THE JOURNAL OF TURKISH EDUCATIONAL SCIENCES
TÜRK EĞİTİM BİLİMLERİ DERGİSİ

Açık Uçlu Maddelerin Puanlanmasında Dereceli Puanlama Anahtarı Türünün Puanlayıcı Davranışlarına Etkisi

The Effect of the Type of Rubric on Rater Behavior in Scoring Open-Ended Items

Umur Öç, Esra Onkun Özgür, İsmail Karakaya

Yazar Bilgileri

Umur Öç
Millî Eğitim Bakanlığı,
umur586@hotmail.com

Esra Onkun Özgür
Millî Eğitim Bakanlığı,
esraonkunozgur@gmail.com

İsmail Karakaya
Gazi Eğitim Fakültesi, Eğitim Bilimleri, Eğitimde Ölçme ve Değerlendirme,
ikarakaya@gazi.edu.tr

ÖZ

Bu araştırma, rutin olmayan açık uçlu matematik maddelerini içeren matematik başarı testinin puanlanmasında analitik ve bütünsel dereceli puanlama anahtarı kullanımının puanlayıcı davranışlarına etkisini çok yüzeyli Rasch modeli ile incelenmesini amaçlamaktadır. Araştırmanın çalışma grubunu, açık uçlu rutin olmayan matematik problemlerinden oluşan başarı testinin uygulandığı, devlet okulunda sekizinci sınıfta öğrenim gören 20 öğrenci ve cevaplanan başarı testini değerlendiren 16 matematik öğretmeni oluşturmaktadır. Bu çalışmada, betimsel araştırma yöntemlerinden tarama modeli kullanılmıştır. Bu çalışmada, Onkun-Özgür (2024) tarafından hazırlanmış, 15 farklı rutin olmayan açık uçlu matematik probleminden oluşan başarı testi, iki farklı oturum şeklinde, iki günde öğrencilere uygulanmıştır. Puanlayıcılardan elde edilen veriler çok yüzeyli Rasch modeli ile değerlendirilmiştir. Çalışmada, puanlayıcılara ait merkeze eğilim, yanlılık ve halo etkisi davranışları incelenmiştir. Çalışmanın bulguları incelendiğinde yapılan tüm puanlamalarda puanlayıcı, birey ve madde yüzeylerinde model veri uyumunun sağlandığı; bununla birlikte, bütünsel dereceli puanlama anahtarı kullanan puanlayıcılarda, analitik dereceli puanlama anahtarı kullanan puanlayıcılara göre daha az puanlayıcı etkisi olduğu belirlenmiştir.

Makale Bilgileri

Anahtar Kelimeler
Analitik dereceli puanlama anahtarı
Bütünsel dereceli puanlama anahtarı
Puanlayıcı davranışları
Puanlayıcı yanlılığı
Çok yüzeyli Rasch modeli

Keywords
Analytical rubric
Holistic rubric
Rater behaviors
Rater bias
Many facet Rasch model

Makale Geçmişi
Geliş: 01.07.2024
Kabul: 17.10.2024

ABSTRACT

This research aims to examine the effects of using analytical and holistic rubrics in scoring a mathematics achievement test containing non-routine open-ended mathematics items on rater behaviors using the many-facet Rasch model. The study group of the research consists of 20 eighth-grade students studying in a public school to whom the achievement test consisting of open-ended non-routine mathematics problems was applied and 16 mathematics teachers who evaluated the answered achievement test. In this study, the survey model, one of the descriptive research methods, was used. In this study, the achievement test consisting of 15 different non-routine open-ended mathematics problems prepared by Onkun-Özgür (2024) was applied to the students in two different sessions and on two days. The data obtained from the raters were evaluated using the many-facet Rasch model. In the study, the central tendency, bias and halo effect behaviors of the raters were examined. When the findings of the study were examined, it was seen that model data fit was provided in all scorings on the rater, individual and item surfaces, while raters who used holistic rubrics had less rater effects than raters who used analytical rubrics.

Makale Türü Araştırma

Önerilen Atıf Öç, U., Onkun-Özgür E., & Karakaya, İ. (2024). Açık uçlu maddelerin puanlanmasında dereceli puanlama anahtarı türünün puanlayıcı davranışlarına etkisi. *TEBD*, *22*(3), 2123-2151. https://doi.org/10.37217/tebd.1501178

Giriş

Eğitimde ölçme ve değerlendirme, öğrencilerin öğrenme süreçlerini anlamak, gelişimlerini değerlendirmek ve eğitim programlarının etkililiğini belirlemek için kritik bir rol oynar. Öğrenci başarısını ölçmek, uzun yıllar boyunca öğrencilerin öğretim sırasında edindikleri bilgileri ölçmek olarak değerlendirilmiştir ancak günümüzde üst düzey zihinsel becerilerin ölçülmesi, eğitim hedeflerinin, ekonomi ve iş gücünün doğasının değişmesi gibi nedenlerden dolayı daha önemli hâle gelmiştir (Aslanoğlu, 2022, s. 2). Klasik ölçme ve değerlendirme yöntemlerinin öğrencilerden beklenen üst düzey zihinsel becerilerin ölçülmesi sürecinde yetersiz kalması sebebiyle performans değerlendirmeye dayalı yaklaşımlar önem kazanmıştır. Performans değerlendirme temelli yaklaşımda kullanılan madde türlerinden biri de açık uçlu maddelerdir. Açık uçlu maddeler, öğrencilere kendi cevaplarını oluşturma (Haladyna ve Rodriguez, 2013) fırsatı sunarken öğretmenlere de öğrencilerin üst düzey zihinsel becerilerini ölçme fırsatı sunar (Popham, 2001). Açık uçlu maddeler, öğrencilerin derinlemesine düşünme ve yaratıcı problem çözme becerilerini değerlendirmenin önemli bir yolu olarak kabul edilir (Karakaya ve Şata, 2022, s. 31). Bu tür maddeler, öğrencilere kavramları anlama ve gerçek dünya problemlerine analitik bir yaklaşımla çözüm üretme fırsatı sağlar (Anderson ve Krathwohl, 2001; Crooks, 1988; Haladyna vd., 2002). Açık uçlu maddelerin sağlamış olduğu avantajlar düşünüldüğünde açık uçlu maddelerin kullanım alanlarından biri de rutin olmayan matematik problemleridir.

Rutin olmayan problemler, işlem becerisinden ziyade, verileri sınıflandırma, organize etme, ilişkileri görme gibi becerileri, gerektiğinde arka arkaya kullanmayı gerektiren, ispat fikrini geliştiren, gerçek dünya ile bağlantılar kurarak eleştirel düşünmeyi teşvik eden ve birden çok çözüm yolu olan problemlerdir (Altun, 2020). Matematik öğretiminde, öğrencilerin sadece temel kavramları öğrenmeleri değil, aynı zamanda derin anlayış geliştirmeleri ve yaratıcı problem çözme becerilerini kazanmaları önemlidir. Matematik öğretiminde bu hedeflere ulaşmak için açık uçlu rutin olmayan matematik problemleri kullanılarak öğrencilerin analitik ve eleştirel düşünme becerileri teşvik edilmektedir (Kilpatrick ve Lerman, 2020). Ancak bu tür problemlerin değerlendirilmesi ve puanlanması geleneksel çoktan seçmeli veya doğru-yanlış maddelerine göre daha karmaşık bir süreç gerektirmektedir (Crooks, 1988; Romagnano, 2001; Tuckman, 1991). Bu sebeple açık uçlu maddelerin nesnel olarak değerlendirilmesini sağlayacak alternatif yöntemlerin incelenmesi, öğrenci başarısının geçerli ve güvenilir şekilde kestirilebilmesi için önemlidir. Ancak açık uçlu maddelerin sınırlılıklarından bir tanesi de puanlayıcıların objektifliğini sağlamaktır. Puanlayıcıların objektif olamaması ölçüm sonuçlarının geçerliğini ve güvenirliğini olumsuz etkilemektedir. İlgili alanyazında puanlayıcıdan kaynaklanan varyansa puanlayıcı etkisi, puanlayıcı hatası, puanlayıcı yanlılığı, puanlayıcı davranışları ve problemli puanlayıcı davranışları adı verilmektedir (Farrokhi vd., 2011;

Haladyna, 1997, s. 139; Myford ve Wolfe, 2004; Royal ve Hecker, 2016). Araştırma kapsamında puanlayıcı etkisi kavramı kullanılmıştır.

Puanlayıcı etkisi, değerlendirme sürecindeki puanlayıcıların sergiledikleri tutum, davranış ve karar alma süreçlerini ifade eder. Puanlayıcı etkisi, puanlayıcıların katılığı veya cömertliği, ranj sınırlaması, halo etkisi, merkeze yönelme etkisi, yanlılık, tutarsızlık gibi farklı şekillerde ortaya çıkabilmektedir (Myford ve Wolfe, 2004; Saal vd., 1980). Değerlendirme sürecinin kalitesini artırmak için puanlayıcı etkisinin belirlenmesi ve gerekli önlemlerin alınması önemlidir. Açık uçlu maddelerin puanlanmasında puanlayıcı etkisini azaltmak ya da ortadan kaldırmak için dereceli puanlama anahtarı (Dunbar vd., 2006) ya da birden fazla puanlayıcı kullanılabilir (Şata, 2019). Puanlama anahtarının belirlenen öğrenme amaçlarına göre oluşturulması ve sınav uygulamasından önce hazırlanması geçerlik ve güvenirlik için önemlidir (Atılgan, 2009; Nitko ve Brookhart, 2014). Puanlama anahtarları, puanlama yapan puanlayıcıya ve zamana bakılmadan puanlama işleminin gerçekleştirilmesini sağlayarak puanlayıcı etkilerinin azalmasına destek olur (Moskal ve Leydens, 2000). Açık uçlu maddelerin puanlanmasında puanlayıcı etkisini azaltmak için puanlama anahtarının doğru şekilde kullanılması gerekmektedir. Açık uçlu maddeleri puanlarken maddelerin her biri için planlanan puan değerlerine uygun puanlama anahtarları kullanılmalıdır. Değerlendirme sürecinde amaç genel bir yargıya ulaşmaksa bütünsel puanlama anahtarı, performansın birbirinden farklı yönleri hakkında bilgi almaksa analitik puanlama anahtarı kullanılmalıdır (Nitko ve Brookhart, 2014). Bu kapsamda rutin olmayan matematik problemlerinin puanlanmasında analitik dereceli puanlama anahtarı da bütünsel dereceli puanlama anahtarı da kullanılabilir. Ancak hangi dereceli puanlama anahtarının kullanımında puanlayıcı etkisinin az olacağının belirlenmesi ölçüm sonuçlarının geçerliği ve güvenirliği açısından önemlidir.

Bu çalışmada rutin olmayan matematik problemlerinin puanlanmasında kullanılan analitik dereceli puanlama anahtarı ve bütünsel dereceli puanlama anahtarı türlerinin puanlayıcıların merkeze eğilim, yanlılık ve halo etkisi davranışlarına etkisi çok yüzeyli Rasch modeliyle belirlenmesi amaçlanmıştır. Linacre (1989) tarafından geliştirilen çok yüzeyli Rasch modeli, eğitim ve psikometri alanlarında yaygın olarak kullanılan bir modeldir. Bu model, puanlayıcıların cevapları nasıl değerlendirdiğini, hangi faktörlerin değerlendirme kararlarını etkilediğini ve puanlama sürecindeki doğruluk ve tutarlılığı nasıl etkilediğini belirlememize olanak tanır (Abu Kassim, 2007; DeMars, 2010; Wainer ve Thissen, 1993).

Bu çalışma, açık uçlu maddelerin puanlanmasında kullanılan analitik ve bütünsel dereceli puanlama anahtarlarının puanlayıcı davranışları üzerindeki etkisini belirleyerek değerlendirme süreçlerine ışık tutmayı amaçlamaktadır. Araştırmada puanlayıcıların halo etkisi, merkeze eğilim gibi davranışların yanında puanlayıcıların yanlılık davranışlarına da odaklanılmıştır. Bir diğer ifadeyle

araştırma puanlama anahtarı türünün puanlayıcı davranışlarına etkisine odaklanmaktadır. Bu yönüyle de öğretmenlerin öğrenciler arası göstermiş oldukları puanlayıcı davranışları hakkında ayrıntılı bilgi sunulmaktadır. Bu durum araştırmanın özgünlüğünü ortaya koymaktadır. Bu nedenle araştırma sonuçlarının matematik eğitimi ve değerlendirme süreçleri alanında önemli bir katkı sağlaması beklenmektedir. Tüm bu bilgiler ışığında rutin olmayan matematik problemlerini içeren açık uçlu maddelerin puanlanmasında analitik ve bütünsel dereceli puanlama anahtarı kullanımının puanlayıcı davranışlarına etkisi çok yüzeyli Rasch modeli ile araştırılmıştır. Bu amaç doğrultusunda aşağıdaki sorulara cevaplar aranmıştır:

1) Analitik dereceli puanlama anahtarı kullanan puanlayıcılarda;
    a) Birey, madde ve puanlayıcı yüzeylerine ait uyum istatistikleri ve model ölçüm sonuçları nasıldır?
    b) Merkeze eğilim davranışları nedir?
    c) Halo etkisi davranışları nedir?
    d) Yanlılık davranışları nedir?
2) Bütünsel dereceli puanlama anahtarı kullanan puanlayıcılarda;
    a) Birey, madde ve puanlayıcı yüzeylerine ait uyum istatistikleri ve model ölçüm sonuçları nasıldır?
    b) Merkeze eğilim davranışları nedir?
    c) Halo etkisi davranışları nedir?
    d) Yanlılık davranışları nedir?

Yöntem

Bu bölümde; araştırmanın deseni, örneklem/çalışma grubu/araştırma grubu, veri toplama araçları ve verilerin analizi, işlem bölümleri yer almaktadır.

Araştırmanın Deseni

Bu araştırma, rutin olmayan problemlere yönelik hazırlanan açık uçlu matematik maddelerini içeren başarı testinin puanlanmasında analitik ve bütünsel dereceli puanlama anahtarı kullanımının puanlayıcı davranışlarına etkisini çok yüzeyli Rasch modeli ile incelenmesini amaçladığı için tarama modellerinden betimsel tarama modeli ile desenlenmiştir. Betimsel tarama modeli, gruplar üzerinden yürütülen ve gruptaki bireylere ait özelliklerin betimlenmeye çalışıldığı araştırmalardır (Karakaya, 2012, s. 59). Puanlayıcıların tamamı tüm bireyleri puanladığı için araştırma tümüyle çaprazlanmış desen ile desenlenmiştir. Araştırma kapsamında birey, madde ve puanlayıcı olmak üzere üç yüzey yer almaktadır.

Çalışma Grubu

Araştırmanın çalışma grubu bireyler ve puanlayıcılar olmak üzere iki bölümden oluşmaktadır. Araştırmanın birey grubu 2023-2024 eğitim öğretim yılında Sakarya ili Merkez ilçesine bağlı bir devlet okulunun sekizinci sınıfında öğrenim gören 20 öğrenciden oluşmaktadır. Öğrenciler belirlenirken öğrencilerin gönüllülük esası ve kolay ulaşılabilir olması gibi özellikler göz önünde bulundurulmuştur. Araştırmanın puanlayıcı grubu ise farklı devlet okullarında görev yapan 16 matematik öğretmeninden oluşmaktadır. Puanlayıcılar belirlenirken gönüllülük esası ve yazılılarında açık uçlu maddeler kullanmaması gibi özellikler göz önünde bulundurulmuştur. Puanlayıcı grubunda yedi kadın, dokuz erkek puanlayıcı yer almaktadır.

Veri Toplama Araçları

Araştırmada veri toplama aracı olarak Onkun-Özgür (2024) tarafından hazırlanan açık uçlu maddelerden oluşan matematik başarı testi, analitik ve bütünsel dereceli puanlama anahtarı ve puanlayıcı değerlendirme formu kullanılmıştır. Açık uçlu maddeler, 8. sınıf öğrencilerine uygulanmak amacıyla 20 rutin olmayan matematik problemlerinden oluşmaktadır. Açık uçlu sorular hazırlandıktan sonra dört matematik öğretmeni, bir ölçme ve değerlendirme uzmanı ve bir Türkçe öğretmeni olmak üzere altı uzmandan görüş alınmıştır. Uzman görüşleri doğrultusunda üç madde testten çıkarılmış ve 17 maddeden oluşan test gerekli izinler alındıktan sonra önce pilot uygulama yapılmıştır. Pilot uygulamaya 30 öğrenci dâhil edilmiş ve madde istatistikleri hesaplanmıştır. İdeal maddelerin ayırt edicilik indeksi 0,30 ve 0,30'dan büyük, madde güçlük indeksi de 0,40 ile 0,60 arasındadır (Crocker ve Algina, 1986; Karakaya ve Şata, 2022, s. 110-113). Bu kapsamda istenen istatistiksel aralığın dışında olan iki madde tekrar dört matematik öğretmeninden görüş alınarak testten çıkarılmış ve araştırmaya 15 madde ile devam edilmiştir. Dereceli puanlama anahtarları geliştirilirken uzman görüşü alınmış, pilot uygulama yapılmış ve dereceli puanlama anahtarlarına son halleri verilmiştir. Analitik dereceli puanlama anahtarında altı ana kriter (1-Temel işlem becerileri, 2-Sembolik ve teknik dil ve işlemleri kullanma, 3- Modelleme, 4- Muhakeme ve argümantasyon, 5-Problem çözme için stratejiler oluşturma, 6- Temsil ile gösterim) yer alırken her ana kriter dört alt kategoriden oluşmaktadır. Bütünsel dereceli puanlama anahtarında ise dört kriter yer almaktadır.

Verilerin Toplanması

Araştırmanın veri toplama süreci üç aşamadan oluşmaktadır. İlk aşamada ortaokul öğrencilerine rutin olmayan açık uçlu matematik maddelerini içeren başarı testi iki farklı oturum şeklinde, iki günde öğrencilere uygulanmıştır. Uygulama öncesinde öğrencilere araştırma hakkında bilgi verilmiş ve elde edilen verilerin sadece araştırma kapsamında kullanılacağı belirtilmiştir. Araştırmanın ikinci aşamasında puanlayıcılara hem analitik hem de bütünsel dereceli puanlama anahtarı hakkında bilgi verilmiş ve örnek uygulama yapılmıştır. Araştırmanın üçüncü aşamasında

puanlayıcılardan öğrenci kâğıtlarını değerlendirmeleri istenmiştir ve bu sayede araştırma kapsamında ele alınan veriler elde edilmiştir. Araştırma; Onkun-Özgür'ün (2024) yüksek lisans tezi çalışması kapsamında toplanan veriler üzerinde yürütülmüştür.

Verilerin Analizi

Araştırma kapsamında elde edilen verilerin analizinde çok yüzeyli Rasch ölçme modeli kullanılmıştır. Analizler yapılırken FACET paket programından faydalanılmıştır. Çok yüzeyli Rasch modeli analizine başlamadan önce çok yüzeyli Rasch modelinin varsayımları incelenmiştir. Çok yüzeyli Rasch modeli, Rasch modellerinden olduğu için Rasch modeli varsayımlarını karşılamalıdır (Baker ve Kim, 2004, s. 115; Eckes, 2015, s. 124; Embretson ve Reise, 2000, s. 65; Farrokhi vd., 2011; Farrokhi vd., 2012). Bu varsayımlar tek boyutluluk, yerel bağımsızlık ve model veri uyumudur. Tek boyutluluk, testin bir özelliği ölçmesini ifade etmektedir (Harvey ve Hammer, 1999). Veri toplama araçlarının geliştirilmesi sürecinde veri toplama araçlarının tek bir yapıyı ölçtüğü uzman görüşleri ile belirlenmiştir. Bu kapsamda tek boyutluluk varsayımının karşılandığı belirlenmiştir. Yerel bağımsızlık varsayımı, testte yer alan bir maddeye verilen cevabın diğer maddeye verilen cevaptan etkilenmemesidir (Crocker ve Algina, 1986, s. 343). Yerel bağımsızlık varsayımı, tek boyutluluk varsayımı ile ilişkili olup tek boyutluluk varsayımı karşılandığında yerel bağımsızlık varsayımı da karşılanmaktadır (Hambleton vd., 1991). Model veri uyumu, ±2 aralığı dışında kalan standartlaştırılmış artıkların oranının toplam etkileşim sayısının %5'ini ve ±3 aralığı dışında kalan standartlaştırılmış artık oranının da toplam etkileşim sayısının %1'ini geçmemesini ifade etmektedir (Linacre, 2014, s. 181). Araştırma kapsamında ±2 ve ±3 aralığı dışında kalan standartlaştırılmış artık oranları Tablo 1'de verilmiştir.

Tablo 1. Dereceli Puanlama Anahtarı Türüne Göre Standartlaştırılmış Artıklar

Dereceli puanlama anahtarı	*Aralık*	*Artık (n)*	*Artık (%)*
Analitik	±2	100	2,08
	±3	62	1,29
Bütünsel	±2	100	2,08
	±3	52	1,08

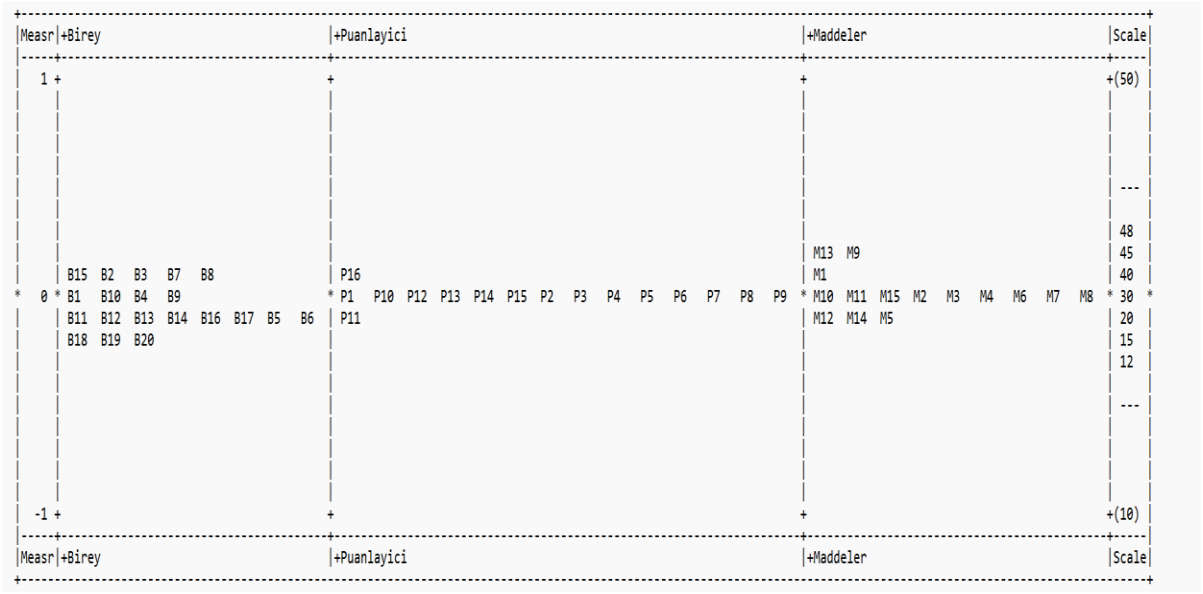
Tablo 1 incelendiğinde her iki dereceli puanlama anahtarı türünün ±2 aralığında artık yüzdelerinin uygun aralıkta olduğu ancak ±3 aralığında artık yüzdelerinin uygun aralığın dışında olduğu görülmektedir. McNamara (1996) ±2 ve ±3 aralığı dışında kalan artık yüzdesinde çok büyük bir sapma olmadığı sürece analize devam edilmesi gerektiğini ifade etmiştir. Çünkü Rasch analizinde uygun model seçme imkânı varken çok yüzeyli Rasch modelinde alternatif bir model bulunmamaktadır. Çok yüzeyli Rasch modeli varsayımları incelendikten sonra analize devam edilmiştir.

Bulgular

Araştırmanın amacı, rutin olmayan problemlere yönelik hazırlanan açık uçlu matematik maddelerini içeren başarı testinin puanlanmasında analitik ve bütünsel dereceli puanlama anahtarı kullanımının puanlayıcı davranışlarına etkisini incelemektir. Bu kapsamda puanlayıcılardan öğrencilerin cevaplarını analitik ve bütünsel dereceli puanlama anahtarı kullanarak puanlamaları istenmiştir. Araştırmada elde edilen bulgular, analitik dereceli puanlama anahtarından elde edilen bulgular ve bütünsel dereceli puanlama anahtarından elde edilen bulgular olmak üzere iki farklı başlıkta sunulmuştur.

Analitik Dereceli Puanlama Anahtarı

Analitik dereceli puanlayıcı anahtarı kullanan puanlayıcılardan elde edilen verilere ait kalibrasyon haritası Şekil 1'de verilmiştir.



Şekil 1. Analitik dereceli puanlama anahtarı kullanımına ait kalibrasyon haritası

Şekil 1'deki analitik dereceli puanlama anahtarı kullanan puanlayıcılara ait logit cetveli incelendiğinde bireylerin, maddelerin ve puanlayıcıların logit cetvelinde sıfır noktası etrafında kümelendiği görülmektedir. Bu durum, birey performanslarının benzer olduğu, maddelerin benzer güçlükte olduğu ve puanlayıcıların analitik dereceli puanlama anahtarı kullanırken benzer davranışlar sergilemiş olabileceklerini göstermektedir. Logit cetvelinde artan logit değeri birey yüzeyi için yetenek seviyesinin arttığı, madde yüzeyi için madde güçlüğünün arttığı anlamına gelmektedir. Örneğin B15, B2, B3, B7 ve B8 numaralı bireyler grup içinde yetenek seviyesi en yüksek bireylerken B18, B19 ve B20 numaralı bireyler de grup içinde yetenek seviyesi en düşük bireylerdir. Benzer şekilde M13 ve M9 numaralı maddelerin madde güçlüğü yüksekken yani maddeler kolayken M12, M14 ve M5 numaralı maddelerin madde güçlüğü düşüktür yani maddeler zordur. Logit cetvelinde puanlayıcı

yüzeyi için artan logit değeri ise puanlayıcı cömertliği davranışına işaret ederken azalan logit değeri puanlayıcı katılığı davranışına işaret etmektedir. Puanlayıcı davranışlarının detaylı incelenmesi için her bir yüzeye ait model ölçüm sonuçları incelenmelidir. Birey yüzeyine ait model ölçüm sonuçları Tablo 2'de verilmiştir.

Tablo 2. Analitik Dereceli Puanlama Anahtarı Birey Ölçüm Raporu

Birey	*Gözlenen Ortalama*	*Düzeltilmiş Ortalama*	*Logit Değeri*	*Standart Hata*	*Uyum İçi*	*Uyum Dışı*
B1	26,62	25,90	-0,03	0,01	0,57	0,56
B2	36,93	38,86	0,08	0,01	1,12	1,17
B3	39,75	41,77	0,11	0,01	1,38	1,39
B4	26,87	26,18	-0,03	0,01	1,48	1,61
B5	20,21	18,29	-0,12	0,01	0,81	0,85
B6	22,61	20,92	-0,08	0,01	0,87	0,89
B7	36,83	38,74	0,08	0,01	1,24	1,19
B8	35,69	37,43	0,06	0,01	0,96	0,87
B9	30,93	31,54	0,01	0,01	0,73	0,79
B10	28,55	28,41	-0,01	0,01	1,19	1,15
B11	18,95	17,05	-0,14	0,01	0,68	0,59
B12	19,13	17,19	-0,13	0,01	0,95	1,02
B13	19,38	17,46	-0,13	0,01	0,85	0,95
B14	21,07	19,18	-0,10	0,01	1,15	1,37
B15	37,06	38,97	0,08	0,01	1,10	1,06
B16	21,57	19,76	-0,10	0,01	1,09	0,94
B17	20,25	18,34	-0,11	0,01	0,90	0,99
B18	17,62	15,82	-0,16	0,01	0,93	0,98
B19	17,43	15,67	-0,16	0,01	1,55	1,59
B20	17,48	15,72	-0,16	0,01	0,74	0,67
Ortalama	25,75	25,16	-0,05	0,01	1,01	1,03
S (Evren)	7,59	9,15	0,09	0,00	0,26	0,29
S (Örneklem)	7,78	9,39	0,09	0,00	0,27	0,30

Model, Evren RMSE=0,01; Düzeltilmiş S.H.=0,09; Ayırma Oranı=9,15; Ayırma İndeksi=12,54; Güvenirlik 0,99
Model, Örneklem RMSE=0,01; Düzeltilmiş S.H.=0,09; Ayırma Oranı=9,39; Ayırma İndeksi=12,86; Güvenirlik=0,99
Model, Ki-kare (Sabit etkili): 1702,0; sd=19; p=0,00
Model, Ki-kare (Normal): 18,8; sd=18; p=0,41

Bireylere ait logit aralığı bireylerin yetenek düzeylerindeki değişimi göstermektedir. Birey yüzeyine ait model ölçüm sonuçlarının yer aldığı Tablo 2 incelendiğinde bireylerin yetenek düzeyleri arasında 0,27 [0,11-(-0,16)] logit değişim olduğu görülmektedir. Bu durum birey performanslarının geniş sayılabilecek bir aralığa sahip olduğunu göstermektedir. Birey yüzeyine ait logit değerleri ortalaması -0,05 iken standart sapma 0,09'dur. Bu kapsamda B19 numaralı birey en düşük yetenek seviyesine sahipken B3 numaralı birey en yüksek yetenek seviyesine sahiptir. Birey yüzeyine ait uyum istatistikleri incelendiğinde uyum içi ve uyum dışı istatistiklerinin ortalamaları bire yakındır. Bu durum birey yüzeyi için model veri uyumunun iyi olduğunu göstermektedir. Birey yüzeyine ait ayırma oranı (9,39), ayırma indeksi (12,86) ve güvenirlik (0,99) istatistikleri incelendiğinde bu istatistiklerin büyük olduğu görülmektedir. Bu durum birey yeteneklerinin istatistiksel açıdan farklı olduğunu göstermektedir. Ancak birey yetenekleri arasındaki farkın istatistiksel olarak anlamlı olup olmadığının kontrol için sabit etkili ki kare değeri incelenmelidir. Sabit etkili ki kare değeri

incelendiğinde bu değerin istatistiksel olarak anlamlı olduğu görülmektedir ($\chi^2(sd)$=1702,0(19); p=0,00<0,01). Bu durum birey yeteneklerinin istatistiksel açıdan anlamlı olarak farklılaştığını göstermektedir.

Birey yüzeyine ait ölçüm sonuçları incelendikten sonra madde yüzeyine ait ölçüm sonuçları incelenmiştir. Madde yüzeyine ait ölçüm sonuçları Tablo 3'te yer almaktadır.

Tablo 3. Analitik Dereceli Puanlama Anahtarı Madde Ölçüm Raporu

Maddeler	*Gözlenen Ortalama*	*Düzeltilmiş Ortalama*	*Logit Değeri*	*Standart Hata*	*Uyum İçi*	*Uyum Dışı*
M1	32,67	33,97	0,08	0,01	0,54	0,57
M2	26,76	25,76	0,02	0,01	0,74	0,66
M3	24,05	22,15	-0,02	0,01	0,83	0,79
M4	24,23	22,35	-0,01	0,01	0,60	0,56
M5	16,83	14,94	-0,13	0,01	1,20	1,27
M6	23,71	21,76	-0,02	0,01	1,02	0,87
M7	28,37	28,01	0,04	0,01	0,56	0,53
M8	23,62	21,61	-0,02	0,01	1,16	1,23
M9	39,12	41,53	0,16	0,01	1,40	1,29
M10	21,96	19,68	-0,04	0,01	1,63	1,83
M11	22,09	19,83	-0,04	0,01	1,16	1,09
M12	18,23	16,06	-0,10	0,01	1,62	1,47
M13	39,82	42,24	0,17	0,01	0,89	0,94
M14	19,21	16,92	-0,09	0,01	1,64	1,54
M15	32,67	33,97	0,00	0,01	0,84	0,82
Ortalama	25,75	24,73	0,00	0,01	1,06	1,03
S (Evren)	6,60	8,16	0,08	0,00	0,37	0,38
S (Örneklem)	6,83	8,45	0,09	0,00	0,39	0,40

Model, Evren RMSE=0,01; Düzeltilmiş S.H.=0,08; Ayırma Oranı=9,74; Ayırma İndeksi=13,32; Güvenirlik=0,99
Model, Örneklem RMSE=0,01; Düzeltilmiş S.H.=0,09; Ayırma Oranı=10,08; Ayırma İndeksi=13,78; Güvenirlik=0,99
Model, Ki-kare (Sabit etkili): 1359,4; sd=14; p=0,00
Model, Ki-kare (Normal): 13,9; sd=13; p=0,38
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi

Analitik dereceli puanlama anahtarı madde yüzeyine ait ölçüm sonuçlarının yer aldığı Tablo 3 incelendiğinde maddelerin logit değerlerinin -0,13 ile 0,17 arasında değiştiği görülmektedir. Bu durum madde güçlüklerinin arasında 0,30 [0,17-(-0,13)] logit değişim olduğunu göstermektedir. Testte yer alan maddelerden en küçük logit değerine sahip M5 en zor madde iken en büyük logit değerine sahip M13 ise en kolay maddedir. Madde yüzeyine ait uyum istatistiklerinin ortalaması beklenen değer olan bire oldukça yakındır. Bu durum model veri uyumunun iyi olduğunun göstergesidir. Madde yüzeyine ait ayırma oranı (10,08), ayırma indeksi (13,78) ve güvenirlik (0,99) istatistikleri incelendiğinde bu istatistiklerin büyük olduğu görülmektedir. Bu durum madde güçlüklerinin istatistiksel açıdan farklı olduğunu göstermektedir. Ancak madde güçlükleri arasındaki farkın istatistiksel olarak anlamlı olup olmadığının kontrol için sabit etkili ki kare değeri incelenmelidir. Sabit etkili ki kare değeri incelendiğinde bu değerin istatistiksel olarak anlamlı olduğu görülmektedir ($\chi^2(sd)$=1359,4(14); p=0,00<0,01). Bu durum madde güçlüklerinin istatistiksel açıdan anlamlı olarak farklılaştığını göstermektedir.

Birey ve madde yüzeylerine ait ölçüm sonuçları incelendikten sonra puanlayıcı yüzeyine ait ölçüm sonuçları incelenmiştir. Puanlayıcı yüzeyine ait ölçüm sonuçları Tablo 4'te yer almaktadır.

Tablo 4. Analitik Dereceli Puanlama Anahtarı Puanlayıcı Ölçüm Raporu

Puanlayıcı	*Gözlenen Ortalama*	*Düzeltilmiş Ortalama*	*Logit Değeri*	*Standart Hata*	*Uyum İçi*	*Uyum Dışı*
P1	24,29	22,06	-0,02	0,01	0,69	0,63
P2	28,62	28,30	0,04	0,01	0,81	0,84
P3	26,43	24,95	0,01	0,01	0,80	0,68
P4	23,90	21,47	-0,02	0,01	1,06	1,02
P5	25,72	23,96	0,00	0,01	0,93	0,84
P6	26,18	24,57	0,01	0,01	0,80	0,73
P7	26,49	25,07	0,01	0,01	0,77	0,68
P8	26,45	25,04	0,01	0,01	1,76	2,74
P9	28,92	28,75	0,04	0,01	1,12	1,09
P10	29,11	29,09	0,05	0,01	1,29	1,31
P11	17,90	15,21	-0,12	0,01	0,78	0,71
P12	23,48	20,94	-0,03	0,01	1,05	1,00
P13	23,41	20,86	-0,03	0,01	1,10	1,05
P14	24,87	22,75	-0,01	0,01	0,88	0,90
P15	25,82	24,04	0,00	0,01	1,00	0,90
P16	30,35	31,02	0,06	0,01	1,36	1,40
Ortalama	25,75	24,25	0,00	0,01	1,01	1,03
S (Evren)	2,86	3,77	0,04	0,00	0,27	0,49
S (Örneklem)	2,96	3,89	0,04	0,00	0,28	0,51

Model, Evren RMSE=0,01; Düzeltilmiş S.H.=0,04; Ayırma Oranı=4,60; Ayırma İndeksi=6,46; Güvenirlik=0,95
Model, Örneklem RMSE=0,01; Düzeltilmiş S.H.=0,04; Ayırma Oranı=4,76; Ayırma İndeksi=6,67; Güvenirlik=0,96
Model, Ki-kare (Sabit etkili): 299,8; sd=15; p=0,00
Model, Ki-kare (Normal): 14,3; sd=14; p=0,43
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi

Analitik dereceli puanlama anahtarı puanlayıcı yüzeyine ait ölçüm sonuçlarının yer aldığı Tablo 4 incelendiğinde puanlayıcıların logit değerlerinin -0,12 ile 0,06 arasında değiştiği görülmektedir. Bu durum puanlayıcıların katılık ve cömertlik durumları arasında 0,18 [0,06-(-0,12)] logit değişim olduğunu göstermektedir. Bu bilgiler ışığında logit değeri en küçük olan yani en katı puanlayıcının P11, logit değeri en büyük olan yani cömert puanlayıcının da P16 olduğu görülmektedir. Puanlayıcı yüzeyine ait uyum istatistiklerinin ortalaması bire yakındır. Bu durum puanlayıcı yüzeyine ait model veri uyumunun iyi olduğunun göstergesidir. Puanlayıcı yüzeyine ait ayırma oranı (4,76), ayırma indeksi (6,67) ve güvenirlik (0,96) istatistiklerinin büyük olduğu görülmektedir. Bu durum puanlayıcıların puanlama esnasında birbirlerinden farklı puanlayıcı davranışı sergilemiş olabileceklerinin göstergesidir. Bu durumun istatistiksel açıdan değerlendirilmesi için sabit etkili ki-kare değeri incelenmelidir. Sabit etkili ki-kare değeri incelendiğinde bu değerin istatistiksel açıdan anlamlı olduğu ($\chi^2(sd)$=299,8(15); p=0,00<0,01) ve dolayısıyla puanlayıcıların puanlama esnasında farklı davranışlar sergilediği görülmektedir.

Araştırma kapsamında oluşturulan modeldeki yüzeylere ait ölçüm sonuçları incelendikten sonra puanlayıcı davranışları incelenmiştir. Araştırmada puanlayıcı davranışlarından merkeze eğilim, halo etkisi ve yanlılık davranışları incelenmiştir.

Merkeze Eğilim

Puanlayıcıların merkeze eğilim davranışları incelenirken önce grup düzeyindeki davranışlar sonrasında da bireysel düzeydeki davranışlar incelenmiştir. Analitik dereceli puanlama anahtarı ile puanlama yapan puanlayıcıların grup düzeyinde merkeze eğilim davranışları incelenirken kategori istatistikleri, birey ve madde yüzeylerine ait istatistikler incelenmiştir. Kategori istatistikleri Tablo 5'te verilmiştir.

Tablo 5. Analitik Dereceli Puanlama Anahtarına Ait Kategori İstatistikleri

Kategoriler	*Frekans (f)*	*Yüzde (%)*	*Yığılmalı yüzde (%)*	*Ortalama logit*	*Beklenen logit*	*Uyum dışı*
10	1570	33	33	-0,14	-0,14	1,10
12	155	3	36	-0,15	-0,13	0,70
13	54	1	37	-0,10	-0,12	2,00
15	47	1	38	-0,12	-0,11	0,90
17	44	1	39	-0,11	-0,11	0,80
18	143	3	42	-0,11	-0,10	0,70
20	464	10	52	-0,08	-0,09	0,70
22	129	3	54	-0,06	-0,08	1,20
23	96	2	56	-0,05	-0,07	0,80
25	63	1	58	-0,04	-0,06	1,00
27	59	1	59	-0,05	-0,05	0,90
28	100	2	61	-0,04	-0,04	0,80
30	221	5	66	-0,04	-0,03	0,70
32	73	2	67	-0,03	-0,02	0,80
33	61	1	68	-0,01	-0,01	0,70
35	57	1	70	-0,02	0,00	1,30
37	65	1	71	0,00	0,01	1,10
38	99	2	73	0,03	0,02	0,70
40	157	3	76	0,02	0,03	1,20
42	51	1	77	0,05	0,04	0,70
43	62	1	79	0,03	0,05	1,90
45	55	1	80	0,05	0,06	1,70
47	61	1	81	0,06	0,07	1,00
48	136	3	84	0,08	0,08	1,00
50	778	16	100	0,09	0,09	1,20

Kategori istatistiklerinin olduğu Tablo 5 incelendiğinde 10, 20 ve 50 numaralı kategorilerin çok kullanıldığı, diğer kategorilerin ise benzer sayıda kullanıldığı görülmektedir. Tablo 5'ten hareketle puanlayıcıların merkeze eğilim davranışları olmadığı söylenebilir. Ancak merkeze eğilim davranışın belirlenmesinde kategori istatistikleri tek başına yeterli değildir. Merkeze eğilim davranışı belirlenirken birey yüzeyine ait istatistikler ve madde yüzeyine ait istatistikler de incelenmelidir. Birey yüzeyine ait istatistikler incelendiğinde (Tablo 2), ayırma oranı (9,39), ayırma indeksi (12,86) ve güvenirlik (0,99) istatistiklerinin büyük olduğu ve ki-kare istatistiğinin de ($\chi^2(sd)$=1702,0(19); p=0,00<0,01) anlamlı olduğu görülmüştür. Bu durum bireylerin yeteneklerinin doğru ve istatistiksel olarak anlamlı bir şekilde ayrıldığını göstermektedir. Madde yüzeyine ait uyum istatistikleri de incelendiğinde (Tablo 3) maddelere ait uyum istatistiklerinin kabul edilebilir aralıkta olduğu görülmektedir. Tüm göstergeler bir arada ele alındığında analitik dereceli puanlama anahtarı

kullanan puanlayıcı grubunda, grup düzeyinde merkeze eğilim davranışının olmadığı söylenebilir. Grup düzeyinde merkeze eğilim davranışı incelendikten sonra bireysel düzeyde merkeze eğilim davranışı incelenmiştir. Bireysel düzeyde merkezi eğilim davranışı belirlenirken puanlayıcılara ait kategori istatistikleri ve uyum istatistikleri incelenmektedir. Puanlayıcılara ait uyum istatistikleri incelendiğinde P8 numaralı puanlayıcının uyum dışı istatistiğinin beklenen aralık dışında olduğu görülmektedir. Bu durum P8 numaralı puanlayıcının merkeze eğilim davranışı gösterdiğinin göstergesidir. Merkeze eğilim davranışının belirlenmesindeki bir diğer yöntem de puanlayıcıya ait kategori istatistiklerinden uyum dışı istatistiğinin incelenmesidir. Araştırma kapsamında maddeler değerlendirilirken kriter ve madde etkileşimi fazla olduğu için her birini tabloda vermek mümkün olmamıştır. Her bir puanlayıcıya ait kategori istatistikleri ve uyum dışı istatistikleri incelendiğinde analitik dereceli puanlama anahtarı kullanan puanlayıcıların tamamında merkeze eğilim davranışının olduğu belirlenmiştir.

Halo Etkisi

Puanlayıcıların halo etkisi davranışları incelenirken önce grup düzeyindeki davranışlar sonrasında da bireysel düzeydeki davranışlar incelenmiştir. Analitik dereceli puanlama anahtarı ile puanlama yapan puanlayıcıların grup düzeyinde halo etkisi davranışları incelenirken madde yüzeyine ait istatistikler incelenmiştir. Madde yüzeyine ait istatistikler incelendiğinde (Tablo 3) ayırma oranının (10,08), ayırma indeksinin (13,78) ve güvenirlik istatistiğinin (0,99) yüksek olduğu ve ki kare istatistiğinin de istatistiksel olarak anlamlı ($\chi^2(sd)$=1359,4(14); p=0,00<0,01) olduğu görülmektedir. Bu durum puanlayıcıların grup düzeyinde halo etkisi davranışına sahip olmadıklarının göstergesidir. Grup düzeyindeki istatistikler incelendikten sonra bireysel istatistikler incelenmiştir. Halo etkisi bireysel istatistiklerde incelenirken öncelikle madde güçlükleri arasındaki farka bakılmaktadır. Madde güçlükleri arasındaki fark büyükse puanlayıcıların uyum istatistiklerinin 1'den çok büyük olması, madde güçlükleri arasındaki fark küçükse puanlayıcıların uyum istatistiklerinin 1'den çok küçük olması halo etkisi davranışına işaret etmektedir (Myford ve Wolfe, 2004). Madde güçlüklerinin yer aldığı Tablo 3 incelendiğinde madde güçlükleri arasındaki farkın küçük (0,30) olduğu görülmektedir. Bu kapsamda Tablo 4'te yer alan puanlayıcılara ait uyum istatistikler incelendiğinde P1 (0,63), P7 (0,68) ve P11 (0,71) numaralı puanlayıcılarda halo etkisi davranışının olduğu söylenebilir. Halo etkisi davranışının belirlenmesinde kullanılan bir diğer yöntem de madde güçlüklerinin eşitlenerek çok yüzeyli Rasch analizini tekrar etmektir (Linacre, 2023, s. 360). Linacre (2023), madde güçlüklerinin eşitlenerek oluşturulduğu modelde uyum istatistikleri 1'e eşit olan yani modele mükemmel uyum gösteren puanlayıcıların halo etkisi davranışına sahip olduklarını ifade etmektedir. Bu kapsamda madde güçlükleri eşitlenerek yeni birçok yüzeyli Rasch

analizi yapılmıştır. Madde güçlükleri eşitlendikten sonra analitik dereceli puanlama anahtarı kullanan puanlayıcılara ait ölçüm raporu Tablo 6'da verilmiştir.

Tablo 6. Madde Güçlükleri Eşitlendikten Sonra Puanlayıcı Yüzeyine Ait Ölçüm Raporu

Puanlayıcı	*Uyum İçi*	*Uyum Dışı*
P1	0,77	0,76
P2	0,78	0,8
P3	0,98	0,98
P4	1,05	1,03
P5	0,96	0,95
P6	0,84	0,83
P7	0,9	0,9
P8	1,57	1,65
P9	1,03	1,03
P10	1,11	1,12
P11	0,69	0,69
P12	1,09	1,11
P13	1,13	1,14
P14	0,86	0,85
P15	1,03	1,03
P16	1,16	1,17

Tablo 6'da yer alan her bir puanlayıcıya ait uyum istatistikleri incelendiğinde modele mükemmel uyum sağlayan puanlayıcının olmadığı dolayısıyla halo etkisi davranışı sergileyen puanlayıcının olmadığı söylenebilir.

Tüm bireysel istatistikler göz önüne alındığında analitik dereceli puanlama anahtarı kullanan puanlayıcılardan üç tanesinde (P1, P7, P11) halo etkisi davranışı belirlenmiştir.

Yanlılık

Analitik dereceli puanlama anahtarı kullanan puanlayıcıların yanlılık davranışının tespiti için *puanlayıcı x birey (pxb)* etkileşimi incelenmiştir. Grup düzeyindeki yanlılık davranışına ait ki kare istatistiği incelendiğinde ki kare istatistiğinin anlamlı olmadığı gözlenmiştir ($\chi^2(sd)$=286,8(320); p>0,01). Bu durum, analitik dereceli puanlama anahtarı kullanan puanlayıcıların grup düzeyinde yanlılık davranışı olmadığının göstergesidir. Grup düzeyinde istatistikler incelendikten sonra bireysel istatistikler incelenmiştir.

Çok yüzeyli Rasch analizinde bireysel düzeyde yanlılık davranışı incelenirken t istatistiği kullanılmaktadır. Linacre (2023), ±2 aralığı dışındaki t değerlerinin istatistiksel olarak anlamlı etkileşimler olduğunu ifade etmiştir (s. 190). Araştırma kapsamında 20 birey 16 puanlayıcı olmak üzere 320 etkileşim yer almaktadır. Her bir etkileşimi tabloda vermek mümkün olmadığı için sadece anlamlı etkileşimler raporlanmıştır. Analitik dereceli puanlama anahtarı kullanımına ait anlamlı etkileşimler Tablo 7'de yer almaktadır.

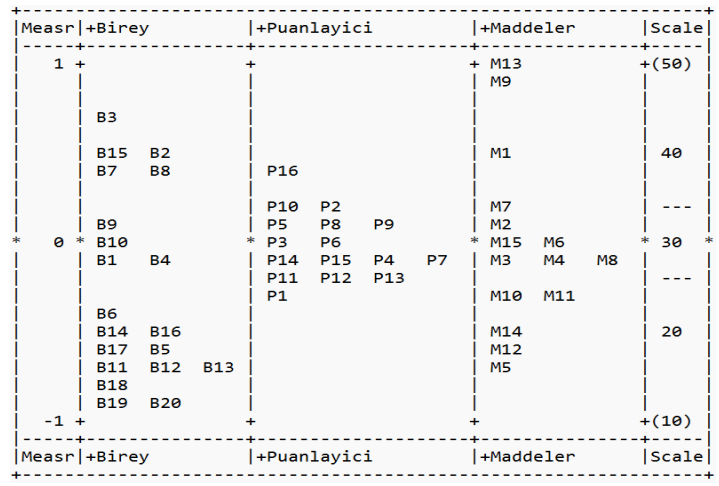
Tablo 7. Analitik Dereceli Puanlama Anahtarı Kullanımına Ait Anlamlı Etkileşimler

Puanlayıcı	*Birey*	*Gözlenen Puan*	*Beklenen Puan*	*Bias (logit)*	*Standart Hata*	*t*
P8	B17	513,00	390,81	0,15	0,03	4,47
P8	B19	426,00	330,13	0,14	0,03	4,11
P8	B12	450,00	355,98	0,13	0,03	3,74
P8	B5	444,00	363,35	0,10	0,03	3,12
P8	B18	373,00	310,78	0,10	0,04	2,73
P8	B13	403,00	338,54	0,09	0,03	2,61
P13	B2	628,00	559,93	0,09	0,04	2,20
P14	B4	487,00	426,38	0,07	0,03	2,09
P12	B13	178,00	228,18	-0,19	0,10	-2,03
P8	B8	433,00	504,91	-0,08	0,03	-2,43
P8	B15	431,00	516,22	-0,10	0,03	-2,92
P8	B2	422,00	512,35	-0,10	0,03	-3,07
P8	B7	415,00	508,35	-0,11	0,03	-3,15
P8	B3	463,00	553,05	-0,11	0,03	-3,32

Analitik dereceli puanlama anahtarı kullanan puanlayıcıların yanlılık davranışlarına ait istatistiklerin yer aldığı Tablo 7 incelendiğinde dört puanlayıcının toplam 14 yanlılık davranışı sergilediği görülmektedir. Puanlayıcılardan P13 numaralı puanlayıcının B2 numaralı bireye ve P14 numaralı puanlayıcının da B4 numaralı bireye olması gerekenden fazla puan verdiği görülmektedir. Puanlayıcılardan P12 numaralı puanlayıcının da B13 numaralı bireye olması gerekenden az puan verdiği görülmektedir. P8 numaralı puanlayıcının ise B5, B12, B13, B17, B18 ve B19 numaralı bireylere olması gerekenden fazla ve B2, B3, B7, B8 ve B15 numaralı bireylere de olması gerekenden az puan verdiği görülmektedir.

Bütünsel Dereceli Puanlama Anahtarı

Analitik dereceli puanlama anahtarı kullanan puanlayıcıların puanlayıcı davranışları incelendikten sonra bütünsel dereceli puanlama anahtarı kullanan puanlayıcıların puanlayıcı davranışları incelenmiştir. Bütünsel dereceli puanlama anahtarı kullanan puanlayıcılara ait kalibrasyon haritası Şekil 2'de verilmiştir.



Şekil 2. Bütünsel dereceli puanlama anahtarı kullanımına ait kalibrasyon haritası

Şekil 2'deki bütünsel dereceli puanlama anahtarı kullanan puanlayıcılara ait logit cetveli incelendiğinde bireylerin, maddelerin ve puanlayıcıların logit cetvelinde uçtan uca sıralandığı görülmektedir. Bu durum, birey performanslarının birbirlerinden ve madde güçlüklerinin birbirlerinden farklı olduğunu ve puanlayıcıların bütünsel dereceli puanlama anahtarı kullanırken farklı davranışlar sergilemiş olabileceklerini göstermektedir. Logit cetvelinde artan logit değeri birey yüzeyi için yetenek seviyesinin arttığı, madde yüzeyi için madde güçlüğünün arttığı anlamına gelmektedir. Örneğin, B3 numaralı birey grup içinde yetenek seviyesi en yüksek bireyken B19 ve B20 numaralı bireyler de grup içinde yetenek seviyesi en düşük bireylerdir. Benzer şekilde M13 numaralı maddenin madde güçlüğü yüksekken yani madde kolayken M5 numaralı maddenin madde güçlüğü düşüktür yani madde zordur. Logit cetvelinde puanlayıcı yüzeyi için artan logit değeri ise puanlayıcı cömertliği davranışına işaret ederken azalan logit değeri puanlayıcı katılığı davranışına işaret etmektedir. Puanlayıcı davranışlarının detaylı incelenmesi için her bir yüzeye ait model ölçüm sonuçları incelenmelidir. Birey yüzeyine ait model ölçüm sonuçları Tablo 8'de verilmiştir.

Tablo 8. Bütünsel Dereceli Puanlama Anahtarı Birey Ölçüm Raporu

Birey	*Gözlenen Ortalama*	*Düzeltilmiş Ortalama*	*Logit Değeri*	*Standart Hata*	*Uyum İçi*	*Uyum Dışı*
B1	27,50	27,24	-0,12	0,05	0,58	0,57
B2	38,00	39,90	0,51	0,05	1,08	1,09
B3	40,75	42,59	0,71	0,06	1,38	1,35
B4	27,92	27,78	-0,10	0,05	1,40	1,45
B5	21,00	19,28	-0,56	0,06	0,78	0,83
B6	23,33	21,99	-0,39	0,05	0,89	0,90
B7	36,62	38,44	0,42	0,05	1,32	1,22
B8	36,58	38,39	0,42	0,05	0,97	0,87
B9	31,33	32,17	0,10	0,05	0,78	0,82
B10	28,83	28,97	-0,04	0,05	1,22	1,23
B11	19,67	17,85	-0,67	0,06	0,74	0,69
B12	19,58	17,77	-0,68	0,06	0,91	0,88
B13	19,42	17,60	-0,69	0,06	0,78	0,83
B14	21,46	19,79	-0,53	0,06	1,10	1,25
B15	37,17	39,02	0,46	0,05	1,25	1,39
B16	22,21	20,65	-0,47	0,06	1,13	1,05
B17	20,58	18,82	-0,60	0,06	0,91	0,97
B18	18,25	16,45	-0,80	0,06	0,89	0,85
B19	17,50	15,75	-0,88	0,07	1,40	1,46
B20	17,42	15,67	-0,88	0,07	0,74	0,70
Ortalama	26,26	25,81	-0,24	0,06	1,01	1,02
S (Evren)	7,68	9,18	0,51	0,01	0,25	0,26
S (Örneklem)	7,88	9,42	0,52	0,01	0,25	0,27

Model, Evren RMSE=0,06; Düzeltilmiş S.H.=0,50; Ayırma Oranı=8,94; Ayırma İndeksi=12,26; Güvenirlik=0,99
Model, Örneklem RMSE=0,06; Düzeltilmiş S.H.=0,52; Ayırma Oranı=9,18; Ayırma İndeksi=12,57; Güvenirlik=0,99
Model, Ki-kare (Sabit etkili): 1585,6; sd=19; p=0,00
Model, Ki-kare (Normal): 18,8; sd=18; p=0,41

Bireylere ait logit aralığı bireylerin yetenek düzeylerindeki değişimi göstermektedir. Birey yüzeyine ait model ölçüm sonuçlarının yer aldığı Tablo 8 incelendiğinde bireylerin yetenek düzeyleri arasında 1,59 [0,71-(-0,88)] logit değişim olduğu görülmektedir. Bu durum birey performanslarının

geniş bir aralığa sahip olduğunu göstermektedir. Birey yüzeyine ait logit değerleri ortalaması -0,24 iken standart sapma 0,52'dir. Bu kapsamda B19 ve B21 numaralı bireyler en düşük yetenek seviyesine sahipken B3 numaralı birey en yüksek yetenek seviyesine sahiptir. Birey yüzeyine ait uyum istatistikleri incelendiğinde uyum içi ve uyum dışı istatistiklerinin ortalamaları bire yakındır. Bu durum birey yüzeyi için model veri uyumunun iyi olduğunu göstermektedir. Birey yüzeyine ait ayırma oranı (9,18), ayırma indeksi (12,57) ve güvenirlik (0,99) istatistikleri incelendiğinde bu istatistiklerin büyük olduğu görülmektedir. Bu durum birey yeteneklerinin istatistiksel açıdan farklı olduğunu göstermektedir. Ancak birey yetenekleri arasındaki farkın istatistiksel olarak anlamlı olup olmadığının kontrol için sabit etkili ki kare değeri incelenmelidir. Sabit etkili ki kare değeri incelendiğinde bu değerin istatistiksel olarak anlamlı olduğu görülmektedir ($\chi^2(sd)$=1585,6(19); p=0,00<0,01). Bu durum birey yeteneklerinin istatistiksel açıdan anlamlı olarak farklılaştığını göstermektedir.

Birey yüzeyine ait ölçüm sonuçları incelendikten sonra madde yüzeyine ait ölçüm sonuçları incelenmiştir. Madde yüzeyine ait ölçüm sonuçları Tablo 9'da yer almaktadır.

Tablo 9. Bütünsel Dereceli Puanlama Anahtarı Madde Ölçüm Raporu

Maddeler	*Gözlenen Ortalama*	*Düzeltilmiş Ortalama*	*Logit Değeri*	*Standart Hata*	*Uyum İçi*	*Uyum Dışı*
M1	33,50	35,08	0,48	0,04	0,59	0,66
M2	27,56	27,09	0,11	0,04	0,85	0,79
M3	24,88	23,49	-0,07	0,05	0,96	0,93
M4	24,97	23,61	-0,06	0,05	0,73	0,69
M5	17,47	15,48	-0,67	0,06	1,06	1,01
M6	25,28	24,02	-0,04	0,05	0,98	0,88
M7	28,81	28,82	0,19	0,04	0,60	0,57
M8	23,75	22,06	-0,15	0,05	1,14	1,21
M9	39,94	42,15	0,91	0,05	1,53	1,44
M10	22,31	20,33	-0,25	0,05	1,58	1,77
M11	22,31	20,33	-0,25	0,05	1,12	1,08
M12	17,88	15,83	-0,63	0,06	1,49	1,35
M13	40,50	42,66	0,96	0,05	0,82	0,91
M14	18,97	16,82	-0,52	0,05	1,47	1,22
M15	25,72	24,59	-0,01	0,05	0,81	0,79
Ortalama	26,26	25,49	0,00	0,05	1,05	1,02
S (Evren)	6,80	8,26	0,47	0,00	0,32	0,32
S (Örneklem)	7,04	8,55	0,49	0,00	0,33	0,33

Model, Evren RMSE=0,05; Düzeltilmiş S.H.=0,47; Ayırma Oranı=9,64; Ayırma İndeksi=13,19; Güvenirlik=0,99
Model, Örneklem RMSE=0,05; Düzeltilmiş S.H.=0,48; Ayırma Oranı=9,98; Ayırma İndeksi=13,64; Güvenirlik=0,99
Model, Ki-kare (Sabit etkili): 1297,9; sd=14; p=0,00
Model, Ki-kare (Normal): 13,8; sd=13; p=0,38
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi

Bütünsel dereceli puanlama anahtarı madde yüzeyine ait ölçüm sonuçlarının yer aldığı Tablo 9 incelendiğinde maddelerin logit değerlerinin -0,67 ile 0,96 arasında değiştiği görülmektedir. Bu durum madde güçlüklerinin arasında 1,63 [0,96-(-0,67)] logit değişim olduğunu göstermektedir. Testte yer alan maddelerden en küçük logit değerine sahip M5 en zor madde iken en büyük logit değerine

sahip M13 ise en kolay maddedir. Madde yüzeyine ait uyum istatistiklerinin ortalaması beklenen değer olan bire oldukça yakındır. Bu durum model veri uyumunun iyi olduğunun göstergesidir. Madde yüzeyine ait ayırma oranı (9,98), ayırma indeksi (13,64) ve güvenirlik (0,99) istatistikleri incelendiğinde bu istatistiklerin büyük olduğu görülmektedir. Bu durum madde güçlüklerinin istatistiksel açıdan farklı olduğunu göstermektedir. Ancak madde güçlükleri arasındaki farkın istatistiksel olarak anlamlı olup olmadığının kontrol için sabit etkili ki kare değeri incelenmelidir. Sabit etkili ki kare değeri incelendiğinde bu değerin istatistiksel olarak anlamlı olduğu görülmektedir ($\chi^2(sd)$=1297,9(14); p=0,00<0,01). Bu durum madde güçlüklerinin istatistiksel açıdan anlamlı olarak farklılaştığını göstermektedir.

Birey ve madde yüzeylerine ait ölçüm sonuçları incelendikten sonra puanlayıcı yüzeyine ait ölçüm sonuçları incelenmiştir. Puanlayıcı yüzeyine ait ölçüm sonuçları Tablo 10'da yer almaktadır.

Tablo 10. Bütünsel Dereceli Puanlama Anahtarı Puanlayıcı Ölçüm Raporu

Puanlayıcı	*Gözlenen Ortalama*	*Düzeltilmiş Ortalama*	*Logit Değeri*	*Standart Hata*	*Uyum İçi*	*Uyum Dışı*
P1	22,83	20,26	-0,26	0,05	1,10	1,01
P2	28,70	28,56	0,18	0,05	0,78	0,79
P3	26,17	24,77	0,00	0,05	0,75	0,64
P4	24,67	22,65	-0,11	0,05	0,84	0,74
P5	27,97	27,45	0,13	0,05	0,85	0,83
P6	26,53	25,30	0,02	0,05	0,77	0,74
P7	25,37	23,62	-0,06	0,05	0,74	0,79
P8	27,43	26,64	0,09	0,05	1,78	2,67
P9	28,00	27,50	0,13	0,05	1,05	0,95
P10	29,53	29,84	0,24	0,05	1,21	1,14
P11	24,10	21,88	-0,16	0,05	1,22	1,06
P12	23,33	20,88	-0,22	0,05	1,01	0,97
P13	23,47	21,05	-0,21	0,05	1,04	1,00
P14	24,97	23,06	-0,09	0,05	0,83	0,82
P15	25,40	23,67	-0,06	0,05	1,03	0,93
P16	31,63	33,00	0,38	0,05	1,28	1,23
Ortalama	26,26	25,01	0,00	0,05	1,02	1,02
S (Evren)	2,39	3,47	0,17	0,00	0,26	0,45
S (Örneklem)	2,47	3,59	0,18	0,00	0,27	0,47

Model, Evren RMSE=0,05; Düzeltilmiş S.H.= 0,17; Ayırma Oranı=3,36; Ayırma İndeksi=4,82; Güvenirlik=0,92
Model, Örneklem RMSE=0,05; Düzeltilmiş S.H.=0,17; Ayırma Oranı=3,48; Ayırma İndeksi=4,98; Güvenirlik=0,92
Model, Ki-kare (Sabit etkili): 198,2; sd=15; p=0,00
Model, Ki-kare (Normal): 13,9; sd=14; p=0,45
S: Standart sapma, RMSE: Hata kareleri ortalamasının karekökü, S.H.: Standart hata, sd: Serbestlik derecesi

Bütünsel dereceli puanlama anahtarı puanlayıcı yüzeyine ait ölçüm sonuçlarının yer aldığı Tablo 10 incelendiğinde puanlayıcıların logit değerlerinin -0,26 ile 0,38 arasında değiştiği görülmektedir. Bu durum puanlayıcıların katılık ve cömertlik durumları arasında 0,64 [0,38-(-0,26)] logit değişim olduğunu göstermektedir. Bu bilgiler ışığında logit değeri en küçük olan yani en katı puanlayıcının P1, logit değeri en büyük olan yani cömert puanlayıcının da P16 olduğu görülmektedir. Puanlayıcı yüzeyine ait uyum istatistiklerinin ortalaması bire yakındır. Bu durum puanlayıcı yüzeyine ait model veri uyumunun iyi olduğunun göstergesidir. Puanlayıcı yüzeyine ait ayırma oranı (3,48),

ayırma indeksi (4,98) ve güvenirlik (0,92) istatistiklerinin büyük olduğu görülmektedir. Bu durum puanlayıcıların puanlama esnasında birbirlerinden farklı puanlayıcı davranışı sergilemiş olabileceklerinin göstergesidir. Bu durumun istatistiksel açıdan değerlendirilmesi için sabit etkili ki-kare değeri incelenmelidir. Sabit etkili ki-kare değeri incelendiğinde bu değerin istatistiksel açıdan anlamlı olduğu ($\chi^2(sd)$=198,2(15); p=0,00<0,01) ve dolayısıyla puanlayıcıların puanlama esnasında farklı davranışlar sergilediği görülmektedir.

Araştırma kapsamında oluşturulan modeldeki yüzeylere ait ölçüm sonuçları incelendikten sonra puanlayıcı davranışları incelenmiştir. Araştırmada puanlayıcı davranışlarından merkeze eğilim, halo etkisi ve yanlılık davranışları incelenmiştir.

Merkeze Eğilim

Puanlayıcıların merkeze eğilim davranışları incelenirken önce grup düzeyindeki davranışlar sonrasında da bireysel düzeydeki davranışlar incelenmiştir. Bütünsel dereceli puanlama anahtarı ile puanlama yapan puanlayıcıların grup düzeyinde merkeze eğilim davranışları incelenirken kategori istatistikleri, birey ve madde yüzeylerine ait istatistikler incelenmiştir. Kategori istatistikleri Tablo 11'de verilmiştir.

Tablo 11. Bütünsel Dereceli Puanlama Anahtarına Ait Kategori İstatistikleri

Kategoriler	*Frekans (f)*	*Yüzde (%)*	*Yığılmalı yüzde (%)*	*Ortalama logit*	*Beklenen logit*	*Uyum dışı*
10	1815	38	38	-0,72	-0,71	1,10
20	777	16	54	-0,43	-0,45	0,80
30	634	13	67	-0,15	-0,15	0,90
40	538	11	78	0,13	0,16	1,00
50	1036	22	100	0,49	0,49	1,10

Bütünsel dereceli puanlama anahtarı kullanımına ait kategori istatistiklerinin olduğu Tablo 11 incelendiğinde 10 ve 50 numaralı kategorilerin çok kullanıldığı, diğer kategorilerin ise benzer sayıda kullanıldığı görülmektedir. Tablo 11'den hareketle puanlayıcıların merkeze eğilim davranışları olmadığı söylenebilir. Ancak merkeze eğilim davranışın belirlenmesinde kategori istatistikleri tek başına yeterli değildir. Merkeze eğilim davranışı belirlenirken birey yüzeyine ait istatistikler ve madde yüzeyine ait istatistikler de incelenmelidir. Birey yüzeyine ait istatistikler incelendiğinde (Tablo 8), ayırma oranı (9,18), ayırma indeksi (12,57) ve güvenirlik (0,99) istatistiklerinin büyük olduğu ve ki- kare istatistiğinin de ($\chi^2(sd)$=1585,6(19); p=0,00<0,01) anlamlı olduğu görülmüştür. Bu durum bireylerin yeteneklerinin doğru ve istatistiksel olarak anlamlı bir şekilde ayrıldığını göstermektedir. Madde yüzeyine ait uyum istatistikleri de incelendiğinde (Tablo 9), maddelere ait uyum istatistiklerinin kabul edilebilir aralıkta olduğu görülmektedir. Tüm göstergeler bir arada ele alındığında bütünsel dereceli puanlama anahtarı kullanan puanlayıcı grubunda, grup düzeyinde merkeze eğilim davranışının olmadığı söylenebilir. Grup düzeyinde merkeze eğilim davranışı

incelendikten sonra bireysel düzeyde merkeze eğilim davranışı incelenmiştir. Bireysel düzeyde merkezi eğilim davranışı belirlenirken puanlayıcılara ait kategori istatistikleri ve uyum istatistikleri incelenmektedir. Puanlayıcılara ait uyum istatistikleri incelendiğinde P8 numaralı puanlayıcının uyum dışı istatistiğinin beklenen aralık dışında olduğu görülmektedir. Bu durum P8 numaralı puanlayıcının merkeze eğilim davranışı gösterdiğinin göstergesidir. Merkeze eğilim davranışının belirlenmesindeki bir diğer yöntem de puanlayıcıya ait kategori istatistiklerinden uyum dışı istatistiğinin incelenmesidir. Araştırma kapsamında puanlayıcıların madde kategori istatistikleri, uyum dışı istatistiği ve eşik değerleri Tablo 12'de verilmiştir.

Tablo 12. Her Bir Puanlayıcıya Ait Kategori İstatistikleri

Puanlayıcı	*Kategori İstatistiği (%)*					*Uyum Dışı*					*Eşik Değer*				
	10	20	30	40	50	10	20	30	40	50	10	20	30	40	50
P1	58	2	13	10	18	1,60	0,70	0,90	1,00	0,60		2,51	-2,33	0,16	-0,34
P2	24	18	22	17	18	1,30	0,90	1,10	1,10	0,70		-0,52	-0,55	0,41	0,66
P3	38	16	13	11	21	1,80	0,70	0,60	0,70	0,80		-0,36	-0,30	0,37	0,29
P4	38	21	14	10	17	0,90	1,10	1,20	1,00	0,90		-0,39	-0,01	0,42	-0,02
P5	26	25	14	13	22	1,30	0,70	1,10	1,30	0,80		-0,67	0,28	0,26	0,14
P6	25	30	16	12	17	1,10	1,00	0,70	0,60	1,10		-1,05	0,26	0,42	0,37
P7	27	34	11	15	13	1,40	0,80	0,80	1,00	0,80		-1,23	0,71	-0,18	0,70
P8	36	10	53	45	63	0,90	2,00	1,20	1,20	1,00		0,98	-0,80	0,06	-0,25
P9	32	22	9	9	28	1,30	0,90	0,60	0,70	1,00		-0,21	0,66	0,11	-0,57
P10	34	10	15	8	33	1,30	0,60	0,50	1,00	1,00		0,76	-0,55	0,79	-1,01
P11	52	4	12	5	27	1,00	0,40	0,60	0,40	1,30		1,82	-1,34	0,79	-1,27
P12	52	12	8	8	21	0,90	0,60	0,70	0,70	2,60		0,53	0,06	0,01	-0,60
P13	50	16	6	6	22	0,90	0,60	0,70	1,10	2,00		0,26	0,62	0,09	-0,97
P14	33	25	15	13	14	1,10	0,70	0,80	0,90	1,10		-0,64	0,04	0,21	0,39
P15	44	12	12	12	20	1,30	1,10	0,50	0,90	0,90		0,52	-0,44	-0,02	-0,06
P16	33	2	15	16	34	1,40	0,20	0,40	0,70	1,10		2,60	-2,33	0,04	-0,31

Tablo 12'de her bir puanlayıcının bütünsel dereceli puanlama anahtarını kullanım örüntüsü incelendiğinde P11, P12 ve P16 numaralı puanlayıcılara ait uyum dışı değerlerinin istenen aralık dışında olduğu görülmektedir. Bu durum P11, P12 ve P16 numaralı puanlayıcılarda olmak üzere toplam üç puanlayıcıda merkeze eğilim davranışı olduğunun göstergesidir. Tüm istatistikler incelendiğinde P8, P11, P12 ve P16 numaralı puanlayıcılarda merkeze eğilim davranışı belirlenmiştir. Puanlayıcılar tecrübesiz oldukları için puanlayıcıların tamamında merkeze eğilim davranışı meydana gelmiş olabilir (Baird vd., 2013).

Halo Etkisi

Puanlayıcıların halo etkisi davranışları incelenirken önce grup düzeyindeki davranışlar sonrasında da bireysel düzeydeki davranışlar incelenmiştir. Bütünsel dereceli puanlama anahtarı ile puanlama yapan puanlayıcıların grup düzeyinde halo etkisi davranışları incelenirken madde yüzeyine ait istatistikler incelenmiştir. Madde yüzeyine ait istatistikler incelendiğinde (Tablo 9) ayırma oranının (9,64), ayırma indeksinin (13,19) ve güvenirlik istatistiğinin (0,99) yüksek olduğu ve ki kare istatistiğinin de istatistiksel olarak anlamlı ($\chi^2(sd)$=1297,9(14); p=0,00<0,01) olduğu

görülmektedir. Bu durum puanlayıcıların grup düzeyinde halo etkisi davranışına sahip olmadıklarının göstergesidir. Grup düzeyindeki istatistikler incelendikten sonra bireysel istatistikler incelenmiştir. Halo etkisi bireysel istatistiklerde incelenirken öncelikle madde güçlükleri arasındaki farka bakılmaktadır. Madde güçlükleri arasındaki fark büyükse puanlayıcıların uyum istatistiklerinin 1'den çok büyük olması, madde güçlükleri arasındaki fark küçükse puanlayıcıların uyum istatistiklerinin 1'den çok küçük olması halo etkisi davranışına işaret etmektedir (Myford ve Wolfe, 2004). Madde ölçüm raporunun yer aldığı Tablo 9 incelendiğinde madde güçlükleri arasındaki farkın büyük (1,63) olduğu görülmektedir. Bu kapsamda Tablo 10'da yer alan puanlayıcılara ait uyum istatistikler incelendiğinde P8 (uyum içi: 1,78 uyum dışı: 2,67) numaralı puanlayıcıda halo etkisi davranışının olduğu söylenebilir. Halo etkisi davranışının belirlenmesinde kullanılan bir diğer yöntem de madde güçlüklerinin eşitlenerek çok yüzeyli Rasch analizini tekrar etmektir (Linacre, 2023, s. 360). Linacre (2023), madde güçlüklerinin eşitlenerek oluşturulduğu modelde uyum istatistikleri 1'e eşit olan yani modele mükemmel uyum gösteren puanlayıcıların halo etkisi davranışına sahip olduklarını ifade etmektedir. Bu kapsamda madde güçlükleri eşitlenerek yeni birçok yüzeyli Rasch analizi yapılmıştır. Madde güçlükleri eşitlendikten sonra bütünsel dereceli puanlama anahtarı kullanan puanlayıcılara ait ölçüm raporu Tablo 13'te verilmiştir.

Tablo 13. Madde Güçlükleri Eşitlendikten Sonra Puanlayıcı Yüzeyine Ait Ölçüm Raporu

Puanlayıcı	*Uyum İçi*	*Uyum Dışı*
P1	1,15	1,17
P2	0,76	0,77
P3	0,95	0,95
P4	0,90	0,88
P5	0,88	0,89
P6	0,76	0,77
P7	0,71	0,69
P8	1,57	1,63
P9	0,97	0,97
P10	1,04	1,04
P11	1,28	1,27
P12	1,06	1,06
P13	1,07	1,08
P14	0,83	0,81
P15	1,05	1,05
P16	1,12	1,15

Tablo 13'te yer alan her bir puanlayıcıya ait uyum istatistikleri incelendiğinde modele mükemmel uyum sağlayan puanlayıcının olmadığı dolayısıyla halo etkisi davranışı sergileyen puanlayıcının olmadığı söylenebilir.

Tüm bireysel istatistikler göz önüne alındığında bütünsel dereceli puanlama anahtarı kullanan puanlayıcılardan bir tanesinde (P8) halo etkisi davranışı belirlenmiştir.

Yanlılık

Bütünsel dereceli puanlama anahtarı kullanan puanlayıcıların yanlılık davranışının tespiti için *puanlayıcı x birey (pxb)* etkileşimi incelenmiştir. Grup düzeyindeki yanlılık davranışına ait ki kare istatistiği incelendiğinde ki kare istatistiğinin anlamlı olmadığı gözlenmiştir ($\chi^2(sd)$=271,6(320); p>0,01). Bu durum, bütünsel dereceli puanlama anahtarı kullanan puanlayıcıların grup düzeyinde yanlılık davranışı olmadığının göstergesidir. Grup düzeyinde istatistikler incelendikten sonra bireysel istatistikler incelenmiştir.

Çok yüzeyli Rasch analizinde bireysel düzeyde yanlılık davranışı incelenirken t istatistiği kullanılmaktadır. Linacre (2023), ±2 aralığı dışındaki t değerlerinin istatistiksel olarak anlamlı etkileşimler olduğunu ifade etmiştir (s. 190). Araştırma kapsamında 20 birey 16 puanlayıcı olmak üzere 320 etkileşim yer almaktadır. Her bir etkileşimi tabloda vermek mümkün olmadığı için sadece anlamlı etkileşimler raporlanmıştır.

Tablo 14. Bütünsel Dereceli Puanlama Anahtarı Kullanımına Ait Anlamlı Etkileşimler

Puanlayıcı	*Birey*	*Gözlenen Puan*	*Beklenen Puan*	*Bias (logit)*	*Standart Hata*	*t*
P8	P2	430	446,21	-0,65	0,20	-3,31
P8	P3	480	495,04	-0,66	0,19	-3,39
P8	P5	460	447,01	0,54	0,19	2,77
P8	P7	440	453,24	-0,52	0,20	-2,67
P8	P8	430	444,18	-0,56	0,20	-2,84
P8	P9	380	391,36	-0,44	0,20	-2,16
P8	P12	420	408,72	0,50	0,20	2,54
P8	P13	440	426,45	0,59	0,20	3,04
P8	P15	430	445,02	-0,60	0,20	-3,03
P8	P17	530	509,34	0,84	0,20	4,22
P8	P18	390	379,56	0,50	0,20	2,49
P8	P19	430	414,34	0,74	0,20	3,76

Bütünsel dereceli puanlama anahtarı kullanan puanlayıcıların yanlılık davranışlarına ait istatistiklerin yer aldığı Tablo 14 incelendiğinde bir puanlayıcının toplam 12 yanlılık davranışı sergilediği görülmektedir. Puanlayıcılardan P8 numaralı puanlayıcının B5, B12, B13, B17, B18 ve B19 numaralı bireylere olması gerekenden fazla ve B2, B3, B7, B8, B9 ve B15 numaralı bireylere de olması gerekenden az puan verdiği görülmektedir.

Tartışma, Sonuç ve Öneriler

Bu çalışma, rutin olmayan problemlere yönelik hazırlanan açık uçlu matematik maddelerini içeren başarı testinin puanlanmasında analitik ve bütünsel dereceli puanlama anahtarı kullanımının puanlayıcı davranışlarına etkisini incelemek amacıyla yapılmıştır. Araştırma kapsamında merkeze eğilim, halo etkisi ve yanlılık davranışları incelenmiştir. Puanlayıcı davranışları belirlenirken çok yüzeyli Rasch modeli kullanılmıştır. Sonuçlar belirtilirken analitik dereceli puanlama anahtarı ve bütünsel dereceli puanlama anahtarı kullanımına göre farklı başlıklar altında sunulmuştur.

Analitik Dereceli Puanlama Anahtarı

Analitik dereceli puanlama anahtarı kullanan puanlayıcılardan elde edilen verilerin analizi sonucunda araştırma kapsamında yer alan birey, madde ve puanlayıcı yüzeyleri incelendiğinde tüm yüzeylerin model veri uyumunun iyi olduğu ve dolayısıyla birey performanslarının, madde güçlüklerinin ve puanlayıcı davranışlarının istatistiksel olarak anlamlı bir şekilde farklılaştığı belirlenmiştir.

Puanlayıcıların merkeze eğilim davranışları incelendiğinde analitik dereceli puanlama anahtarı kullanan puanlayıcı grubunda grup düzeyinde merkeze eğilim davranışı gözlenmezken bireysel düzeyde tüm puanlayıcılarda merkeze eğilim davranışının yer aldığı belirlenmiştir. Analitik dereceli puanlama anahtarında düzey sayısı çok fazla olduğu için puanlayıcılarda merkeze eğilim davranışı gözlenmiş olabilir. Ayrıca araştırma kapsamında puanlama yapan puanlayıcıların deneyimsiz olmaları puanlayıcıların tamamında merkeze eğilim davranışı oluşmasında etkili olmuş olabilir (Baird vd., 2013).

Puanlayıcıların halo etkisi davranışları incelendiğinde analitik dereceli puanlama anahtarı kullanan puanlayıcı grubunda grup düzeyinde halo etkisi davranışı olmadığı belirlenmiştir. Puanlayıcılara ait bireysel istatistikler incelendiğinde üç puanlayıcıda halo etkisi davranışı belirlenmiştir.

Puanlayıcıların yanlılık davranışları incelendiğinde merkeze eğilim ve halo etkisi davranışlarında olduğu gibi yanlılık davranışı da grup düzeyinde gözlenmemiştir. Bireysel düzeyde yanlılık davranışları incelendiğinde dört puanlayıcıda yanlılık davranışı belirlenmiştir. Yanlılık davranışının belirlendiği dört puanlayıcının *puanlayıcı x birey* etkileşimlerinden 14 tanesi istatistiksel olarak anlamlı çıkmıştır. Anlamlı etkileşimlerden sekiz tanesinde puanlayıcılar cömert davranırken altı tanesinde de katı davranmışlardır.

Bütünsel Dereceli Puanlama Anahtarı

Bütünsel dereceli puanlama anahtarı kullanan puanlayıcılardan elde edilen verilerin analizi sonucunda araştırma kapsamında yer alan birey, madde ve puanlayıcı yüzeyleri incelendiğinde tüm yüzeylerin model veri uyumunun iyi olduğu ve dolayısıyla birey performanslarının, madde güçlüklerinin ve puanlayıcı davranışlarının istatistiksel olarak anlamlı bir şekilde farklı olduğu belirlenmiştir.

Puanlayıcıların merkeze eğilim davranışları incelendiğinde bütünsel dereceli puanlama anahtarı kullanan puanlayıcı grubunda grup düzeyinde merkeze eğilim davranışı gözlenmezken bireysel düzeyde dört puanlayıcıda merkeze eğilim davranışının yer aldığı belirlenmiştir.

Puanlayıcıların halo etkisi davranışları incelendiğinde bütünsel dereceli puanlama anahtarı kullanan puanlayıcı grubunda grup düzeyinde halo etkisi davranışı olmadığı belirlenmiştir. Puanlayıcılara ait bireysel istatistikler incelendiğinde bir puanlayıcıda halo etkisi davranışı belirlenmiştir.

Puanlayıcıların yanlılık davranışları incelendiğinde, merkeze eğilim ve halo etkisi davranışlarında olduğu gibi yanlılık davranışı da grup düzeyinde gözlenmemiştir. Bireysel düzeyde yanlılık davranışları incelendiğinde, bir puanlayıcıda yanlılık davranışı belirlenmiştir. Yanlılık davranışının belirlendiği bir puanlayıcının *puanlayıcı x birey* etkileşimlerinden 12 tanesi istatistiksel olarak anlamlı çıkmıştır. Anlamlı etkileşimlerden altı tanesinde puanlayıcı cömert davranırken altı tanesinde de katı davranmışlardır.

Bütünsel dereceli puanlama anahtarında merkeze eğilim, halo etkisi ve yanlılık davranışını sergileyen puanlayıcı sayısının az olması kategori sayısının az olması ile açıklanabilir. Kategori sayısı az olduğu için puanlayıcılar ölçülen özellikleri arasında ayrımı daha kolay yapmış olabilirler. Ayrıca kategori sayısı az olduğu için puanlayıcılar yorgunluk ve bıkkınlık gibi puanlayıcı davranışlarından da daha az etkilenmiş olabilirler.

Özetle, aynı amaç için hazırlanmış analitik ve bütünsel dereceli puanlama anahtarları ile yapılan puanlamalarda bütünsel dereceli puanlama anahtarı kullanımında puanlayıcı davranışlarının daha az olduğu söylenebilir. Analitik dereceli puanlama anahtarında ölçülecek özellik parçalara ayrıldığı için daha objektif sonuçlar sağlasa da bütünsel dereceli puanlama anahtarında ölçülecek özellik bir bütün olarak ele alındığı için puanlayıcıları benzer şekilde puanlamaya yöneltiyor olabilir. Araştırma sonuçları alanyazın ile kıyaslandığında Kutlu vd. (2017), Jonsson ve Svingby (2007) ile Bıkmaz-Bilgen ve Doğan'ın (2017) yapmış olduğu çalışma sonuçları ile çelişmektedir. Hem analitik dereceli puanlama anahtarı hem de bütünsel dereceli puanlama anahtarı kullanımında puanlayıcılarda grup düzeyinde olmasa da bireysel düzeyde merkeze eğilim, halo etkisi ve yanlılık davranışları ortaya çıkmıştır. Bu durum Şata vd. (2020), Esfandiari (2015, 2021), Jones ve Bergin (2019), Linlin (2020), Tursynbayeva vd. (2024), Yılmaz'ın (2017) yapmış oldukları çalışmaları desteklemektedir.

Araştırma sonuçlarından hareketle rutin olmayan problemlere yönelik hazırlanan açık uçlu matematik maddelerinin kullanıldığı çalışmalarda puanlayıcı davranışlarını en aza indirmek için bütünsel dereceli puanlama anahtarı kullanımı önerilmektedir. Araştırmada puanlayıcılara ait demografik bilgilere yer verilmemiştir. Yapılacak çalışmalarda puanlayıcıların yaş, cinsiyet, kıdem yılı gibi değişkenler göz önüne alınarak benzer çalışmalar yapılabilir. Ayrıca farklı öğrenim düzeyindeki öğrencilerle de benzer çalışmalar yapılıp mevcut araştırma sonuçları ile kıyaslanabilir.

Kaynaklar

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

Altun, M. (2020). *Matematik okuryazarlığı el kitabı.* Aktüel Alfa Akademi.

Anderson, L. W. & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: Complete edition*. Addison Wesley Longman.

Aslanoğlu, A. E. (2022). Açık uçlu maddelerin hazırlanması ve incelenmesi. İ. Karakaya (Ed.), *Açık uçlu soruların hazırlanması, uygulanması ve değerlendirilmesi* (1. b.) içinde (s. 2-27). Pegem Akademi.

Atılgan, H. (2009). Test geliştirme. H. Atılgan, A. Kan, & N. Doğan (Ed.), *Eğitimde ölçme ve değerlendirme* (4. b.) içinde (s. 315-348). Anı.

Baird, J., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). Marker effects and examination reliability: A comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling (Research Report No. 5261). http://dera.ioe.ac.uk/17683/1/2013-01-21-marker-effects-and-examination-reliability.pdf sayfasından erişilmiştir.

Baker, F. B. & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques.* Marcel Dekker.

Bıkmaz-Bilgen, Ö. & Doğan, N. (2017). Puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması. *Journal of Measurement and Evaluation in Education and Psychology, 8*(1), 63-78. https://doi.org/10.21031/epod.294847

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* Cengage Learning.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438-481. https://doi.org/10.3102/00346543058004438

DeMars, C. (2010). *Item response theory*. Oxford University.

Dunbar, N. E., Brooks, C. F., & Kubicka-Miller, T. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education, 31*(2), 115-128. https://doi.org/10.1007/s10755-006-9012-x

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

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists.* Multivariate Applications Books Series. Lawrence Erlbaum.

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

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. https://doi.org/10.22049/jalda.2021.27032.1234

Farrokhi, F., Esfandiari, R., & Dalili, M. V. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal, 15*(11), 76-83.

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

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking.* Allyn and Bacon.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5

Haladyna, T. M. & Rodriguez, M. C. (2013). *Developing and validating test items*. Taylor & Francis.

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

Harvey, R. J. & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist, 27*(3), 353-383.

Jones, E. & Bergin, C. (2019). Evaluating teacher effectiveness using classroom observations: A Rasch analysis of the rater effects of principals. *Educational Assessment, 24*(2), 91-118.

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

Karakaya, İ. (2012). Bilimsel araştırma yöntemleri. A. Tanrıöğen (Ed), *Bilimsel araştırma yöntemleri* içinde (s. 57-83). Anı.

Karakaya, İ. & Şata, M. (2022). Açık uçlu maddelerin hazırlanması ve incelenmesi. İ. Karakaya (Ed.), *Açık uçlu soruların hazırlanması, uygulanması ve değerlendirilmesi* (1. b.) içinde (s. 28-39). Pegem Akademi.

Kilpatrick, J. & Lerman, S. (2020). Education of professional development providers (for educators of practicing teachers). S. Lerman (Ed.), *Encyclopedia of mathematics education* içinde (s. 262-262). Springer International.

Kutlu, Ö., Doğan, D. C., & Karakaya, İ. (2017). *Ölçme ve değerlendirme performansa ve portfolyoya dayalı durum belirleme*. Pegem Akademi.

Linacre, J. M. (1989). *Many-faceted Rasch measurement.* (Doktora Tezi). ProQuest Dissertations & Theses Global database. (T-30889)

Linacre, J. M. (2014). *A user's guide to FACETS: Rasch-model computer programs.* Winsteps.

Linacre, J. M. (2023). *Facets computer program for many-facet Rasch measurement* [version 3.85.1].

Linlin, C. (2020). Comparison of automatic and expert teachers' rating of computerized English listening-speaking test. *English Language Teaching, 13*(1), 18-30.

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

Moskal, B. M. & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, and Evaluation, 7*(10), 71-81. https://doi.org/10.7275/q7rm-gg74

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.

Nitko, A. J. & Brookhart, S. M. (2014). *Educational assessment of students* (6. b.). Pearson Education.

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.

Popham, W. J. (2001). *Clasroom assesment: What teachers need to know*. Allyn and Bacon.

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Principles and Standards for School Mathematics (NCTM), 94*(1), 22.

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. https://doi.org/10.1037/0033-2909.88.2.413

Şata, M. (2019). *Performans değerlendirme sürecinde puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisinin incelenmesi* (Doktora Tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

Şata, M., Karakaya, İ., & Erman-Aslanoğlu, A. (2020). Evaluation of university students' rating behaviors in self and peer rating process via many facet Rasch model. *Eurasian Journal of Educational Research, 20*(89), 25-46.

Tuckman, B. W. (1991). Evaluating the alternative to multiple-choice testing for teachers. *Contemporary Education, 62*(4), 299.

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. https://doi.org/10.21449/ijate.1438798

Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103-118. https://doi.org/10.1207/s15324818ame0602_1

Yılmaz, F. N. (2017). Analysis of the rater effects in the rating of diagnostic trees prepared by teacher candidates by the many-facet Rasch model. *JEP, 8*(18).

Extended Summary

Measurement and evaluation play a critical role in the education process to understand students' learning processes, evaluate their development, and determine the effectiveness of curricula. Measuring student success has been considered for many years as measuring the knowledge students acquire during education, but today measuring high-level mental skills has become more important due to reasons such as changing educational goals, economy and the nature of the workforce (Karakaya, 2012). The fact that students create their own answers (Haladyna and Rodriguez, 2013) and receive high-level cognitive responses from students (Popham, 2001) makes the place of open-ended items in the measurement and evaluation process important. In teaching mathematics, it is important that students not only learn basic concepts, but also develop deep understanding and acquire creative problem-solving skills. In order to achieve these and similar goals in mathematics teaching, students' analytical and critical thinking skills are encouraged by using open-ended, non-routine mathematics problems (Kilpatrick and Lerman, 2020). However, evaluating and scoring such problems requires a more complex process. For this reason, examining alternative methods that will enable objective evaluation of open-ended items is important for valid and reliable estimation of student success. This study aims to shed light on evaluation processes by determining the effect of analytical and holistic rubrics used in scoring open-ended questions on rater behavior. The results of the research are expected to make a significant contribution to the field of mathematics education and evaluation processes.

This research was designed with a survey model, one of the best types of research. The study group of this research consisted of 20 students and 16 mathematics teachers studying in the eighth grade of a public school. As data collection tools in the study, a mathematics achievement test consisting of non-routine open-ended items prepared by the researchers, an analytical and holistic rubric developed by the researchers, and a rater evaluation form were used. The data collected within the scope of Onkun-Özgür's (2024) master's thesis study was used in the research.

The results obtained in the research were presented under two different headings including findings obtained from the analytical rubric and findings obtained from the holistic rubric.

For the analytical rubric, when the logit scale was examined, it showed that the individual performances were similar, the items were of similar difficulty, and the raters may have exhibited similar behaviors when using the analytical rubric. When the analytical rubric measurement results were examined, it was determined that the model data fit of all surfaces was good. While the central tendency behaviors of the raters who scored with the analytical rubric were examined at the group level, category statistics, statistics of individual and item surfaces were examined. When all indicators were taken together, it can be said that there was no central tendency behavior at the group level in the rater group using the analytical rubric. When the conformity statistics of the raters were examined, it was seen that the non-conformity statistic of rater R8 was outside the expected range. This was an indication that rater R8 showed a tendency towards the center. While examining the halo effect behavior of the raters at the group level, the statistics of the item surface were examined. When the fit statistics of each rater were examined, it can be said that there was no rater that fits the model perfectly, and therefore there was no rater that exhibited halo effect behavior. Considering all individual statistics, halo effect behavior was determined in three of the raters who used the analytical rubric. Rater x individual (rxp) interaction was examined to detect bias behavior of raters using analytical rubrics. When the chi-square statistic of bias behavior at the group level was examined, it was seen that the chi-square statistic was not significant, that is, there was no bias behavior at the group level ($\chi 2(sd)$=286,8(320); p>0,01). When the statistics of the bias behaviors of the raters using the analytical rubric were examined, it was seen that four raters exhibited a total of 14 bias behaviors.

For the holistical rubric when the logit scale for the holistic rubric was examined, it was seen that individual performances were similar, the items were of similar difficulty, and the raters may have exhibited similar behaviors when using the holistic rubric. When the holistical rubric measurement results were examined, it was determined that the model data fit of all surfaces was good. While the central tendency behaviors of the raters who scored with the holistical rubric were examined at the group level, category statistics, statistics of individual and item surfaces were examined. When all indicators were taken together, it can be said that there was no central tendency behavior at the group level in the rater group using the holistical rubric. While examining the halo effect behavior of the raters who scored with the holistic rubric at the group level, the statistics of the item surface were examined. Considering all individual statistics, halo effect behavior was detected in one of the raters (R8) who used a holistic rubric. Rater x individual (rxp) interaction was examined to detect bias behavior of raters using holistic rubrics. When the chi-square statistic of bias behavior at the group level was examined, it was observed that the chi-square statistic was not significant

($\chi 2(sd)$=271.6(320); p>0.01). This indicated that raters using holistic rubrics did not exhibit bias at the group level. When the statistics of bias behaviors of raters using holistic rubrics were examined, it was figured out that one rater exhibited a total of 12 bias behaviors.

It can be said that rater behavior is less when using holistic rubrics when scoring is done with analytical and holistic rubrics prepared for the same purpose. Although the feature to be measured in the analytical rubric provides more objective results because it is divided into parts, it may lead the raters to score in a similar way since the feature to be measured in the holistic rubric is considered as a whole. Based on the research results, the use of holistic rubrics is recommended to minimize rater behavior in studies where open-ended mathematical items prepared for non-routine problems are used.

Araştırmacıların Katkı Oranı Beyanı

Bu araştırmanın planlanması, yürütülmesi ve yazılı hâle getirilmesinde araştırmacılar eşit oranda katkı sağlamıştır.

Destek ve Teşekkür Beyanı

Bu araştırmada herhangi bir kurum, kuruluş ya da kişiden destek alınmamıştır.

Çatışma Beyanı

Araştırmacıların araştırma ile ilgili diğer kişi ve kurumlarla herhangi bir kişisel ve finansal çıkar çatışması yoktur.

Etik Kurul Beyanı

Bu araştırma, Gazi Üniversitesi Etik Komisyonunun 14.05.2024 tarih ve E-77082166-604.01-957439 sayılı onayı ile yürütülmüştür.