriği		Hedefler	Materyaller	Yöntem
Referans Çerçevesi Eğitimi 1. Puanlayıcılara her bir puanlama boyutunun tanımının yapısı 2. Her bir boyutun hangi davranışların göstergesi olduğunu belirlemek 3. Sınıflandırma 4. Ağırlıklandırma 5. Değerlendirme		Değerlendiriciler her bir puanlama boyutunun tanımını ve boyutlara ilişkin hangi davranışların göstergesi olduğunu bilirler. Değerlendiriciler ilgili performansla ilişkin basit davranışların sıralanmasını ağırlıklandırmasını ve değerlendirmesini bilirler.	Konu ile ilgili yurt dışı ve yurt içi çalışmalardan örnekler Sunum	Anlatım Soru Cevap

Performans Boyutu Eğitim			Değerlendiriciler ilgili	Konu ile ilgili yurt	Anlatım
1. Puanlayıcılar	ilgili		performansa ilişkin	dışı ve yurt içi	Soru
yeterliğin/performansın			basit davranışları	çalışmalardan	Cevap
boyutlarını, göstergeleri veya alt			düzgün sıralayarak	örnekler	
boyutlarını ve esas/temel			değerlendirmeyi	Sunum	
kavramlarını eksiksiz bir şekilde			yapabilir.		
öğretilmesi			Puanlayıcılar ilgili		
2. Puanlayıcılar ilgili performans			performansa ilişkin		
ilişkin örnek davranışların			örnek davranışlar		
açıklanması			söyleyebilir		
Uygulama -3					
Ön	Uygulama	Sonuçlarının	Puanlayıcılar	FACETS	Anlatım
Puanlayıcılarla Paylaşılması			puanlama sürecinde	İstatistiksel	Soru
			puanlayacağı	Program Çıktıları	Cevap
			performansa hâkim		
			olmayı ve puanlamada		
			karşılaşılan yaygın		
			hata davranışlarını		
			bilirler.		
Deney Sürecine 1 Hafta Ara Verilmesi					
4. Gün Son Uygulamanın Alınması Süre 60 Dakika					
Uygulama- 4 Süre 30 Dakika					
Son	Uygulama	Sonuçlarının	Puanlayıcılar	FACETS	Anlatım
Puanlayıcılarla Paylaşılması ve Ön			puanlama sürecinde	İstatistiksel	Soru
Uygulama Sonucu ile Karşılaştırılması			puanlayacağı	Program Çıktıları	Cevap
			performansa hâkim		
			olmayı ve puanlamada		
			karşılaşılan yaygın		
			hata davranışlarını		
			bilirler.		

Verilerin Analizi

Verilerin Çözümlemesi iki alt başlıkta ele alınabilir. Birincisi veri toplama aracının geliştirmesi sürecinde yapılanlar, ikincisi de puanlayıcıların öğrencilerin performanslarına yönelik verdikleri puanların analizleridir.

Araştırmanın problem durumu ve sorularına göre deneysel uygulama sürecine katılan çalışma grubunda yer alan öğretim görevlilerinden yarı deneysel desene göre veriler toplanmıştır. Puanlayıcı eğitimi öncesinde başlangıç düzey piyano performansına yönelik eserin puanlanması sürecinde; birinci sınıf öğrencilerinin piyano başlangıç düzeyine uygun etüt (czerny 12. etüt) performans olarak seçilmiştir. Seçilen performansın video kaydı alındıktan sonra puanlama esnasında objektif puanlamanın sağlanması için çalışma grubunda yer alan 10 öğrencinin isimleri kodlamış ve video kayıtlar araştırmacı tarafından öğrencilerin yüzleri blurlaştırılmıştır. Blurlaştırılan öğrenci videoları Deney A grubunda yer alan 10 puanlayıcıya çevrim içi ortamında izletilerek puanlama işlemi yapılmıştır. Puanlama işlemi araştırmacılar tarafından geliştirilmiş analitik dereceli puanlama anahtarındaki 10 ölçüt dikkate alınarak puanlanmıştır. Devamında puanlayıcılardan elde edilen veriler Çok

Yüzeyli Rasch ölçme modeline göre düzenlenerek analizi gerçekleştirilmiştir. Analiz öncesinde elde edilen her iki testteki verilerin analize uygunluğu (eksik, kayıp veriler) incelenmiştir. Verilerin analizi yapılırken ilk önce ön uygulamaya ait sonuçlar raporlanmış akabinde de son uygulamaya ait bulgular sunulmuştur.

Verilerin çözümlenmesinde FACET 3.70.1 istatistiksel paket programından yararlanılmıştır. Bu çalışmadaki kullanılan FACET programı, öğrencinin veya öğrenci performansının herhangi bir ölçüt referans alınarak puanlayıcı ya da değerlendirici tarafından bir öğrencinin aldığı görev üzerindeki performansının değerlendirmesine olanak sağlayan bir istatistiksel paket programdır (Linacre, 2023).

Üç yüzeyden oluşan ilgili araştırmada yüzeyler sırası ile öğrenci performansı, ölçme aracındaki ölçütler ve puanlayıcılar şeklindedir. ÇYRÖM’ne göre çalışma grubunda yer alan on öğrenci seçilen eseri/etüt çaldığı ve tüm puanlayıcı ölçme aracında bulunan ölçütleri kullanarak on öğrenci performansının hepsini değerlendirdiği için çaprazlanmış desen tercih edilmiştir. Böylece modeldeki tüm parametreler kesin ve doğru ölçülme olasılığı artar (Linacre, 1989, s. 14).

BÖLÜM IV

BULGULAR VE YORUMLAR

Bu bölümde, araştırmanın problem ifadesinden türetilerek belirlenen araştırma sorularına ilişkin veriler analiz edilerek elde edilen bulgular; sırasıyla sunulmuş ve literatürdeki ilgili araştırma sonuçlarıyla yorumlanarak sunulmuştur.

Birinci araştırma sorusuna ilişkin araştırma bulguları

Deney A, Deney B ve Kontrol gruplarından toplanan ön uygulama ve son uygulamaya ilişkin verilerin modele uygunluk analizleri

1. Deney A grubundaki öğretim görevlilerinden oluşan puanlayıcı grubunun ön uygulamadan ve son uygulamadan elde edilen *öğrenci performansı*, *puanlayıcı*, *ölçüt*, *yüzeylerine* ilişkin toplanan verilerin uygunluk değerleri ile model ölçümleri nasıldır? Deney A grubundaki puanlayıcıların ön uygulama *öğrenci performansı* yüzeyine ait ölçüm raporu Tablo 9’da verilmiştir.

Tablo 9

Deney A Grubu Öğrenci Performansı Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu

Öğrenci Performansı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	4,38	4,46	2,57	0,16	0,97	1,01
2	4,24	4,31	2,25	0,15	1,46	1,45
3	4,74	4,79	3,72	0,22	0,92	1,07
4	4,00	4,06	1,78	0,14	0,97	0,97
5	4,08	4,14	1,93	0,14	0,77	0,78
6	4,73	4,79	3,68	0,21	0,88	1,06
7	4,06	4,12	1,89	0,14	0,86	0,94
8	4,01	4,07	1,79	0,14	0,70	0,68
9	3,39	3,40	0,74	0,13	1,23	1,21
10	4,61	4,68	3,21	0,18	0,95	1,05
Ortalama	4,22	4,28	2,36	0,16	0,97	1,02
S (Evren)	0,39	0,40	0,90	0,03	0,21	0,20
S (Örneklem)	0,41	0,42	0,95	0,03	0,22	0,21
Model, Evren	RMSE= 0,16 Düzeltilmiş S.H.= 0,88 Ayırma Oranı= 5,46 Ayırma İndeksi= 7,61 Güvenirlilik= 0,97					
Model, Örneklem	RMSE= 0,16 Düzeltilmiş S.H.= 0,93 Ayırma Oranı= 5,76 Ayırma İndeksi= 8,02 Güvenirlilik= 0,97					
Model, Ki-kare (Sabit etkili):	288,0 sd= 9 p= 0,00					
Model, Ki-kare (Normal):	8,7 sd= 8 p= 0,37					
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi						

ÇYRÖM, grup düzeyindeki istatistiksel analizlerde iki farklı model sonucu üretir. Evren için hesaplanan model, herhangi bir yüzeyin tüm potansiyel unsurlarının dahil edilmesini dikkate alırken, örneklem için hesaplanan model yalnızca belirli bir yüzey için seçilen unsurları dikkate alır (Linacre, 2017). Bir çalışmadaki yüzeylerin tüm bileşenlerine erişim, ekonomik açıdan zorlu olduğundan, genellikle yüzeylerden seçilen örneklem tercih edilmektedir. Bu nedenle, herhangi bir yüzeye ait grup düzeyindeki istatistikler raporlandığında, örneklem için hesaplanan model değerlerinin dikkate alınması, sonuçların geçerliği ve güvenirliliğine katkıda bulunmaktadır.

Deney A grubundaki 10 öğrencinin ön uygulamadaki performans yüzeyinin ölçümlerine ait Tablo 9 incelendiğinde, öğrenci performansı düzeylerinin yorumlanması için referans olarak logit değerinin 0,74’ten başlayıp 3,72’e doğru arttığı görülmektedir.

Öğrenci performans yüzeyine ait logit değerlerinin ortalaması 2,36 iken standart sapması ise 0,95 olarak hesaplanmıştır. Bu hesaplama kapsamında piyano çalarken performansını en iyi derecede icra edebilen kişinin 3’nolu öğrenci olduğu görülmektedir. Benzer şekilde piyano çalma esnasında performansını iyi derecede icra edemeyen ise 9’nolu öğrencidir.

Öğrencilerin performanslarının değerlendirildiği bu yüzeyde uyum içi ve uyum dışı değerleri de incelenmiştir. Bu bağlamda uygunluk içi 0,70 ile 1,46 arasında değişim gösterirken ortalama değeri 0,97 olarak bulunmuştur. Benzer şekilde uygunluk dışı değerleri de incelendiğinde 0,68 ile 1,45 arasında değer alarak ortalama değerinin ise 1,02 şeklinde olduğu hesaplanmıştır. Ön uygulama Deney A grubunda performans gösteren öğrencilerin performans yüzeylerindeki hem uygunluk içi hem de uygunluk dışındaki değerlerin 1'e yakın olması model veri uyumunun iyi olduğuna işaret etmektedir.

Öğrencilerin performanslarının değerlendirildiği ilgili yüzeydeki diğer bir bakılacak olan kısım ayırma oranıdır. Bu oranın 5,76 ve güvenirlik indeks değerinin ise 0,97 olarak hesaplanmıştır. Bu iki değer yüksek çıkması öğrencilerin başlangıç düzey olarak piyano çalma sürecinde öğrencilerin birbirinden farklı yetenek düzeyine sahip olduklarını vurgulamaktadır. Elde edilen bulgulardaki farklılıkların istatistiksel anlamda anlamlı olup olmamasına karar vermek için sabit etkili ki-kare değerine bakılmıştır. İlgili değer istatistiksel anlamda anlamlı olduğu belirlenmiştir ($\chi^2(sd) = 288,0 (9)$, $p = 0,00 < 0,01$). Bu bulgular piyano çalarken performans gösteren öğrencilerin yetenekleri arasında istatistiksel anlamda anlamlı bir farka sahip olduğuna işaret etmektedir.

Deney A grubu ön uygulama öğrenci performansı yüzeyine ait ölçüm sonuçları incelendikten sonra Deney A grubu ön uygulama puanlayıcı yüzeyi ölçüm sonuçları incelenmiştir. Deney A grubu ön uygulama puanlayıcı yüzeyi ölçüm sonuçları Tablo 10'da yer almaktadır.

Tablo 10

Deney A Grubu Puanlayıcı Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu

Puanlayıcılar	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı	t değeri
1	3,78	3,85	-0,93	0,13	0,83	0,85	-1,55
2	4,36	4,48	0,27	0,16	0,61	0,67	0,45
3	4,37	4,49	0,29	0,16	1,14	1,85	0,48
4	4,24	4,36	-0,02	0,15	0,99	0,92	-0,03
5	3,86	3,94	-0,78	0,14	0,94	0,90	-1,3
6	4,06	4,16	-0,40	0,14	0,71	0,69	-0,67
7	4,54	4,65	0,77	0,18	1,26	1,20	1,28
8	4,14	4,25	-0,23	0,14	0,86	0,78	-0,38
9	4,57	4,68	0,86	0,18	0,83	0,84	1,43
10	4,32	4,44	0,17	0,15	1,79	1,50	0,28
Ortalama	4,22	4,33	0,00	0,15	0,99	1,02	
S (Evren)	0,25	0,27	0,57	0,01	0,32	0,37	
S (Örneklem)	0,27	0,28	0,60	0,02	0,34	0,39	
Model, Evren	RMSE= 0,15 Düzeltilmiş S.H.= 0,54 Ayırma Oranı= 3,53 Ayırma İndeksi= 5,04 Güvenirlilik= 0,93						
Model, Örneklem	RMSE= 0,15 Düzeltilmiş S.H.= 0,58 Ayırma Oranı= 3,73 Ayırma İndeksi= 5,04 Güvenirlilik= 0,93						
Model, Ki-kare (Sabit etkili):	135,7	sd= 9	p= 0,00				
Model, Ki-kare (Normal):	8,4,0	sd= 8	p= 0,39				
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi							

Deney A grubundaki puanlayıcıların ön uygulamaya ait ölçümleri Tablo 10 incelendiğinde, puanlayıcılara ait logit, standart hata, uygunluk içi ve dışı değerleri genel olarak istenilen ölçütler aralığında olduğu söylenebilir. Bu değerlere ek olarak *t* değerleri de elle hesaplanarak Tablo 10'da verilmiştir. *t* değerinin elle hesaplanması bir örnek üzerinde şu şekilde açıklanır; ön uygulama Deney A grubunda yer alan birinci puanlayıcının logit ölçüsü olan -0,93'den tabloda bulunan tüm puanlayıcılara ait logit ölçümlerinin ortalaması 0,00'dan çıkarıldığındaki ortaya çıkan farkı yine aynı şekilde tüm puanlayıcıların logit ölçümlerine ait standart hatası olan 0,60'e bölündüğünde ortaya çıkan -1,55 sonucu birinci puanlayıcının *t* değeridir $(-0,93)-0,00/0,60=-1,55$). Diğer puanlayıcıların hesaplanan *t* değerleri ise yukarıdaki Tablo 10'da yer almaktadır.

Deney A grubu ön uygulama puanlayıcı yüzeyine ait ölçüm sonuçları incelendikten sonra Deney A grubu ön uygulama ölçüt yüzeyi ölçüm sonuçları incelenmiştir. Deney A grubu ön uygulama ölçüt yüzeyi ölçüm sonuçları Tablo 11'de yer almaktadır.

Tablo 11

Deney A Grubu Ölçüt Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu

Ölçütler	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	4,44	4,55	0,45	0,16	1,48	1,38
2	4,33	4,44	0,17	0,16	1,19	1,09
3	4,35	4,46	0,22	0,16	1,00	0,92
4	3,95	4,03	-0,63	0,14	0,85	0,90
5	4,73	4,81	1,45	0,22	1,04	1,27
6	3,88	3,95	-0,76	0,14	0,96	1,24
7	4,18	4,28	-0,17	0,15	0,89	0,91
8	4,43	4,54	0,43	0,16	1,19	0,98
9	3,97	4,05	-0,59	0,14	0,77	0,82
10	3,98	4,06	-0,57	0,14	0,69	0,70
Ortalama	4,22	4,32	0,00	0,16	1,01	1,02
S (Evren)	0,26	0,27	0,65	0,02	0,22	0,21
S (Örneklem)	0,28	0,29	0,69	0,02	0,23	0,22
Model, Evren	RMSE= 0,16 Düzeltilmiş S.H.= 0,63 Ayırma Oranı= 4,04 Ayırma İndeksi= 5,71 Güvenirlik= 0,94					
Model, Örneklem	RMSE= 0,16 Düzeltilmiş S.H.= 0,67 Ayırma Oranı= 4,27 Ayırma İndeksi= 6,02 Güvenirlik= 0,95					
Model, Ki-kare (Sabit etkili):	142,6	sd= 9	p= 0,00			
Model, Ki-kare (Normal):	8,4	sd= 8	p= 0,39			

S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi

Tablo 11’de Deney A grubundaki analitik dereceli puanlama anahtarındaki ölçütler yüzeyinin ön uygulama ölçümlerine ait logit değerlerinin -0,63 ile 1,45 aralığında değer aldığı ve bu değerlerin ölçme aracında bulunan ölçütlerin güçlük düzeyleri arasında 2,08’lik bir logit değişim ortaya çıktığı görülmüştür. Ölçütlerin ortalaması 0,00 ve standart sapması da 0,69 logit şeklinde hesaplanmıştır. Bu bağlamda ölçme aracında yer alan on ölçütün öğrencilere hangisinin daha kolay hangisinin ise daha zor olduğu incelenebilir. Deney A grubunda bulunan öğrencilerin ön uygulama sonuçlarına göre yedinci ölçüt çok zor olmuş, en kolay olan ölçüt ise de beşinci ölçüt olarak belirtilmiştir.

Ölçütler yüzeyi için kestirilen uygunluk içi ve uygunluk dışı değerleri incelendiğinde; uyum içi değerlerinin 0,69 ile 1,48 arasında değer aldığı ve ortalama değerinin de 1,01 olarak hesaplanmıştır. Aynı şekilde uyum dışı değerleri de 0,70 ile 1,38 arasında değer almış ve ortalama değeri 1,02 olarak bulunmuştur. Bulunan değerler ışığında ön uygulama Deney A grubundaki öğrencilerin değerlendirilmesi için geliştirilen ölçme aracında yer alan ölçütlerin amaca hizmet edecek düzeyde geliştirildiği ve işlendiğini hem uygunluk içi hem de uygunluk dışı değerlerin 1’e yakın olması model veri uyumunun da iyi olduğuna işaret etmektedir. Ölçüt yüzeyinin ayırma indeks değerinin 6,02 olarak bulunması Nakamura’nın (2002) belirlediği ayırma indeksi sınırının

üzerinde değer aldığı için; iyi derecede ayırıcı, geliştirilen ölçütlerin öğrenciler arasında belirgin farklılıklar oluşturma yeteneğine sahip olduğu ve yetenekler bazında birbirinden farklılık gösteren öğrencilere de uygulanabileceğine dair işaret etmektedir. Ölçüt yüzeyinin ayırma oranı 4,27 ve güvenilirlik değerinin de 0,95 olarak bulunması bu değerlerin farklı yetenek düzeylerinde de güvenilir bir şekilde test edilebileceği ve öğrenci performansları birbirinden ayırt edebileceği sonucuna varılabilir. “Başlangıç düzey Piyano performansına ilişkin geliştirilen analitik dereceli puanlama anahtarında yer alan ölçütlerin zorluk ve kolaylık açısından anlamlı bir farklılık vardır” yani H_1 varlık hipotezinin test edilmesi için ki-kare istatistiği sonucunda çıkan değerler incelenmiştir. İnceleme sonucunda ($\chi^2(sd) = 142,6 (9)$, $p = 0,00 < 0,01$) H_0 yokluk hipotezi reddedilmiştir. Bu sonuç ön testteki kullanılan ölçme aracında yer alan on ölçütün zorluk veya kolaylık düzeyleri arasında istatistiksel açıdan anlamlı bir farklılıklar bulunmaktadır.

Deney A grubu ön uygulama yüzeylerine ait ölçüm sonuçları incelendikten sonra Deney A grubu son uygulama öğrenci performansı yüzeyi ölçüm sonuçları incelenmiştir. Deney A grubu son uygulama öğrenci performansı yüzeyi ölçüm sonuçları Tablo 12’de yer almaktadır.

Tablo 12

Deney A Grubu Öğrenci Performansı Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu

Öğrenci Performansı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	4,57	4,74	2,78	0,19	0,88	0,95
2	4,40	4,60	2,25	0,17	1,14	1,29
3	4,62	4,78	2,69	0,19	0,74	0,71
4	4,16	4,36	1,62	0,16	0,98	0,87
5	4,10	4,30	1,48	0,15	0,96	0,83
6	4,72	4,85	3,38	0,22	0,70	0,61
7	4,12	4,32	1,52	0,15	1,14	1,02
8	4,15	4,35	1,60	0,16	1,06	1,08
9	3,55	3,64	0,26	0,15	1,10	1,04
10	4,64	4,79	3,04	0,20	1,13	1,42
Ortalama	4,30	4,47	2,09	0,17	0,98	0,98
S (Evren)	0,34	0,35	0,91	0,02	0,15	0,23
S (Örneklem)	0,36	0,36	0,96	0,02	0,16	0,25
Model, Evren	RMSE= 0,17 Düzeltilmiş S.H.= 0,90 Ayırma Oranı= 5,13 Ayırma İndeksi= 7,17 Güvenirlik= 0,96					
Model, Örneklem	RMSE= 0,17 Düzeltilmiş S.H.= 0,95 Ayırma Oranı= 5,42 Ayırma İndeksi= 7,55 Güvenirlik= 0,97					
Model, Ki-kare (Sabit etkili):	280,2	sd= 9	p= 0,00			
Model, Ki-kare (Normal):	8,7	sd= 8	p= 0,37			
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi						

Deney A grubundaki on öğrencinin performans yüzeyinin son uygulamadaki ölçümlerine ait Tablo 12 incelendiğinde, öğrenci performansı düzeylerine ait logit değerlerinin 0,26'dan başlayıp 3,38'e doğru arttığı ve ortalaması 2,09 iken standart sapması ise 0,96 olduğu görülmektedir. Bu hesaplama kapsamında piyano çalarken performansını en iyi derecede icra edebilen öğrenci 6'nolu öğrencidir. Benzer şekilde piyano çalma esnasında performansını iyi derecede icra edemeyen ise 9'nolu öğrencidir. Bu bulgular ise ön uygulamadaki bulgularla aynı olduğu görülmektedir. Öğrencilerin performanslarının değerlendirildiği bu yüzeyde uyum içi ve uyum dışı değerleri ise sırasıyla 0,70 ile 1,14 arasında değişim gösterirken ortalamasında 0,98 olarak bulunmuştur. Aynı şekilde uygunluk dışındaki değerler de incelendiğinde 0,61 ile 1,42 arasında değer alarak ortalama değerinin ise 0,98 olarak hesaplanmıştır.

Son uygulama Deney A grubunda performans gösteren öğrencilerin performans yüzeylerindeki hem uygunluk içi hem de uygunluk dışındaki değerlerin 1'e yakın olması model veri uyumunun iyi olduğuna işaret etmektedir. Öğrencilerin performanslarının değerlendirildiği ilgili yüzeydeki diğer bir bakılacak kısım ayırma oranıdır. Bu oran 5,42 ve güvenilirlik indeks değerinin ise 0,97 olarak hesaplanmıştır. Bu iki değer yüksek çıkması öğrencilerin başlangıç düzey olarak piyano çalma sürecinde öğrencilerin birbirinden farklı yetenek düzeyine sahip olduklarını vurgulamaktadır. Elde edilen bulgulardaki farklılıkların istatistiksel anlamda anlamlı olup olmamasına karar vermek için sabit etkili ki-kare değerine bakılmıştır. İlgili değer istatistiksel anlamda anlamlı olduğu belirlenmiştir ($\chi^2(sd) = 280,2$ (9), $p = 0,00 < 0,01$). Bu bulgular piyano çalarken performans gösteren öğrencilerin yetenekleri arasında istatistiksel anlamda anlamlı bir farka sahip olduğuna işaret etmektedir.

Deney A grubu son uygulama öğrenci performansı yüzeyine ait ölçüm sonuçları incelendikten sonra Deney A grubu son uygulama puanlayıcı yüzeyi ölçüm sonuçları incelenmiştir. Deney A grubu son uygulama puanlayıcı yüzeyi ölçüm sonuçları Tablo 13'te yer almaktadır.

Tablo 13

Deney A Grubu Puanlayıcı Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu

Puanlayıcı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı	t-değeri
1	4,77	4,88	1,52	0,24	1,08	1,22	2,14
2	4,82	4,91	1,83	0,26	1,16	0,84	2,58
3	4,29	4,47	-0,19	0,16	1,28	1,21	-0,27
4	4,27	4,45	-0,24	0,16	0,98	0,84	-0,34
5	3,81	3,94	-1,30	0,15	0,63	1,08	-1,83
6	4,09	4,26	-0,68	0,15	1,19	1,01	-0,96
7	3,86	4,00	-1,19	0,15	0,74	0,69	-1,68
8	4,14	4,32	-0,56	0,15	1,18	1,29	-0,79
9	4,36	4,54	0,00	0,17	1,22	1,07	0,00
10	4,62	4,77	0,83	0,20	0,75	0,58	1,17
Ortalama	4,30	4,46	0,00	0,18	1,02	0,98	
S (Evren)	0,33	0,32	1,02	0,04	0,22	0,22	
S (Örneklem)	0,35	0,34	1,07	0,04	0,23	0,24	
Model, Evren	RMSE= 0,18 Düzeltilmiş S.H.= 1,00 Ayırma Oranı= 5,49 Ayırma İndeksi= 7,65 Güvenirlilik= 0,97						
Model, Örneklem	RMSE= 0,18 Düzeltilmiş S.H.= 1,06 Ayırma Oranı= 5,79 Ayırma İndeksi= 8,08 Güvenirlilik= 0,97						
Model, Ki-kare (Sabit etkili):	254,6 sd= 9 p= 0,00						
Model, Ki-kare (Normal):	8,7 sd= 8 p= 0,37						
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi							

Tablo 13 incelendiğinde, son uygulama Deney A grubunda yer alan puanlayıcıların uygunluk içi ve uygunluk dışı değerlerin 0.50 – 1.50 aralığında olduğu anlaşılmaktadır. Ayırma indeksinin ise 7.65 olması ve bu ayırma indeksine ait güvenirlilik değerinin 0.97 çıkması, puanlayıcıların farklı puanlamalar yaptıkları şeklinde yorumlanabilir. Son uygulamadaki ayırma indeksinin ön uygulamaya ait puanlayıcı yüzeyine ait ayırma indeksinden büyük çıkması istenilen durumun tersi olarak ortaya çıkmıştır. Bu durum, verilen eğitimin etkili olmadığı şeklinde yorumlanabileceği gibi son uygulamada puanlama davranışlarını etkileyen kontrol edilemeyen başka bir değişkenin olabileceğini göstermektedir.

Deney A grubu son uygulama puanlayıcı yüzeyine ait ölçüm sonuçları incelendikten sonra Deney A grubu son uygulama ölçüt yüzeyi ölçüm sonuçları incelenmiştir. Deney A grubu son uygulama ölçüt yüzeyi ölçüm sonuçları Tablo 14’te yer almaktadır.

Tablo 14

Deney A Grubu Ölçüt Yüzeylerinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu

Ölçütler	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	4,32	4,46	-0,23	0,16	1,70	1,76
2	4,17	4,30	-0,61	0,16	1,41	1,22
3	4,56	4,70	0,51	0,19	0,82	0,93
4	4,11	4,23	-0,75	0,15	1,05	1,24
5	4,97	4,99	3,79	0,59	1,02	0,84
6	4,01	4,11	-0,98	0,15	0,78	0,78
7	4,22	4,35	-0,49	0,16	0,75	0,69
8	4,57	4,71	0,55	0,19	0,92	0,74
9	4,10	4,22	-0,78	0,15	0,88	0,84
10	4,00	4,10	-1,01	0,15	0,75	0,77
Ortalama	4,30	4,42	0,00	0,21	1,01	0,98
S (Evren)	0,29	0,28	1,37	0,13	0,30	0,32
S (Örneklem)	0,31	0,29	1,44	0,14	0,31	0,33
Model, Evren	RMSE= 0,24 Düzeltilmiş S.H.= 1,35 Ayırma Oranı= 5,54 Ayırma İndeksi= 7,72 Güvenirlik= 0,97					
Model, Örneklem	RMSE= 0,24 Düzeltilmiş S.H.= 1,42 Ayırma Oranı= 5,85 Ayırma İndeksi= 8,13 Güvenirlik= 0,97					
Model, Ki-kare (Sabit etkili):	142,3	sd= 9	p= 0,00			
Model, Ki-kare (Normal):	8,0	sd= 8	p= 0,44			
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi						

Son uygulama Deney A grubundaki analitik dereceli puanlama anahtarındaki ölçütler yüzeyinin ölçümlerine ait Tablo 14 incelendiğinde, logit değerlerinin -1,01 ile 3,79 aralığında değer aldığı ve bu değerlerin ölçme aracında bulunan ölçütlerin güçlük düzeyleri arasında 4,80'lik bir logit değişim ortaya çıktığı ve ölçütlerin ortalaması 0,00 ve standart sapması da 1,44 logit olduğu görülmüştür. Deney A grubunda bulunan öğrencilerin son test sonuçlarına göre birinci ölçüt çok zor olmuş, en kolay olan ölçüt ise de beşinci ölçüt olarak belirtilmiştir.

Ölçüt yüzeyi için kestirilen uygunluk içi ve uygunluk dışı değerleri incelendiğinde; uyum içi değerlerinin 0,75 ile 1,70 arasında değer aldığı ve ortalama değeri de 1,01 olarak hesaplanmıştır. Aynı şekilde uyum dışı değerleri de 0,69 ile 1,76 arasında değer almış ve ortalama değeri 0,98 olarak bulunmuştur. Bulunan değerler ışığında son uygulama Deney A grubundaki öğrencilerin değerlendirilmesi için geliştirilen ölçme aracında yer alan ölçütlerin amaca hizmet edecek düzeyde geliştirildiği ve işlendiğini hem uygunluk içi hem de uygunluk dışındaki değerlerin 1'e yakın olması model veri uyumunun da iyi olduğuna işaret etmektedir. Ölçüt yüzeyinin ayırma oranı 5,85 ve güvenirlilik değerinin de 0,97 olarak bulunması bu değerlerin farklı yetenek düzeylerinde de güvenilir bir şekilde test

edilebileceği ve öğrenci performansları birbirinden ayırt edebileceği sonucuna varılabilir. “Piyano başlangıç düzeyi için geliştirilen analitik dereceli puanlama anahtarında yer alan ölçütlerin zorluk ve kolaylık açısından anlamlı bir farklılık vardır” yani H_1 varlık hipotezinin test edilmesi için ki-kare istatistiği sonucunda çıkan değerler incelenmiştir. İnceleme sonucunda ($\chi^2(sd) = 142,3 (9), p = 0,00 < 0,01$) H_0 yokluk hipotezi reddedilmiştir. Bu sonuca göre son testteki kullanılan ölçme aracında yer alan on ölçütün zorluk veya kolaylık düzeyleri arasında istatistiksel açıdan anlamlı bir farklılıklar bulunmaktadır.

Deney A grubu son uygulama yüzeylerine ait ölçüm sonuçları incelendikten sonra Deney B grubu ön uygulama öğrenci performansı yüzeyi ölçüm sonuçları incelenmiştir. Deney B grubu ön uygulama öğrenci performansı yüzeyi ölçüm sonuçları Tablo 15’te yer almaktadır.

Tablo 15.

Deney B Grubu Öğrenci Performansı Yüzeyinin Ön Test Ölçümlerine Ait Ölçüm Raporu

Öğrenci Performansı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	4,31	4,44	2,92	0,22	0,98	1,08
2	3,99	4,11	1,90	0,20	1,09	0,94
3	4,64	4,78	4,28	0,27	0,90	0,64
4	3,89	4,03	1,64	0,19	1,27	1,66
5	3,73	3,90	1,27	0,18	0,72	0,74
6	4,27	4,39	2,77	0,22	0,95	0,70
7	4,10	4,22	2,22	0,21	0,77	0,74
8	3,93	4,06	1,75	0,19	1,30	1,06
9	3,20	3,47	0,26	0,16	0,86	0,88
10	4,49	4,62	3,57	0,24	0,96	0,82
Ortalama	4,05	4,20	2,26	0,21	0,98	0,93
S (Evren)	0,39	0,36	1,11	0,03	0,18	0,28
S (Örneklem)	0,41	0,38	1,17	0,03	0,19	0,30
Model, Evren	RMSE= 0,21 Düzeltilmiş S.H.= 1,09 Ayırma Oranı= 5,19 Ayırma İndeksi= 7,26 Güvenirlilik= 0,96					
Model, Örneklem	RMSE= 0,21 Düzeltilmiş S.H.= 1,15 Ayırma Oranı= 5,48 Ayırma İndeksi= 7,64 Güvenirlilik= 0,97					
Model, Ki-kare (Sabit etkili):	291,2	sd= 9	p= 0,00			
Model, Ki-kare (Normal):	8,7	sd= 8	p= 0,37			
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi						

Ön uygulama Deney B grubundaki on öğrencinin performans yüzeyinin ölçümlerine ait Tablo 15 incelendiğinde, öğrenci performansı düzeylerinin yorumlanması için referans olarak logit değerine bakılacaktır. Tablo 15’teki ilgili değer 0,26’dan başlayıp 4,28’e doğru arttığı ve öğrenci performans yüzeyine ait logit değerlerinin ortalaması 2,26 iken standart sapması ise 1,17 olduğu görülmüştür. Bu hesaplama kapsamında piyano çalarken performansını en iyi derecede icra edebilen öğrenci, 3’nolu öğrencidir. Benzer şekilde

piyano çalma esnasında performansını iyi derecede icra edemeyen ise 9'nolu öğrencidir. Öğrencilerin performanslarının değerlendirildiği bu yüzeyde uyum içi ve uyum dışı değerleri de incelenmiştir. Bu bağlamda uygunluk içi 0,72 ile 1,30 arasında değişim gösterirken ortalama değeri 0,98 olarak bulunmuştur. Aynı şekilde uygunluk dışındaki değerler de incelendiğinde 0,70 ile 1,66 arasında değer alarak ortalama değerinin ise 0,93 olarak hesaplanmıştır. Ön uygulama Deney B grubunda performans gösteren öğrencilerin performansı yüzeylerindeki hem uygunluk içi hem de uygunluk dışındaki değerlerin 1'e yakın olması model veri uyumunun iyi olduğuna işaret etmektedir. Öğrencilerin performanslarının değerlendirildiği ilgili yüzeydeki diğer bir bakılacak kısım ayırma oranıdır. Bu oran 7,26 ve güvenilirlik indeks değerinin ise 0,96 olarak hesaplanmıştır. Bu iki değer yüksek çıkması öğrencilerin başlangıç düzey olarak piyano çalma sürecinde öğrencilerin birbirinden farklı yetenek düzeyine sahip olduklarını vurgulamaktadır. Elde edilen bulgulardaki farklılıkların istatistiksel anlamda anlamlı olup olmamasına karar vermek için sabit etkili ki-kare değerine bakılmıştır. İlgili değerin istatistiksel anlamda anlamlı olduğu belirlenmiştir ($\chi^2(sd) = 291,2 (9), p = 0,00 < 0,01$). Bu bulgular piyano çalarken performans gösteren öğrencilerin yetenekleri arasında istatistiksel anlamda anlamlı bir farka sahip olduğuna işaret etmektedir.

Deney B grubu ön uygulama öğrenci performansı yüzeyine ait ölçüm sonuçları incelendikten sonra Deney B grubu ön uygulama puanlayıcı yüzeyi ölçüm sonuçları incelenmiştir. Deney B grubu ön uygulama puanlayıcı yüzeyi ölçüm sonuçları Tablo 16'da yer almaktadır.

Tablo 16

Deney B Grubu Puanlayıcı Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu

Puanlayıcı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı	t değeri
1	4,03	4,14	-0,26	0,17	0,50	0,45	-0,19
2	3,85	3,99	-0,73	0,16	0,82	0,76	-0,54
3	3,95	4,07	-0,48	0,16	0,67	0,65	-0,35
4	4,27	4,38	0,47	0,18	0,79	0,77	0,34
5	4,22	4,33	0,30	0,18	0,95	1,07	0,22
6	3,31	3,54	-1,85	0,13	1,64	1,70	-1,36
7	4,75	4,86	2,55	0,25	1,34	1,08	1,87
Ortalama	4,05	4,19	0,00	0,18	0,96	0,93	
S (Evren)	0,41	0,37	1,26	0,03	0,37	0,38	
S (Örnekleme)	0,44	0,40	1,36	0,04	0,40	0,41	
Model, Evren RMSE= 0,18 Düzeltilmiş S.H.= 1,24 Ayırma Oranı= 6,96 Ayırma İndeksi= 9,61 Güvenirlilik= 0,98							
Model, Örneklem RMSE= 0,18 Düzeltilmiş S.H.= 1,35 Ayırma Oranı= 7,53 Ayırma İndeksi= 10,37 Güvenirlilik= 0,98							
Model, Ki-kare (Sabit etkili):		298,9	sd= 6	p= 0,00			
Model, Ki-kare (Normal):		5,9	sd= 5	p= 0,32			
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi							

Ön uygulama Deney B grubundaki puanlayıcı yüzeyinin ölçümlerine ait Tablo 16 incelendiğinde, uygunluk içi ve uygunluk dışı değerlere bakıldığında 1 numaralı puanlayıcının uygunluk dışı değerinin 0,50'den düşük olması ve uygunluk içi değerinin de 0,50 olması, 6 numaralı puanlayıcının hem uygunluk içi hem de uygunluk dışı değerleri 1,50 üzeri olması bu kişilerin puanlamalarının sıkıntılı olduğu şeklinde yorumlanabilir. Söz konusu bu kişilere ait logit değerlerinde desteklemektedir.

Deney B grubu ön uygulama puanlayıcı yüzeyine ait ölçüm sonuçları incelendikten sonra Deney B grubu ön uygulama ölçüt yüzeyi ölçüm sonuçları incelenmiştir. Deney B grubu ön uygulama ölçüt yüzeyi ölçüm sonuçları Tablo 17'da yer almaktadır.

Tablo 17

Deney B Grubu Ölçüt Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu

Ölçütler	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	3,81	3,95	-0,85	0,18	1,07	1,14
2	4,23	4,32	0,27	0,21	2,05	2,22
3	4,14	4,23	0,01	0,21	1,26	1,06
4	4,03	4,13	-0,31	0,20	1,16	1,07
5	4,93	4,97	4,27	0,50	0,86	0,65
6	3,81	3,95	-0,85	0,18	0,64	0,56
7	4,01	4,12	-0,35	0,20	0,67	0,61
8	4,09	4,18	-0,16	0,20	0,87	0,70
9	3,69	3,84	-1,14	0,18	0,68	0,63
10	3,80	3,93	-0,89	0,18	0,74	0,61
Ortalama	4,05	4,16	0,00	0,22	1,00	0,93
S (Evren)	0,33	0,31	1,49	0,09	0,41	0,48
S (Örneklem)	0,35	0,32	1,57	0,10	0,43	0,51
Model, Evren	RMSE= 0,24 Düzeltilmiş S.H.= 1,47 Ayırma Oranı= 6,08 Ayırma İndeksi= 8,44 Güvenirlik= 0,97					
Model, Örneklem	RMSE= 0,24 Düzeltilmiş S.H.= 1,55 Ayırma Oranı= 6,41 Ayırma İndeksi= 8,89 Güvenirlik= 0,98					
Model, Ki-kare (Sabit etkili):	140,6	sd= 9	p= 0,00			
Model, Ki-kare (Normal):	8,3	sd= 8	p= 0,40			
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi						

Ön uygulama Deney B grubundaki analitik dereceli puanlama anahtarındaki ölçütler yüzeyinin ölçümlerine ait Tablo 17 incelendiğinde, logit değerlerinin -1,14 ile 4,27 aralığında değer aldığı ve bu değerlerin ölçme aracında bulunan ölçütlerin güçlük düzeyleri arasında 5,41'lik bir logit değişim ortaya çıktığı ve ölçütlerin ortalaması 0,00 ve standart sapması da 1,57 logit olduğu görülmüştür. Bu bağlamda ölçme aracında yer alan on ölçütün öğrencilere hangisinin daha kolay hangisinin ise daha zor olduğu incelenebilir. Deney B grubunda bulunan öğrencilerin ön uygulama sonuçlarına göre dokuzuncu ölçüt çok zor olmuş, en kolay olan ölçüt ise de beşinci ölçüt olarak belirtilmiştir.

Ölçütler yüzeyi için kestirilen uygunluk içi ve uygunluk dışı değerleri incelendiğinde; uyum içi değerlerinin 0,64 ile 2,05 arasında değer aldığı ve ortalama değeri de 1,00 olarak hesaplanmıştır. Aynı şekilde uyum dışı değerleri de 0,56 ile 2,22 arasında değer almış ve ortalama değeri 0,93 olarak bulunmuştur. Bulunan değerler ışığında son uygulama Deney B grubundaki öğrencilerin değerlendirilmesi için geliştirilen ölçme aracında yer alan ölçütlerin amaca hizmet edecek düzeyde geliştirildiği ve işlendiğini hem uygunluk içi hem de uygunluk dışındaki değerlerin 1'e yakın olması model veri uyumunun da iyi olduğuna işaret etmektedir. Ölçüt yüzeyinin ayırma oranı 6,41 ve güvenirlilik değerinin de 0,98 olarak

bulunması bu değerlerin farklı yetenek düzeylerinde de güvenilir bir şekilde test edilebileceği ve öğrenci performansları birbirinden ayırt edebileceği sonucuna varılabilir.

Deney B grubu ön uygulama yüzeylerine ait ölçüm sonuçları incelendikten sonra Deney B grubu son uygulama öğrenci performansı yüzeyi ölçüm sonuçları incelenmiştir. Deney B grubu son uygulama öğrenci performansı yüzeyi ölçüm sonuçları Tablo 18’de yer almaktadır.

Tablo 18

Deney B Grubu Öğrenci Performansı Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu

Öğrenci Performansı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	4, 24	4,25	3,00	0,21	1,20	1,21
2	3,96	3,97	2,14	0,20	0,97	0,96
3	4,24	4,25	3,00	0,21	0,89	0,91
4	3,73	3,74	1,49	0,20	0,91	0,90
5	3,74	3,76	1,53	0,20	0,93	0,93
6	4,23	4,23	2,95	0,21	1,26	1,25
7	3,64	3,66	1,26	0,19	0,74	0,74
8	3,70	3,71	1,42	0,20	1,05	1,04
9	3,03	3,04	-0,22	0,18	1,03	1,02
10	4,13	4,13	2,65	0,21	1,00	1,02
Ortalama	3,86	3,87	1,92	0,20	1,00	1,00
S (Evren)	0,36	0,36	0,98	0,01	0,14	0,14
S (Örneklem)	0,38	0,38	1,03	0,01	0,15	0,15
Model, Evren	RMSE= 0,20 Düzeltilmiş S.H.= 0,96 Ayırma Oranı= 4,75 Ayırma İndeksi= 6,67 Güvenirlilik= 0,96					
Model, Örneklem	RMSE= 0,20 Düzeltilmiş S.H.= 1,01 Ayırma Oranı= 5,02 Ayırma İndeksi= 7,03 Güvenirlilik= 0,96					
Model, Ki-kare (Sabit etkili):	257,8	sd= 9	p= 0,00			
Model, Ki-kare (Normal):	8,7	sd= 8	p= 0,37			
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi						

Son uygulama Deney B grubundaki on öğrencinin performans yüzeyinin ölçümlerine ait Tablo 18 incelendiğinde, öğrenci performansı düzeylerinin yorumlanması için referans olarak logit değerine bakılacaktır. Tablo 18’deki ilgili değer -0,22’den başlayıp 3,00’e doğru arttığı ve öğrenci performans yüzeyine ait logit değerlerinin ortalaması 1,92 iken standart sapması ise 0,98 olduğu belirlenmiştir. Bu hesaplama kapsamında piyano çalarken performansını en iyi derecede icra edebilen iki öğrenci 1 ve 3’ü öğrencilerdir. Benzer şekilde piyano çalma esnasında performansını iyi derecede icra edemeyen ise 9’ü öğrencidir. Öğrencilerin performanslarının değerlendirildiği bu yüzeyde uyum içi ve uyum dışı değerleri de incelenmiştir. Bu bağlamda uygunluk içi 0,74 ile 1,26 arasında değişim

gösterirken ortalama değeri 1,00 olarak bulunmuştur. Aynı şekilde uygunluk dışı değerler de incelendiğinde 0,74 ile 1,25 arasında değer olarak ortalama değerinin ise 1,00 olarak hesaplanmıştır. Son uygulama Deney B grubunda performans gösteren öğrencilerin performansı yüzeylerindeki hem uygunluk içi hem de uygunluk dışındaki değerlerin 1'e yakın olması model veri uyumunun iyi olduğuna işaret etmektedir. Öğrencilerin performanslarının değerlendirildiği ilgili yüzeydeki diğer bir bakılacak kısım ayırma oranıdır. Bu oran 5,02 ve güvenilirlik indeks değerinin ise 0,96 olarak hesaplanmıştır. Bu iki değer yüksek çıkması öğrencilerin başlangıç düzey olarak piyano çalma sürecinde öğrencilerin birbirinden farklı yetenek düzeyine sahip olduklarını vurgulamaktadır.

Deney B grubu son uygulama öğrenci performansı yüzeyine ait ölçüm sonuçları incelendikten sonra Deney B grubu son uygulama puanlayıcı yüzeyi ölçüm sonuçları incelenmiştir. Deney B grubu son uygulama puanlayıcı yüzeyi ölçüm sonuçları Tablo 19'da yer almaktadır.

Tablo 19

Deney B Grubu Puanlayıcı Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu

Puanlayıcı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı	t değeri
1	4,03	4,05	0,46	0,17	0,59	0,60	0,71
2	3,74	3,77	-0,36	0,16	1,38	1,33	-0,55
3	3,99	4,01	0,34	0,17	0,83	0,84	0,52
4	3,85	3,87	-0,06	0,17	0,86	0,86	-0,09
5	4,19	4,21	0,94	0,18	1,23	1,27	1,45
6	3,78	3,80	-0,25	0,17	1,08	1,11	-0,38
7	3,47	3,50	-1,07	0,16	0,97	0,98	-1,64
Ortalama	3,86	3,89	0,00	0,17	0,99	1,00	
S (Evren)	0,22	0,21	0,60	0,01	0,25	0,24	
S (Örnekleme)	0,23	0,23	0,65	0,01	0,27	0,26	
Model, Evren	RMSE= 0,17 Düzeltilmiş S.H.= 0,58 Ayırma Oranı= 3,45 Ayırma İndeksi= 4,93 Güvenirlik= 0,92						
Model, Örnekleme	RMSE= 0,17 Düzeltilmiş S.H.= 0,63 Ayırma Oranı= 3,75 Ayırma İndeksi= 5,33 Güvenirlik= 0,93						
Model, Ki-kare (Sabit etkili):	91,6	sd= 6	p= 0,00				
Model, Ki-kare (Normal):	5,6	sd= 5	p= 0,34				

S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi

Tablo 19 Deney B grubunda yer alan puanlayıcıların son uygulamaya ilişkin uygunluk içi ve uygunluk dışı değerlerin ölçüt değerleri arasında olduğu görülmektedir. Ayırma indeksinin ise 5,33 olması puanlayıcıların görece görece benzer düzeyde puanlama yaptıkları şeklinde yorumlanabilir. Bu değer ise ön uygulamaya ait ayırma indeksinin yarısına inmesi eğitimin etkisi açısından yorumlanabilir.

Deney B grubu son uygulama puanlayıcı yüzeyine ait ölçüm sonuçları incelendikten sonra Deney B grubu son uygulama ölçüt yüzeyi ölçüm sonuçları incelenmiştir. Deney B grubu son uygulama ölçüt yüzeyi ölçüm sonuçları Tablo 20’da yer almaktadır.

Tablo 20

Deney B Grubu Ölçüt Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu

Ölçütler	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	3,71	3,74	-0,42	0,20	1,53	1,51
2	4,14	4,16	0,81	0,21	1,23	1,23
3	4,00	4,02	0,38	0,20	0,75	0,74
4	3,94	3,97	0,21	0,20	0,59	0,60
5	3,94	3,97	0,21	0,20	2,34	2,42
6	3,81	3,84	-0,15	0,20	0,86	0,87
7	3,90	3,92	0,09	0,20	0,56	0,56
8	3,89	3,91	0,05	0,20	0,71	0,68
9	3,67	3,70	-0,54	0,19	0,75	0,74
10	3,63	3,66	-0,65	0,19	0,65	0,63
Ortalama	3,86	3,89	0,00	0,20	1,00	1,00
S (Evren)	0,15	0,15	0,42	0,00	0,53	0,55
S (Örnekleme)	0,16	0,15	0,45	0,00	0,56	0,58
Model, Evren	RMSE= 0,20 Düzeltilmiş S.H.= 0,37 Ayırma Oranı= 1,87 Ayırma İndeksi= 2,82 Güvenirlilik= 0,78					
Model, Örnekleme	RMSE= 0,20 Düzeltilmiş S.H.= 0,40 Ayırma Oranı= 2,00 Ayırma İndeksi= 3,00 Güvenirlilik= 0,80					
Model, Ki-kare (Sabit etkili):	44,6 sd= 9 p= 0,00					
Model, Ki-kare (Normal):	7,5 sd= 8 p= 0,48					
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi						

Son uygulama Deney B grubundaki analitik dereceli puanlama anahtarındaki ölçütler yüzeyinin ölçümlerine ait Tablo 20 incelendiğinde, logit değerlerinin -0,65 ile 0,81 aralığında değer aldığı ve bu değerlerin ölçme aracında bulunan ölçütlerin güçlük düzeyleri arasında 1,46’lık bir logit değişim ortaya çıktığı ve ölçütlerin ortalaması 0,00 ve standart sapması da 0,45 logit olduğu belirlenmiştir. Bu bağlamda ölçme aracında yer alan on ölçütün öğrencilere hangisinin daha kolay hangisinin ise daha zor olduğu incelenebilir. Deney B grubunda bulunan öğrencilerin son uygulama sonuçlarına göre onuncu ölçüt çok zor olmuş, en kolay olan ölçüt ise de ikinci ölçüt olarak belirtilmiştir.

Ölçütler yüzeyi için kestirilen uygunluk içi ve uygunluk dışı değerleri incelendiğinde; uyum içi değerlerinin 0,56 ile 2,34 arasında değer aldığı ve ortalama değeri de 1,00 olarak hesaplanmıştır. Aynı şekilde uyum dışı değerleri de 0,56 ile 2,42 arasında değer almış ve ortalama değeri 1,00 olarak bulunmuştur. Bulunan değerler ışığında son uygulama Deney B grubundaki öğrencilerin değerlendirilmesi için geliştirilen ölçme aracında yer alan ölçütlerin amaca hizmet edecek

düzeyde geliştirildiği ve işlendiğini hem uygunluk içi hem de uygunluk dışındaki değerlerin 1'e yakın olması model veri uyumunun da iyi olduğuna işaret etmektedir. Ölçüt yüzeyinin ayırma oranı 2,00 ve güvenilirlik değerinin de 0,80 olarak bulunması bu değerlerin farklı yetenek düzeylerinde de güvenilir bir şekilde test edilebileceği ve öğrenci performansları birbirinden ayırt edebileceği sonucuna varılabilir. "Piyano başlangıç düzeyi için geliştirilen analitik dereceli puanlama anahtarında yer alan ölçütlerin zorluk ve kolaylık açısından anlamlı bir farklılık vardır" yani H_1 varlık hipotezinin test edilmesi için ki-kare istatistiği sonucunda çıkan değerler incelenmiştir. İnceleme sonucunda ($\chi^2(sd) = 44,6 (9), p = 0,00 < 0,01$) H_0 yokluk hipotezi reddedilmiştir. Bu sonuç ön testteki kullanılan ölçme aracında yer alan on ölçütün zorluk veya kolaylık düzeyleri arasında istatistiksel açıdan anlamlı bir farklılıklar bulunmaktadır.

Deney B grubu son uygulama yüzeylerine ait ölçüm sonuçları incelendikten sonra kontrol grubu ön uygulama öğrenci performansı yüzeyi ölçüm sonuçları incelenmiştir. Kontrol grubu ön uygulama öğrenci performansı yüzeyi ölçüm sonuçları Tablo 21'de yer almaktadır.

Tablo 21

Kontrol Grubu Öğrenci Performansı Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu

Öğrenci Performansı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	4,30	4,43	2,48	0,18	1,02	1,03
2	3,92	4,05	1,63	0,16	0,98	0,93
3	4,49	4,61	3,02	0,20	0,90	0,82
4	3,91	4,04	1,60	0,16	0,92	1,05
5	3,74	3,86	1,26	0,15	0,75	0,89
6	4,03	4,15	1,83	0,16	1,03	0,92
7	3,55	3,66	0,92	0,15	1,26	1,31
8	3,76	3,89	1,31	0,15	1,01	0,97
9	3,09	3,13	0,14	0,14	1,06	1,07
10	4,34	4,46	2,58	0,18	1,07	1,09
Ortalama	3,91	4,03	1,68	1,16	1,00	1,01
S (Evren)	0,39	0,41	0,81	0,02	0,13	0,13
S (Örnekleme)	0,41	0,43	0,85	0,02	0,13	0,14
Model, Evren	RMSE= 0,17 Düzeltilmiş S.H.= 0,79 Ayırma Oranı= 4,79 Ayırma İndeksi= 6,72 Güvenirlik= 0,96					
Model, Örnekleme	RMSE= 0,17 Düzeltilmiş S.H.= 0,84 Ayırma Oranı= 5,06 Ayırma İndeksi= 7,08 Güvenirlik= 0,96					
Model, Ki-kare (Sabit etkili):	234,6 sd= 9 p= 0,00					
Model, Ki-kare (Normal):	8,7 sd= 8 p= 0,37					

S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi

Ön uygulama kontrol grubundaki on öğrencinin performans yüzeyinin ölçümlerine ait Tablo 21 incelendiğinde, öğrenci performansı düzeylerinin yorumlanması için referans olarak logit

değerine bakılacaktır. Tablo 21'deki ilgili değer 0,14'den başlayıp 3,02'e doğru arttığı ve öğrenci performans yüzeyine ait logit değerlerinin ortalaması 1,16 iken standart sapması ise 0,02 olduğu belirlenmiştir. Bu hesaplama kapsamında piyano çalarken performansını en iyi derecede icra edebilen öğrenci 3'nolu öğrencidir. Benzer şekilde piyano çalma esnasında performansını iyi derecede icra edemeyen ise 9'nolu öğrencidir. Öğrencilerin performanslarının değerlendirildiği bu yüzeyde uyum içi ve uyum dışı değerleri de incelenmiştir. Bu bağlamda uygunluk içi 0,75 ile 1,26 arasında değişim gösterirken ortalama değeri 1,00 olarak bulunmuştur. Aynı şekilde uygunluk dışı değerler de incelendiğinde 0,89 ile 1,31 arasında değer olarak ortalama değerinin ise 1,00 olarak hesaplanmıştır. Ön uygulama kontrol grubunda performans gösteren öğrencilerin performansı yüzeylerindeki hem uygunluk içi hem de uygunluk dışı değerlerin 1'e yakın olması model veri uyumunun iyi olduğuna işaret etmektedir. Öğrencilerin performanslarının değerlendirildiği ilgili yüzeydeki diğer bir bakılacak kısım ayırma oranıdır. Bu oranın 5,06 ve güvenilirlik indeks değerinin ise 0,96 olarak hesaplanmıştır. Bu iki değer yüksek çıkması öğrencilerin başlangıç düzey olarak piyano çalma sürecinde öğrencilerin birbirinden farklı yetenek düzeyine sahip olduklarını vurgulamaktadır. Elde edilen bulgulardaki farklılıkların istatistiksel anlamda anlamlı olup olmamasını karar vermek için sabit etkili ki-kare değerine bakılmıştır. İlgili değer istatistiksel anlamda anlamlı olduğu belirlenmiştir ($\chi^2(sd) = 234,6 (9), p = 0,00 < 0,01$). Bu bulgular piyano çalarken performans gösteren öğrencilerin yetenekleri arasında istatistiksel anlamda anlamlı bir farka sahip olduğuna işaret etmektedir.

Kontrol grubu ön uygulama öğrenci performansı yüzeyine ait ölçüm sonuçları incelendikten sonra kontrol grubu ön uygulama puanlayıcı yüzeyi ölçüm sonuçları incelenmiştir. Kontrol grubu ön uygulama puanlayıcı yüzeyi ölçüm sonuçları Tablo 22'de yer almaktadır.

Tablo 22

Kontrol Grubu Puanlayıcı Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu

Puanlayıcı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı	t değeri
1	4,56	4,65	1,49	0,19	1,12	1,07	1,33
2	4,49	4,58	1,26	0,18	1,34	1,18	1,12
3	4,29	4,39	0,69	0,16	0,67	0,80	0,62
4	3,90	4,00	-0,16	0,14	1,87	1,77	-0,14
5	3,67	3,76	-0,58	0,13	0,74	0,82	-0,52
6	3,00	3,02	-1,69	0,13	0,93	0,94	-1,51
7	4,03	4,13	0,10	0,14	0,56	0,59	0,09
8	3,36	3,43	-1,11	0,13	0,93	0,90	-0,99
Ortalama	3,91	3,99	0,00	0,15	1,02	1,01	
S (Evren)	0,51	0,53	1,04	0,02	0,40	0,33	
S (Örnekleme)	0,55	0,57	1,12	0,02	0,43	0,36	
Model, Evren	RMSE= 0,15 Düzeltilmiş S.H.= 1,03 Ayırma Oranı= 6,85 Ayırma İndeksi= 9,47 Güvenirlilik= 0,98						
Model, Örneklem	RMSE= 0,15 Düzeltilmiş S.H.= 1,11 Ayırma Oranı= 7,34 Ayırma İndeksi= 10,12 Güvenirlilik= 0,98						
Model, Ki-kare (Sabit etkili):	381,6	sd= 7	p= 0,00				
Model, Ki-kare (Normal):	6,9	sd= 6	p= 0,33				

S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi

Tablo 22 incelendiğinde, logit değerlerinin -1,69 ile 1,49 aralığında değer aldığı, puanlayıcılara ait uygunluk içi ve uygunluk dışı değerlerine göre 4 numaralı puanlayıcının ölçüt değerler dışında olduğu görülmektedir. Puanlayıcılara ait ayırma oranının 7,34, ayırma indeksinin ise 10,32 olması ve ayırma indeksine ait güvenirlilik katsayısının 0,98 olması puanlayıcıların birbirinden önemli düzeyde farklı bir puanlama davranışı sergilediği şeklinde yorumlanabilir.

Kontrol grubu ön uygulama puanlayıcı yüzeyine ait ölçüm sonuçları incelendikten sonra kontrol grubu ön uygulama ölçüt yüzeyi ölçüm sonuçları incelenmiştir. Kontrol grubu ön uygulama ölçüt yüzeyi ölçüm sonuçları Tablo 23'de yer almaktadır.

Tablo 23

Kontrol Grubu Ölçüt Yüzeyinin Ön Uygulama Ölçümlerine Ait Ölçüm Raporu

Ölçütler	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	4,09	4,22	0,29	0,17	1,86	1,69
2	4,19	4,32	0,52	0,17	1,28	1,15
3	3,84	3,96	-0,23	0,16	0,98	1,11
4	3,60	3,71	-0,68	0,15	0,89	0,98
5	4,65	4,76	1,96	0,23	0,95	0,94
6	3,38	3,45	-1,07	0,15	0,88	0,86
7	3,88	4,00	-0,16	0,16	0,66	0,68
8	4,09	4,22	0,29	0,17	1,12	1,06
9	3,75	3,87	-0,40	0,15	0,81	0,81
10	3,67	3,79	-0,54	0,15	0,79	0,78
Ortalama	3,91	4,03	0,00	0,17	1,02	1,01
S (Evren)	0,34	0,35	0,80	0,02	0,32	0,27
S (Örneklem)	0,36	0,37	0,85	0,02	0,34	0,28
Model, Evren	RMSE= 0,17 Düzeltilmiş S.H.= 0,78 Ayırma Oranı= 4,68 Ayırma İndeksi= 6,58 Güvenirlik= 0,96					
Model, Örneklem	RMSE= 0,17 Düzeltilmiş S.H.= 0,83 Ayırma Oranı= 4,95 Ayırma İndeksi= 6,93 Güvenirlik= 0,96					
Model, Ki-kare (Sabit etkili):	172,8	sd= 9	p= 0,00			
Model, Ki-kare (Normal):	8,5	sd= 8	p= 0,38			

S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi

Kontrol grubunun ön uygulamaya ait analitik dereceli puanlama anahtarındaki ölçütler yüzeyinin ölçümlerine ait Tablo 23 incelendiğinde, logit değerlerinin -1,07 ile 1,96 aralığında değer aldığı ve bu değerlerin ölçme aracında bulunan ölçütlerin güçlük düzeyleri arasında 3,03'lük bir logit değişim ortaya çıktığı ve ölçütlerin ortalaması 0,00 ve standart sapması da 0,85 logit olarak belirlenmiştir. Bu bağlamda ölçme aracında yer alan on ölçüt içerisinde altıncı ölçüt en zor olmuş, beşinci ölçütte en kolay olarak görülmektedir.

Ölçütler yüzeyi için kestirilen uygunluk içi ve uygunluk dışı değerleri incelendiğinde; uyum içi değerlerinin 0,66 ile 1,86 arasında değer aldığı ve ortalama değeri de 1,02 olarak hesaplanmıştır. Aynı şekilde uyum dışı değerleri de 0,68 ile 1,69 arasında değer almış ve ortalama değeri 1,01 olarak bulunmuştur. Bulunan değerler ışığında ölçme aracında yer alan ölçütlerin amaca hizmet edecek düzeyde geliştirildiği ve işlendiğini hem uygunluk içi hem de uygunluk dışı değerlerin 1'e yakın olması model veri uyumunun da iyi olduğuna işaret etmektedir. Ölçüt yüzeyinin ayırma oranı 4,95 ve güvenilirlik değerinin de 0,96 olarak bulunması bu değerlerin farklı yetenek düzeylerinde de güvenilir bir şekilde test edilebileceği ve öğrenci performanslarını birbirinden ayırt edebileceği sonucuna varılabilir. "Piyano başlangıç düzeyi için geliştirilen analitik dereceli puanlama anahtarında yer alan

ölçütlerin zorluk ve kolaylık açısından anlamlı bir farklılık vardır” yani H_1 varlık hipotezinin test edilmesi için ki-kare istatistiği sonucunda çıkan değerler incelenmiştir. İnceleme sonucunda ($\chi^2(sd) = 172,8 (9)$, $p = 0,00 < 0,01$) H_0 yokluk hipotezi reddedilmiştir. Bu sonuç ön testteki kullanılan ölçme aracında yer alan on ölçütün zorluk veya kolaylık düzeyleri arasında istatistiksel açıdan anlamlı bir farklılıklar bulunmaktadır.

Kontrol grubu ön uygulama yüzeylerine ait ölçüm sonuçları incelendikten sonra kontrol grubu son uygulama öğrenci performansı yüzeyi ölçüm sonuçları incelenmiştir. Kontrol grubu son uygulama öğrenci performansı yüzeyi ölçüm sonuçları Tablo 24’de yer almaktadır.

Tablo 24

Kontrol Grubu Öğrenci Performansı Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu

Öğrenci Performansı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	4,24	4,38	2,95	0,19	1,00	1,00
2	4,09	4,21	2,52	0,19	0,83	0,72
3	4,45	4,61	3,65	0,21	0,86	0,68
4	4,11	4,24	2,59	0,19	1,09	1,24
5	4,04	4,16	2,38	0,18	0,78	1,19
6	4,24	4,38	2,95	0,19	0,96	0,83
7	3,75	3,84	1,66	0,17	1,02	1,19
8	3,84	3,94	1,87	0,18	0,94	0,93
9	3,22	3,26	0,48	0,16	1,23	1,24
10	4,43	4,59	3,56	0,21	1,33	1,40
Ortalama	4,04	4,16	2,46	0,19	1,00	1,04
S (Evren)	0,35	0,38	0,90	0,02	0,16	0,23
S (Örnekleme)	0,36	0,40	0,95	0,02	0,17	0,24
Model, Evren	RMSE= 0,19 Düzeltilmiş S.H.= 0,88 Ayırma Oranı= 4,65 Ayırma İndeksi= 6,53 Güvenirlilik= 0,96					
Model, Örnekleme	RMSE= 0,19 Düzeltilmiş S.H.= 0,93 Ayırma Oranı= 4,91 Ayırma İndeksi= 6,88 Güvenirlilik= 0,96					
Model, Ki-kare (Sabit etkili):	245,0 sd= 9 p= 0,00					
Model, Ki-kare (Normal):	8,7 sd= 8 p= 0,37					

S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi

Tablo 24 incelendiğinde, öğrenci performansı düzeylerine ait logit değerinin 0,16’dan başlayıp 0,21’e doğru arttığı ve öğrenci performans yüzeyine ait logit değerlerinin ortalaması 2,46 iken standart sapması ise 0,19 olduğu görülmüştür. Bu hesaplama kapsamında piyano çalarken performansını en iyi derecede icra edebilen öğrenci 3’nolu öğrencidir. Benzer şekilde performansını iyi derecede icra edemeyen ise 9’nolu öğrencidir. Öğrencilerin performanslarının değerlendirildiği bu yüzeyde uyum içi ve uyum dışı değerleri de

incelenmiştir. Bu bağlamda uygunluk içi 0,78 ile 1,33 arasında değişim gösterirken ortalama değeri 1,00 olarak bulunmuştur. Aynı şekilde uygunluk dışı değerler de incelendiğinde 0,68 ile 1,40 arasında değer alarak ortalama değerinin ise 1,04 şeklinde olduğu hesaplanmıştır. Öğrencilerin performans yüzeylerindeki hem uygunluk içi hem de uygunluk dışındaki değerlerin 1'e yakın olması model veri uyumunun iyi olduğuna işaret etmektedir. Öğrencilerin performanslarının değerlendirildiği ilgili yüzeydeki diğer bir bakılacak kısım ayırma oranıdır. Bu oranın 4,91 ve güvenirlik indeks değerinin ise 0,96 olarak hesaplanmıştır. Bu iki değer yüksek çıkması öğrencilerin başlangıç düzey olarak piyano çalma sürecinde öğrencilerin birbirinden farklı yetenek düzeyine sahip olduklarını vurgulamaktadır. Elde edilen bulgulardaki farklılıkların istatistiksel anlamda anlamlı olup olmamasını karar vermek için sabit etkili ki-kare değerine bakılmıştır. İlgili değer istatistiksel anlamda anlamlı olduğu belirlenmiştir ($\chi^2(sd) = 245,0$ (9), $p = 0,00 < 0,01$). Bu bulgular piyano çalarken performans gösteren öğrencilerin yetenekleri arasında istatistiksel anlamda anlamlı bir farka sahip olduğuna işaret etmektedir.

Kontrol grubu son uygulama öğrenci performansı yüzeyine ait ölçüm sonuçları incelendikten sonra kontrol grubu son uygulama puanlayıcı yüzeyi ölçüm sonuçları incelenmiştir. Kontrol grubu son uygulama puanlayıcı yüzeyi ölçüm sonuçları Tablo 25'de yer almaktadır.

Tablo 25

Kontrol Grubu Puanlayıcı Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu

Puanlayıcı	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı	t değeri
1	4,77	4,84	2,27	0,25	1,85	1,43	1,36
2	4,75	4,82	2,15	0,24	1,12	1,17	1,29
3	4,11	4,16	-0,07	0,16	0,73	0,74	-0,04
4	4,34	4,41	0,55	0,17	1,63	1,62	0,33
5	3,49	3,51	-1,48	0,14	0,70	0,72	-0,87
6	3,13	3,14	-2,22	0,14	1,27	1,26	-1,33
7	4,25	4,31	0,30	0,17	0,49	0,49	0,18
8	3,48	3,50	-1,50	0,14	0,89	0,90	-0,90
Ortalama	4,04	4,09	0,00	0,18	1,08	1,04	
S (Evren)	0,57	0,59	1,56	0,04	0,45	0,37	
S (Örneklem)	0,61	0,64	1,67	0,04	0,48	0,39	
Model, Evren	RMSE= 0,18 Düzeltilmiş S.H.= 1,55 Ayırma Oranı= 8,57 Ayırma İndeksi= 11,76 Güvenirlik= 0,99						
Model, Örneklem	RMSE= 0,18 Düzeltilmiş S.H.= 1,66 Ayırma Oranı= 9,17 Ayırma İndeksi= 12,56 Güvenirlik= 0,99						
Model, Ki-kare (Sabit etkili):	557,2	sd= 7	p= 0,00				
Model, Ki-kare (Normal):	6,9	sd=6	p= 0,33				

S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi

Tablo 25 incelendiğinde, uygunluk içi ve dışı değerlerine bakıldığında; 4 ve 6 numaralı puanlayıcıların 0,50-1,50 değerlerinin dışında oldukları 1 numaralı puanlayıcının ise uygunluk içinin ölçüt dışı, uygunluk dışı ölçütünde ise sınırda olduğu görülmektedir. Puanlayıcılara ait ayırma oranının 9,17 ayırma indeksinin ise 12,56 olması ve bu ayırma indeksine ait güvenilirlik katsayısının 0,99 olması kontrol grubundaki puanlayıcıların birbirlerinden oldukça farklı puanlama yaptıkları söylenebilir. Bu değerlerin ön uygulama bulgularıyla karşılaştırıldığında kısmen yüksek olduğu görülmektedir.

Kontrol grubu son uygulama puanlayıcı yüzeyine ait ölçüm sonuçları incelendikten sonra kontrol grubu son uygulama ölçüt yüzeyi ölçüm sonuçları incelenmiştir. Kontrol grubu son uygulama ölçüt yüzeyi ölçüm sonuçları Tablo 26’da yer almaktadır.

Tablo 26

Kontrol Grubu Ölçüt Yüzeyinin Son Uygulama Ölçümlerine Ait Ölçüm Raporu

Ölçütler	Gözlenen Ortalama	Düzeltilmiş Ortalama	Logit Değeri	Standart Hata	Uygunluk İçi	Uygunluk Dışı
1	4,22	4,38	0,48	0,20	1,52	1,45
2	4,11	4,25	0,15	0,19	1,59	2,38
3	4,04	4,17	-0,06	0,18	0,94	0,92
4	3,85	3,96	-0,55	0,18	1,13	1,09
5	4,56	4,73	1,68	0,23	0,91	0,62
6	3,56	3,63	-1,23	0,17	0,74	0,80
7	4,07	4,21	0,04	0,19	0,72	0,63
8	4,15	4,29	0,26	0,19	0,82	0,67
9	3,90	4,01	-0,42	0,18	0,92	0,95
10	3,92	4,04	-0,36	0,18	0,84	0,89
Ortalama	4,04	4,17	0,00	0,19	1,01	1,04
S (Evren)	0,25	0,28	0,73	0,29	0,29	0,51
S (Örneklem)	0,26	0,29	0,77	0,31	0,31	0,53
Model, Evren	RMSE= 0,19 Düzeltilmiş S.H.= 0,70 Ayırma Oranı= 3,72 Ayırma İndeksi= 5,29 Güvenirlik= 0,93					
Model, Örneklem	RMSE= 0,19 Düzeltilmiş S.H.= 0,74 Ayırma Oranı= 3,93 Ayırma İndeksi= 5,57 Güvenirlik= 0,94					
Model, Ki-kare (Sabit etkili):	130,2	sd= 9	p= 0,00			
Model, Ki-kare (Normal):	8,4	sd= 8	p= 0,39			

S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi

Kontrol grubunun son uygulamaya ait ölçüt yüzeyine ait bulgular Tablo 26 incelendiğinde, logit değerlerinin -1,23 ile 1,68 aralığında değer aldığı ve bu değerlerin ölçme aracında bulunan ölçütlerin güçlük düzeyleri arasında 2,91’lik bir logit değişim ortaya çıktığı ve ölçütlerin ortalaması 0,00 ve standart sapması da 0,77 logit olduğu belirlenmiştir. Bu bağlamda ölçme aracında yer alan

on ölçütün öğrencilere hangisinin daha kolay hangisinin ise daha zor olduğu incelenebilir. Kontrol grubunda bulunan öğrencilerin son uygulama sonuçlarına göre altıncı ölçüt diğerlerine göre zor, beşinci ölçüt ise diğerlerine göre daha kolay olarak görülmüştür.

Ölçütler yüzeyi için kestirilen uyum içi ve uyum dışı değerleri incelendiğinde; uyum içi değerlerinin 0,72 ile 1,59 arasında değer aldığı ve ortalama değeri de 1,01 olarak hesaplanmıştır. Aynı şekilde uyum dışı değerleri de 0,62 ile 2,38 arasında değer almış ve ortalama değeri 1,04 olarak bulunmuştur. Bulunan değerler ışığında son uygulamaya ilişkin kontrol grubundaki öğrencilerin değerlendirilmesi için geliştirilen ölçme aracında yer alan ölçütlerin amaca hizmet edecek düzeyde geliştirildiği ve işlendiğini hem uygunluk içi hem de uygunluk dışındaki değerlerin 1'e yakın olması model veri uyumunun da iyi olduğuna işaret etmektedir. Ölçüt yüzeyinin ayırma oranı 3,93 ve güvenilirlik değerinin de 0,94 olarak bulunması bu değerlerin farklı yetenek düzeylerinde de güvenilir bir şekilde test edilebileceği ve öğrenci performanslarını birbirinden ayırt edebileceği sonucuna varılabilir. Bu sonuç son testteki kullanılan ölçme aracında yer alan on ölçütün zorluk veya kolaylık düzeyleri arasında istatistiksel açıdan anlamlı bir farklılıklar bulunmaktadır.

Deney A grubu uygulama öncesi ve sonrası öğrenci performansı yüzeyine ait ölçüm sonuçları karşılaştırıldığında, öğrencilerin piyano performans düzeylerinde genel bir gelişim olduğu görülmektedir. Ön uygulamada logit değerleri 0,74 ile 3,72 arasında değişirken, son uygulamada bu aralık 0,26 ile 3,38 arasına gerilemiştir. Ancak ortalama logit değerleri deney öncesinde 2,36 iken deney sonrasında 2,09 olarak hesaplanmıştır. Bu durum, bazı öğrencilerde ilerleme görülürse de genel performans seviyelerindeki dağılımın daha dengeli hale geldiğini düşündürmektedir. Ayırma oranı ve güvenilirlik indeksinin her iki aşamada da yüksek çıkması, öğrencilerin yetenek düzeylerindeki farklılıklarının devam ettiğini göstermektedir. Ki-kare testi sonuçları ise, ön ve son uygulamada öğrenciler arasında anlamlı bir fark olduğunu doğrulamaktadır. Özetle, öğrencilerin piyano çalma performanslarında bireysel farklılıklar korunmakla beraber, genel dağılımda bir dengeleme olduğu söylenilebilir. Deney A grubu uygulama öncesi ve sonrası puanlayıcı ölçümleri karşılaştırıldığında, son uygulamada puanlayıcıların değerlendirme tutarlılıklarının değiştiği görülmektedir. Ön uygulamada ayırma indeksi 5,04 iken, son uygulamada bu değer 7,65 olarak artış göstermiştir. Bu durum, puanlayıcılar arasındaki değerlendirme farklılıklarının arttığını ve dolayısıyla eğitim sürecinin puanlama tutarlılığını artırma noktasında etkili olmayabileceğini düşündürmektedir. Ayrıca, güvenilirlik değeri her iki uygulamada da yüksek olsa da puanlayıcıların değerlendirme ölçütleri arasında daha fazla farklılık göstermesi, puanlama süreçlerini etkileyen kontrol edilemeyen başka bir değişkenin varlığına işaret edebilir. Özetle, Referans Çerçevesi ile

Puanlayıcı Hatası eğitimlerinin puanlayıcılar üzerindeki etkisinin istenilen şekilde olmadığı ve değerlendirme tutarlılığının sağlanması için önlemler alınması gerektiği söylenilebilir. Bu bulgu tıp alanında yapılan Weitz vd., (2014) tarafından araştırılan çalışmada, fiziksel muayene becerileri değerlendirmesinde bir derecelendirici eğitiminin derecelendirme doğruluğu üzerinde etkisi Referans çerçevesi eğitim desenine göre incelenmiştir. 21 değerlendirici 242 üçüncü sınıf tıp öğrencisinin performansı değerlendirilmiştir. Kısa bir referans çerçevesi eğitime tabi tutulmak üzere yirmi bir puanlayıcının on biri puanlayıcı eğitime rastgele atanmıştır. Toplamda 218 karşılaşma başarıyla videoya kaydedilerek üç ek gözlemci tarafından bağımsız olarak yeniden değerlendirildi. Araştırmanın sonuçlarına göre puanlayıcı eğitiminin ölçülebilir düzeyde etkili olmadığı ortaya çıkmıştır.

Görgülü Öztürk tarafından (2023) yapmış olduğu lise ikinci sınıf öğrencilerinin matematik açık uçlu maddelerin puanlama davranışlarına etkisinin incelendiği çalışmanın sonucu ile örtüşmektedir. Deney A grubu uygulama öncesi ve sonrası ölçüt yüzeyi karşılaştırıldığında, ölçütlerin güçlük düzeylerinde belirgin bir değişim olduğu görülmektedir. Ön uygulamada ölçütlerin logit değerleri 0,63 ile 1,45 arasında değişirken, son uygulamada bu aralık -1,01 ile 3,79 şeklinde genişlemiştir. Bu durum, ölçütler arasındaki ayırmanın arttığını ve öğrencilerin belirli ölçütler üzerinde daha fazla zorlandığını göstermektedir. Ayırma oranı ve güvenilirlik değerleri hem ön hem de son uygulamada yüksek çıkmış olup, ölçme aracının öğrencileri farklı yetenek düzeylerine göre güvenilir bir şekilde değerlendirildiğini göstermektedir. Ki-kare testi sonuçları da ölçütlerin zorluk ve kolaylık düzeyleri arasında anlamlı farklılıklar olduğunu doğrulamaktadır. Bu bulgular, geliştirilen analitik dereceli puanlama anahtarının öğrencilerin piyano performanslarını değerlendirme açısından yeterli bir hassasiyete sahip olduğunu göstermektedir.

Deney B grubu uygulama öncesi ve sonrası öğrenci performansı yüzeyinin karşılaştırıldığında, öğrencilerin piyano performans düzeylerinde belirgin değişiklikler olduğu görülmektedir. Ön uygulamada logit değerleri 0,26 ile 4,28 arasında değişirken, son uygulamada bu aralık -0,22 ile 3,00 arasında seyretmiştir. Ortalama logit değeri ön uygulamada 2,26 iken, son uygulamada 1,92'ye düşmüştür, bu da genel performans seviyelerinde hafif bir düşüş yaşandığını göstermektedir. Ayırma oranı ve güvenilirlik değerleri her iki uygulamada da yüksek çıkmış olup, öğrencilerin yetenek düzeylerinin farklı olduğu ve ölçme aracının bu farklılıkları belirlemede güvenilir olduğu sonucuna varılabilir. Özetle, bireysel öğrenci performanslarının değişim gösterdiğini ancak genel olarak yetenek farklılıklarının sürdüğünü ortaya koymaktadır. Deney B grubu puanlayıcı yüzeyi uygulama öncesi ve uygulama sonrası ölçümleri karşılaştırıldığında, puanlayıcılar arasındaki tutarlılıkta önemli değişiklikler olduğu görülmektedir. Ön uygulamada

ayırma indeksi 10,37 iken, son uygulamada bu değer 5,33'e düşmüştür, bu da puanlayıcılar arasındaki farklılıkların azaldığını göstermektedir. Ayrıca, ön uygulamada bazı puanlayıcıların uygunluk içi ve uygunluk dışı değerleri belirlenen sınırların dışına çıkarken, son uygulamada bu değerlerin ölçüt aralığında olduğu tespit edilmiştir. Bu durum, eğitim sürecinin puanlayıcılar üzerinde etkili olduğunu ve puanlamalar arasındaki uyumu arttırdığını göstermektedir. Özetle, deney öncesine kıyasla puanlayıcılar daha tutarlı ve benzer puanlama eğilimleri göstermiştir, bu da değerlermenin güvenilirliğini artıran olumlu bir gelişme olarak yorumlanabilir. Deney B grubunun ölçüt yüzeyine ait ön uygulama ve son uygulama ölçümleri karşılaştırıldığında, ölçütlerin güçlük düzeyinde önemli değişimler olduğu görülmektedir. Ön uygulama logit değerleri -1,14 ile 4,27 arasında değişirken, son uygulamada bu aralık -0,65 ile 0,81 olarak daralmıştır. Bu durum, ölçütler arasındaki güçlük farkının azaldığını ve değerlendirme ölçütlerinin daha dengeli hale geldiğini göstermektedir. Ayırma oranı ve güvenilirlik değeri son uygulamada düşmüş olup, bu durum ölçme aracının başlangıçtaki yüksek ayırt ediciliğinin azaldığını ve öğrenciler arasındaki farklılıkları belirlemede daha sınırlı hale geldiğini işaret etmektedir. Ki-kare test sonuçları hem ön uygulama hem de son uygulama ölçütlerin zorluk düzeyleri arasında anlamlı farklılıklar olduğunu doğrulamaktadır. Özetle, değerlendirme aracının kullanım sürecinde ölçütler arasındaki farkların azaldığı ve değerlendirme kriterlerinin daha homojen hale geldiği söylenebilir.

Kontrol grubunun öğrenci performansı yüzeyine ait ön uygulama ve son uygulama ölçümleri karşılaştırıldığında, genel performans düzeylerinde artış olduğu gözlemlenmektedir. Ön uygulamada logit değerleri 0,14 ile 3,02 arasında değişirken, son uygulamada bu aralık 0,16 ile 3,65 olarak genişlemiştir. Ortalama logit değeri ön uygulamada 1,16 iken son uygulamada 2,46'ya yükselmiş, bu da öğrencilerin piyano performansında genel bir iyileşme olduğunu göstermektedir. Ayırma oranı ve güvenilirlik değerleri her iki uygulamada da yüksek çıkmış olup, öğrencilerin yetenek farklılıklarının sürdüğünü ve ölçme aracının bu farklılıkları güvenilir bir şekilde belirlediğini göstermektedir. Özetle, kontrol grubundaki öğrencilerin piyano performanslarında gelişim olduğu, ancak bireysel farklılıkların belirgin şekilde devam ettiği söylenebilir. Kontrol grubundaki puanlayıcı yüzeyine ait ön uygulama ve son uygulama ölçümleri karşılaştırıldığında, puanlayıcılar arasındaki farklılıkların arttığı görülmektedir. Ön uygulamada ayırma indeksi 10,12 iken, son uygulamada bu değer 12,56'ya yükselmiş ve bu da puanlayıcıların değerlendirme tutarsızlıklarının arttığını göstermektedir. Ayrıca, son uygulamada bazı puanlayıcıların uygunluk içi ve uygunluk dışı değerlerinin belirlenen sınırların dışında olduğu tespit edilmiştir. Güvenirlik katsayısı her iki uygulamada da oldukça yüksek çıkmış olsa da puanlama davranışlarındaki farklılıkların artması, değerlendirmenin tutarlılığı açısından sorun teşkil edebilir. Özetle, kontrol

grubundaki puanlayıcıların puanlama eğilimleri deney sürecinde daha da farklılaşmış, bu da değerlendirme sürecinin homojenliğini olumsuz yönde etkilemiştir. Kontrol grubunun ölçüt yüzeyine ait ön uygulama ve son uygulama ölçümleri karşılaştırıldığında, ölçütlerin güçlük seviyelerinde bazı değişiklikler olduğu görülmektedir. Ön uygulamada logit değerleri -1,07 ile 1,96 arasında değişirken, son uygulamada bu aralık -1,23 ile 1,68 olarak hesaplanmıştır. Bu durum, ölçütlerin zorluk düzeylerinde küçük değişiklikler olduğunu ve ölçme aracının dengeli bir değerlendirme yapmaya devam ettiğini göstermektedir. Ayırma oranı ve güvenilirlik değerleri her iki uygulamada da yüksek olmakla birlikte, son uygulamada ayırma oranının 3,93'e düşmesi, ölçütlerin öğrenciler arasındaki farklılıkları belirlemede başlangıça kıyasla biraz daha az etkili olduğunu göstermektedir. Ki-kare testi sonuçları, her iki aşamada da ölçütlerin zorluk düzeyleri arasında istatistiksel olarak anlamlı farklılıklar olduğunu doğrulamaktadır. Özetle, kontrol grubundaki ölçüt yüzeylerinin genel geçerliğini koruduğu ancak ölçütler arasındaki ayırım gücünün biraz azaldığı söylenebilir.

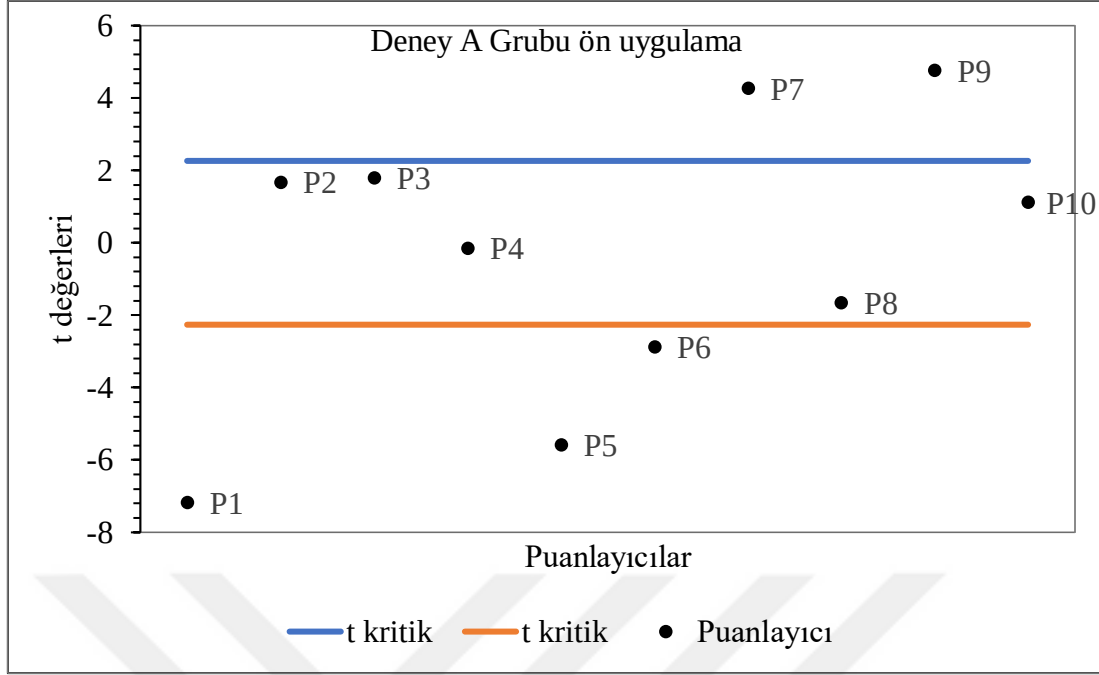
İkinci araştırma sorusuna ilişkin araştırma bulguları

Katılık ve Cömertlik, Halo Etkisi ile Yanlılık analizleri

a) *Deney A, Deney B ve Kontrol grubu ön uygulamada yer alan puanlayıcıların katılık ve cömertlik davranışları nedir?*

İlgili araştırmanın sorusuna göre Deney A, Deney B ve Kontrol grubu ön uygulamada yer alan puanlayıcıların öncelikle grup düzeyinde katılık ve cömertlik durumları incelenmiştir. Bireysel düzeyde puanlayıcı davranışları Şekil 2 Şekil 3 ve Şekil 4'te verilmiştir.

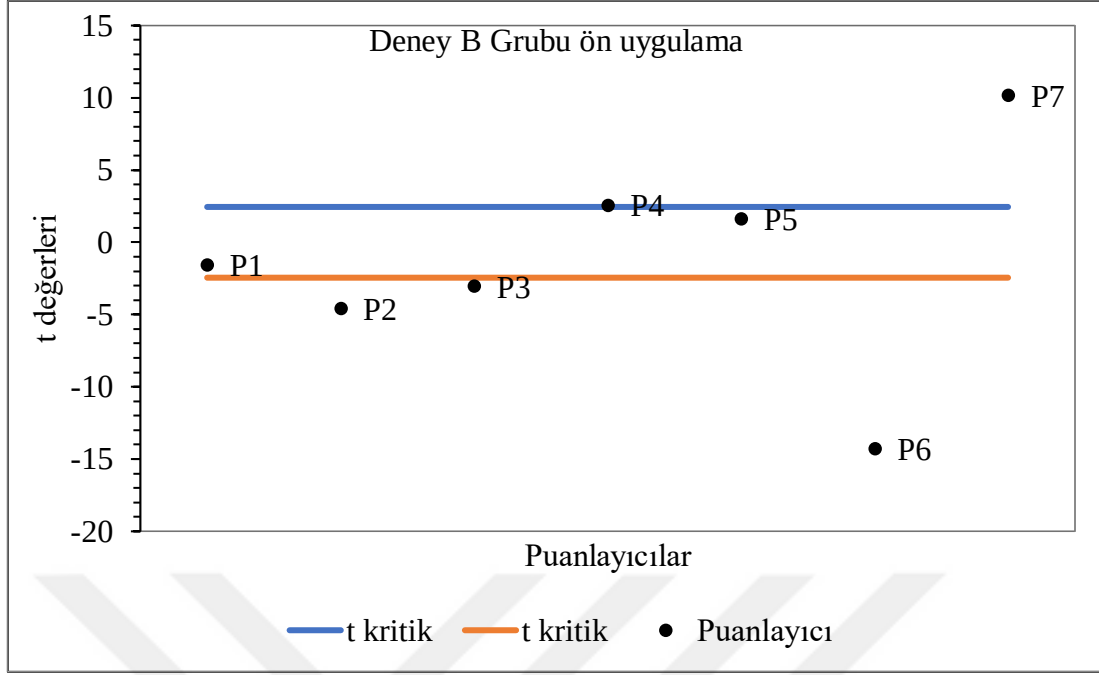
Grup düzeyinde katılık ve cömertlik davranışı incelendiğinde Deney A, Deney B ve Kontrol ön uygulama gruplarında ilgili istatistiklerin yüksek olduğu gözlenmiştir. Bu durumun istatistiksel açıdan anlamlılığına bakmak için sabit etkili ki kare değeri incelenmiş ve bu değer tüm gruplar için anlamlı olduğu görülmüştür. Sonuç olarak Deney A (Tablo 10), Deney B (Tablo 16) ve Kontrol (Tablo 22) ön uygulama gruplarında grup düzeyinde katılık ve cömertlik davranışının olduğu söylenebilir.



Şekil 2. Deney A Grubu ön uygulama t değerleri

Şekil 2 incelendiğinde Deney A grubu ön uygulamada yer alan puanlayıcılardan P7 ve P9 numaralı puanlayıcılarda cömertlik, P1, P5 ve P6 numaralı puanlayıcılarda da katılık davranışı olduğu görülmektedir.

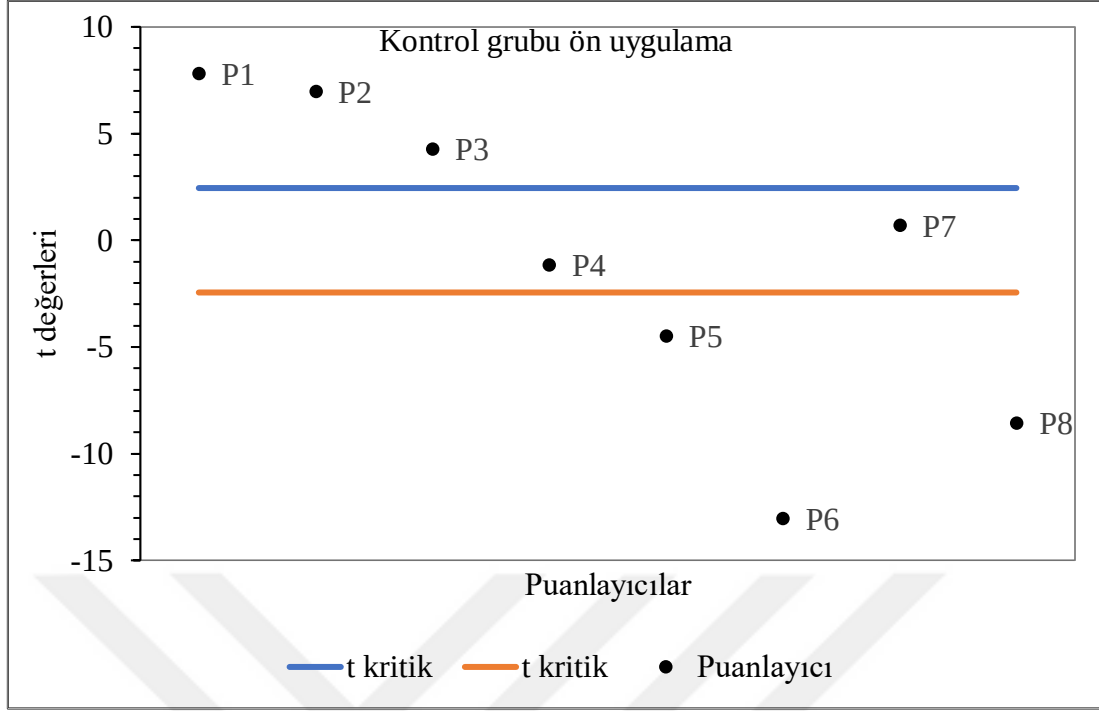
Öğrencilerin performanslarının değerlendirilmesi için Deney A ön uygulama grubunda yer alan 10 puanlayıcının ölçülmek istenilen yapıyı değerlendirirken puanlama davranışlarının puanlamaya (katılık ve cömertlik) karışıp karışmadığı incelenmiştir. Buna göre iki (%20) puanlayıcının cömert, üç (%30) puanlayıcının ise katı davranış gösterdiği bulunmuştur. Herhangi bir puanlayıcı katılığı ve cömertliği davranışında bulunmayan beş (%50) puanlayıcı mevcuttur.



Şekil 3. Deney B grubu ön uygulama t değerleri

Şekil 3 incelendiğinde Deney B grubu ön uygulamada yer alan puanlayıcılardan P4 ve P7 numaralı puanlayıcılarda cömertlik, P2, P3 ve P6 numaralı puanlayıcılarda da katılık davranışı olduğu görülmektedir.

Öğrencilerin performanslarının değerlendirilmesi için Deney B ön uygulama grubunda yer alan 7 puanlayıcının ölçülmek istenilen yapıyı değerlendirirken puanlama davranışlarının (katılık ve cömertlik) karışık karışmadığı incelenmiştir. Buna göre iki (%28,57) puanlayıcının cömert, üç (%42,86) puanlayıcının ise katı davranış göstermiştir. Puanlayıcı katılığı ve cömertliği davranışında bulunmayan iki (%28,57) puanlayıcı mevcuttur.



Şekil 4. Kontrol grubu ön uygulama t değerleri

Şekil 4 incelendiğinde kontrol grubu ön uygulamada yer alan puanlayıcılardan P1, P2 ve P3 numaralı puanlayıcılarda cömertlik, P5, P6 ve P8 numaralı puanlayıcılarda da katılık davranışı olduğu görülmektedir.

Öğrencilerin performanslarının değerlendirilmesi için Kontrol ön uygulama grubunda yer alan 8 puanlayıcının ölçülmek istenilen yapıyı değerlendirirken puanlama davranışlarının (katılık ve cömertlik) karışık karışmadığı incelenmiştir. Buna göre üç (%37,5) puanlayıcının cömert, üç (%37,5) puanlayıcının ise katı davranış gösterdiği bulunmuştur. Puanlayıcı katılığı ve cömertliği davranışında bulunmayan iki (%25) puanlayıcı bulunmaktadır.

Her üç grupta yer alan puanlayıcıların puanlama davranışları sonucunda öğrencinin performanslarını değerlendirirken ölçülmek istenilen yapıya karıştığı bulunmuştur. Bu araştırmanın araştırma sorusundan elde edilen sonuca göre öğrencilerin performanslarının değerlendirildiğinde puanlayıcılardan kaynaklı katılık ve cömertlik davranışları meydana gelmiştir. Alanyazında yürütülen farklı disiplinlerdeki performans değerlendirmenin incelendiği araştırmaların sonucunda da benzer bulgular elde edilmiştir (Abu Kassım, 2007; Farrokhi, Esfandiari ve Schaefer, 2012; Karakaya, 2015; Köse, Sözbir ve Kalender, 2016; Çetin ve İlhan, 2017; Knock vd., 2018; Şata, 2019; Esfandiari, 2021; Görgülü Öztürk, 2023; ÖÇ, 2024).

b) Deney A, Deney B ve Kontrol grubu ön uygulamada yer alan puanlayıcıların halo davranışları nedir?

Puanlayıcılarda grup düzeyinde halo etkisi olup olmadığını incelemek amacıyla tüm gruplar için ölçüt yüzeyine ait ölçüm sonuçları incelenmiştir. Deney A grubu, Deney B grubu ve Kontrol grubu ön uygulama ölçüt yüzeyine ait ölçüm sonuçları incelendiğinde ayırma oranı (4,27; 6,41; 4,95) ayırma indeksi (6,02; 8,89; 6,93) ve güvenilirlik (0,95; 0,98; 0,96) değerlerinin yüksek olduğu ve sabit etkili ki kare istatistiklerinin anlamlı olduğu ($p<0,05$) belirlenmiştir. Bu durum Deney A, Deney B ve Kontrol grubu ön uygulamada puanlayıcılarda grup düzeyinde halo etkisinin olmadığını göstergesidir.

Bireysel düzeyde halo etkisinin incelenmesi için madde güçlüklerine bakılmaktadır. Deney A, Deney B ve Kontrol grubu ön uygulamada (2,08; 5,41; 3,03) yer alan madde güçlükleri yüksektir. Bu yüzden puanlayıcıların uygunluk istatistiklerinden 1'den çok büyük değer bulunduğu halo etkisi vardır (Myford & Wolfe, 2004). Tüm grupların ön uygulama puanlayıcı yüzeyine ait sonuçlara bakıldığında Deney A grubunda P3 ve P10 numaralı puanlayıcılarda, Deney B grubunda P6 numaralı puanlayıcıda ve kontrol grubunda P2 ve P4 numaralı puanlayıcılarda halo etkisi davranışı belirlenmiştir.

Halo etkisi davranışının diğer belirlenme yöntemi de madde güçlüklerini eşitleyip mükemmel uyum sağlayan puanlayıcıların belirlenmesidir (Linacre, 2023, s. 360). Tablo 27'da madde güçlükleri eşitlendikten sonraki uyum istatistikleri verilmiştir.

Tablo 27

Madde Güçlükleri Eşitlendikten Sonra Uyum İstatistikleri

Grup	Puanlayıcılar	Uygunluk içi	Uygunluk dışı
Deney A	P1	0,76	0,75
	P2	0,73	0,89
	P3	1,05	1,14
	P4	0,74	0,69
	P5	0,78	0,77
	P6	0,92	0,90
	P7	1,42	1,77
	P8	1,06	0,95
	P9	0,79	0,78
	P10	1,69	1,48
Deney B	P1	0,53	0,51
	P2	0,92	0,89
	P3	0,69	0,72
	P4	0,77	0,80
	P5	0,76	0,76
	P6	1,85	1,80
	P7	1,15	0,88
Kontrol	P1	0,85	0,77
	P2	1,20	1,25
	P3	0,60	0,63
	P4	1,93	1,90
	P5	0,48	0,50
	P6	1,07	1,08
	P7	0,70	0,74
	P8	1,10	1,10

Tablo 27 incelendiğinde tüm gruplarda mükemmel uyum gösteren puanlayıcının olmadığı dolayısıyla halo etkisi gösteren puanlayıcının olmadığı görülmektedir.

Tüm grupların bireysel istatistiklerinin ön uygulama puanlayıcı yüzeyine ait sonuçlarına bakıldığında Deney A Referans çerçevesi eğitimi ile Puanlayıcı Hatası eğitimini alan gruptaki 10 puanlayıcıdan ikisinde (P3 ve P10) numaralı puanlayıcılarda, benzer şekilde Deney B Referans çerçevesi eğitimi ile Performans Boyutu eğitimlerini alan grupta P6 numaralı puanlayıcıda ve hiçbir müdahale yapılmayan Kontrol grubunda ise P2, P4, P6 ve P8 numaralı puanlayıcılarda halo etkisi davranışı belirlenmiştir. Alanyazında Öç, (2024) tarafından yapılan çalışmada, açık uçlu soruların puanlanmasında puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisi araştırılmıştır. Puanlayıcı eğitimi öncesinde hem Türkçe hem de matematik gruplarındaki puanlayıcıların benzer şekilde halo davranışının sadece bireysel bazında ortaya çıktığı tespit edilmiştir. Benzer bir bulgu Şata (2019) tarafından sunulan doktora tezinde üniversite öğrencilerinin yazılı performans olarak kompozisyon değerlendirme sürecindeki nesnelliği engelleyen sorunlu puanlama davranışlarını belirlemeyi ve bu davranışları ölçümlerde

kontrol altına almak için deneysel olarak puanlayıcı eğitimlerinin etkisini belirlemeyi amaçlamıştır. Çalışmada hem deney hem de kontrol grubunda yer alan puanlayıcıların puanlama eğitimi öncesinde halo davranışı sergilemişlerdir. Esfandiari (2015) ve Engelhard (1994) tarafından araştırılan çalışmalarda performansın değerlendirilmesinde ölçülmesi hedeflenen yapıya halo davranışının bireysel düzeyde karıştığına ilişkin bulgular bulunmuştur. Wu ve Tan (2016) yapılan performans değerlendirme çalışmasında da değerlendirme sonuçlarının belli bir kısmında halo davranışları bulunmaktadır. Genel olarak benzer çalışmalarda hem ulusal düzeyde hemde uluslararası çalışmalarda halo davranışlarının puanlayıcılar tarafından sergilendiği görülmektedir.

c) *Deney A, Deney B ve Kontrol grubu ön uygulamada yer alan puanlayıcıların farklılaşan Puanlayıcı Katılığı ve Cömertliği (Yanlılık) nasıldır?*

Grupların yanlılık davranışlarının incelenmesi için grup düzeyindeki istatistikler incelenmiştir. Grup düzeyinde istatistikler incelendiğinde Deney A ve Kontrol son uygulama gruplarının sabit etkili ki kare istatistiğinin anlamlı olduğu ($p < 0,01$) ve Deney B grubu son uygulama grubunun ise sabit etkili ki kare istatistiğinin anlamlı olmadığı ($p < 0,01$) bulunmuştur. Bu durumda Deney A ve Kontrol son uygulama grupları için H_0 hipotezi reddedilerek, puanlayıcıların grup düzeyinde yanlılık davranışı sergiledikleri şeklinde yorumlanabilir. Ayrıca Deney B son uygulama grubu içinse H_0 hipotezinin kabul edilerek, puanlayıcıların grup düzeyinde yanlılık davranışı sergilemedikleri şeklinde yorumlanabilir. Grup düzeyindeki istatistikler incelendikten sonra bireysel istatistikler incelenmiştir.

Farklılaşan puanlayıcı katılığı ve cömertliği davranışının incelenirken modelde yer alan yüzeyler arasındaki etkileşimler incelenmektedir. İlgili çalışma kapsamında *öğrenci performansı (öp)*, *puanlayıcı (p)* ve *ölçüt (ö)* yüzeyi olmak üzere üç yüzey bulunmaktadır.

Bu yüzeyler arasındaki etkileşimler de *puanlayıcı x öğrenci performansı (pxöp)*, *puanlayıcı x ölçüt (pxö)*, *öğrenci performansı x ölçüt (öpxö)* ve *öğrenci performansı x puanlayıcı x ölçüt (öpxpxö)* şeklindedir. İlgili araştırmada puanlayıcıların ölçme aracındaki ölçütlerle ilgili davranışlarını incelemek amaçlanmadığı için *puanlayıcı x ölçüt (pxö)* etkileşimi çalışmaya dahil edilmemiştir. Benzer şekilde çalışmada puanlayıcı davranışları incelendiği ve öğrenci performanslarının ölçme aracındaki ölçütleri cevaplandırma seviyeleri de incelenmediği için *öğrenci performansı x ölçüt (öpxö)* etkileşimi de çalışma kapsamında değildir. Ayrıca puanlayıcıların hem ölçüt hem de öğrenci performansına göre davranışları araştırma kapsamında olmadığı için *öğrenci performansı x puanlayıcı x ölçüt (öpxpxö)* etkileşimi de çalışmaya dahil edilmemiştir. Çalışma kapsamında başlangıç düzey piyano performansın puanlanması sürecinde

meydana gelen yanlışlık davranışlarının tespiti için *puanlayıcı x öğrenci performansı (pxöp)* etkileşimi incelenmiştir.

Puanlayıcı x öğrenci performansı (pxöp) etkileşiminde yer alan istatistiksel olarak anlamlı olan etkileşimlerin tespiti için t istatistiğinden faydalanılmaktadır. Linacre (2023, s. 190), ± 2 aralığı dışında kalan t değerlerinin istatistiksel olarak anlamlı olduğunu ifade etmektedir. Mevcut çalışmada 10 öğrencinin başlangıç düzey piyano performansın puanlanması sürecinde vermiş oldukları tüm cevaplar Deney A (n=10); Deney B (n=7); Kontrol (n=8) olmak üzere toplam 25 puanlayıcı tarafından puanlanmıştır. Dolayısıyla üç yüzey arasında 2500 (25x 10x 10) olası etkileşim varken Deney A grubunda 1000 (10x 10x 10) etkileşim; Deney B grubunda, 700 (7x 10x 10) etkileşim ve Kontrol grubunda ise 800 (8x 10x 10) etkileşim meydana gelmiştir. Analiz sonuçları incelenirken; **öncelikle grup düzeyindeki** istatistikler incelenmiş ardından bireysel istatistiklere geçilmiştir. **Grup düzeyindeki** istatistikler incelendiğinde Deney A, Deney B ve Kontrol ön uygulama gruplarının sabit etkili ki kare istatistiğinin anlamlı olduğu bulunmuştur ($p<0,01$). Bu durumda H_0 hipotezi reddedilerek, puanlayıcıların grup düzeyinde yanlışlık davranışı sergilemedikleri şeklinde yorumlanabilir. Grup düzeyindeki istatistikler incelendikten sonra bireysel istatistikler incelenmiştir.

Deney A ön uygulama puanlayıcı x öğrenci performansı anlamlı etkileşimleri Tablo 28’de verilmiştir.

Tablo 28

Deney A Ön Uygulama Puanlayıcı x Öğrenci Performansı Etkileşimleri

Grup	Puanlayıcı	Öğrenci Performansı	Gözlenen Puan	Beklenen Puan	Yanlışlık (logit)	Standart Hata	t
DAÖT	P1	9	21,00	28,11	-1,30	0,45	-2,86
DAÖT	P1	1	34,00	39,44	-0,89	0,40	-2,24
DAÖT	P1	2	45,00	37,56	1,47	0,52	2,84
DAÖT	P2	3	45,00	48,14	-1,19	0,52	-2,31
DAÖT	P2	6	45,00	48,06	-1,15	0,52	-2,22
DAÖT	P3	6	44,00	48,11	-1,43	0,49	-2,93
DAÖT	P3	3	45,00	48,18	-1,22	0,52	-2,36
DAÖT	P4	4	35,00	40,13	-0,85	0,40	-2,14
DAÖT	P5	4	27,00	35,53	-1,37	0,41	-3,32
DAÖT	P5	10	50,00	43,52	3,00	1,47	2,05
DAÖT	P5	1	46,00	40,27	1,30	0,56	2,31
DAÖT	P8	2	48,00	41,52	1,89	0,75	2,52
DAÖT	P10	2	27,00	43,47	-2,80	0,41	-6,79
DAÖT	P10	7	48,00	41,73	1,85	0,75	2,46

χ^2 (sd)=211,5 (100), $p< 0,05$

Tablo 28’de Deney A grubunun ön uygulamasına ait puanlayıcıların ve performans gösteren öğrencilerin gözlenen, beklenen, yanlılık, standart hata ve t değerleri yer almaktadır. Gözlenen puan; öğrencilerin piyano çalmadaki sergilemiş oldukları performanslarına ilişkin gerçek performansını temsil eden puanlardır. Beklenen puan ise ÇYRÖM tarafından tahmin edilen değerler veya puanlardır. Yanlılık (logit) değeri ise puanlayıcının öğrenci performansını değerlendirme sürecinde gösterdiği puanlama davranışlarıdır ve bunlar olduğundan fazla bol veya olduğundan fazla az puan vermesidir. Bu değer pozitif olması cömertliğe, negatif olması ise katılığı işaret etmektedir. Standart hata değeri ise ÇYRÖM ile tahmin edilecek ölçümün güvenilir ve doğru yorumlanmasına yardımcı olan istatistiksel bir kriterdir. t değeri istatistiksel testlerde, bir gözlemin beklenen değerden ne kadar saptığını ve bu sapmanın rastgele varyansa göre önemini ifade eden bir ölçüttür. Bu ölçütte eşik değer olarak (+/- 2) aralığı kabul edilmektedir.

Deney A grubu ön uygulamasına göre *puanlayıcı x öğrenci performansı* etkileşimine ait 14 anlamlı etkileşim hesaplanmıştır. Bu gruba ait diğer etkileşim tablosunun tamamını sunmak mümkün olmadığı için Ek 5’de ayrıntılı olarak her bir puanlayıcıya ve öğrenciye ilişkin gözlenen puan, beklenen puan, yanlılık, standart hata, t değeri, serbestlik derecesi, olasılık iç ve dış uyumlara yer verilmiştir.

Buna göre puanlayıcılardan (P1, P2, P3, P4, P5, P8 ve P10) diğer puanlayıcılardan farklı olarak hem farklılaşan katılık hem de farklılaşan cömertlik davranışlarını sergilemişlerdir. Dolayısıyla (P1, P2, P3, P4, P5, P8 ve P10) puanlayıcıların puanlama davranışları dikkate alınarak yorumlanmıştır.

P1’in puanlama davranışı incelendiğinde; öğrenci performanslarından 9 ve 1 numaralıları değerlendirme esnasında farklılaşan katılık ve 2 numaralı öğrenci performansına da farklılaşan cömertlik davranışlarını gösterdiği görülmüştür. Örneğin P1’in, 9 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($21 < 28,11$) benzer şekilde 1 numaralı öğrenci performansı dikkate alındığında da gözlenen puanı beklenen puandan küçük olmasından ($34 < 39,44$) dolayı farklılaşan katılık davranışını sergilemiştir. Aynı puanlayıcının (P1) 2 numaralı öğrencinin performansına gözlenen puanı beklenen puandan büyük şeklinde puan vermiş ($45 > 37,56$) dolayısıyla farklılaşan cömertlik davranışını sergilemiştir.

P2’nin puanlama davranışı incelendiğinde; öğrenci performanslarından 3 ve 6 numaralıları değerlendirme esnasında farklılaşan katılık davranışlarını gösterdiği görülmüştür. Örneğin P2’nin, 3 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan

küçük olduğu için ($45 < 48,14$) benzer şekilde aynı puanlayıcının 6 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($45 < 48,06$) dolayı farklılaşan katılık davranışını sergilemiştir.

P3'in puanlama davranışı incelendiğinde; öğrenci performanslarından 6 ve 3 numaralıları değerlendirme esnasında farklılaşan katılık davranışlarını gösterdiği görülmüştür. Örneğin P3'in, 6 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($44 < 48,11$) benzer şekilde aynı puanlayıcının 3 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($45 < 48,18$) dolayı farklılaşan katılık davranışını sergilemiştir.

P4'ün puanlama davranışı incelendiğinde; öğrenci performansı olarak 4 numarayı değerlendirme esnasında farklılaşan katılık davranışlarını gösterdiği görülmüştür. Örneğin P4'ün, 4 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($35 < 40,13$) farklılaşan katılık davranışını sergilemiştir.

P5'in puanlama davranışı incelendiğinde; öğrenci performanslarından 10 ve 1 numaralıları değerlendirme esnasında farklılaşan cömertlik ve 4 numaralı öğrenci performansına da farklılaşan katılık davranışlarını gösterdiği görülmüştür. Örneğin P5'in, 10 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan büyük olduğu için ($50 > 43,52$) benzer şekilde 1 numaralı öğrenci performansı dikkate alındığında da gözlenen puanı beklenen puandan büyük olmasından ($46 > 40,27$) dolayı farklılaşan cömertlik davranışını sergilemiştir. Aynı puanlayıcının (P5) 4 numaralı öğrencinin performansına gözlenen puanı beklenen puandan küçük şeklinde puan vermiş ($27 < 35,53$) dolayısıyla farklılaşan katılık davranışını sergilemiştir.

P8'in puanlama davranışı incelendiğinde; öğrenci performansı olarak 2 numarayı değerlendirme esnasında farklılaşan cömertlik davranışlarını gösterdiği görülmüştür. Örneğin P8'in, 2 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan büyük olduğu için ($48 > 41,52$) farklılaşan cömertlik davranışını sergilemiştir.

P10'un puanlama davranışı incelendiğinde; öğrenci performanslarından 2 numaralı değerlendirme esnasında farklılaşan katılık ve 7 numaralı öğrenci performansına da farklılaşan cömertlik davranışlarını gösterdiği görülmüştür. Örneğin P10'un, 9 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($27 < 43,47$) farklılaşan katılık davranışını sergilemiştir. Aynı puanlayıcının (P10) 7 numaralı

öğrencinin performansına gözlenen puanı beklenen puandan büyük şeklinde puan vermiş (48>41,43) dolayısıyla farklılaşan cömertlik davranışını sergilemiştir.

Dolayısıyla tablo 28'deki Deney A ön uygulama *puanlayıcı x öğrenci performansı* etkileşimleri özetlendiğinde diğer puanlayıcılardan farklı olarak üç defa etkileşim yapan P1 ile P5 nolu puanlayıcılardır. P1 nolu puanlayıcı 1, 2 ve 9 nolu öğrenci performanslarını değerlendirirken iki katı ve bir cömert puanlama davranışı sergilemiştir. Benzer şekilde P5 nolu puanlayıcı da 1, 4 ve 10 nolu öğrenci performanslarını değerlendirirken bir katı ve iki cömert puanlama davranışında bulunmuştur. Her iki puanlayıcı 1 nolu öğrenci performansını değerlendirmişler ve P1 1 nolu öğrencinin performansına katı davranış göstermişken, P5 ise cömert davranış göstermiştir.

Deney B ön uygulama puanlayıcı x öğrenci performansı anlamlı etkileşimleri Tablo 29'da verilmiştir.

Tablo 29

Deney B Ön Uygulama Puanlayıcı x Öğrenci Performansı Etkileşimleri

Grup	Puanlayıcı	Öğrenci Performansı	Gözlenen Puan	Beklenen Puan	Yanlılık (logit)	Standart Hata	t
DBÖT	P2	3	42,00	45,31	-1,16	0,57	-2,03
DBÖT	P3	3	42,00	45,94	-1,42	0,57	-2,46
DBÖT	P4	10	40,00	46,51	-2,28	0,54	-4,20
DBÖT	P4	8	38,00	41,48	-1,00	0,50	-2,01
DBÖT	P4	1	49,00	44,88	2,25	1,06	2,12
DBÖT	P5	1	41,00	44,43	-1,16	0,56	-2,07
DBÖT	P6	4	36,00	30,22	0,96	0,45	2,12
DBÖT	P6	10	45,00	39,91	1,70	0,61	2,78
DBÖT	P7	9	39,00	43,25	-1,33	0,52	-2,56

χ^2 (sd)=115,4 (70), $p < 0,05$

Tablo 29'da Deney B grubunun ön uygulamasına ait puanlayıcıların ve performans gösteren öğrencilerin gözlenen, beklenen, yanlılık, standart hata ve t değerleri yer almaktadır. Gözlenen puan; öğrencilerin piyano çalmadaki sergilemiş oldukları performanslarına ilişkin gerçek performansını temsil eden puanlardır. Beklenen puan ise ÇYRÖM tarafından tahmin edilen değerler veya puanlardır. Yanlılık (logit) değeri ise puanlayıcının öğrenci performansını değerlendirme sürecinde olduğundan fazla bol veya olduğundan fazla az puan vermesidir. Standart hata değeri ise ÇYRÖM ile tahmin edilecek ölçümün güvenilir ve doğru yorumlanmasına yardımcı olan istatistiksel bir kriterdir. t değeri istatistiksel testlerde, bir gözlemin beklenen değerden ne kadar saptığını ve bu sapmanın rastgele varyansa göre

önemini ifade eden bir ölçüttür. Bu ölçütte eşik değer olarak (+/- 2) aralığı kabul edilmektedir.

Deney B grubu ön uygulamasına göre puanlayıcı x öğrenci performansı etkileşimine ait 9 anlamlı etkileşim hesaplanmıştır. Bu gruba ait diğer etkileşim tablosunun tamamını sunmak mümkün olmadığı için Ek 7’de ayrıntılı olarak her bir puanlayıcıya ve öğrenciye ilişkin gözlenen puan, beklenen puan, yanlışlık, standart hata, t değeri, serbestlik derecesi, olasılık iç ve dış uyumlara yer verilmiştir.

Buna göre puanlayıcılardan (P4, P6) diğer puanlayıcılardan farklı olarak hem farklılaşan katılık hem de farklılaşan cömertlik davranışlarını sergilemişlerdir. P2, P3, P5 ve P7 ise farklılaşan katılık davranışlarını sergilemişlerdir. Dolayısıyla P2, P3, P4, P5, P6 ve P7 ile puanlayıcıların puanlama davranışları dikkate alınarak yorumlanmıştır.

P2’nin puanlama davranışı incelendiğinde; öğrenci performanslarından sadece 3 numaralı performansı değerlendirme esnasında farklılaşan katılık davranışı sergilemiştir. Örneğin P2’nin 3 nolu öğrenci performansına vermiş olduğu puan gözlenen puanı beklenen puandan küçük olduğu için ($42 < 45,31$) farklılaşan katılık davranışını sergilemiştir.

P3’ün puanlama davranışı incelendiğinde; öğrenci performanslarından sadece 3 numaralı performansı değerlendirme esnasında farklılaşan katılık davranışı sergilemiştir. Örneğin P3’ün 3 nolu öğrenci performansına vermiş olduğu puan gözlenen puanı beklenen puandan küçük olduğu için ($42 < 45,94$) farklılaşan katılık davranışını sergilemiştir.

P4’in puanlama davranışı incelendiğinde; öğrenci performanslarından 10 ve 8 numaralıları değerlendirme esnasında farklılaşan katılık ve 1 numaralı öğrenci performansına da farklılaşan cömertlik davranışlarını gösterdiği görülmüştür. Örneğin P4’in, 10 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($40 < 46,51$) benzer şekilde 8 numaralı öğrenci performansı dikkate alındığında da gözlenen puanı beklenen puandan küçük olmasından ($38 < 41,48$) dolayı farklılaşan katılık davranışını sergilemiştir. Aynı puanlayıcının (P4) 1 numaralı öğrencinin performansına gözlenen puanı beklenen puandan büyük şeklinde puan vermiş ($49 > 44,88$) dolayısıyla farklılaşan cömertlik davranışını sergilemiştir.

P5’in puanlama davranışı incelendiğinde; öğrenci performanslarından sadece 1 numaralı performansı değerlendirme esnasında farklılaşan katılık davranışı sergilemiştir. Örneğin P5’in 1 nolu öğrenci performansına vermiş olduğu puan gözlenen puanı beklenen puandan küçük olduğu için ($41 < 44,43$) farklılaşan katılık davranışını sergilemiştir.

P6'nın puanlama davranışı incelendiğinde; öğrenci performanslarından 4 ve 10 numaralıları değerlendirme esnasında farklılaşan cömertlik davranışını gösterdiği görülmüştür. Örneğin P6'nın, 4 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan büyük olduğu için ($36 > 30,22$) benzer şekilde 10 numaralı öğrenci performansı dikkate alındığında da gözlenen puanı beklenen puandan büyük olmasından ($45 > 39,91$) dolayı farklılaşan cömertlik davranışını sergilemiştir.

P7'nin puanlama davranışı incelendiğinde; öğrenci performanslarından sadece 9 numaralı performansı değerlendirme esnasında farklılaşan katılık davranışı sergilemiştir. Örneğin P7'nin 9 nolu öğrenci performansına vermiş olduğu puan gözlenen puanı beklenen puandan küçük olduğu için ($39 < 43,25$) farklılaşan katılık davranışını sergilemiştir.

Dolayısıyla tablo 29'daki Deney B ön uygulama *puanlayıcı x öğrenci performansı* etkileşimleri özetlendiğinde diğer puanlayıcılardan farklı olarak üç defa etkileşim yapan P4 ile iki defa etkileşim yapan P6 nolu puanlayıcıların davranışları ele alınmıştır. P4 nolu puanlayıcı 1, 8 ve 10 nolu öğrenci performanslarını değerlendirirken üç değerlendirmenin tamamında katı puanlama davranışı sergilemiştir. Benzer şekilde P6 nolu puanlayıcı da 4 ve 10 nolu öğrenci performanslarını değerlendirirken iki cömert puanlama davranışında bulunmuştur. Her iki puanlayıcı 10 nolu öğrenci performansını değerlendirmişler ve P4 10 nolu öğrenci performansını katı davranış göstermişken, P6 ise cömert davranış göstermiştir.

Kontrol ön uygulama puanlayıcı x öğrenci performansı anlamlı etkileşimleri Tablo 30'da verilmiştir.

Tablo 30

Kontrol Ön Uygulama Puanlayıcı x Öğrenci Performansı Etkileşimleri

Grup	Puanlayıcı	Öğrenci Performansı	Gözlenen Puan	Beklenen Puan	Yanlılık (logit)	Standart Hata	t
KÖNT	P1	5	39,00	44,77	-1,32	0,43	-3,06
KÖNT	P2	1	42,00	46,07	-1,13	0,48	-2,38
KÖNT	P3	6	33,00	40,27	-1,24	0,39	-3,18
KÖNT	P4	2	45,00	39,20	1,35	0,55	2,46
KÖNT	P5	3	48,00	43,63	1,61	0,78	2,08
KÖNT	P6	9	13,00	20,40	-1,96	0,67	-2,94
KÖNT	P7	8	21,00	27,52	-1,06	0,43	-2,49
KÖNT	P8	1	39,00	44,17	-1,15	0,43	-2,67

χ^2 (sd)=130,8 (80), $p < 0,05$

Tablo 30'da Kontrol grubunun ön uygulamasına ait puanlayıcıların ve performans gösteren öğrencilerin gözlenen, beklenen, yanlılık, standart hata ve t değerleri yer almaktadır. Gözlenen puan; öğrencilerin piyano çalmadaki sergilemiş oldukları performanslarına ilişkin

gerçek performansını temsil eden puanlardır. Beklenen puan ise ÇYRÖM tarafından tahmin edilen değerler veya puanlardır. Yanlılık (logit) değeri ise puanlayıcının öğrenci performansını değerlendirme sürecinde olduğundan fazla bol veya olduğundan fazla az puan vermesidir. Standart hata değeri ise ÇYRÖM ile tahmin edilecek ölçümün güvenilir ve doğru yorumlanmasına yardımcı olan istatistiksel bir kriterdir. t değeri istatistiksel testlerde, bir gözlemin beklenen değerden ne kadar saptığını ve bu sapmanın rastgele varyansa göre önemini ifade eden bir ölçüttür. Bu ölçütte eşik değer olarak (+/- 2) aralığı kabul edilmektedir.

Kontrol grubu ön uygulamasına göre puanlayıcı x öğrenci performansı etkileşimine ait 8 anlamlı etkileşim hesaplanmıştır. Bu gruba ait diğer etkileşim tablosunun tamamını sunmak mümkün olmadığı için Ek 9'da ayrıntılı olarak her bir puanlayıcıya ve öğrenciye ilişkin gözlenen puan, beklenen puan, yanlılık, standart hata, t değeri, serbestlik derecesi, olasılık iç ve dış uyumlara yer verilmiştir.

Bunlar P1, P2, P3, P4, P5, P6, P7, P8 şeklindedir. Dolayısıyla sekiz puanlayıcının puanlama davranışları dikkate alınarak yorumlanmıştır.

P1'in puanlama davranışı incelendiğinde; öğrenci performanslarından 5 numaralıya değerlendirme esnasında farklılaşan katılık davranışlarını gösterdiği görülmüştür. Örneğin P1'in, 5 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($39 < 44,77$) dolayı farklılaşan katılık davranışını sergilemiştir.

P2'nin puanlama davranışı incelendiğinde; öğrenci performanslarından 1 numaralıya değerlendirme esnasında farklılaşan katılık davranışlarını gösterdiği görülmüştür. Örneğin P2'in, 1 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($42 < 46,07$) dolayı farklılaşan katılık davranışını sergilemiştir.

P3'in puanlama davranışı incelendiğinde; öğrenci performanslarından 6 numaralıya değerlendirme esnasında farklılaşan katılık davranışlarını gösterdiği görülmüştür. Örneğin P3'in, 6 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($33 < 40,27$) dolayı farklılaşan katılık davranışını sergilemiştir.

P4'ün puanlama davranışı incelendiğinde; öğrenci performanslarından 2 numaralıya değerlendirme esnasında farklılaşan cömertlik davranışlarını gösterdiği görülmüştür. Örneğin P4'ün, 2 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($45 > 39,20$) dolayı farklılaşan cömertlik davranışını sergilemiştir.

P5'in puanlama davranışı incelendiğinde; öğrenci performanslarından 3 numaralıya değerlendirme esnasında farklılaşan cömertlik davranışlarını gösterdiği görülmüştür. Örneğin P5'in, 3 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($48 > 43,63$) dolayı farklılaşan cömertlik davranışını sergilemiştir.

P6'nın puanlama davranışı incelendiğinde; öğrenci performanslarından 9 numaralıya değerlendirme esnasında farklılaşan katılık davranışlarını gösterdiği görülmüştür. Örneğin P6'in, 9 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($13 < 20,40$) dolayı farklılaşan katılık davranışını sergilemiştir.

P7'nin puanlama davranışı incelendiğinde; öğrenci performanslarından 8 numaralıya değerlendirme esnasında farklılaşan katılık davranışlarını gösterdiği görülmüştür. Örneğin P7'nin, 8 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($21 < 27,52$) dolayı farklılaşan katılık davranışını sergilemiştir.

P8'in puanlama davranışı incelendiğinde; öğrenci performanslarından 1 numaralıya değerlendirme esnasında farklılaşan katılık davranışlarını gösterdiği görülmüştür. Örneğin P7'nin, 8 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($21 < 27,52$) dolayı farklılaşan katılık davranışını sergilemiştir.

Dolayısıyla tablo 30'daki Kontrol ön uygulama *puanlayıcı x öğrenci performansı* etkileşimleri özetlendiğinde P2 ile P8 nolu puanlayıcı birinci numaralı öğrenci performansını puanlamışlar. Buna göre P2 Ö1'e performansını değerlendirirken katı; P8 de ise Ö1'in performansını değerlendirirken katı davranış sergilemiştir.

Özetle, araştırmanın ikinci araştırma sorusuna ilişkin Deney A, Deney B ve Kontrol gruplarındaki toplamdaki 25 puanlayıcının öğrenci performanslarını değerlendirilirken puanlayıcı eğitimi öncesinde grup düzeyinde yanlışlık davranışının olduğu belirlenmiştir. Deney A grubundaki puanlayıcıdan kaynaklı üç yanlışlık (P1, P5 ve P10); Deney B grubundaki iki puanlayıcıdan kaynaklı yanlışlık (P4 ve P6) ve Kontrol grubunda ise sadece bir puanlayıcıdan kaynaklı yanlışlık (P4) davranışı belirlenmiştir. Mevcut çalışmanın sonuçları literatürde yer alan diğer çalışmalarla benzerlik göstermektedir. Performans değerlendirmedeki yanlışlık davranışının tespit edilmesi ile ilgili gerçekleştirilen çalışmalar; Şata (2019) tarafından yapılan performans değerlendirme sürecinde ölçümlere karışan problemli puanlayıcı davranışlarının belirlenerek, puanlayıcı eğitiminin bu problemler üzerindeki etkisi araştırılmıştır. Araştırma sonucunda puanlayıcı eğitimi öncesi deney ve kontrol gruplarında yer alan 45 puanlayıcının farklılaşan puanlayıcı cömertliği ile farklılaşan

puanlayıcı katılığı gibi davranışları sergiledikleri bulunmuştur. Benzer bir çalışmada ise; Esfandiari, (2021) tarafından yapılan çalışmada İngilizce öğretmen değerlendiricilerin değerlendirme ölçütlerine ilişkin yanlılık derecesinde ne tür farklılık olduğunu araştırmıştır. Araştırmada 150 İngilizce dil öğretmeni, öğrenciler tarafından yazılan ve beş paragraftan oluşan 10 makaleyiiki farklı derecelendirme ölçeği kullanarak değerlendirmiştir. Araştırma bulgularına göre, bazı öğretmenler makaleyi puanlarken değerlendirme ölçütlerinde yanlı davrandığı bulunmuştur. Öç (2024) yaptığı çalışmada hem Türkçe hem de Matematik puanlayıcı grubundaki öğretmen adaylarının grup düzeyinde yanlılık davranışının olmadığı, fakat bireysel düzeyde yanlılık davranışının ortaya çıktığı bulunmuştur.

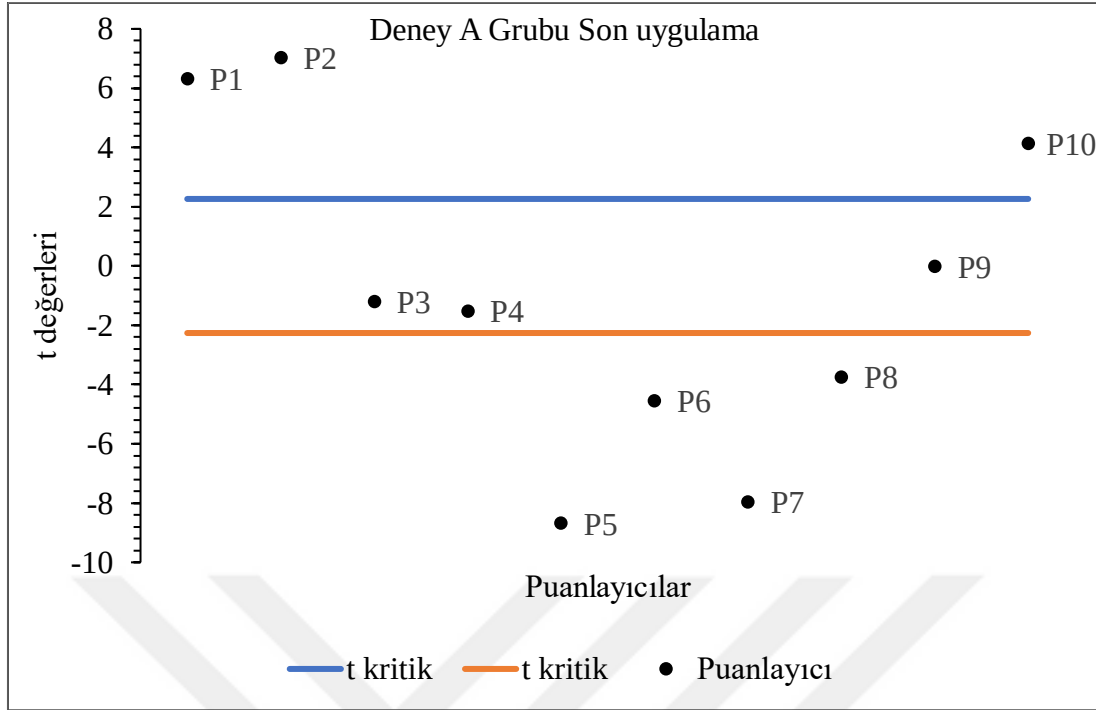
Üçüncü araştırma sorusuna ilişkin araştırma bulguları

Katılık ve Cömertlik, Halo Etkisi ile Yanlılık analizleri

a) *Deney A, Deney B ve Kontrol grubu son uygulamada yer alan puanlayıcıların katılık ve cömertlik davranışları nedir?*

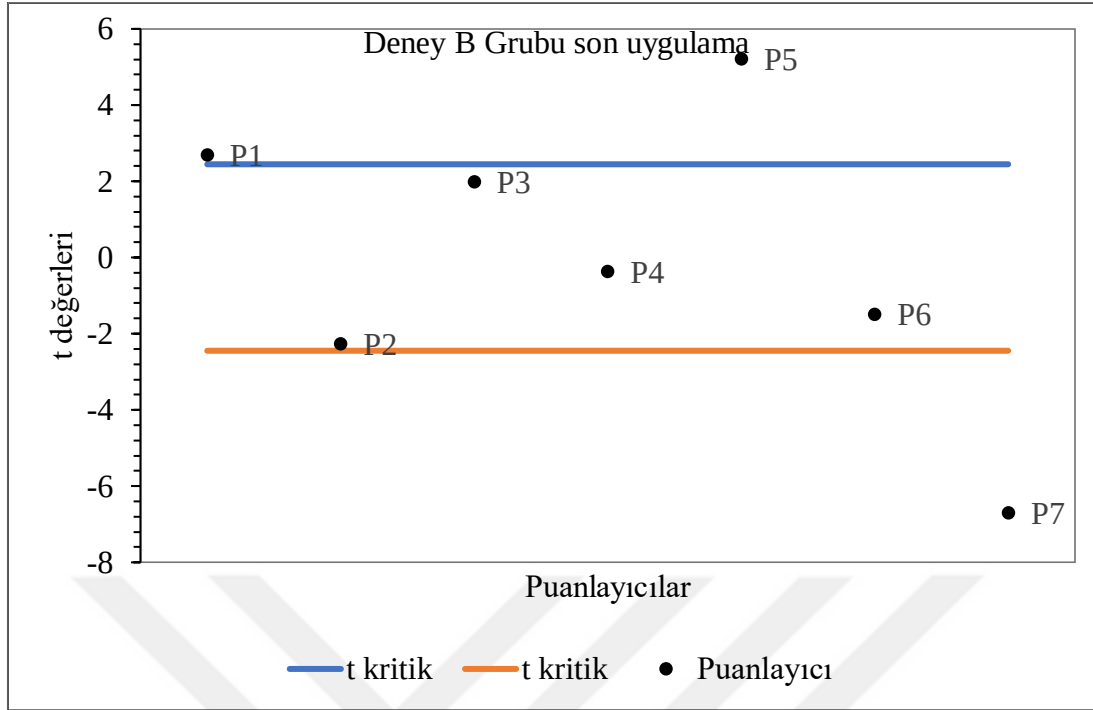
İlgili araştırmanın sorusuna göre Deney A, Deney B ve Kontrol grubu son uygulamada yer alan puanlayıcıların öncelikle grup düzeyinde katılık ve cömertlik durumları incelenmiştir. Bireysel düzeyde puanlayıcı davranışları Şekil 5, Şekil 6 ve Şekil 7’de verilmiştir.

Grup düzeyinde katılık ve cömertlik davranışı incelendiğinde Deney A, Deney B ve Kontrol son uygulama gruplarında ilgili istatistiklerin yüksek olduğu gözlenmiştir. Bu durumun istatistiksel açıdan anlamlılığına bakmak için sabit etkili ki kare değerleri incelenmiş ve bu değerlerin tüm gruplar için anlamlı olduğu görülmüştür. Sonuç olarak Deney A (Tablo 13), Deney B (Tablo 19) ve Kontrol (Tablo 25) son uygulama gruplarında grup düzeyinde katılık ve cömertlik davranışının olduğu söylenebilir.



Şekil 5. Deney A Grubu son uygulama t değerleri

Şekil 5 incelendiğinde Deney A grubu son uygulamada yer alan puanlayıcılardan P1, P2 ve P10 numaralı puanlayıcılarda cömertlik, P5, P6, P7 ve P8 numaralı puanlayıcılarda da katılık davranışı olduğu görülmektedir. Öğrencilerin performanslarının değerlendirilmesi için Deney A grubunun son uygulamasında yer alan 10 puanlayıcının ölçülmek istenilen yapıyı değerlendirirken puanlama davranışlarının puanlamaya (katılık ve cömertlik) karışıp karışmadığı incelenmiştir. Buna göre üç (%30) puanlayıcının cömert, dört (%40) puanlayıcının ise katı davranış göstermiştir. Puanlayıcı katılığı ve cömertliği davranışında bulunmayan üç (%30) puanlayıcı mevcuttur. Aynı Deney A grubunun ön uygulamadaki puanlayıcıların cömertlik davranışları ile son uygulamadaki davranışlar karşılaştırıldığında P7 ve P9 cömert puanlama davranışında bulunurken; son uygulamada ise sadece P7 katı davranış sergilemiştir. P9 ise herhangi bir katılık veya cömertlik davranışta bulunmayıp istenilen davranışta puanlama yapmıştır. Deney A ön uygulamada sadece iki puanlayıcı cömertlik davranışta bulunmuşken; son uygulamada ise cömert davranışta bulunan puanlayıcı sayısı artmış durumdadır. Bunlar P1, P2 ve P3 nolu puanlayıcılardır. Benzer şekilde Deney A ön ve son uygulamada yer alan puanlayıcıların katılık davranışı incelendiğinde ön uygulamada P1, P5 ve P6 nolu puanlayıcılar katılık davranışta bulunmuşken; son uygulamada ise P5, P6, P7, P8 nolu puanlayıcılarda puanlama davranışlarında katılık görülmüştür. Buna göre her iki uygulamada yer alan P5 ile P6 nolu puanlayıcılarda Referans Çerçevesi ile Puanlayıcı Hatası eğitiminin yeterli düzeyde etkili olmadığı şeklinde yorumlanabilir.



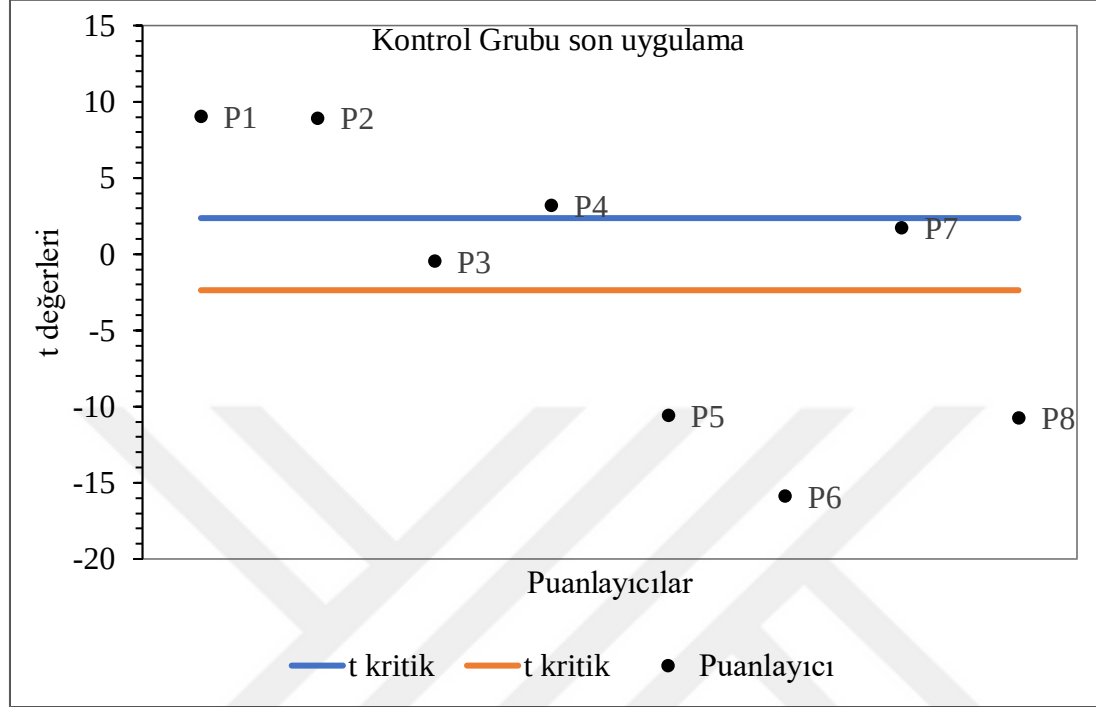
Şekil 6. Deney B grubu son uygulama t değerleri

Şekil 6 incelendiğinde Deney B grubu son uygulamada yer alan puanlayıcılardan P1 ve P5 numaralı puanlayıcılarda cömertlik, P7 numaralı puanlayıcıda da katılık davranışı olduğu görülmektedir.

Öğrencilerin performanslarının değerlendirilmesi için Deney B grubunun son uygulamasında yer alan 7 puanlayıcının ölçülmek istenilen yapıyı değerlendirirken puanlama davranışlarının puanlamaya (katılık ve cömertlik) karışıp karışmadığı incelenmiştir. Buna göre iki (%28, 57) puanlayıcının cömert, bir (%14, 28) puanlayıcının ise katı davranış göstermiştir. Puanlayıcı katılığı ve cömertliği davranışında bulunmayan dört (%57,14) puanlayıcı mevcuttur.

Aynı Deney B grubunun ön uygulamadaki puanlayıcıların cömertlik davranışları ile son uygulamadaki davranışlar karşılaştırıldığında P4 ve P7 cömert puanlama davranışında bulunurken; son uygulamada ise sadece P1 ve P5 katı davranış sergilemiştir. Katılık açısından incelendiğinde ise P7 nolu puanlayıcı ön uygulamada cömert davranış göstermişken son uygulamada ise katı davranış sergilemiştir. Yine ön uygulamada katı puanlama davranışında olup (P2, P3 ve P6) puanlayıcılar son uygulamada istenilen puanlama davranışında bulunmuşlar. Herhangi bir puanlama davranışında (katılık ve cömertlik) bulunmayan puanlayıcı ön uygulamada iki puanlayıcı (%28,57) olmuşken son uygulamada ise bu oran dört puanlayıcıya artmıştır (%57,14). Dolayısıyla Deney B grubu için verilen Referans Çerçevesi ile Performans Boyutu eğitimlerinin birey bazında etkili olduğu sonucuna varılabilir. Burada elde edilen bulgular yönelik

çeşitli çalışmalarda da puanlayıcılar tarafından segilendiği görülmektedir (Abu Kassım, 2007; Farrokhi, Esfandiari ve Schaefer, 2012; Karakaya, 2015; Çetin ve İlhan, 2017; Knock vd., 2018; Şata, 2019; Esfandiari, 2021; Görgülü Öztürk, 2023; Öç, 2024).



Şekil 7. Kontrol Grubu son uygulama t değerleri

Şekil 7 incelendiğinde Kontrol grubu son uygulamada yer alan puanlayıcılardan P1, P2 ve P4 numaralı puanlayıcılarda cömertlik, P5, P6 ve P8 numaralı puanlayıcılarda da katılık davranışı olduğu görülmektedir.

Öğrencilerin performanslarının değerlendirilmesi için Kontrol grubunun son uygulamasında yer alan 8 puanlayıcının ölçülmek istenilen yapıyı değerlendirirken puanlama davranışlarının puanlamaya (katılık ve cömertlik) karışıp karışmadığı incelenmiştir. Buna göre üç (%37,5) puanlayıcının cömert, üç (%37,5) puanlayıcının ise katı davranış göstermiştir. Puanlayıcı katılımı ve cömertliği davranışında bulunmayan iki (%25) puanlayıcı mevcuttur.

Aynı Kontrol grubunun ön uygulamadaki puanlayıcıların cömertlik davranışları ile son uygulamadaki davranışlar karşılaştırıldığında P1, P2 ve P3 cömert puanlama davranışında bulunurken; son uygulamada ise bu sayı aynı kalmış sadece P3'ün yerine P4 katı puanlama davranışı sergilemiştir. Katılık açısından incelendiğinde her iki uygulamada P5, P6 ve P8 nolu puanlayıcılar aynı katılık davranışını sürdürmüşlerdir. Her iki uygulamada puanlayıcı katılımı ve cömertliği davranışında bulunmayan iki (%25) puanlayıcı mevcuttur. Bu durum istenilen bir

durumdur. Çünkü kontrol grubunda herhangi bir müdahale veya eğitim söz konusu olmadığı için davranışların değişikliği beklenmemektedir.

a) *Deney A, Deney B ve Kontrol grubu son uygulamada yer alan puanlayıcıların halo etkisi davranışları nedir?*

Puanlayıcılarda grup düzeyinde halo etkisi olup olmadığını incelemek amacıyla tüm gruplar için ölçüt yüzeyine ait ölçüm sonuçları incelenmiştir. Deney A Grubu, Deney B Grubu ve Kontrol Grubu son uygulama ölçüt yüzeyine ait ölçüm sonuçları incelendiğinde ayırma oranı (5,85; 2,00; 3,93) ayırma indeksi (8,13; 3,00; 5,57) ve güvenirlik (0,97; 0,80; 0,94) değerlerinin yüksek olduğu ve sabit etkili ki kare istatistiklerinin anlamlı olduğu ($p<0,01$) belirlenmiştir. Bu durum Deney A, Deney B ve Kontrol Grubu son uygulamada puanlayıcılarda grup düzeyinde halo etkisinin olmadığını göstergesidir.

Bireysel düzeyde halo etkisinin incelenmesi için madde güçlüklerine bakılmalıdır. Deney A, Deney B ve Kontrol Grubu son uygulamada (4,80; 1,46; 2,91) yer alan madde güçlükleri yüksektir. Bu yüzden puanlayıcıların uygunluk istatistiklerinden 1'den çok büyük olanda halo etkisi vardır (Myford & Wolfe, 2004). Tüm grupların son uygulama puanlayıcı yüzeyine ait sonuçlara bakıldığında Deney A grubunda P8 numaralı puanlayıcıda, Deney B grubunda P2 ve P5 numaralı puanlayıcılarda ve Kontrol grubunda P1 ve P4 numaralı puanlayıcılarda halo etkisi davranışı belirlenmiştir.

Halo etkisi davranışının diğer belirlenme yöntemi de madde güçlüklerini eşitleyip mükemmel uyum sağlayan puanlayıcıların belirlenmesidir (Linacre, 2023, s. 360). Tablo 31'de madde güçlükleri eşitlendikten sonraki uyum istatistikleri verilmiştir.

Tablo 31

Madde Güçlükleri Eşitlendikten Sonra Uygun İstatistikleri

Grup	Puanlayıcılar	Uygunluk içi	Uygunluk dışı
Deney A	P1	1,00	1,00
	P2	1,15	1,02
	P3	1,26	1,38
	P4	0,90	0,83
	P5	0,67	0,67
	P6	1,27	1,21
	P7	0,96	0,98
	P8	1,10	1,02
	P9	1,15	1,04
	P10	0,75	0,64
Deney B	P1	0,56	0,57
	P2	1,43	1,36
	P3	0,79	0,80
	P4	0,90	0,89
	P5	1,13	1,16
	P6	1,23	1,25
	P7	0,89	0,92
Kontrol	P1	1,73	1,40
	P2	0,96	0,94
	P3	0,79	0,78
	P4	1,66	1,79
	P5	0,50	0,50
	P6	1,28	1,28
	P7	0,68	0,68
	P8	1,00	1,01

Tablo 31 incelendiğinde Deney A son uygulama grubunda P1 numaralı puanlayıcıda, Kontrol son uygulama grubunda P8 numaralı puanlayıcıda halo etkisi yer alırken Deney B son uygulama grubunda halo etkisi yer almadığı bulunmuştur.

Sonuç olarak tüm grupların son uygulama puanlayıcı yüzeyine ait sonuçlara bakıldığında Deney A Referans çerçevesi eğitimi ile Puanlayıcı Hatası eğitimi alan gruptaki 10 puanlayıcıdan ikisinde P1 ve P8 numaralı puanlayıcılarda, Deney B Referans çerçevesi eğitimi ile Performans Boyutu eğitimlerini alan grupta 7 puanlayıcının ikisinde P2 ve P5 numaralı puanlayıcılarda ve Kontrol grubundaki 8 puanlayıcının üçünde P1, P4 ve P8 numaralı puanlayıcılarda halo etkisi davranışı belirlenmiştir. Uygulama öncesi ve sonrasındaki üç grubun durumu ele alındığında öncelikle Deney A grubu ön uygulamada P1 ile P8 nolu puanlayıcıların öğrenci performansını değerlendirirken puanlama davranışlarına halo davranışı karışmıştır. Diğer bir ifade ile Deney A grubuna uygulama öncesinde ve sonrasındaki halo davranışını gösteren puanlayıcı aynı değildir. Dolayısıyla Deney A grubuna verilen Referans Çerçevesi ile Puanlayıcı Hatası eğitimi etkili olmamıştır.

Alanyazında bu sonucu destekleyen çalışmalardanbiri de İlhan (2015) tarafından yazılan Standart ve Solo Taksonomisine Dayalı Rubrikler ile Performans Değerlendirmelerin karşılaştırılması başlıklı çalışmasında puanlayıcılar arasında anlamlı farklar bulunduğu ve puanlayıcı güvenilirliğinin düşük olduğu tespit edilmiştir. Bu durum, Referasn Çerçevesi eğitiminin puanlayıcı hatalarını azaltmada her zaman yeterli olmadığını göstermektedir.

Deney B grubundaki ön uygulamada 7 puanlayıcıdan sadece P6 nolu puanlayıcıda halo davranışı mevcut iken, son uygulamada ise yine 7 puanlayıcıdan P2 ile P5 nolu punalayıcı davranışında da halo davranışı görülmüştür. Alanyazındaki kaynaklar incelendiğinde performans değerlendirme sürecine halo davranışının karışabileceğine ilişkin çalışmalar mevcuttur (Engelhard, 1994; Esfandiari, 2015). Dolayısıyla Referans Çerçevesi ile Performans Boyutu eğitimi alındıktan sonra punalyacıların halo davranışlarında çok az artış görülmüştür. Bu bulgu Öç (2024) tarafından yapılan çalışmada deney grubundaki Türkçe açık uçlu soruların değerlendirilmesi sürecinde son test ölçümlerine ait halo etkisi ile diğer puanlayıcılardan kaynaklı davranışları incelenmiştir. İnceleme sonucunda halo etkisi grup düzeyinde ortaya çıkmadığı, ancak bireysel düzeyde ortaya çıktığı bulunmuştur.

Kontrol grubundaki durum ele alındığında ise 8 puanlayıcıdan P2, P4, P6 ve P8 numaralı puanlayıcılarda halo etkisi davranışı belirlenmişken; son uygulamada ise P1, P4 ve P8 nolu puanlayıcılarda halo davranışları görülmüştür. Kontrol grubundaki davranışların uygulama öncesi veya sonrasında çok büyük bir miktarda olmaması beklen bir durumdur.

c) *Deney A, Deney B ve Kontrol grubu son uygulamada yer alan puanlayıcıların farklılaşan katılık ve cömertlik (yanlılık) davranışları nedir?*

Deney A son uygulama puanlayıcı x öğrenci performansı anlamlı etkileşimleri Tablo 32’de verilmiştir.

Tablo 32

Deney A Son Uygulama Puanlayıcı x Öğrenci Performansı Etkileşimleri

Grup	Puanlayıcı	Öğrenci Performansı	Gözlenen Puan	Beklenen Puan	Yanlılık (logit)	Standart Hata	t
DAST	P1	2	44,00	48,44	-1,72	0,51	-3,39
DAST	P3	10	39,00	46,61	-1,90	0,45	-4,26
DAST	P4	4	37,00	41,10	-1,37	-1,41	-3,16
DAST	P5	10	48,00	42,73	1,75	0,76	2,30
DAST	P6	7	32,00	38,48	-1,29	0,46	-2,80
DAST	P6	4	34,00	38,96	-0,97	0,45	-2,18
DAST	P6	9	37,00	32,11	0,97	0,44	2,20
DAST	P7	1	36,00	42,06	-1,23	0,44	-2,78
DAST	P7	9	34,00	29,78	0,90	0,45	2,01
DAST	P8	8	29,00	39,43	-2,15	0,50	-4,34
DAST	P8	9	28,00	32,68	-1,07	0,52	-2,08
DAST	P8	7	44,00	39,07	1,09	0,51	2,14
DAST	P9	4	47,00	42,19	1,39	0,64	2,15
DAST	P10	2	44,00	47,18	-1,03	0,51	-2,03

χ^2 (sd)=187,6 (100), $p < 0,05$

Tablo 32’de Deney A grubunun son uygulamasına ait puanlayıcıların ve performans gösteren öğrencilerin gözlenen, beklenen, yanlılık, standart hata ve t değerleri yer almaktadır. Gözlenen puan; öğrencilerin piyano çalmadaki sergilemiş oldukları performanslarına ilişkin gerçek performansını temsil eden puanlardır. Beklenen puan ise ÇYRÖM tarafından tahmin edilen değerler veya puanlardır. Yanlılık (logit) değeri ise puanlayıcının öğrenci performansını değerlendirme sürecinde olduğundan fazla bol veya olduğundan fazla az puan vermesidir. Standart hata değeri ise ÇYRÖM ile tahmin edilecek ölçümün güvenilir ve doğru yorumlanmasına yardımcı olan istatistiksel bir kriterdir. t değeri istatistiksel testlerde, bir gözlemin beklenen değerden ne kadar saptığını ve bu sapmanın rastgele varyansa göre önemini ifade eden bir ölçüttür. Bu ölçütte eşik değer olarak (+/- 2) aralığı kabul edilmektedir.

Deney A grubu son uygulamasına göre puanlayıcı x öğrenci performansı etkileşimine ait 14 anlamlı etkileşim hesaplanmıştır. Bu gruba ait diğer etkileşim tablosunun tamamını sunmak mümkün olmadığı için Ek 6’da ayrıntılı olarak her bir puanlayıcıya ve öğrenciye ilişkin gözlenen puan, beklenen puan, yanlılık, standart hata, t değeri, serbestlik derecesi, olasılık iç ve dış uyumlara yer verilmiştir.

Buna göre puanlayıcılar öğrenci performanslarını değerlendirirken (P1, P3, P4, P5, P6, P7, P8, P9 ve P10) hem farklılaşan katılık hem de farklılaşan cömertlik davranışlarını sergilemişlerdir. Dolayısıyla dokuz puanlayıcının puanlama davranışları dikkate alınarak yorumlanmıştır.

P1'in puanlama davranışı incelendiğinde; öğrenci performansından 2 numarayı değerlendirme esnasında farklılaşan katılık davranışı gösterdiği görülmüştür. Örneğin P1'in, 2 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($44 < 48,44$) farklılaşan katılık davranışını sergilemiştir.

P3'in puanlama davranışı incelendiğinde; öğrenci performansından 10 numarayı değerlendirme esnasında farklılaşan katılık davranışı gösterdiği görülmüştür. Örneğin P3'in, 10 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($39 < 46,61$) farklılaşan katılık davranışını sergilemiştir.

P4'in puanlama davranışı incelendiğinde; öğrenci performansından 4 numarayı değerlendirme esnasında farklılaşan katılık davranışı gösterdiği görülmüştür. Örneğin P4'ün, 4 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($37 < 41,10$) farklılaşan katılık davranışını sergilemiştir.

P5'in puanlama davranışı incelendiğinde; öğrenci performansından 10 numarayı değerlendirme esnasında farklılaşan cömertlik davranışı gösterdiği görülmüştür. Örneğin P5'in, 10 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($48 > 42,72$) farklılaşan cömertlik davranışını sergilemiştir.

P6'nın puanlama davranışı incelendiğinde; öğrenci performansından 7 numarayı değerlendirme esnasında farklılaşan katılık davranışı gösterdiği görülmüştür. Örneğin P6'in, 7 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($32 < 38,48$) farklılaşan katılık davranışını sergilemiştir. Benzer şekilde aynı puanlayıcının P6'nın 4 numaralı öğrenci performansı dikkate alındığında da gözlenen puanı beklenen puandan küçük olmasından ($34 < 38,96$) dolayı farklılaşan katılık davranışını sergilemiştir. Aynı puanlayıcının P6'nın 9 numaralı öğrencinin performansına gözlenen puanı beklenen puandan büyük şeklinde puan vermiş ($37 > 32,11$) dolayısıyla farklılaşan cömertlik davranışını sergilemiştir.

P7'nin puanlama davranışı incelendiğinde; öğrenci performansından 1 numarayı değerlendirme esnasında farklılaşan katılık davranışı gösterdiği görülmüştür. Örneğin P7'nin, 1 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($36 < 42,06$) farklılaşan katılık davranışını sergilemiştir. Benzer şekilde aynı puanlayıcının P7'nin 9 numaralı öğrenci performansı dikkate alındığında da gözlenen puanı beklenen puandan büyük olmasından ($34 > 29,78$) dolayı farklılaşan cömertlik davranışını sergilemiştir.

P8'in puanlama davranışı incelendiğinde; öğrenci performansından 8 numarayı değerlendirme esnasında farklılaşan katılık davranışı gösterdiği görülmüştür. Örneğin P8'in, 8 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($29 > 39,43$) farklılaşan katılık davranışını sergilemiştir. Benzer şekilde aynı puanlayıcının P8'in 9 numaralı öğrenci performansı dikkate alındığında da gözlenen puanı beklenen puandan küçük olmasından ($28 < 32,68$) dolayı farklılaşan katılık davranışını sergilemiştir. Aynı puanlayıcının P8'in 7 numaralı öğrencinin performansına gözlenen puanı beklenen puandan büyük şeklinde puan vermiş ($44 > 39,07$) dolayısıyla farklılaşan cömertlik davranışını sergilemiştir.

P9'in puanlama davranışı incelendiğinde; öğrenci performansından 4 numarayı değerlendirme esnasında farklılaşan cömertlik davranışı gösterdiği görülmüştür. Örneğin P9'in, 4 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan büyük olduğu için ($47 > 42,19$) farklılaşan cömertlik davranışını sergilemiştir.

P10'ın puanlama davranışı incelendiğinde; öğrenci performansından 2 numarayı değerlendirme esnasında farklılaşan katılık davranışı gösterdiği görülmüştür. Örneğin P10'ın, 2 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($44 < 47,18$) farklılaşan katılık davranışını sergilemiştir.

Dolayısıyla tablo 32'deki Deney A son uygulama *puanlayıcı x öğrenci performansı* etkileşimleri özetlendiğinde diğer puanlayıcılardan farklı olarak üç defa etkileşim yapan P6 ile P8 ve iki defa etkileşim yapan P7 nolu puanlayıcılardır. P6 nolu puanlayıcı 4, 7 ve 9 nolu öğrenci performanslarını değerlendirirken iki katı ve bir cömert puanlama davranışı sergilemiştir. P8 nolu puanlayıcı da 7, 8 ve 9 nolu öğrenci performanslarını değerlendirirken iki katı ve bir cömert puanlama davranışında bulunmuştur. Benzer şekilde P7 nolu puanlayıcı da 1 ve 9 nolu öğrenci performanslarını değerlendirirken bir katı ve bir cömert puanlama davranışında bulunmuştur.

Her üç puanlayıcı 9 nolu öğrenci performansını değerlendirmişler ve P6 ile P7 9 nolu öğrenci performansını puanlarken cömert davranış ve P7 ise katı davranış göstermişlerdir.

Deney A ön ve son uygulama *puanlayıcı x öğrenci performansı* etkileşimleri özetlendiğinde, her iki aşamada da bazı puanlayıcıların diğerlerine kıyasla daha fazla etkileşimde bulunduğu görülmektedir. Son uygulamada P6, P7 ve P8 öne çıkarken, ön uygulamada P1 ve P5 daha fazla etkileşim sergilemiştir. Her iki aşamada da bazı puanlayıcılar hem katı hem de cömert puanlama davranışları göstermiş, ancak ön

uygulamada P1 ve P5 özellikle 1 nolu öğrenci performansı üzerinde zıt değerlendirmeler yapmıştır, benzer şekilde son uygulamada da P6 ve P7 9 nolu öğrenciye farklı tutumlar sergilemiştir. Bu durum, puanlayıcıların tutarlılığını ve değerlendirme süreçlerindeki değişkenliği göstermektedir.

Deney B son uygulama puanlayıcı x öğrenci performansı anlamlı etkileşimleri Tablo 33’de verilmiştir.

Tablo 33

Deney B Son Uygulama Puanlayıcı x Öğrenci Performansı Etkileşimleri

Grup	Puanlayıcı	Öğrenci Performansı	Gözlenen Puan	Beklenen Puan	Yanlılık (logit)	Standart Hata	t
DBST	P1	10	39,00	42,81	-1,15	0,54	-2,14
DBST	P2	6	37,00	41,15	-1,20	0,52	-2,31
DBST	P2	10	45,00	40,13	1,54	0,59	2,60
DBST	P3	1	40,00	43,55	-1,09	0,54	-2,01

χ^2 (sd)=84,0 (70), $p > 0,05$

Tablo 33’de Deney B grubunun son uygulamasına ait puanlayıcıların ve performans gösteren öğrencilerin gözlenen, beklenen, yanlılık, standart hata ve t değerleri yer almaktadır. Gözlenen puan; öğrencilerin piyano çalmadaki sergilemiş oldukları performanslarına ilişkin gerçek performansını temsil eden puanlardır. Beklenen puan ise ÇYRÖM tarafından tahmin edilen değerler veya puanlardır. Yanlılık (logit) değeri ise puanlayıcının öğrenci performansını değerlendirme sürecinde olduğundan fazla bol veya olduğundan fazla az puan vermesidir. Standart hata değeri ise ÇYRÖM ile tahmin edilecek ölçümün güvenilir ve doğru yorumlanmasına yardımcı olan istatistiksel bir kriterdir. T değeri istatistiksel testlerde, bir gözlemin beklenen değerden ne kadar saptığını ve bu sapmanın rastgele varyansa göre önemini ifade eden bir ölçüttür. Bu ölçütte eşik değer olarak (+/- 2) aralığı kabul edilmektedir.

Deney B grubu son uygulamasına göre *puanlayıcı x öğrenci* performansı etkileşimine ait 4 anlamlı etkileşim hesaplanmıştır. Bu gruba ait diğer etkileşim tablosunun tamamını sunmak mümkün olmadığı için Ek 8’de ayrıntılı olarak her bir puanlayıcıya ve öğrenciye ilişkin gözlenen puan, beklenen puan, yanlılık, standart hata, t değeri, serbestlik derecesi, olasılık iç ve dış uyumlara yer verilmiştir.

Buna göre puanlayıcılardan P2’de hem farklılaşan katılık hem de farklılaşan cömertlik davranışlarını sergilemiştir.

P2'nin puanlama davranışı incelendiğinde; öğrenci performanslarından 6 numaralıya değerlendirme esnasında farklılaşan katılık ve 10 numaralı öğrenci performansına da farklılaşan cömertlik davranışlarını gösterdiği görülmüştür. Örneğin P2'nin, 6 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($37 < 41,15$) benzer şekilde 10 numaralı öğrenci performansı dikkate alındığında da gözlenen puanı beklenen puandan büyük olmasından ($45 > 40,13$) dolayı farklılaşan cömertlik davranışını sergilemiştir.

Sadece farklılaşan katılık davranışını sergileyen puanlayıcılar ise P1 ile P3'tür. Buna göre P1'in puanlama davranışı incelendiğinde; öğrenci performanslarından 10 numaralı öğrenci performansını değerlendirme esnasında farklılaşan katılık davranışlarını gösterdiği görülmüştür. Örneğin P1'in, 10 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($39 < 42,81$) farklılaşan katılık davranışını sergilemiştir.

P3'ün puanlama davranışı incelendiğinde; öğrenci performanslarından 1 numaralı öğrenci performansını değerlendirme esnasında farklılaşan katılık davranışlarını gösterdiği görülmüştür. Örneğin P3'ün, 1 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($40 < 43,55$) farklılaşan katılık davranışını sergilemiştir. Dolayısıyla 4 anlamlı etkileşim içerisinde en çok yanlı davranan puanlayıcı P2 olmuş ve aynı zamanda bu puanlayıcı 6 ve 10 nolu öğrenci performanslarını değerlendirirken yanlı davranış göstermiştir. Deney B ön ve son uygulamalarda *puanlayıcı x öğrenci performansı* etkileşimleri özetlendiğinde, P2 en çok yanlı davranış sergileyen puanlayıcı olarak öne çıkarken, özellikle 6 ve 10 nolu öğrenci performanslarında taraflı puanlama yapmıştır. Ön uygulamada ise P4 ve P6 daha fazla etkileşimde bulunmuş, ancak farklı değerlendirme eğilimleri göstermiştir. P4, 1,8 ve 10 nolu öğrenci performanslarını değerlendirirken her seferinde katı davranış sergilemiş, buna karşın P6, 4 ve 10 nolu öğrenci performanslarını değerlendirirken cömert puanlama eğiliminde olmuştur. Özellikle 10 nolu öğrenci performansında P4'ün katı, P6'nın ise cömert davranış göstermesi, puanlayıcılar arasındaki tutarsızlığın önemli bir örneği olarak dikkat çekmektedir.

Kontrol son uygulama puanlayıcı x öğrenci performansı anlamlı etkileşimleri Tablo 34'te verilmiştir.

Tablo 34

Kontrol Son Uygulama Puanlayıcı x Öğrenci Performansı Etkileşimleri

Grup	Puanlayıcı	Öğrenci Performansı	Gözlenen Puan	Beklenen Puan	Yanlılık (logit)	Standart Hata	t
KSNT	P3	1	39,00	43,28	-1,05	0,47	-2,21
KSNT	P4	10	40,00	46,96	-2,06	0,48	-4,27
KSNT	P5	4	30,00	35,64	-1,13	0,44	-2,55
KSNT	P5	7	36,00	31,01	1,00	0,46	2,20
KSNT	P6	9	15,00	21,42	-1,60	0,56	-2,84
KSNT	P6	6	29,00	33,79	-0,95	0,44	-2,14
KSNT	P6	1	41,00	33,79	1,57	0,49	3,18
KSNT	P6	4	40,00	31,97	1,69	0,48	3,51

χ^2 (sd)= 136,7 (80), $p < 0,05$

Tablo 34'te kontrol grubunun son uygulamasına ait puanlayıcıların ve performans gösteren öğrencilerin gözlenen, beklenen, yanlılık, standart hata ve t değerleri yer almaktadır. Gözlenen puan; öğrencilerin piyano çalmadaki sergilemiş oldukları performanslarına ilişkin gerçek performansını temsil eden puanlardır. Beklenen puan ise ÇYRÖM tarafından tahmin edilen değerler veya puanlardır. Yanlılık (logit) değeri ise puanlayıcının öğrenci performansını değerlendirme sürecinde olduğundan fazla bol veya olduğundan fazla az puan vermesidir. Standart hata değeri ise ÇYRÖM ile tahmin edilecek ölçümün güvenilir ve doğru yorumlanmasına yardımcı olan istatistiksel bir kriterdir. t değeri istatistiksel testlerde, bir gözlemin beklenen değerden ne kadar saptığını ve bu sapmanın rastgele varyansa göre önemini ifade eden bir ölçüttür. Bu ölçütte eşik değer olarak (+/- 2) aralığı kabul edilmektedir.

Kontrol grubu son uygulamasına göre *puanlayıcı x öğrenci* performansı etkileşimine ait 8 anlamlı etkileşim hesaplanmıştır. Bu gruba ait diğer etkileşim tablosunun tamamını sunmak mümkün olmadığı için Ek 10'da ayrıntılı olarak her bir puanlayıcıya ve öğrenciye ilişkin gözlenen puan, beklenen puan, yanlılık, standart hata, t değeri, serbestlik derecesi, olasılık iç ve dış uyumlara yer verilmiştir.

Buna göre puanlayıcılardan (P5, P6) diğerlerinden farklı olarak hem farklılaşan katılık hem de farklılaşan cömertlik davranışlarını sergilemişlerdir. Buna göre puanlayıcılar öğrenci performanslarını değerlendirirken (P3, P4, P5, P6,) farklılaşan katılık davranışlarını sergilemişlerdir. Dolayısıyla sekiz puanlayıcının puanlama davranışları dikkate alınarak yorumlanmıştır.

P3'ün puanlama davranışı incelendiğinde; öğrenci performanslarından 1 numaralı öğrenci performansını değerlendirme esnasında farklılaşan katılık davranışlarını gösterdiği

görülmüştür. Örneğin P3'ün, 1 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($39 < 43,28$) farklılaşan katılık davranışını sergilemiştir.

P4'ün puanlama davranışı incelendiğinde; öğrenci performanslarından 10 numaralı öğrenci performansını değerlendirme esnasında farklılaşan katılık davranışlarını gösterdiği görülmüştür. Örneğin P4'ün, 10 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($40 < 46,96$) farklılaşan katılık davranışını sergilemiştir.

P5'in puanlama davranışı incelendiğinde; öğrenci performanslarından 4 numaralıya değerlendirme esnasında farklılaşan katılık ve 7 numaralı öğrenci performansına da farklılaşan cömertlik davranışlarını gösterdiği görülmüştür. Örneğin P5'in, 4 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($30 < 35,64$) benzer şekilde 7 numaralı öğrenci performansı dikkate alındığında da gözlenen puanı beklenen puandan büyük olmasından ($36 > 31,01$) dolayı farklılaşan cömertlik davranışını sergilemiştir.

P6'nın puanlama davranışı (katılık) incelendiğinde; öğrenci performanslarından 6 ve 9 numaralılara değerlendirme esnasında farklılaşan katılık ve 1 ile 4 numaralı öğrenci performanslarına da farklılaşan cömertlik davranışlarını gösterdiği görülmüştür. Örneğin P6'nın, 6 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan küçük olduğu için ($29 < 33,79$) benzer şekilde 9 numaralı öğrenci performansı dikkate alındığında da gözlenen puanı beklenen puandan küçük olmasından ($15 < 21,42$) dolayı farklılaşan katılık davranışını sergilemiştir.

P6'nın puanlama davranışı (cömertlik) incelendiğinde; öğrenci performanslarından 1 ve 4 numaralılara değerlendirme esnasında farklılaşan cömertlik davranışlarını gösterdiği görülmüştür. Örneğin P6'nın, 1 numaralı öğrenci performansına vermiş olduğu gözlenen puanı beklenen puandan büyük olduğu için ($41 > 33,79$) benzer şekilde 4 numaralı öğrenci performansı dikkate alındığında da gözlenen puanı beklenen puandan büyük olmasından ($40 > 31,97$) dolayı farklılaşan cömertlik davranışını sergilemiştir.

Dolayısıyla tablo 35'deki Kontrol son uygulama *puanlayıcı x öğrenci performansı* etkileşimleri özetlendiğinde diğer puanlayıcılardan farklı olarak üç defa etkileşim yapan P6 ve iki defa etkileşim yapan P5 nolu puanlayıcılardır. P6 nolu puanlayıcı 1, 6 ve 9 nolu öğrenci performanslarını değerlendirirken iki katı ve bir cömert puanlama davranışı sergilemiştir. P5

nolu puanlayıcı da 4, 7 nolu öğrenci performanslarını değerlendirirken bir katı ve bir cömert puanlama davranışında bulunmuştur.

Dolayısıyla Kontrol ön uygulama *puanlayıcı x öğrenci performansı* etkileşimleri özetlendiğinde P2 ile P8 nolu puanlayıcı birinci numaralı öğrenci performansını puanlamışlar. Buna göre P2 Ö1'e performansını değerlendirirken katı; P8 de ise Ö1'in performansını değerlendirirken katı davranış sergilemiştir.

Kontrol grubunun son ve ön uygulamalarında *puanlayıcı x öğrenci performansı* etkileşimleri özetlendiğinde, son uygulamada en fazla etkileşimde bulunan puanlayıcılar P6 ve P5 olurken, ön uygulamada P2 ve P8'in 1 nolu öğrenci performansında benzer tutumlar sergilediği görülmektedir. Son uygulamada P6, 1,6, 9 nolu öğrenci değerlendirirken genellikle katı bir tutum sergilemiş ancak bir değerlendirmede cömert davranmıştır. P5 ise daha az etkileşim göstermesine rağmen, 4 ve 7 nolu öğrencileri değerlendirirken hem katı hem de cömert puanlama yapmıştır. Ön uygulamada ise P2 ve P8, 1 nolu öğrenci performansına karşı tutarlı bir şekilde katı puanlama eğilimini göstermiştir. Bu durum, kontrol grubundaki puanlayıcıların zaman içinde puanlama tutumlarında kısmen tutarlı olduklarını ancak bazı farklılıklar gözserdiklerini ortaya koymaktadır.

BÖLÜM V

SONUÇLAR VE ÖNERİLER

Sonuçlar

Bu araştırmada, “Referans çerçevesi ile Puanlayıcı Hatası” ile “Referans Çerçevesi ile Performans Boyutu” Puanlayıcı eğitimleri kullanılarak verilen puanlayıcı eğitiminin; puanlayıcı davranışları üzerine etkisi incelenmiştir. Bu kapsamda bu bölümde araştırma sonucunda elde edilen sonuçlar ve elde edilen sonuçlar doğrultusunda yapılabilecek çalışmalar ile ilgili öneriler yer almaktadır. Araştırma sonuçları sunulurken araştırma sorularının sırasına göre sunulmuştur. Araştırmanın sonuçları şu şekildedir;

Araştırmada öğrenci performansı, ölçütler ve puanlayıcılar olmak üzere üç yüzey bulunduğu için her bir yüzeyin ölçüm sonuçları ve uyum istatistikleri incelenmiştir. Deney A, Deney B ve Kontrol grubu ön ve son uygulama sonuçları incelendiğinde tüm yüzeylerin model veri uyumunun iyi olduğu belirlenmiştir. Bu durum araştırmada elde edilen sonuçların geçerliğine ve güvenilirliğine katkı sağlamaktadır.

İkinci alt problemine ilişkin sonuçlar: Puanlayıcı eğitimi öncesinde puanlayıcıların katılık, cömertlik, halo etkisi ve yanlılık davranışlarının hem grup düzeyinde hem de bireysel düzeyde ortaya çıkma durumları incelendiğinde;

Deney A, Deney B ve Kontrol ön uygulama gruplarında grup düzeyinde katılık, cömertlik ve yanlılık davranışlarının olduğu belirlenmiştir.

Gruplarda grup düzeyinde halo etkisi davranışının olmadığı belirlenmiştir. Her bir grubun bireysel düzeyde puanlayıcı davranışları incelendiğinde Deney A ön uygulama grubunda iki cömert (P7 ve P9), üç katı (P1, P5, P6), iki halo etkisi (P3 ve P10) ve üç tane de yanlılık (P1, P5 ve P10) davranışlarının olduğu belirlenmiştir. Deney B grubu ön uygulama grubunda iki cömert (P4 ve P7), üç katı (P2, P3 ve P6), bir halo etkisi (P6) ve iki tane de yanlılık (P4 ve P6) davranışlarının olduğu belirlenmiştir. Kontrol grubu ön uygulama grubunda üç cömert

(P1, P2 ve P3), üç katı (P5, P6 ve P8), iki halo etkisi (P2 ve P4) ve bir tane de yanlışlık (P4) davranışı belirlenmiştir.

Üçüncü alt problemine ilişkin sonuçlar: Puanlayıcı eğitimi sonrasında puanlayıcıların katılık, cömertlik, halo etkisi ve yanlışlık davranışlarının hem grup düzeyinde hem de bireysel düzeyde ortaya çıkma durumları incelendiğinde;

Deney A ve Kontrol son uygulama gruplarında grup düzeyinde katılık, cömertlik ve yanlışlık davranışlarının olduğu belirlenirken gruplarda grup düzeyinde halo etkisi davranışının olmadığı belirlenmiştir. Deney B grubu son uygulama grubunda ise grup düzeyinde katılık ve cömertlik davranışları yer alırken halo etkisi ve yanlışlık davranışlarının yer almadığı belirlenmiştir. Bu durum Deney B grubuna verilen puanlayıcı eğitiminin halo etkisi ve yanlışlık davranışları üzerinde daha etkili olduğu ile açıklanabilir. Deney A grubuna verilen eğitimde eğitim süresi, eğitimin zamanı, eğitmen kaynaklı sınırlılıklar gibi durumlar eğitimin etkisi sınırlandırmış olabilir. Her bir grubun bireysel düzeyde puanlayıcı davranışları incelendiğinde Deney A son uygulama grubunda üç cömert (P1, P2 ve P10), dört katı (P5, P6, P8 ve P8), iki halo etkisi (P1 ve P8) ve iki tane de yanlışlık (P6 ve P8) davranışlarının olduğu belirlenmiştir. Deney B grubu son uygulama grubunda iki cömert (P1 ve P5), bir katı (P7), iki halo etkisi (P2 ve P5) ve bir tane de yanlışlık (P2) davranışlarının olduğu belirlenmiştir. Kontrol grubu son uygulama grubunda üç cömert (P1, P2 ve P4), üç katı (P5, P6 ve P8), üç halo etkisi (P1, P4 ve P8) ve iki tane de yanlışlık (P5 ve P6) davranışı belirlenmiştir.

Tüm bireysel sonuçlar grup içinde değerlendirildiğinde;

Deney A grubuna verilen “Referans çerçevesi ile Puanlayıcı Hatası” puanlayıcı eğitiminin puanlayıcıların grup düzeyindeki davranışlarını değiştirmedeği görülmektedir. Deney A grubunda ön uygulamada katılık (P7 ve P9) davranışı gösteren puanlayıcıların puanlayıcı eğitiminden sonra davranışlarının ortadan kalktığı belirlenmiştir. Ayrıca Deney A grubunda halo etkisi davranışı gösteren puanlayıcıların (P3 ve P10) ve yanlışlık davranışı gösteren puanlayıcıların (P1, P5 ve P10) da puanlayıcı eğitiminden sonra bu davranışları ortadan kalkmıştır. Bu durum Deney A grubuna verilen eğitimin bireysel olarak puanlayıcılara göre farklılıklar göstermektedir. Grup üzerinde ise etkili olmadığını göstermektedir.

Deney B grubuna verilen “Referans Çerçevesi ile Performans Boyutu” puanlayıcı eğitiminin puanlayıcıların yanlışlık davranışlarını ortadan kaldırdığı belirlenmiştir. Ayrıca Deney B grubunda ön uygulamada katılık (P4 ve P7), cömertlik (P2, P3 ve P6), halo etkisi (P6) ve

yanlılık (P2) davranışlarının ortadan kalktığı belirlenmiştir. Bu durum Deney B grubuna verilen eğitimin puanlayıcıların katılıkları, halo etkisi ve yanlılık davranışları üzerinde etkili olduğunu göstermektedir. Ayrıca Deney B grubuna verilen puanlayıcı eğitiminin Deney A grubuna göre daha etkili olduğu da söylenebilir. Çünkü Deney B grubuna verilen puanlayıcı eğitimi sonrasında ön uygulamada farklı puanlayıcı davranışlarına sahip olan tüm puanlayıcılar benzer davranışlara sahip olmuşlardır.

Kontrol grubunda ise her iki uygulamada da puanlayıcı davranışları genel olarak benzer kalmıştır. Bu sonuç ise beklenen bir durum olarak görülmektedir.

Öneriler

Bu araştırmada, bir devlet üniversitesinin müzik bölümündeki üniversite birinci sınıf öğrencilerinin piyano başlangıç düzeyini ölçmeye ilişkin seçilen bir eseri çalma esnasındaki gösterdikleri performansların değerlendirilmesi kapsamında öğretim elemanlarının bir diğer ifade ile puanlayıcıların puanlama sürecinde ortaya çıkan puanlayıcı davranışlarının incelenmesi ile ilgili verilen puanlayıcı eğitiminin bu davranışlara olan etkisini belirlemek amacıyla incelenmiştir.

Araştırmadan elde edilen sonuçlardan hareketle hem araştırmacılara hem de uygulayıcılara yönelik çeşitli öneriler sunulabilir. Bu öneriler;

- 1) Bu çalışmada verilen puanlayıcı eğitimi gruplarında yer alan öğretim üyeleri sayısı (Deney-A, n=10; Deney-B, n=7; Kontrol n=8) uygulanmıştır. Bundan sonraki yapılacak olan çalışmalarda daha büyük puanlayıcı grupları ile çalışılıp puanlayıcı eğitiminin etkisi daha ayrıntılı olarak incelenebilir.
- 2) Gelecekteki örnek çalışmalar için puanlayıcı eğitiminin diğer türleri de dikkat alındığında sadece müzik başlangıç düzeyi için değil diğer (orta ve üst) düzeylerine de uygun puanlayıcı eğitimi verilerek puanlayıcı grubunun hem akademisyenler hem de öğrenciler (akran) olmak üzere puanlayıcı davranışları incelenebilir.
- 3) Diğer çalgı aletleri üzerinde (vurmalı, yaylı, üfleli ve telli çalgılar) ve diğer disiplinlerde; resim, spor, şarkıcılık, dans, işe alımlar (insan kaynakları) gibi alanlarda da puanlayıcıların puanlama davranışları ele alınarak özellikle burada ele alınan iki farklı eğitim uygulaması verilerek farklı çalışmalar yapılabilir.
- 4) Başlangıç düzey piyano performansının ölçülmesine ilişkin puanlayıcı eğitimi esnasında deney ve kontrol gruplarında yer alan puanlayıcıların puanlayıcı katılımları,

puanlayıcı cömertliği, merkeze eğilim, halo etkisi ve yanlılık davranışları ile ilgili olan etkileri ele alınmıştır. Örnek çalışmalarda piyano orta ve üst düzeylerine ilişkin puanlayıcı gruplarının davranışları ayrıntılı olarak incelenebilir.

- 5) Bu araştırmada verilen puanlayıcı eğitimleri (Referans Çerçevesi ve Puanlayıcı Hatası; Referans Çerçevesi ile Performans Boyutu) çevrim içinde verilmiştir. Araştırma sonucunda; elde edilen sonuçlar üzerinde eğitimin çevrim içi verilmesinin etkisi olabilir. Bu nedenle her iki eğitim deseni yüz yüze verilerek etkisine bakılarak sonunun kaynağına yönelik daha ayrıntılı sonuçlara ulaşılabilir.
- 6) Sözkonusu müzik performansına ilişkin puanlayıcı davranışlarının azaltılmasına yönelik diğer (davranış gözlem eğitimi, öz liderlik gibi) eğitim desenleri kullanılarak puanlayıcı davranışlarına etkisi ayrıntılı olarak ele alınabilir.
- 7) Piyano başlangıç düzeyi için geliştirilmiş analitik dereceli puanlama anahtarını kullanarak diğer çalgı aletleri için uyarlanarak benzer çalışma orta ve üst düzey içinde test edilebilir.

KAYNAKÇA

- Aguinis, H., Mazurkiewicz, M. D., & Heggstad, E. D. (2009). Using web-based frame of reference training to decrease biases in personality-based job analysis: An experimental field study. *Personnel Psychology*, 62(2), 405-438.
- Akın, Ö., & Baştürk, R. (2012). Keman eğitiminde temel becerilerin Rasch ölçme modeli ile değerlendirilmesi. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 31(31), 175-187.
- Ames, A. J., & Luecht, R. (2018). Item development. In B. B. Frey (Ed.) *The SAGE Encyclopedia of educational research, measurement and evaluation* (pp. 894-898). Thousand Oaks, CA: Sage.
- Amhag, L. (2015). *Learner centered experiences with flipped classroom and mobile online webinars in distance higher education programs*. Paper presented at the 11th International Conference on Mobile Learning, Madeira, Portugal.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573.
- Antonioni, D., & Woehr, D.J. (2001). Improving the quality of multisource rater performance. In D.W. Bracken, C.W. Timmreck and A.H. Church (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes* (pp. 114- 129). California: Jossey Bass.
- Arsan, N. (2012). *Buz pateninde hakem değerlendirmelerinin genellenebilirlik kuramı ve rasch modeli ile incelenmesi*. Doktora Tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.
- Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology*, 72(4), 567-572.

- Atılğan, H. (2004). *Genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch modelinin karşılaştırılmasına ilişkin bir araştırma*. (Doktora Tezi). <https://tez.yok.gov.tr> sayfasından erişilmiştir.
- Atılğan, H. (2005). Müzik öğretmenliği özel yetenek seçme sınavının çok Yüzeyli Rasch modeli ile analizi (İnönü Üniversitesi örneği). *Eurasian Journal of Educational Measurement*, 20, 62-73.
- Atılğan, H., Kan, A., & Doğan, N. (2007). *Eğitimde ölçme ve değerlendirme*. Ankara: Anı.
- Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Journal of Second Language Writing*, 19(1), 60-79.
- Baştürk, R. (2010). Bilimsel araştırma ödevlerinin çok Yüzeyli Rasch ölçme modeli ile değerlendirilmesi. *Journal of Measurement and Evaluation in Education and Psychology*, 1(1), 51-57.
- Baştürk, R., & Işıkoğlu, N. (2008). Okul öncesi eğitim kurumlarının işlevsel kalitelerinin çok-Yüzeyli Rasch ölçme modeli ile analizi. *Kurum ve Uygulamada Eğitim Bilimleri*, 8(1), 7-32.
- Baykul, Y. (2010). *Eğitimde ve psikolojide ölçme: klasik test teorisi ve uygulaması* (2. Baskı). Ankara: Pegem Akademi.
- Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, 2(1), 49-58.
- Bech, P. (1993). *Rating scales for psychopathology, health status, and quality of life*. New York, NY: Springer.
- Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, 31(3), 2-9.
- Bell, B. S., Tannenbaum, S. I., Ford, J. K., Noe, R. A., & Kraiger, K. (2017). 100 years of training and development research: what we know and where we should go. *Journal of Applied Psychology*, 102, 305-23.
- Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: New response sets and decreasing accuracy. *Journal of Applied Psychology*, 65, 60-66.
- Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology*, 60(5), 556-560.

- Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Alexandria, Virginia: ASCD.
- Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32(4), 653-675.
- Büyüköztürk, Ş. (2021). *Deneysel desenler: öntest-sontest kontrol grubu deseni ve veri analizi*. Ankara: Pegem Akademi.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. London: Ravenio.
- Castro, M.D.B., & Tumibay, G.M. (2021). A literature review: efficacy of online learning courses for higher education institution using meta-analysis. *Educ Inf Technol*, 26, 1367–1385
- Cohen, R. J., & Swerdlik, M. E. (2013). *Psikolojik Test ve değerlendirme testlere ve ölçmeye giriş*. (G. Gözen, Çev). E. Tavşancıl (Çev. Ed.). Ankara: Nobel.
- Colwell R. (2007). Music assessment in an increasingly politicized, accountability-driven educational environment. In Brophy T. S. (Ed.), *Proceedings of the 2007 Symposium on Assessment in Music Education* (pp. 3–16). Chicago, IL: GIA.
- Cook, D. A., Garside, S., Levinson, A. J., Dupras, D. M., & Montori, V. M. (2010). What do we mean by web-based learning?'. A Systematic Review of the Variability of Interventions', *Medical Education*, 44, 765–74.
- Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218-244.
- Cornelius, S. (2014), 'Facilitating in a demanding environment: experiences of teaching in virtual classrooms using web conferencing', *British Journal of Educational Technology*, 45, 260–71.
- Cornelius, S., & Gordon, C. (2013). 'Facilitating learning with web conferencing recommendations based on learners' experiences. *Education and Information Technologies*, 18, 275–85.
- Cox, G., Morrison, J., & Brathwaite, B. (2015, June). *The rubric: an assessment tool to guide students and markers*. In 1st International Conference on Higher Education Advances (Head'15) (pp. 26-32). Editorial Universitat Politècnica de València.

- Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, Ohio: Cengage Learning.
- Çetin, B., & İlhan, M. (2017). Standart ve solo taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı katılığı ve cömertliğinin incelenmesi. *Eğitim ve Bilim*, 42(189), 10-1.
- Dierdorff, E. C., Surface, E. A., & Brown, K. G. (2010). Frame-of-reference training effectiveness: Effects of goal orientation and self-efficacy on affective, cognitive, skill-based, and transfer outcomes. *Journal of Applied Psychology*, 95(6), 1181-1191.
- Downing, S. M. (2006). Twelve steps for effective test development. In *Handbook of Test Development* Lawrence Erlbaum Associates, 3, 1-25.
- Downing, S. M., & Yudkowsky, R. (2009). *Assessment in health professions education*. New York: Taylor & Francis.
- Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika*, 16(4), 407-424.
- Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31(2), 93-112.
- Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal and T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261-287). Mahway, NJ: Lawrence Erlbaum Associates.
- Eckes, T. (2005b). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2, 197–221.
- Eckes, T. (2015). Introduction to many-facet Rasch measurement. Analyzing and evaluating rater-mediated assessments. Frankfurt am Main: Peter Lang.
- Erman Aslanoğlu, A., & Şata, M. (2023). Examining the rater drift in the assessment of presentation skills in secondary school context. *Journal of Measurement and Evaluation in Education and Psychology*, 14(1), 62-75.

- Ezike, N. C., & Ames, A. J. (2021). The diagnostic rating system: Rater behavior for an alternative performance assessment rating method. *Psychological Test and Assessment Modeling*, 63(3), 273–304.
- Esfandiari, R. (2021). Rater-mediated assessment of Iranian undergraduate students' college essays: Many-facet Rasch modelling. *Journal of Applied Linguistics and Applied Literature: Dynamics and Advances*, 9(1), 93-119.
- Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of raters. *Theory & Practice in Language Studies (TPLS)*, 1(11), 1531-1540.
- Farrokhi F., Esfandiari R. & Schaefer E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, 34(1), 79-102.
- Feldman, M., Lazzara, E. H., Vanderbilt, A. A., & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions*, 32(4), 279-286.
- Fox, J.-P. (2010). *Rater bias models for ordinal evaluations*. New York, NY: Springer.
- Furtner, M. R., Sachse, P., & Exenberger, S. (2012). Learn to influence yourself: Full range self-leadership training. *Journal of the Indian Academy of Applied Psychology*, 38(2), 294-304.
- Furtner, M. R., Tutzer, L., & Sachse, P. (2018). The mindful self-leader: Investigating the relationships between self-leadership and mindfulness. *Social Behavior and Personality: an International Journal*, 46(3), 353-360.
- Gegenfurtner, A., & Ebner, C. (2019). Webinars in higher education and professional training: a meta-analysis and systematic review of randomized controlled trials. *Educational Research Review*, 28, 100293.
- Gegenfurtner, A. (2018). Testing the gender similarities hypothesis: differences in subjective task value and motivation to transfer training. *Human Resource Development International*, 23(3), 309–320.
- Gegenfurtner, A., Festner, D., Gallenberger, W., Lehtinen, E. & Gruber, H. (2009). Predicting autonomous and controlled motivation to transfer training. *International Journal of Training and Development*, 13(2), 124–138.

- Gegenfurtner, A., Quesada-Pallarès, C. & Knogler, M. (2014). Digital simulation-based training: a meta-analysis. *British Journal of Educational Technology*, 45, 1097–114.
- Gegenfurtner, A., Schwab, N. & Ebner, C. (2018). There's no need to drive from A to B": exploring the lived experience of students and lecturers with digital learning in higher education. *Bavarian Journal of Applied Sciences*, 4, 310–22.
- Gegenfurtner, A., Spagert, L., Weng, G., Bomke, C., Fisch, K., Oswald, A., Reitmaier-Krebs, M., Resch, C., Schwab, N., Stern, W., & Zitt, A. (2017). LernCenter: Ein Konzept für die Digitalisierung berufsbegleitender Weiterbildungen an Hochschulen. *Bavarian Journal of Applied Sciences*, 3, 234–43.
- Gegenfurtner, A., Vauras, M., Gruber, H., & Festner, D. (2010). Motivation to transfer revisited. in K. Gomez, L. Lyons and J. Radinsky (eds), *Learning in the disciplines*, (pp. 452–59). Chicago, IL: ISLS.
- Gegenfurtner, A., Veermans, K., & Vauras, M. (2013). Effects of computer support, collaboration, and time lag on performance self-efficacy and transfer of training: a longitudinal meta-analysis. *Educational Research Review*, 8, 75–89.
- Goe, R., Ipsen, C., & Bliss, S. (2018). Pilot testing a digital career literacy training for vocational rehabilitation professionals. *Rehabilitation Counseling Bulletin*, 61, 236–43.
- Goodrich, H. (1997). Understanding rubrics: The dictionary may define "rubric," but these models provide more clarity. *Educational Leadership*, 54(4), 14-17.
- Görgülü Öztürk, B. (2023). *Acık uçlu maddelerinin puanlanmasında öğrencilere verilen puanlayıcı eğitiminin puanlayıcı davranışlarına etkisinin çok yuzeyli Rasch olcme modeli ile incelenmesi*. Yüksek Lisans Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.
- Guilford, J. P. (1954). *Psychometric methods*. 2nd ed. New York: McGraw-Hill.
- Güler, N., Kaya Uyanık, G., & Taşdelen Teker, G. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi.
- Haladyna, T. M. (1997). *Writing test items in order to evaluate higher order thinking*. USA: Allyn & Bacon.

- Hambleton, R. K. (2004). Theory, methods, and practices in testing for the 21st century. *Psicothema*, 16(4), 696–701.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage.
- Hamp-Lyons, L. (1990). Second language writing: assessment issues. In B. Kroll (Ed.), *Second Language Writing: Research insights for the classroom*, (69-87). Cambridge: Cambridge University.
- Hauenstein, N. M. A., & McCusker, M. E. (2017). Rater training: Understanding effects of training content, practice ratings, and feedback. *International Journal of Selection and Assessment*, 25(3), 253-266.
- Houghton, J. D., & Neck, C. P. (2002). The revised self-leadership questionnaire. *Journal of Managerial Psychology*, 17(8), 672-691.
- Idris, M., & Zakaria, M. H. (2016). Gauging ESL learners' CEFR ratings on oral proficiency in rater training. *Man in India*, 96(6), 1675-1682.
- İlhan, M., & Gezer, M. (2023). Araştırmaların sınıflandırılması. In B. Çetin, M. İlhan, & M. G. Şahin (Eds.), *Eğitimde araştırma yöntemleri: Temel kavramları, ilkeler ve süreçler* (2. ed., s. 101-133). Ankara: Pegem Akademi.
- Johnson, C. L. & Schumacher, J. B. (2016). Does webinar-based financial education affect knowledge and behavior? *Journal of Extension*, 54, 1–10.
- Johnson, C. M., Corazzini, K. N. & Shaw, R. (2011), 'Assessing the feasibility of using virtual environments in distance education', *Knowledge Management & E-Learning*, 3, 5–16.
- Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: designing, scoring, and validating performance tasks*. New York: Guilford.
- Karakaya, İ. & Fişne, F. N. (Ed.). (2023). *Dereceli puanlama anahtarlarının hazırlanması ve uygulanması*. Ankara: Pegem A.
- 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.

- Kaynak, T. (2011). *Piyano eğitimi yarıyıl sonu sınavlarında öğrenci performansının rubrik ile ölçülmesi*. Doktora Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.
- Kenyon, D. (1992, February). Introductory remarks at symposium on development and use of rating scales in language testing. In *14th language testing research colloquium, Vancouver* (Vol. 27).
- Kim, H. (2020). Effects of rating criteria order on the halo effect in L2 writing assessment: a many-facet Rasch measurement analysis. *Language Testing in Asia*, 10(1), 16.
- Kim, S., Copeland, L., Cohen, E., Galt, J., Terregino, C. A., & Pradhan, A. (2022). Frame-of-Reference Training for Students: Promoting a Shared Mental Model for Clerkship Performance with an Online, Interactive Training Module. *Journal of General Internal Medicine*, 37(6), 1575–1577.
- Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, 12(1), 26-43.
- Knoch, U., Fairbairn, J., Myford, C., & Huisman, A. (2018). Evaluating the relative effectiveness of online and face-to-face training for new writing raters. *Papers in Language Testing and Assessment*, 7(1), 61-86.
- Korkmaz, H. (2004). *Fen ve teknoloji eğitiminde alternatif değerlendirme yaklaşımları*. Ankara: Yeryüzü.
- Kubiszyn, T., & Borich, G. D. (2013). *Educational testing and measurement*. New Jersey: John Wiley & Sons.
- Kumar, D. (2005). Performance appraisal: The importance of rater training. *Journal of the Kuala Lumpur Royal Malaysia Police College*, 4(1), 1-17.
- Kutlu, Ö., Doğan, C. D., & Karakaya, İ. (2017). *Ölçme ve değerlendirme performansa ve portfolyoya dayalı durum belirleme* (5. Baskı). Ankara: Pegem Akademi.
- Landy, E J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87, 72-107.
- Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology*, 60(5), 550-555.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563-575.

- Linacre, J. M. (2017). *A user's guide to FACETS: Rasch-model computer programs*. Chicago: MESA.
- Linacre, J. M. (2023). *Facets computer program for many-facet Rasch measurement* [version 3.85.1]. Chicago, IL Mesa.
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15–21.
- Liu, Q., Peng, W., Zhang, F., Hu, R., Li, Y., & Yan, W. (2016). The effectiveness of blended learning in health professions: systematic review and meta-analysis. *Journal of Medical Internet Research*, 18(1), e2.
- Lunz, Mary, E., & Wright, Benjamin, D. (1997). "Latent Trait Models for Performance Examinations" *Application of latent trait and latent class models in social science*. Jurgen Rost, Rolf Langeheine (Eds.) Waxmann.
- May, G. L. (2008). The Effect of rater training on reducing social style bias in peer evaluation. *Business Communication Quarterly*, 71(3), 297-313.
- Martin, B. O., Kolomitro, K., & Lam, T. C. M. (2014). 'Training methods: a review and analysis', *Human Resource Development Review*, 13, 11–35.
- McKinney, W. P. (2017). Assessing the evidence for the educational efficacy of webinars and related internet-based instruction. *Pedagogy in Health Promotion: The Scholarship of Teaching and Learning*, 3, 475–515.
- McNamara, T.F. (1996). *Measuring second language performance*. London: Longman.
- Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research, and Evaluation*, 7(1), 25.
- Moser, K., Kemter, V., Wachsmann, K., Köver, N. Z., & Soucek, R. (2018). Evaluating rater training with double-pretest one-posttest designs: an analysis of testing effects and the moderating role of rater self-efficacy. *The International Journal of Human Resource Management*, 29(18), 2609-2631.
- Moura, N., Dias, P., Veríssimo, L., Oliveira-Silva, P., & Serra, S. (2024). Solo music performance assessment criteria: A systematic review. *Frontiers in Psychology*, 15, 1467434.

- Murphy, K. R., & Davidshofer, C. O. (1988). The halo effect and the impact of rater personality. *Journal of Applied Psychology*, 73(4), 744–750.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386-422.
- Myford, C.M., & Wolfe, E.W. (2004). Detecting and measuring rater effects using many facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189-227.
- Nakamura, Y. (2002). Teacher assessment and peer assessment in practice. *Educational Studies*, 44, 203-215.
- Neck, C. P. (1996). Thought self-leadership: a self-regulatory approach towards overcoming resistance to organizational change. *The International Journal of Organizational Analysis*, 4(2), 202-216.
- Neck, C. P., Stewart, G. L., & Manz, C. C. (1995). Thought self-leadership as a framework for enhancing the performance of performance appraisers. *The journal of Applied Behavioral Science*, 31(3), 278-302.
- Neşe ÖZTÜRK GÜBEŞ. (2021). *Investigating musical aptitude examination with a many-facet Rasch model* (Vol. 8). Vol. 8. Participatory Educational Research.
- Newman, L. R., Brodsky, D., Jones, R. N., Schwartzstein, R. M., Atkins, K. M., & Roberts, D. H. (2016). frame-of-reference training: Establishing reliable assessment of teaching effectiveness. *The Journal of Continuing Education in the Health Professions*, 36(3), 206–210.
- Nicklen, P., Keating, J. L., Paynter, S., Storr, M., & Maloney, S. (2016). Remote–online case–based learning: a comparison of remote–online and face–to–face, case–based learning –a randomized controlled trial. *Education for Health*, 29, 195–202.
- Nitko, A. J., & Brookhart, S. M. (2007). *Education assessment of students*. New Jearsey: Merrill Prentice Hall.
- Noonan, L.E., & Sulsky, L.M. (2001). Impact of frame-of- reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance*, 14(1), 3-26.

- Onkun-Özgür, E. (2024). *Dereceli puanlama anahtarı türünün rutin olmayan matematik problemlerinin puanlanmasında puanlayıcı davranışları üzerine etkisinin incelenmesi* (Yüksek Lisans Tezi). <https://tez.yok.gov.tr> sayfasından erişilmiştir.
- Öç U. (2024). *Açık uçlu soruların puanlanmasında puanlayıcı eğitiminin puanlayıcı davranışlarına etkisinin incelenmesi*. (Doktora Tezi). <https://tez.yok.gov.tr> sayfasından erişilmiştir.
- Payne, P. D., Burrack, F., Parkes, K. A., & Wesolowski, B. (2019). An emerging process of assessment in music education. *Music Educators Journal*, 105(3), 36-44.
- Perlman, C.C. (2003). *Performance assessment: designing appropriate performance tasks and scoring rubrics*. <https://eric.ed.gov/?id=ED480070>, sayfasından erişilmiştir.
- Popham, W. J. (1997). What's wrong-and what's right-with rubrics. *Educational Leadership*, 55, 72-75.
- Pulakos, E. (1991). Rater training for performance appraisal. In J.W. Jones, B. Steffy, D. Bray, B.D. Steffy., and D.W. Bray (Eds.), *Applying psychology in business: The handbook for managers and human resources professionals* (pp. 326-332). New York: Lexington.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. (Original work published 1960).
- Robertson, R. L., Park, J., Gillman, L., & Vergis, A. (2020). The impact of rater training on the psychometric properties of standardized surgical skill assessment tools. *The American Journal of Surgery*, 220(3), 610-615.
- Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczyńska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, 85(2), 370-395.
- Rosales Sánchez, C., Díaz-Cabrera, D., & Hernández-Fernaudo, E. (2019). Does effectiveness in performance appraisal improve with rater training? *PloS one*, 14(9), e0222694.
- Royal, K. D., & Hecker, K. G. (2016). Rater errors in clinical performance assessments. *Journal of Veterinary Medical Education*, 43(1), 5-8.
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413-428.

- Schmid, R. F., Bernard, R. M., Borokhovski, E., Tamim, R. M., Abrami, P. C., Surkes, M. A., Wade, C. A., & Woods, J. (2014). The effects of technology use in postsecondary education: a meta-analysis of classroom applications. *Computers & Education*, 72, 271–91.
- Sever, S., Dogan, C. D., Kamis, O., & Sever, G. (2023). Psychometric properties and standard-setting study of the Piano Performance Test for prospective teachers: Propiedades psicométricas y estudio de establecimiento de normas del Test de Interpretación Pianística para futuros docentes. *Electronic Journal of Music in Education*, 1(52), 18-37.
- Siewiorek, A., & Gegenfurtner, A. (2010). Leading to win: the influence of leadership style on team performance during a computer game training. in K. Gomez, L. Lyons and J. Radinsky (eds), *Learning in the disciplines*, (Vol. 1 pp. 524–31). (Chicago, IL: ISLS),
- Sireci, S. G., & Forte, L. A. (1999). Detecting and correcting rater effects in performance assessment. *Educational Measurement: Issues and Practice*, 18(4), 24–30.
- Stamoulis, D.T. & Hauenstein, N.M.A. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology*, 78(6), 994-1003.
- Stout, J. W., Smith, K., Zhou, C., Solomon, C., Dozor, A. J., Garrison, M. M. & Mangione-Smith, R. (2012). Learning from a distance: Effectiveness of online spirometry training in improving asthma care. *Academic Pediatrics*, 12, 88–95.
- 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.
- Şahan, Ö. (2018). *The impact of rater experience and essay quality on rater behavior and scoring*. Doktora Tezi, Çanakkale Onsekiz Mart Üniversitesi Eğitim Bilimleri Enstitüsü, Çanakkale.
- Şata, M. (2019). *Performans değerlendirme sürecinde puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisinin incelenmesi*. Doktora Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

- Tajeddin, Z., & Alemi, M. (2014). Pragmatic rater training: does it affect non-native L2 teachers' rating accuracy and bias? *International Journal of Language Testing*, 4(1), 66-83.
- Testers, L., Gegenfurtner, A. & Brand-Gruwel, S. (2015), 'Motivation to Transfer Learning to Multiple Contexts', in L. Das, S. Brand-Gruwel, K. Kok and J. Walhout (eds), *The School Library Rocks: Living It, Learning It, Loving It* (pp. 473–87), Heerlen: IASL.
- Testers, L., Gegenfurtner, A., Van Geel, R. & Brand-Gruwel, S. (2019), 'From monocontextual to multicontextual transfer: organizational determinants of the intention to transfer generic information literacy competences to multiple contexts', *Frontline Learning Research*, 7, 23–42.
- Tobaş, C. (2020). *Performansın değerlendirilmesinde farklılaşan puanlayıcı davranışlarının çok Yüzeyle rasch ölçme modeli ile incelenmesi*. Yüksek Lisans Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.
- Turgut, S., & Dalkıran, E. (2021). Piyano eğitimi dersi performans ölçümüne yönelik dereceli puanlama anahtarının geliştirilmesi: geçerlik ve güvenirlik çalışması. *Sanat ve Tasarım Dergisi*, 11(2), 438-449.
- Tursynbayeva, N., Öç, U., & Karakaya, İ. (2024). The effect of rater training on rating behaviors in peer assessment among secondary school students. *International Journal of Assessment Tools in Education*, 11(3), 507-523.
- van der Linden, L. A. F. M., & Hambleton, R. K. (1997). Rater effects in performance assessment. In L. A. F. M. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 43–57). New York, NY: Springer.
- Wang, S.-K., & Hsu, H.-Y. (2008), 'Use of the webinar tool (Elluminate) to support training: the effects of webinar–learning implementation from student–trainers' perspective', *Journal of Interactive Online Learning*, 7, 175–94.
- Wesolowski B. C. (2014). Documenting student learning in music performance: A framework. *Music Educators Journal*, 101, 77–85.
- Wesolowski, B. C. (2016). Assessing jazz big band performance: The development, validation, and application of a facet-factorial rating scale. *Psychology of Music*, 44(3), 324-339.

- Wesolowski, B. C., & Wind, S. A. (2019). Investigating rater accuracy in the context of secondary-level solo instrumental music performance. *Musicae Scientiae*, 23(2), 157-176.
- Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, 19(2), 147-170.
- Weitz, G., Vinzentius, C., Twesten, C., Lehnert, H., Bonnemeier, H., & König, I. R. (2014). Effects of a rater training on rating accuracy in a physical examination skills assessment. *GMS Zeitschrift für Medizinische Ausbildung*, 31(4), 1-17.
- 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.
- Wind, S. A., Engelhard, G., & Wesolowski, B. (2016). Exploring the Effects of Rater Linking Designs and Rater Fit on Achievement Estimates Within the Context of Music Performance Assessments. *Educational Assessment*, 21(4), 278-299.
- Woehr, D.J., & Huffcutt, A.I. (1994). Rater training for performance appraisal. A quantitative review. *Journal of Occupational and Organizational Psychology*, 67(3), 189-205.
- Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, 4(1), 83-106.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.
- Yurdugül, H. (2005). Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması. Paper presented at the XIV. Ulusal Eğitim Bilimleri Kongresi, Pamukkale Üniversitesi Eğitim Fakültesi.
- Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology*, 67(6), 752-758.
- Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections*, 21(2), 1-11.

Zomenou, V., Sigman-Grant, M., Coleman, G., Malekian, F., Zee, J. M. K., Fountain, B. J., & Marsh, A. (2015). Identifying best practices for an interactive webinar. *Journal of Family & Consumer Services*, 107, 62–69.



EKLER



EK 1. Gazi Üniversitesi Etik Komisyon İzin Belgesi

Evrak Tarih ve Sayısı: 13.11.2023-E.795640



T.C.
GAZİ ÜNİVERSİTESİ REKTÖRLÜĞÜ
Etik Komisyonu



Sayı : E-77082166-302.08.01-795640
Konu : Bilimsel ve Eğitim Amaçlı

13.11.2023

Dağıtım Yerlerine

Daha önce 30.12.2022 tarihli ve E.548805 sayılı yazımız ile onay alan 2022-1513 kod numaralı, Üniversitemiz Eğitim Bilimleri Enstitüsü Eğitim Bilimleri Ana Bilim Dalı Doktora Öğrencisi Nazira TURSUNBAYEVA, Prof.Dr.İsmail KARAKAYA'nın danışmanlığında yürüttüğü "*Farklı Puanlama Eğitim Desenlerinin Puanlama Davranışları Üzerine Etkisinin İncelenmesi*" başlıklı tez çalışması hakkında ilgililerden alınan 06.11.2023 tarihli dilekçe Komisyonumuzun 07.11.2023 tarih ve 19 sayılı toplantısında görüşülmüştür.

Dilekçede; Güney Kazakistan'da yapılması planlanan çalışmada yer ve çalışma grubu için gerekli izinlerin ve koşulların sağlanamadığı, çalışmada çalışma grubu üzerinde değişikliğe gidilmesi nedeniyle, çalışmanın Türkiye/ Gazi Üniversitesi Gazi Eğitim Fakültesi Güzel Sanatlar Eğitim Bölümü, Müzik Eğitimi Ana Bilim Dalı'ndan toplanması 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

Bu belge, güvenli elektronik imza ile imzalanmıştır.

Bu belge, güvenli elektronik imza ile imzalanmıştır.

EK 2. Başlangıç Düzey Piyano Performansının Ölçülmesine İlişkin Analitik Dereceli Puanlama Anahtarı

PİYANO DERSİ PERFORMANS DEĞERLENDİRMeye İLİŞKİN ÖLÇME ARACI GELİŞTİRME ÇALIŞMASI

Doç. Dr. Çağla Serin ÖZPARLAK Prof. Dr. İsmail KARAKAYA

Nazira TURSYNBAYEVA

Başlangıç Seviye: Beyaz ve siyah tuşlarda sabit el pozisyonu ve basit düzeyde parmak geçişi içeren etüt ve eserleri; her iki elin tüm parmaklarında eşit güç ve tını elde etmeyi; staccato, legato gibi temel teknikleri uygulamayı kapsar.

Orta Seviye: Orta seviye piyano öğrencilerinin yeterli durumu çeşitli müzik eğitimcileri ve müzik bilimciler tarafından parmak geçişi, çift ses ve akor çalma, ses tutma, pedal kullanımı, oktavlarda rahatlık, ellerde bağımsızlık, temel süslemeleri çalabilme, klavyede geniş alan kullanabilme gibi teknikleri içeren etüt ve eserleri kapsar.

İleri seviye: Üç veya daha çok partili yapı ve motifler, bir elde hızlı aralık değişimi, bir elde dört veya beş sesli akorlar, sık değişen el pozisyonları, aynı elde çalınan iki partinin farklı dinamiklere sahip olması, poliritim, ajilite, virtüözüte gerektiren teknikleri içeren etüt ve eserleri kapsar.

Başlangıç Seviye Piyano İçin Performans Ölçme Aracı Ölçütleri

- Duruş-oturuş (postür) ve el pozisyonu
- Notasyon doğruluğu (Doğru nota ile seslendirme)
- Ritmik boyut ifadesi
- Hız terimine uygunluk ve sabitlik
- Parmak numarası kullanımı
- Dinamikleri sergileme (Eser içerisindeki legato, staccato, vurgu, bağ ayrımı vb... ifadeleri uygulayarak seslendirme)
- Ton / tını bütünlüğü
- Sağ el-sol el koordinasyonu
- Cümleme ve müzikal kimliğini yansıtırma
- Teknik ve müzikal bütünlü

Başlangıç Seviye Piyano İçin Performans Ölçme Aracı

Ölçütler	Çok Zayıf	Zayıf	Orta	Başarılı	Çok Başarılı
Duruş-Oturuş (Postür) ve El Pozisyonu	Eser/etüt icrasında doğru postür ve el pozisyonunda n söz etmek mümkün değildir. Bu durum performansı oldukça olumsuz etkilemektedir.	Eser/etüt icrasında sıklıkla kaybedilen doğru postür ve el pozisyonu performansı olumsuz etkiledi.	Eser/etüt icrasında zaman zaman kaybedilse de genellikle doğru postür ve el pozisyonu ile performans sergilendi.	Eser/etüt icrasının neredeyse tümünde doğru postür ve el pozisyonu ile performans sergilendi.	Eser/etüt icrasının başından sonuna kadar doğru postür ve el pozisyonu ile performans sergilendi.
Notasyon Doğruluğu	Eser/etüt içerisinde sıklıkla yanlış nota seslendirildi. Hatalardan dolayı anlamlı bir performans sergilenemedi.	Eser/etüt içerisindeki yanlış nota icrası, doğru nota icrasından fazladır.	Eser/etüt içerisindeki doğru nota icrası, yanlış nota icrasından fazladır.	Eser/etüt içerisindeki notaların neredeyse tümü doğru seslendirildi	Eser/etüt içerisindeki bütün notalar doğru seslendirildi yanlış nota hiç kullanılmadı.
Ritmik Boyut İfadesi	Eser/etüt içerisindeki ritmik boyutlar sıklıkla yanlış ifade edildi. Hatalardan dolayı anlamlı bir performans sergilenemedi.	Eser/etüt içerisindeki yanlış ritmik ifadeler, doğru ritmik ifadelerden fazladır.	Eser/etüt içerisindeki doğru ritmik ifadeler, yanlış ritmik ifadelerden fazladır.	Eser/etüt içerisindeki ritmik boyutların neredeyse tümü doğru ifade edildi.	Eser/etüt içerisindeki bütün ritmik boyutlar doğru ifade edildi, yanlış ritmik ifade hiç kullanılmadı.
Hız Terimine Uygunluk ve Sabitlik	Verilen hız terimine oldukça uzak bir tempoda performans sergilendi. Performans sırasında tempoda değişimler sıklıkla gözlemlendi.	Verilen hız terimine uzak bir tempoda performans sergilendi. Performans sırasında genellikle sabit olmayan bir tempo gözlemlendi.	Verilen hız terimine kabul edilebilir ölçüde yakın bir tempoda performans sergilendi. Performans sırasında tempoda değişimler birkaç kez gözlemlendi.	Verilen hız terimine yakın bir tempoda performans sergilendi. Performans sırasında tempo genellikle korundu.	Verilen hız terimine tamamen uygun bir tempoda performans sergilendi. Tempo, performansın başından sonuna kadar korundu.
Parmak Numarası Kullanımı	Eser/etüt içerisinde sıklıkla yanlış ve kabul edilemez parmak numaraları kullanıldı.	Eser/etüt içerisinde yanlış parmak numarası kullanımı doğru parmak numarası kullanımından fazladır.	Eser/etüt içerisinde doğru parmak numarası kullanımı yanlış parmak numarası kullanımından fazladır.	Eser/etüt içerisindeki parmak numaralarının neredeyse tümü doğru kullanıldı.	Eser/etüt içerisindeki bütün parmak numaraları doğru kullanıldı.

Dinamikleri Sergileme (Seslendirme)	Eser/etüt içerisindeki dinamikler sıklıkla yanlış seslendirildi.	Eser/etüt içerisinde yanlış seslendirilen dinamikler, doğru seslendirilen dinamiklerden fazladır.	Eser/etüt içerisinde doğru seslendirilen dinamikler, yanlış seslendirilen dinamiklerden fazladır.	Eser/etüt içerisindeki dinamiklerin neredeyse tümü doğru seslendirildi.	Eser/etüt içerisindeki bütün dinamikler doğru seslendirildi.
Ton / Tını Bütünlüğü	Eser/etüt icrasında kaliteli bir ton ve homojen bir tından söz etmek mümkün değildir.	Eser/etüt icrasında sıklıkla kaybedilen ton/tını bütünlüğü performansı olumsuz etkiledi.	Eser/etüt icrasında zaman zaman kaybedilse de genellikle kaliteli bir ton ve homojen bir tını ile performans sergilendi.	Eser/etütün neredeyse tümünde kaliteli bir ton ve homojen bir tını ile performans sergilendi.	Eser/etütün başından sonuna kadar kaliteli bir ton ve homojen bir tını ile performans sergilendi.
Sağ el / Sol el Koordinasyonu	Performansın başından sonuna kadar güvenilir olmayan ve hatalarla dolu bir sağ el-sol el koordinasyonu sergilendi.	Performans içerisinde sağ el-sol koordinasyonunu kontrolsüz ve hatalı olduğu zamanlar daha fazladır.	Performans içerisinde sağ el-sol koordinasyonunu bilinçli ve kontrollü ve kendinden emin olduğu zamanlar daha fazladır.	Performansın neredeyse tümünde bilinçli, kontrollü ve kendinden emin bir sağ el-sol el koordinasyonu sağlandı.	Performansın tümünde bilinçli, kontrollü ve kendinden emin bir sağ el-sol el koordinasyonu sağlandı.
Cümleleme ve Müzikal Kimlik Yansıtma	Eser/etüt icrasında hiçbir şekilde cümleleme ve müzikal kimlikten söz etmek mümkün değildir.	Eser/etüt icrasında sıklıkla kaybedilen cümleleme ve müzikal kimlik performansı olumsuz etkiledi.	Eser/etüt icrasında zaman zaman kaybedilse de cümleleme ve müzikal kimlik genellikle iyi ifade edildi.	Eser/etüt icrasının neredeyse tamamında cümleleme ve müzikal kimlik ön plana çıkarılarak performansa olumlu bir şekilde yansıtıldı.	Eser/etüt icrasının başından sonuna kadar cümleleme ve müzikal kimlik etkileyici bir şekilde ön plana çıkarıldı.
Teknik ve Müzikal Bütünlük	Eser/etüt icrasında teknik ve müzikal bütünlük yoktu. Hatalardan dolayı anlamlı olmayan bir performans sergilendi.	Eser/etüt icrasında teknik ve müzikal anlamda sıklıkla gözlemlenen bütünlük hataları vardı. Performansa olumsuz yansıdı.	Eser/etüt icrasında zaman zaman kaybedilse de genel olarak teknik ve müzikal bütünlük sağlandı.	Eser/etüt icrasının neredeyse tümünde teknik ve müzikal bütünlük sağlandı. Önemli ölçüde etkileyici ve anlamlı bir performans sergilendi.	Eser/etütün başından sonuna kadar teknik ve müzikal bütünlük sağlandı. Oldukça etkileyici ve anlamlı bir performans sergilendi.

EK 3. Puanlayıcı Değerlendirme Formu

Puanlama Formu. (Puanlama 1-5 arasında)

Puan-ı	Öğr	Duruş-Oturuş (Postür) ve El Pozisyonu	Notasyon Doğruluğu	Ritmik Boyut İfadesi	Hız Terimine Uygunluk ve Sabitlik	Parmak Numarası Kullanımı	Artikülasyon ve nüansları Sergileme (Seslendirme)	Ton / Tını Bütünlüğü (Balanı)	Sağ el / Sol el Koordinasyonu ve	Cümleleme ve Müzikal Kimlik Yansıma	Teknik ve Müzikal Bütünlük
1	1										
1	2										
1	3										
1	4										
1	5										
1	6										
1	7										
1	8										
1	9										
1	10										

EK 4 Kapsam Geçerlik Oranına Göre Kriterin Belirlenmesinde Kullanılan Eşik Değerler

Tek kuyruklu hipotezler için anlamlılık düzeyi						
.10	.05	.025	.01	.005	.001	
Çift kuyruklu hipotezler için anlamlılık düzeyi						
N	.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

Not: Bu çalışmada KGO hesaplarırken 16 uzman görüşüne başvurulmuştur. Ek 4'deki tabloda (N) uzman 16 olmadığı için, yerine uzman 15'in değeri dikkate alınmıştır.

EK 5. Deney A Ön Uygulama Puanlayıcı Grubu *Puanlayıcı X Öğrenci Performansı* Etkileşimleri

No	Rater	Stud perform	Observed Score	Expected Score	Bias (logit)	Model S.E.	t	d.f.	Prob.	İnfit	Outfit
1	R4	3	50	47,61	1,68	1,44	1,17	9	0,2718	0,1	0,1
2	R4	6	50	47,52	1,73	1,44	1,20	9	0,2599	0,1	0,1
3	R5	10	50	43,52	2,95	1,44	2,06	9	0,0697	0,1	0,1
4	R8	6	50	47,01	1,94	1,44	1,35	9	0,2088	0,1	0,1
5	R9	6	50	48,87	0,85	1,44	0,59	9	0,5683	0,1	0,1
6	R10	3	50	47,97	1,50	1,44	1,04	9	0,3249	0,1	0,1
	R10	6	50	47,89	1,54	1,44	1,07	9	0,3249	0,1	0,1
7	R10	2	27	43,47	-2,80	0,41	-6,79	9	0,001	0,8	0,8
8	R3	6	44	48,11	-1,43	0,49	-2,93	9	0,167	0,7	1,0
9	R5	4	27	35,53	-1,37	0,41	-3,32	9	0,089	0,5	0,5
10	R1	9	21	28,11	-1,30	0,45	-2,86	9	0,0187	1,4	1,3
11	R3	3	45	48,18	-1,22	0,52	-2,36	9	0,0428	0,7	1,2
12	R2	3	45	48,14	-1,19	0,52	-2,31	9	0,0464	0,3	0,3
13	R2	6	45	48,06	-1,15	0,52	-2,22	9	0,0536	0,3	0,3
14	R7	3	47	48,82	-1,05	0,63	-1,66	9	0,1308	1,3	0,9
15	R7	6	47	48,77	-1,00	0,63	-1,59	9	0,1463	1,3	0,9
16	R9	10	46	48,28	-0,99	0,56	-1,76	9	0,1123	0,7	0,6
17	R1	1	34	39,44	-0,89	0,40	-2,24	9	0,0522	0,8	0,9
18	R8	10	42	45,66	-0,87	0,45	-1,95	9	0,0831	0,2	0,2
19	R4	4	35	40,13	-0,85	0,40	-2,14	9	0,0613	1,1	1,0
20	R2	10	45	47,11	-0,69	0,52	-1,13	9	0,2171	0,5	0,5
21	R5	7	32	36,24	-0,67	0,40	-1,68	9	0,1263	0,6	0,6
22	R4	10	44	46,35	-0,65	0,49	-1,35	9	0,2109	0,3	0,3
23	R9	7	42	44,83	-0,64	0,45	-1,43	9	0,1867	0,8	0,9
24	R10	9	31	34,99	-0,63	0,40	-1,58	9	0,1486	1,6	1,5
25	R8	9	29	32,44	-0,55	0,40	-1,36	9	0,2072	1,3	1,3
26	R7	10	47	48,12	-.54	0,63	-0,86	9	0,4133	1,3	0,9
27	R7	1	45	46,73	-.54	.52	-1,04	9	.3255	.8	.7
28	R4	5	38	40,95	-.52	.41	-1,27	9	.2368	1,0	1,0
29	R6	6	45	46,58	-.48	.52	-.94	9	.3740	.5	.5
30	R4	7	38	40,75	-.48	.41	-1,17	9	.2707	1,2	1,2
31	R6	1	40	42,29	-.43	.42	-1,03	9	.3318	.7	.7
32	R3	10	46	47,17	-.42	.56	-.75	9	.4723	.8	1,4
33	R1	4	32	34,61	-.41	.40	-1,03	9	.3286	.1	.1
34	R10	8	39	41,24	-.40	.41	-.97	9	.3576	1,2	1,1
35	R8	8	37	38,99	-.33	.40	-.82	9	.4323	.8	.8
36	R3	1	44	45,24	-.32	.49	-.65	9	.5306	.7	1,0
37	R8	5	38	39,75	-.30	.41	-.73	9	.4811	.7	.7
38	R6	2	39	40,64	-.29	.41	-.70	9	.5008	.6	.6
39	R9	5	44	44,97	-.24	.49	-.50	9	.6274	.5	.5
40	R6	3	46	46,70	-.24	.56	-.43	9	.6801	.5	.4
41	R8	1	42	43,08	-.22	.45	-.50	9	.6296	.8	.8
42	R4	8	39	40,23	-.22	.41	-.52	9	.6149	.2	.2
43	R2	2	43	43,89	-.20	.46	-.43	9	.6761	.5	.4
44	R5	9	28	29,00	-.16	.41	-.40	9	.6960	1,2	1,2
45	R7	5	44	44,61	-.15	.49	-.31	9	.7664	.8	.7
46	R1	8	34	34,72	-.11	.40	-.29	9	.7801	.6	.5
47	R9	8	44	44,46	-.11	.49	-.23	9	.8233	.5	.5
48	R6	7	38	38,58	-.10	.41	-.24	9	.8174	.7	.7
49	R8	7	39	39,54	-.09	.41	-.22	9	.8278	.5	.5
50	R5	5	36	36,47	-.08	.40	-.19	9	.8535	.8	.8

51	R5	2	38	38.44	-.07	.41	-.18	9	.8596	.4	.4
52	R2	1	45	45.15	-.04	.52	-.08	9	.9387	.5	.5
53	R9	2	46	46.08	-.03	.56	-.05	9	.9634	.7	.7
54	R9	1	47	46.98	.01	.63	.02	9	.9882	1.1	1.2
55	R6	5	39	38.80	.03	.41	.08	9	.9375	.9	.9
56	R2	8	42	41.75	.05	.45	.11	9	.9130	.2	.2
57	R1	5	36	35.56	.07	.40	.18	9	.8647	.6	.6
58	R9	3	49	48.92	.08	1.3	.08	9	.9393	.9	.6
59	R3	7	43	42.34	.14	.46	.30	9	.7741	1.1	1.4
60	R7	7	45	44.46	.14	.52	.27	9	.7920	1.0	.9
61	R5	3	46	45.52	.15	.56	.26	9	.8022	.6	.6
62	R6	4	39	37.90	.18	.41	.45	9	.6663	.8	.8
63	R5	6	46	45.36	.19	.56	.34	9	.7423	.7	.7
64	R8	4	40	38.88	.19	.42	.46	9	.6556	.8	.8
65	R5	8	37	35.65	.22	.40	.54	9	.6023	.7	.7
66	R7	9	40	38.65	.23	.42	.55	9	.5936	1.1	1.0
67	R7	8	45	44.07	.23	.52	.45	9	.6608	1.2	1.0
68	R3	4	43	41.78	.25	.46	.54	9	.6027	1.3	1.5
69	R3	2	45	44.00	.25	.52	.48	9	.6393	.7	1.2
70	R6	10	46	45.08	.27	.56	.48	9	.6450	.6	.6
71	R1	7	37	35.32	.27	.40	.67	9	.5204	.5	.5
72	R9	9	41	39.21	.32	.43	.74	9	.4778	.8	.8
73	R3	5	44	42.53	.32	.49	.67	9	.5225	1.0	1.2
74	R1	6	46	44.82	.34	.56	.60	9	.5640	.6	.6
75	R6	8	40	38.02	.34	.42	.81	9	.4406	.8	.8
76	R2	5	44	42.41	.35	.49	.72	9	.4917	.2	.3
77	R2	7	44	42.22	.39	.49	.80	9	.4467	.3	.3
78	R6	9	34	31.40	.41	.40	1.3	9	.3280	.2	.2
79	R7	2	47	45.78	.42	.63	.67	9	.5203	1.3	.9
80	R3	8	44	41.87	.46	.49	.94	9	.3728	.6	.6
81	R2	4	44	41.65	.50	.49	1.3	9	.3310	.4	.4
82	R4	9	37	33.82	.51	.40	1.26	9	.2404	.9	.9
83	R3	9	39	35.77	.53	.41	1.28	9	.2313	1.6	1.6
84	R2	9	39	35.61	.56	.41	1.34	9	.2118	.9	.9
85	R1	3	47	45.00	.64	.63	1.2	9	.3337	.9	.9
86	R10	5	45	41.92	.70	.52	1.35	9	.2108	.8	.8
87	R1	10	46	42.85	.80	.56	1.42	9	.1891	.8	.8
88	R10	4	45	41.14	.85	.52	1.64	9	.1356	1.1	1.2
89	R4	2	46	42.60	.85	.56	1.51	9	.1651	1.4	1.8
90	R4	1	47	44.04	.89	.63	1.40	9	.1938	.9	1.0
91	R7	4	47	43.99	.90	.63	1.42	9	.1889	1.4	1.1
92	R10	1	48	44.78	1.17	.75	1.56	9	.1535	.9	.9
93	R8	3	49	47.13	1.17	1.3	1.14	9	.2839	.9	.6
94	R9	4	48	44.39	1.27	.75	1.69	9	.1246	.9	.9
95	R10	10	49	46.86	1.28	1.3	1.24	9	.2459	1.1	1.7
96	R5	1	46	40.27	1.30	.56	2.31	9	.0463	.7	.8
97	R1	2	45	37.56	1.47	.52	2.84	9	.0193	.3	.3
98	R10	7	48	41.73	1.85	.75	2.46	9	.0360	.9	.9
99	R8	2	48	41.52	1.89	.75	2.52	9	.0329	.7	.5

EK 6. Deney A Son Uygulama Puanlayıcı Grubu *Puanlayıcı X Öğrenci Performansı* Etkileşimleri

No	Rater	Stud perform	Observed Score	Expected Score	Bias (logit)	Model S.E.	t	d.f.	Prob.	İnfit	Outfit
1	R6	1	50	49.46	.08<	1.44	.06	9	.9543	.1	.1
2	R10	1	50	49.25	.43<	1.44	.30	9	.7725	.1	.1
3	R3	2	50	49.40	.20<	1.44	.14	9	.8951	.1	.1
4	R6	4	50	47.32	1.85<	1.44	1.28	9	.2311	.1	.1
5	R3	10	50	48.47	1.20<	1.44	.83	9	.4272	.1	.1
6	R6	10	50	48.96	.78<	1.44	.54	9	.6031	.1	.1
7	R10	10	50	48.57	1.12<	1.44	.78	9	.4566	.1	.1
8	R8	8	29	39.43	-2.15	.50	-4.34	9	.0019	1.3	1.3
9	R10	3	39	46.61	-1.90	.45	-4.26	9	.0021	.7	.6
10	R2	1	44	48.44	-1.72	.51	-3.39	9	.0080	1.4	1.3
11	R4	4	34	41.10	-1.41	.45	-3.16	9	.0115	.7	.6
12	R7	6	32	38.48	-1.29	.46	-2.80	9	.0206	.4	.4
13	R1	7	36	42.06	-1.23	.44	-2.78	9	.0214	.7	.6
14	R9	8	28	32.68	-1.07	.52	-2.08	9	.0674	1.2	1.1
15	R2	10	44	47.18	-1.03	.51	-2.03	9	.0731	.7	.6
16	R4	6	34	38.96	-.97	.45	-2.18	9	.0571	.9	.8
17	R6	2	49	49.60	-.96	1.04	-.92	9	.3801	.9	.7
17	R10	8	42	45.52	-.90	.47	-1.92	9	.0877	.1	.2
18	R5	6	34	38.24	-.83	.45	-1.86	9	.0959	1.9	1.8
19	R3	3	44	46.40	-.72	.51	-1.42	9	.1885	1.0	.9
20	R3	4	44	46.26	-.67	.51	-1.32	9	.2192	.4	.5
21	R9	1	40	42.93	-.64	.45	-1.41	9	.1931	.4	.4
22	R10	2	49	49.44	-.61	1.04	-.59	9	.5691	.9	.5
23	R9	5	27	29.33	-.61	.54	-1.13	9	.2869	1.2	1.1
24	R7	10	43	45.12	-.55	.49	-1.14	9	.2847	.6	.6
25	R9	10	37	39.73	-.54	.44	-1.22	9	.2528	.8	.8
26	R6	5	42	44.13	-.51	.47	-1.08	9	.3071	.4	.4
27	R7	5	33	35.31	-.46	.45	-1.02	9	.3357	.1	.1
28	R4	2	47	47.94	-.45	.64	-.70	9	.5004	.7	.6
29	R8	9	40	42.09	-.44	.45	-.98	9	.3521	.5	.5
30	R8	7	34	36.22	-.44	.45	-.98	9	.3533	.3	.3
31	R6	6	45	46.21	-.38	.54	-.71	9	.4939	.9	.7
32	R6	9	47	47.81	-.38	.64	-.59	9	.5717	.8	.6
33	R7	9	40	41.76	-.37	.45	-.82	9	.4321	1.6	1.1
34	R1	2	49	49.28	-.35	1.04	-.34	9	.7409	.8	.5
35	R4	5	34	35.80	-.35	.45	-.79	9	.4485	.2	.2
36	R1	5	40	41.57	-.33	.45	-.73	9	.4829	.3	.3
37	R5	9	40	41.54	-.32	.45	-.72	9	.4912	1.0	.9
38	R3	5	41	42.39	-.30	.46	-.66	9	.5240	.1	.2
39	R9	9	34	35.48	-.29	.45	-.65	9	.5294	1.5	1.4
40	R5	4	39	40.41	-.29	.45	-.64	9	.5367	.8	.7
41	R1	3	45	45.88	-.27	.54	-.50	9	.6291	1.1	.9
42	R3	1	49	49.19	-.22	1.04	-.22	9	.8334	1.0	1.2
43	R3	7	42	42.86	-.20	.47	-.42	9	.6877	.7	.6
44	R6	3	47	47.43	-.19	.64	-.30	9	.7738	.6	.5
45	R6	7	44	44.54	-.14	.51	-.28	9	.7859	.6	.5
46	R2	7	39	39.53	-.11	.45	-.24	9	.8168	.6	.6
47	R4	7	36	36.35	-.07	.44	-.15	9	.8811	.2	.2
48	R9	4	34	34.27	-.05	.45	-.12	9	.9086	.7	.6
49	R10	7	43	43.18	-.04	.49	-.09	9	.9298	.5	.4
50	R1	1	49	49.04	-.04	1.04	-.04	9	.9680	1.1	2.4
51	R10	9	47	47.08	-.03	.64	-.05	9	.9591	1.0	1.1

52	R7	3	41	40.89	.02	.46	.05	9	.9602	1.6	1.5
53	R5	8	39	38.83	.03	.45	.08	9	.9415	1.0	.9
54	R3	6	45	44.86	.04	.54	.07	9	.9432	.9	.7
55	R2	4	44	43.82	.05	.51	.09	9	.9311	.9	.8
56	R8	2	48	47.90	.06	.76	.08	9	.9408	.8	.6
57	R5	3	41	40.66	.07	.46	.15	9	.8808	.9	.8
58	R2	9	45	44.73	.08	.54	.14	9	.8883	1.4	1.3
59	R7	2	48	47.76	.13	.76	.17	9	.8682	1.0	.9
60	R5	2	48	47.67	.18	.76	.23	9	.8211	.9	.8
61	R2	2	49	48.82	.18	1.04	.17	9	.8666	.9	.7
62	R5	5	36	35.06	.18	.44	.41	9	.6885	.2	.1
63	R8	10	46	45.36	.20	.58	.35	9	.7344	.6	.5
64	R9	2	45	44.18	.22	.54	.42	9	.6854	2.0	1.5
65	R1	9	47	46.42	.23	.64	.35	9	.7347	.9	1.1
66	R2	6	43	41.97	.24	.49	.49	9	.6386	.8	.7
67	R2	3	45	44.03	.27	.54	.49	9	.6331	1.0	.9
68	R5	10	46	44.96	.32	.58	.56	9	.5918	.5	.4
69	R2	5	41	39.00	.41	.46	.89	9	.3957	.7	.6
70	R1	8	46	44.63	.41	.58	.71	9	.4930	.7	.8
71	R8	4	43	40.98	.45	.49	.92	9	.3800	.4	.4
72	R5	1	48	47.00	.49	.76	.64	9	.5372	1.0	.9
73	R7	4	43	40.64	.52	.49	1.07	9	.3121	2.0	1.9
74	R4	10	47	45.44	.55	.64	.85	9	.4158	.9	1.1
75	R7	7	39	35.86	.62	.45	1.38	9	.2011	.6	.5
76	R4	3	44	41.34	.62	.51	1.22	9	.2530	.7	.6
77	R8	3	44	41.23	.64	.51	1.27	9	.2363	.7	.6
78	R1	10	49	48.20	.65	1.04	.62	9	.5475	.8	.5
79	R8	5	39	35.67	.65	.45	1.46	9	.1786	1.0	1.0
80	R5	7	39	35.61	.66	.45	1.49	9	.1716	.7	.6
81	R6	8	48	46.54	.67	.76	.88	9	.4032	.8	.6
82	R10	4	48	46.47	.69	.76	.91	9	.3872	.9	.8
83	R9	3	39	34.53	.88	.45	1.96	9	.0812	.7	.6
84	R8	6	43	38.84	.89	.49	1.82	9	.1016	1.1	1.0
85	R9	7	34	29.78	.90	.45	2.01	9	.0753	.4	.3
86	R2	8	46	42.49	.95	.58	1.64	9	.1363	1.1	.9
87	R1	4	48	45.72	.95	.76	1.25	9	.2431	.8	.6
88	R9	6	37	32.11	.97	.44	2.20	9	.0549	.4	.4
89	R4	8	44	39.55	.99	.51	1.95	9	.0823	1.0	.9
90	R7	8	44	39.07	1.09	.51	2.14	9	.0606	.4	.4
91	R3	8	48	45.27	1.09	.76	1.43	9	.1862	1.0	1.4
92	R4	1	49	47.33	1.12	1.04	1.08	9	.3099	1.0	.8
93	R10	6	48	45.13	1.13	.76	1.48	9	.1721	.8	.6
94	R8	1	49	47.27	1.14	1.4	1.10	9	.3001	.9	.7
95	R7	1	49	47.11	1.21	1.04	1.17	9	.2726	1.0	1.2
96	R3	9	49	46.89	1.30	1.04	1.25	9	.2414	1.0	1.2
97	R4	9	47	42.19	1.39	.64	2.15	9	.0602	1.5	1.3
98	R1	6	48	44.19	1.39	.76	1.82	9	.1014	.8	.6
99	R10	5	48	42.73	1.75	.76	2.30	9	.0472	1.1	9.0

EK 7. Deney B Ön Uygulama Puanlayıcı Grubu *Puanlayıcı X Öğrenci Performansı* Etkileşimleri

No	Rater	Stud perform	Observed Score	Expected Score	Bias (logit)	Model S.E.	t	d.f.	Prob.	İnfit	Outfit
1	R1	3	50	46.45	2.38<	1.46	1.64	9	.1364	.1	.1
2	R4	3	50	47.88	1.65<	1.46	1.13	9	.2859	.1	.1
3	R7	3	50	49.67	-.43<	1.46	-.30	9	.7737	.1	.1
4	R7	6	50	48.68	1.07<	1.46	.74	9	.4797	.1	.1
5	R7	10	50	49.35	.27<	1.46	.19	9	.8557	.1	.1
6	R4	10	40	46.51	-2.28	.54	-4.20	9	.0023	.3	.2
	R3	3	42	45.94	-1.42	.57	-2.46	9	.0359	.2	.2
7	R7	9	39	43.25	-1.33	.52	-2.56	9	.0309	1.1	1.0
8	R2	3	42	45.31	-1.16	.57	-2.03	9	.0732	.8	.7
9	R5	1	41	44.43	-1.16	.56	-2.07	9	.0682	.6	.6
10	R4	8	38	41.48	-1.00	.50	-2.01	9	.0749	.7	.7
11	R1	10	42	44.67	-.93	.57	-1.61	9	.1415	.3	.3
12	R6	5	23	27.47	-.59	.37	-1.59	9	.1469	.4	.5
13	R6	2	28	32.00	-.56	.36	-1.54	9	.1575	1.6	1.7
14	R3	7	38	39.97	-.53	.50	-1.07	9	.3119	.6	.5
15	R2	4	34	36.72	-.53	.42	-1.27	9	.2344	.6	.5
16	R6	1	35	37.44	-.51	.43	-1.18	9	.2666	1.3	1.6
17	R5	4	39	40.62	-.47	.52	-.90	9	.3900	1.4	2.6
18	R2	8	35	37.20	-.46	.43	-1.06	9	.3180	.5	.7
19	R1	6	41	42.37	-.45	.56	-.80	9	.4464	.1	.1
20	R5	2	40	41.44	-.45	.54	-.82	9	.4322	1.0	1.0
21	R2	9	25	28.23	-.42	.36	-1.16	9	.2744	.9	.9
22	R3	4	36	37.80	-.40	.45	-.89	9	.3962	.7	.7
23	R1	4	37	38.65	-.40	.48	-.85	9	.4189	.7	.7
24	R3	10	43	44.08	-.38	.58	-.64	9	.5373	.5	.5
25	R6	6	35	36.79	-.37	.43	-.84	9	.4223	2.1	2.1
26	R4	9	34	35.82	-.34	.42	-.82	9	.4344	.4	.4
27	R4	7	42	42.91	-.31	.57	-.53	9	.6067	.4	.4
28	R5	8	40	40.97	-.30	.54	-.54	9	.5991	.8	.7
29	R1	1	42	42.82	-.27	.57	-.48	9	.6455	.3	.3
30	R1	5	36	37.15	-.25	.45	-.55	9	.5956	.6	.6
31	R3	6	41	41.72	-.23	.56	-.41	9	.6885	.5	.4
32	R1	7	40	40.68	-.20	.54	-.38	9	.7155	.3	.2
33	R2	2	37	37.84	-.20	.48	-.41	9	.6886	.7	.7
34	R7	2	47	47.37	-.18	.69	-.26	9	.7976	.9	.8
35	R6	9	19	20.06	-.18	.42	-.42	9	.6817	.8	.7
36	R5	7	42	42.43	-.14	.57	-.25	9	.8096	.8	.7
37	R3	1	42	42.18	-.06	.57	-.10	9	.9206	.8	.7
38	R6	3	42	42.15	-.05	.57	-.08	9	.9347	1.2	1.1
39	R3	5	36	36.17	-.04	.45	-.08	9	.9398	.5	.5
40	R1	8	39	39.05	-.01	.52	-.03	9	.9796	.3	.2
41	R5	6	44	44.02	-.01	.60	-.01	9	.9924	.6	.6
42	R7	7	48	47.93	.04	.80	.05	9	.9599	1.1	1.2
43	R7	4	47	46.84	.08	.69	.11	9	.9155	2.3	2.9
44	R7	1	49	48.84	.17	1.06	.16	9	.8767	1.2	1.9
45	R5	5	40	39.37	.18	.54	.33	9	.7458	.9	.9
46	R6	7	35	33.99	.18	.43	.42	9	.6842	1.5	1.5
47	R4	6	45	44.47	.19	.61	.32	9	.7580	.7	.6
48	R6	8	33	30.98	.31	.40	.77	9	.4638	3.3	3.3
49	R4	5	41	39.94	.32	.56	.58	9	.5788	.1	.1

50	R1	2	41	39.59	.42	.56	.76	9	.4683	.1	.1
51	R7	5	47	46.00	.44	.69	.64	9	.5378	2.0	1.7
52	R7	8	48	47.07	.52	.80	.65	9	.5344	1.0	.9
53	R1	9	35	31.58	.57	.43	1.31	9	.2220	.6	.6
54	R2	10	45	43.36	.59	.61	.96	9	.3620	.5	.5
55	R3	9	34	30.09	.60	.42	1.45	9	.1822	.4	.4
56	R4	4	43	41.15	.61	.58	1.05	9	.3220	.7	.6
57	R3	2	41	38.81	.64	.56	1.14	9	.2840	.6	.6
58	R5	9	38	34.98	.65	.50	1.31	9	.2222	.9	.9
59	R2	5	38	34.90	.67	.50	1.34	9	.2128	.8	.8
60	R2	6	43	40.94	.68	.58	1.16	9	.2766	.4	.4
61	R2	1	44	41.41	.88	.60	1.47	9	.1755	.5	.4
62	R2	7	42	39.08	.89	.57	1.55	9	.1561	.2	.2
63	R6	4	36	30.22	.96	.45	2.12	9	.0629	2.4	2.2
64	R5	3	49	47.60	1.06	1.06	1.00	9	.3437	.9	.5
65	R4	2	45	41.95	1.07	.61	1.74	9	.1158	.5	.5
66	R3	8	42	38.24	1.10	.57	1.94	9	.0847	.2	.2
67	R6	10	45	39.91	1.70	.61	2.78	9	.0213	.5	.5
68	R5	10	49	46.12	1.77	1.06	1.66	9	.1307	.9	.7
69	R4	1	49	44.88	2.25	1.06	2.12	9	.0627	.9	.5
70	R1	3	50	46.45	2.38<	1.46	1.64	9	.1364	.1	.1
71	R4	3	50	47.88	1.65<	1.46	1.13	9	.2859	.1	.1
72	R7	3	50	49.67	-.43<	1.46	-.30	9	.7737	.1	.1
73	R7	6	50	48.68	1.07<	1.46	.74	9	.4797	.1	.1
74	R7	10	50	49.35	.27<	1.46	.19	9	.8557	.1	.1
75	R4	10	40	46.51	-2.28	.54	-4.20	9	.0023	.3	.2
76	R3	3	42	45.94	-1.42	.57	-2.46	9	.0359	.2	.2
77	R7	9	39	43.25	-1.33	.52	-2.56	9	.0309	1.1	1.0
78	R2	3	42	45.31	-1.16	.57	-2.03	9	.0732	.8	.7
79	R5	1	41	44.43	-1.16	.56	-2.07	9	.0682	.6	.6
80	R4	8	38	41.48	-1.00	.50	-2.01	9	.0749	.7	.7
81	R1	10	42	44.67	-.93	.57	-1.61	9	.1415	.3	.3
82	R6	5	23	27.47	-.59	.37	-1.59	9	.1469	.4	.5
83	R6	2	28	32.00	-.56	.36	-1.54	9	.1575	1.6	1.7
84	R3	7	38	39.97	-.53	.50	-1.07	9	.3119	.6	.5
85	R2	4	34	36.72	-.53	.42	-1.27	9	.2344	.6	.5
86	R6	1	35	37.44	-.51	.43	-1.18	9	.2666	1.3	1.6
87	R5	4	39	40.62	-.47	.52	-.90	9	.3900	1.4	2.6
88	R2	8	35	37.20	-.46	.43	-1.06	9	.3180	.5	.7
89	R1	6	41	42.37	-.45	.56	-.80	9	.4464	.1	.1
90	R5	2	40	41.44	-.45	.54	-.82	9	.4322	1.0	1.0
91	R2	9	25	28.23	-.42	.36	-1.16	9	.2744	.9	.9
92	R3	4	36	37.80	-.40	.45	-.89	9	.3962	.7	.7
93	R1	4	37	38.65	-.40	.48	-.85	9	.4189	.7	.7
94	R3	10	43	44.08	-.38	.58	-.64	9	.5373	.5	.5
95	R6	6	35	36.79	-.37	.43	-.84	9	.4223	2.1	2.1
96	R4	9	34	35.82	-.34	.42	-.82	9	.4344	.4	.4
97	R4	7	42	42.91	-.31	.57	-.53	9	.6067	.4	.4
98	R5	8	40	40.97	-.30	.54	-.54	9	.5991	.8	.7
99	R1	1	42	42.82	-.27	.57	-.48	9	.6455	.3	.3

EK 8. Deney B Son Uygulama Puanlayıcı Grubu *Puanlayıcı X Öğrenci Performansı* Etkileşimleri

No	Rater	Stud perfor m	Observ ed Score	Expected Score	Bias (logit)	Model S.E.	t	d.f.	Prob.	İnfit	Outfit
1	R12	6	37	41.15	-1.20	.52	-2.31	9	.0464	1.0	1.0
2	R11	10	39	42.81	-1.15	.54	-2.14	9	.0611	.2	.2
3	R13	1	40	43.55	-1.09	.54	-2.01	9	.0757	.7	.7
4	R13	6	40	43.41	-1.05	.54	-1.92	9	.0864	.4	.4
5	R14	2	36	39.43	-.95	.51	-1.86	9	.0961	.5	.5
6	R16	1	39	41.66	-.79	.54	-1.48	9	.1739	1.4	1.4
7	R12	5	33	36.16	-.79	.49	-1.62	9	.1403	.9	.9
8	R15	4	38	40.63	-.77	.53	-1.45	9	.1809	.7	.7
9	R12	9	25	28.72	-.75	.44	-1.70	9	.1226	2.7	2.7
10	R17	7	29	32.22	-.72	.46	-1.56	9	.1541	.2	.2
11	R14	3	40	42.29	-.69	.54	-1.27	9	.2344	.1	.1
12	R14	10	39	41.13	-.63	.54	-1.18	9	.2689	.6	.7
13	R17	4	31	33.20	-.51	.47	-1.07	9	.3110	.8	.8
14	R13	3	42	43.55	-.49	.56	-.88	9	.4025	1.0	1.0
15	R11	8	37	38.72	-.48	.52	-.92	9	.3801	1.1	1.0
16	R13	5	37	38.72	-.48	.52	-.92	9	.3811	.7	.8
17	R15	3	44	45.34	-.46	.57	-.80	9	.4444	1.3	1.3
18	R13	10	41	42.43	-.44	.55	-.80	9	.4450	.7	.8
19	R15	8	39	40.38	-.41	.54	-.75	9	.4701	.8	.8
20	R11	2	40	41.16	-.35	.54	-.64	9	.5400	.5	.5
21	R11	1	43	43.92	-.29	.56	-.52	9	.6140	.7	.8
22	R17	9	24	25.23	-.23	.44	-.54	9	.6026	.4	.4
23	R15	5	40	40.76	-.23	.54	-.42	9	.6847	1.1	1.1
24	R15	2	42	42.73	-.23	.56	-.41	9	.6924	1.0	1.0
25	R17	8	32	32.87	-.20	.48	-.42	9	.6817	1.2	1.2
26	R16	3	41	41.66	-.20	.55	-.36	9	.7237	1.2	1.2
27	R12	8	35	35.72	-.18	.50	-.36	9	.7243	.8	.9
28	R16	5	36	36.57	-.15	.51	-.30	9	.7736	1.0	1.0
29	R16	10	40	40.49	-.15	.54	-.27	9	.7936	.9	.9
30	R15	1	45	45.34	-.12	.59	-.20	9	.8430	1.5	1.5
31	R14	7	36	36.29	-.08	.51	-.15	9	.8846	.8	.9
32	R16	9	29	29.23	-.05	.46	-.11	9	.9177	.4	.4
33	R12	7	35	35.12	-.03	.50	-.06	9	.9533	1.2	1.2
34	R14	9	30	30.12	-.03	.47	-.06	9	.9560	.7	.7
35	R12	4	36	36.01	.00	.51	-.01	9	.9945	1.0	1.0
36	R14	8	37	36.86	.04	.52	.07	9	.9453	.7	.7
37	R15	7	40	39.86	.04	.54	.07	9	.9428	1.1	1.1
38	R17	2	36	35.78	.06	.51	.11	9	.9140	1.6	1.6
39	R16	2	39	38.76	.07	.54	.13	9	.8995	1.1	1.1
40	R17	10	38	37.69	.09	.53	.17	9	.8723	.5	.5
41	R11	9	33	32.40	.14	.49	.29	9	.7770	.5	.6
42	R16	4	37	36.43	.15	.52	.29	9	.7752	1.0	1.0
43	R16	8	37	36.13	.23	.52	.44	9	.6682	1.6	1.6
44	R11	7	39	38.19	.23	.54	.43	9	.6756	.2	.2
45	R14	4	38	37.15	.23	.53	.44	9	.6681	1.3	1.3
46	R11	5	40	39.13	.26	.54	.47	9	.6485	.1	.1
47	R11	4	40	38.99	.30	.54	.54	9	.6004	.1	.1
48	R17	3	40	38.94	.31	.54	.57	9	.5817	.5	.6
49	R13	7	39	37.77	.35	.54	.65	9	.5317	.2	.2
50	R17	6	40	38.78	.36	.54	.65	9	.5295	2.1	2.1

50	R16	7	37	35.55	.38	.52	.74	9	.4802	1.0	1.0
51	R11	6	45	43.78	.41	.59	.69	9	.5054	.9	.9
52	R16	6	43	41.51	.46	.56	.82	9	.4325	1.0	.9
53	R13	9	34	31.89	.50	.49	1.01	9	.3378	.7	.7
54	R12	3	43	41.30	.53	.56	.94	9	.3739	.6	.6
55	R14	6	44	42.14	.59	.57	1.03	9	.3303	.6	.7
56	R17	1	41	38.94	.61	.55	1.10	9	.2955	1.0	1.0
57	R17	5	36	33.36	.66	.51	1.29	9	.2280	.9	.9
58	R15	9	37	34.41	.67	.52	1.29	9	.2306	.8	.9
59	R13	2	43	40.77	.69	.56	1.22	9	.2521	.5	.5
60	R15	6	47	45.21	.71	.68	1.05	9	.3219	1.9	2.0
61	R13	4	41	38.58	.71	.55	1.30	9	.2274	.9	.9
62	R12	2	41	38.38	.77	.55	1.40	9	.1943	.8	.8
63	R14	5	40	37.29	.77	.54	1.42	9	.1895	1.0	1.0
64	R12	1	44	41.30	.85	.57	1.48	9	.1733	.6	.6
65	R14	1	45	42.29	.88	.59	1.49	9	.1700	1.1	1.1
66	R15	10	47	44.31	1.02	.68	1.50	9	.1671	2.2	2.5
67	R13	8	42	38.31	1.10	.56	1.97	9	.0799	.5	.5
68	R11	3	47	43.92	1.15	.68	1.70	9	.1240	1.0	.9
69	R12	10	45	40.13	1.54	.59	2.60	9	.0286	.9	1.0
70	R12	6	37	41.15	-1.20	.52	-2.31	9	.0464	1.0	1.0
71	R11	10	39	42.81	-1.15	.54	-2.14	9	.0611	.2	.2
72	R13	1	40	43.55	-1.09	.54	-2.01	9	.0757	.7	.7
73	R13	6	40	43.41	-1.05	.54	-1.92	9	.0864	.4	.4
74	R14	2	36	39.43	-.95	.51	-1.86	9	.0961	.5	.5
75	R16	1	39	41.66	-.79	.54	-1.48	9	.1739	1.4	1.4
76	R12	5	33	36.16	-.79	.49	-1.62	9	.1403	.9	.9
77	R15	4	38	40.63	-.77	.53	-1.45	9	.1809	.7	.7
78	R12	9	25	28.72	-.75	.44	-1.70	9	.1226	2.7	2.7
79	R17	7	29	32.22	-.72	.46	-1.56	9	.1541	.2	.2
80	R14	3	40	42.29	-.69	.54	-1.27	9	.2344	.1	.1
81	R14	10	39	41.13	-.63	.54	-1.18	9	.2689	.6	.7
82	R17	4	31	33.20	-.51	.47	-1.07	9	.3110	.8	.8
83	R13	3	42	43.55	-.49	.56	-.88	9	.4025	1.0	1.0
84	R11	8	37	38.72	-.48	.52	-.92	9	.3801	1.1	1.0
85	R13	5	37	38.72	-.48	.52	-.92	9	.3811	.7	.8
86	R15	3	44	45.34	-.46	.57	-.80	9	.4444	1.3	1.3
87	R13	10	41	42.43	-.44	.55	-.80	9	.4450	.7	.8
88	R15	8	39	40.38	-.41	.54	-.75	9	.4701	.8	.8
89	R11	2	40	41.16	-.35	.54	-.64	9	.5400	.5	.5
90	R11	1	43	43.92	-.29	.56	-.52	9	.6140	.7	.8
91	R17	9	24	25.23	-.23	.44	-.54	9	.6026	.4	.4
92	R15	5	40	40.76	-.23	.54	-.42	9	.6847	1.1	1.1
93	R15	2	42	42.73	-.23	.56	-.41	9	.6924	1.0	1.0
94	R17	8	32	32.87	-.20	.48	-.42	9	.6817	1.2	1.2
95	R16	3	41	41.66	-.20	.55	-.36	9	.7237	1.2	1.2
96	R12	8	35	35.72	-.18	.50	-.36	9	.7243	.8	.9
97	R16	5	36	36.57	-.15	.51	-.30	9	.7736	1.0	1.0
98	R16	10	40	40.49	-.15	.54	-.27	9	.7936	.9	.9
99	R15	1	45	45.34	-.12	.59	-.20	9	.8430	1.5	1.5

EK 9. Kontrol Ön Uygulama Puanlayıcı Grubu *Puanlayıcı X Öğrenci Performansı* Etkileşimleri

No	Rater	Stud perform	Observed Score	Expected Score	Bias (logit)	Model S.E.	t	d.f.	Prob.	Infit	Outfit
1	R18	1	50	47.86	1.62<	1.45	1.12	9	.2933	.1	.1
2	R18	6	50	46.47	2.26<	1.45	1.56	9	.1529	.1	.1
3	R18	10	50	48.03	1.52<	1.45	1.05	9	.3224	.1	.1
4	R23	9	13	20.40	-1.96	.67	-2.94	9	.0166	.8	.6
5	R18	5	39	44.77	-1.32	.43	-3.06	9	.0135	.2	.3
6	R21	6	33	40.27	-1.24	.39	-3.18	9	.0112	1.5	1.5
7	R24	1	39	44.17	-1.15	.43	-2.67	9	.0257	.3	.3
8	R20	1	42	46.07	-1.13	.48	-2.38	9	.0412	.2	.2
9	R23	8	21	27.52	-1.06	.43	-2.49	9	.0347	.3	.3
91	R19	1	45	47.42	-.91	.55	-1.66	9	.1318	1.6	1.5
10	R21	10	40	43.56	-.80	.44	-1.80	9	.1059	2.0	1.9
11	R18	9	36	40.02	-.72	.40	-1.77	9	.1102	.7	.8
12	R25	2	29	33.52	-.67	.38	-1.75	9	.1146	1.2	1.1
13	R24	10	42	44.52	-.64	.48	-1.34	9	.2124	.5	.5
14	R23	3	35	38.49	-.59	.40	-1.48	9	.1729	.7	.7
15	R18	7	41	43.53	-.59	.46	-1.28	9	.2333	1.6	1.8
16	R22	4	33	36.66	-.58	.39	-1.49	9	.1714	.9	1.1
17	R20	2	41	43.18	-.50	.46	-1.08	9	.3065	.3	.3
18	R21	7	32	35.09	-.47	.39	-1.23	9	.2512	2.6	2.5
19	R25	6	32	34.86	-.44	.39	-1.13	9	.2881	.2	.2
20	R22	5	32	34.56	-.39	.39	-1.01	9	.3399	.7	.9
21	R22	10	40	41.79	-.37	.44	-.84	9	.4224	.5	.5
22	R19	2	44	45.20	-.35	.52	-.67	9	.5202	1.4	1.2
23	R22	6	36	38.01	-.34	.40	-.84	9	.4202	.4	.5
24	R21	3	44	45.13	-.33	.52	-.63	9	.5449	1.2	1.3
25	R19	4	44	45.12	-.32	.52	-.62	9	.5502	1.1	1.0
26	R18	2	45	45.91	-.29	.55	-.53	9	.6071	.6	.6
27	R19	8	43	44.11	-.29	.50	-.58	9	.5751	1.3	1.2
28	R25	7	27	28.78	-.27	.39	-.69	9	.5103	.7	.7
29	R19	5	43	43.94	-.24	.50	-.49	9	.6380	1.3	1.2
30	R19	3	48	48.35	-.23	.78	-.30	9	.7731	.9	.7
31	R22	8	34	34.86	-.13	.39	-.34	9	.7417	.6	.6
32	R20	10	46	46.34	-.13	.59	-.21	9	.8374	.9	2.0
33	R25	9	23	23.69	-.11	.41	-.28	9	.7861	1.3	1.2
34	R24	9	31	31.65	-.10	.38	-.25	9	.8067	.2	.2
35	R25	3	41	41.41	-.09	.46	-.19	9	.8521	.7	.6
36	R24	4	40	40.41	-.08	.44	-.18	9	.8574	.3	.3
37	R21	8	37	37.44	-.08	.41	-.18	9	.8584	1.8	1.6
38	R23	4	29	29.48	-.07	.38	-.18	9	.8578	.7	.6
39	R20	6	44	43.98	.01	.52	.01	9	.9913	.5	.5
40	R24	8	39	38.90	.02	.43	.04	9	.9677	.4	.4
41	R24	3	46	45.93	.02	.59	.04	9	.9675	.6	.5
42	R22	2	37	36.81	.03	.41	.08	9	.9389	.4	.4
43	R20	8	42	41.83	.04	.48	.08	9	.9386	.6	.6
44	R20	5	42	41.62	.08	.48	.18	9	.8627	.4	.4
45	R24	2	41	40.54	.10	.46	.21	9	.8385	.9	.9
46	R19	7	43	42.59	.10	.50	.20	9	.8455	1.2	1.2
47	R22	7	33	32.33	.10	.39	.26	9	.8000	1.4	1.6
48	R25	5	32	31.09	.14	.39	.35	9	.7338	.3	.4
49	R25	10	40	39.22	.15	.44	.34	9	.7413	.6	.5
50	R23	7	26	25.00	.16	.39	.40	9	.6994	.4	.4

51	R21	4	40	39.07	.18	.44	.40	9	.6961	.4	.4
52	R20	7	41	39.98	.21	.46	.45	9	.6620	.8	.7
53	R21	1	44	43.17	.22	.52	.41	9	.6881	1.0	.9
54	R20	4	44	43.08	.24	.52	.46	9	.6573	.7	.7
55	R25	1	40	38.68	.25	.44	.57	9	.5857	1.0	.9
56	R24	5	40	38.65	.26	.44	.58	9	.5777	.4	.4
57	R23	5	29	27.21	.27	.38	.70	9	.5041	.7	.6
58	R20	3	48	47.39	.33	.78	.43	9	.6800	.9	.9
59	R18	3	49	48.65	.33	1.05	.32	9	.7592	.8	.4
60	R23	2	32	29.65	.35	.39	.90	9	.3913	.7	.7
61	R25	4	36	33.35	.42	.40	1.03	9	.3301	1.4	1.5
62	R22	9	30	27.03	.44	.38	1.15	9	.2799	.4	.5
63	R18	4	47	45.83	.45	.66	.68	9	.5132	2.2	3.0
64	R21	9	33	29.89	.46	.39	1.19	9	.2643	2.1	2.0
65	R20	9	39	35.52	.60	.43	1.39	9	.1991	.7	.7
66	R23	6	35	31.06	.60	.40	1.51	9	.1664	1.3	1.3
67	R24	6	44	41.52	.60	.52	1.15	9	.2788	.4	.4
68	R22	1	44	41.33	.64	.52	1.23	9	.2491	.3	.3
69	R19	9	42	38.80	.65	.48	1.37	9	.2045	1.2	1.1
70	R25	8	36	31.40	.71	.40	1.75	9	.1138	1.0	1.0
71	R23	10	40	35.91	.73	.44	1.64	9	.1359	1.0	1.1
72	R24	7	41	36.70	.80	.46	1.75	9	.1148	.5	.5
73	R23	1	40	35.29	.83	.44	1.86	9	.0955	1.0	1.1
74	R21	5	42	37.17	.94	.48	1.98	9	.0796	.8	.9
75	R19	6	48	45.83	.96	.78	1.23	9	.2486	.8	.6
76	R19	10	49	47.62	1.01	1.05	.96	9	.3624	.9	.6
77	R21	2	45	39.20	1.35	.55	2.46	9	.0364	1.2	1.0
78	R22	3	48	43.63	1.61	.78	2.07	9	.0677	.8	.7
79	R18	8	49	44.92	2.04	1.05	1.95	9	.0832	1.1	2.0

EK 10. Kontrol Son Uygulama Puanlayıcı Grubu *Puanlayıcı X Öğrenci Performansı* Etkileşimleri

No	Rater	Stud perform	Observed Score	Expected Score	Bias (logit)	Model S.E.	t	d.f.	Prob.	Infit	Outfit
1	R18	1	50	48.76	.99<	1.45	.68	9	.5132	.1	.1
2	R18	2	50	48.21	1.42<	1.45	.98	9	.3537	.1	.1
3	R18	3	50	49.35	.29<	1.45	.20	9	.8478	.1	.1
4	R18	10	50	49.29	.38<	1.45	.26	9	.8011	.1	.1
5	R19	3	50	49.27	.40<	1.45	.28	9	.7869	.1	.1
6	R19	6	50	48.63	1.11<	1.45	.76	9	.4657	.1	.1
7	R21	10	40	46.96	-2.06	.48	-4.27	9	.0021	2.6	2.5
7	R23	9	15	21.42	-1.60	.56	-2.84	9	.0194	1.3	1.4
8	R22	4	30	35.64	-1.13	.44	-2.55	9	.0314	.3	.3
9	R20	1	39	43.28	-1.05	.47	-2.21	9	.0542	.5	.5
10	R19	1	47	48.63	-.97	.67	-1.44	9	.1824	1.0	.9
11	R23	6	29	33.79	-.95	.44	-2.14	9	.0613	1.4	1.4
12	R20	4	38	41.90	-.91	.47	-1.94	9	.0839	.7	.7
13	R19	5	46	47.80	-.80	.61	-1.33	9	.2173	2.0	2.0
14	R25	3	37	40.39	-.76	.46	-1.64	9	.1350	.9	.9
15	R18	8	45	47.01	-.75	.57	-1.33	9	.2166	1.0	.9
16	R24	1	42	44.56	-.71	.50	-1.40	9	.1944	.2	.2
17	R21	1	43	45.36	-.70	.52	-1.35	9	.2091	.6	.6
18	R25	2	32	35.20	-.64	.44	-1.45	9	.1818	1.4	1.4
19	R20	2	39	41.62	-.62	.47	-1.30	9	.2252	.2	.2
20	R20	7	35	37.90	-.61	.45	-1.35	9	.2096	.4	.4
21	R21	7	38	40.66	-.60	.47	-1.29	9	.2278	1.0	1.0
22	R19	9	40	42.35	-.57	.48	-1.19	9	.2652	.3	.3
23	R18	6	48	48.76	-.57	.78	-.72	9	.4885	1.6	.9
24	R23	8	26	28.31	-.45	.44	-1.02	9	.3333	1.0	1.0
25	R18	9	41	42.80	-.45	.49	-.92	9	.3803	2.4	2.3
26	R23	7	25	27.23	-.44	.45	-.99	9	.3488	.8	.8
27	R25	7	29	30.90	-.37	.44	-.84	9	.4212	.7	.7
28	R24	4	42	43.30	-.34	.50	-.68	9	.5123	.2	.2
29	R22	3	39	40.48	-.34	.47	-.72	9	.4891	.8	.8
30	R24	8	39	40.47	-.34	.47	-.71	9	.4929	.2	.2
31	R23	2	30	31.62	-.32	.44	-.72	9	.4908	.6	.6
32	R19	10	49	49.21	-.25	1.05	-.24	9	.8149	1.1	1.4
33	R22	10	39	40.10	-.25	.47	-.53	9	.6082	.2	.2
34	R22	8	31	32.08	-.21	.44	-.48	9	.6413	.3	.4
35	R18	4	48	48.31	-.20	.78	-.26	9	.8011	2.4	2.8
36	R25	9	24	24.91	-.18	.45	-.41	9	.6920	.7	.7
37	R23	5	30	30.93	-.18	.44	-.41	9	.6903	.3	.3
38	R24	5	42	42.54	-.14	.50	-.28	9	.7895	.2	.2
39	R21	6	45	45.36	-.12	.57	-.21	9	.8390	.5	.6
40	R22	1	37	37.36	-.08	.46	-.17	9	.8718	.4	.4
41	R22	6	37	37.36	-.08	.46	-.17	9	.8718	.7	.7
42	R24	2	43	43.05	-.01	.52	-.03	9	.9800	1.0	.9
43	R24	9	34	34.03	-.01	.45	-.01	9	.9894	.4	.4
44	R18	5	48	48.00	.00	.78	.00	9	.9983	2.4	2.8
45	R25	8	32	31.98	.00	.44	.01	9	.9922	1.0	1.0
46	R22	5	35	34.64	.07	.45	.16	9	.8743	1.0	1.1
47	R24	6	45	44.56	.14	.57	.24	9	.8129	.5	.5
48	R21	5	44	43.49	.14	.54	.27	9	.7941	.5	.5
49	R23	3	38	37.18	.18	.47	.38	9	.7138	.9	.9
50	R24	3	47	46.57	.18	.67	.27	9	.7915	.7	.7

51	R18	7	47	46.51	.21	.67	.31	9	.7653	2.2	2.8
52	R20	5	42	41.07	.23	.50	.46	9	.6571	.4	.4
53	R25	10	41	40.01	.23	.49	.48	9	.6448	.9	.9
54	R21	4	45	44.21	.24	.57	.43	9	.6766	.7	.7
55	R25	5	36	34.54	.30	.46	.66	9	.5260	.6	.6
56	R25	4	37	35.54	.31	.46	.66	9	.5227	.5	.5
57	R19	7	47	46.21	.32	.67	.48	9	.6404	2.2	2.4
58	R22	2	37	35.30	.36	.46	.77	9	.4613	.5	.5
59	R21	8	43	41.53	.38	.52	.73	9	.4843	2.1	1.9
60	R21	3	48	47.16	.45	.78	.57	9	.5831	1.1	1.2
61	R23	10	39	36.76	.49	.47	1.03	9	.3299	.7	.7
62	R20	6	45	43.28	.51	.57	.90	9	.3934	.6	.7
63	R20	3	47	45.59	.55	.67	.82	9	.4312	.8	.8
64	R20	9	35	32.18	.57	.45	1.25	9	.2421	.8	.8
65	R22	9	28	25.02	.59	.44	1.33	9	.2164	.8	.8
66	R25	1	40	37.26	.61	.48	1.26	9	.2377	.5	.5
67	R25	6	40	37.26	.61	.48	1.26	9	.2377	.8	.8
68	R19	8	48	46.74	.63	.78	.81	9	.4407	.9	.7
69	R19	4	49	48.14	.72	1.05	.68	9	.5118	1.1	1.1
70	R19	2	49	48.03	.79	1.05	.75	9	.4732	1.0	.7
71	R24	10	48	46.35	.79	.78	1.01	9	.3382	.9	.7
72	R24	7	43	39.56	.85	.52	1.63	9	.1370	.4	.4
73	R20	8	43	38.86	1.00	.52	1.94	9	.0848	.5	.5
74	R22	7	36	31.01	1.00	.46	2.20	9	.0554	.5	.5
75	R21	2	47	43.97	1.06	.67	1.58	9	.1481	1.0	1.0
76	R20	10	48	45.32	1.16	.78	1.48	9	.1721	.7	.5
77	R21	9	41	35.29	1.26	.49	2.56	9	.0305	1.7	1.7
78	R23	1	41	33.79	1.57	.49	3.18	9	.0111	1.3	1.3
79	R23	4	40	31.97	1.69	.48	3.51	9	.0066	1.1	1.0



GAZİLİ OLMAK AYRICALIKTIR