Uludağ Üniversitesi Eğitim Fakültesi Dergisi / Journal of Uludag University Faculty of Education
e-ISSN: 2667-6788 https://dergipark.org.tr/tr/pub/uefad
Yıl **2024,** Cilt **37**, Sayı 3, **1037-1068** | DOI: https://doi.org/10.19171/uefad.1486010 **Araştırma Makalesi | Research Article**

Dereceli Puanlama Anahtarı Türünün Rutin Olmayan Matematik Problemlerinin Puanlanmasında Puanlayıcı Davranışları Üzerine Etkisi[1]

The Effect of Rubric Type On Raters' Behaviors in Scoring Non-Routine Mathematics Problems

Esra ONKUN ÖZGÜR

Öğretmen ◆ Milli Eğitim Bakanlığı ◆ esraonkunozgur@gmail.com ◆ ORCID: 0000-0002-5179-8847

İsmail KARAKAYA

Prof. Dr. ◆ Gazi Üniversitesi, Eğitim Bilimleri, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı◆ ikarakaya@gazi.edu.tr ◆ ORCID: 0000-0003-4308-6919

Özet

Bu araştırmada, rutin olmayan matematik problemlerinden oluşan matematik başarı testinin analitik ve bütünsel dereceli puanlama anahtarları (DPA) ile puanlanmasının puanlayıcı davranışları üzerine etkileri Çok Yüzeyli Rasch Ölçme Modeli ile incelenmiştir. Çalışma grubu, rutin olmayan matematik problemlerinden oluşan başarı testinin uygulandığı öğrenci grubu ve cevaplanan başarı testini değerlendiren puanlayıcılar olmak üzere iki farklı kısımdan oluşmaktadır. Bu çalışmada, betimsel araştırma yöntemlerinden tarama modeli kullanılmıştır. Bu çalışmada, araştırmacı tarafından hazırlanmış, 15 farklı rutin olmayan matematik probleminden oluşan başarı testi, iki farklı oturum şeklinde, iki günde öğrencilere uygulanmıştır. Çalışmanın bulguları incelendiğinde, yapılan tüm puanlamalarda, puanlayıcı, birey ve madde yüzeylerinde model veri uyumunun sağlandığı görülmüştür. Ayrıca, bireylerin yetenek düzeylerine göre ayrıştığı ve maddelerin güçlük düzeylerinin farklı olduğu görülmüştür. Analitik DPA ile yapılan puanlamanın bütünsel DPA ile yapılan puanlamaya göre, puanlayıcı yüzeyi bakımından göreceli olarak daha güvenilir sonuçlar verdiği; puanlayıcı yüzeyleri karşılaştırıldığında ise, analitik DPA kullanılan puanlamalarda puanlayıcı katılık/cömertlik düzeylerinin, bütünsel DPA kullanılan puanlamalardan daha fazla olduğu belirlenmiştir. Ayrıca, analitik DPA kullanılan puanlamalar arasındaki uyumun, bütünsel DPA kullanılan puanlamalardan daha düşük olduğu sonucuna varılmıştır. Puanlayıcı davranışları incelendiğinde, bütünsel dereceli puanlama anahtarı kullanan puanlayıcılardan, analitik dereceli puanlama anahtarı kullanan puanlayıcılara göre puanlayıcı katılık ve cömertlik davranışları gösteren puanlayıcı sayısının daha fazla olduğu; yanlılık davranışlarının ise daha az olduğu belirlenmiştir.

***Anahtar Kelimeler*:** Analitik dereceli puanlama anahtarı, Bütünsel dereceli puanlama anahtarı, Çok yüzeyli Rasch modeli, Puanlayıcı davranışları

Abstract

This study examined the effects of scoring the mathematics achievement test, which consists of non-routine mathematics problems with analytical and holistic rubrics on rater behavior, with the many-facet Rasch measurement model. The survey model, one of the descriptive research methods, was used in the study. An achievement test comprising 15 different non-routine mathematics problems prepared by the researcher was administered to the students in two different sessions in two days. According to research, model data fit was achieved on the rater, individual and item surfaces in all scoring. Additionally, it was observed that individuals were differentiated according to their ability levels, and the difficulty levels of the items were also different. Scoring with an analytical rubric gives relatively more reliable results in terms of rater surface than scoring with a holistic rubric; when the rater surfaces were compared, it was determined

[1] Bu çalışma 'Dereceli Puanlama Anahtarı Türünün Rutin Olmayan Matematik Problemlerinin Puanlanmasında Puanlayıcı Davranışları Üzerine Etkisinin İncelenmesi' adlı Yüksek Lisans Tezi'nden hareketle oluşturulmuştur.

Kabul Tarihi / Received: 17.05.2024 | Onay Tarihi / Accepted: 21.09.2024

that the rater severity/leniency levels in scoring using an analytical rubric were higher than in scoring using the holistic rubric. When rater behaviors were examined, rater strictness and generosity behaviors were higher in raters using holistic rubrics than raters using analytical rubrics; It was determined that bias behaviors were less.

***Keywords*:** Analytical rubric, Holistic rubric, Many-facet Rasch model, Non-routine math problem, Rater behavior

1. Giriş

Eğitim, bir insanın tüm hayatı boyunca yeni davranışlar kazanmaları ve davranışlarını değiştirmelerini sağlayan süreç (Baykul, 2020); öğretim ise, istenilen davranış değişikliklerinin örgün eğitim kurumlarında, planlı ve programa dayalı olarak, sistematik bir şekilde yapılması sürecini ifade eder (Demirel, 2004) Bireylerin sosyal ve akademik düzeylerinin gelişmesi eğitim kurumlarında yapılan öğretimlerle gerçekleşmektedir; bu nedenle yapılan öğretimin planlanması ve değerlendirilmesi önemlidir (Akgün, 2016).

Bireylerin ön bilgilerinin belirlenmesi, eğitim sürecinde kullanılan yöntem ve tekniklerin belirlenmesi, belirlenen bu yöntem ve tekniklerin etkililiğinin ölçülmesi, eğitim programının uygulanması sırasında öğrenmedeki eksikliklerin ve nedenlerinin bulunması ve giderilmesi, eğitim programının hedeflerine ulaşıp ulaşamadığının ya da ne kadarına ulaşıldığının kontrol edilmesi sürecinde ölçme ve değerlendirmeden faydalanılmaktadır. Nesnelerin herhangi bir özelliğe sahip olma durumunun ya da sahip olma derecesinin gözlemlenmesi ve sonuçlarının sayı, sembol ya da belli sıfatlarla ifade edilmesi ölçme (Stevens, 1951) ölçme sonuçlarının belirlenmiş bir ölçütle karşılaştırarak, yorumlama ve ölçülen nitelikle ilgili bir karara varma süreci ise değerlendirmedir (Turgut ve Baykul, 2010). Ölçülen nitelikle ilgili doğru değerlendirmenin yapılabilmesi için uygun ölçütler kullanılmalı ölçme sonuçları geçerli ve güvenilir olmalıdır.

Öğrenci başarısını ölçmek, uzun yıllar boyunca öğrencilerin öğretim sırasında edindikleri bilgileri ölçmek olarak değerlendirilmiştir. Ancak günümüzde, 21. yüzyıl ihtiyaçları sonucunda oluşan eğitim sisteminin genel amaçları göz önüne alındığında, öğrencilerin teknolojiyi doğru, üretken ve verimli kullanmaları, eleştirel düşünmeleri, bağımsız ve hayat boyu öğrenen, problem çözebilen, sahip olduğu bilgileri kullanabilen bireyler olarak yetişmesi beklenmektedir. Bu sebeple üst düzey zihinsel becerilerin ölçülmesi; eğitim hedeflerinin, ekonomi ve iş gücünün doğasının değişmesi gibi nedenlerden dolayı daha önemli hale gelmiştir (Karakaya, 2022). Klasik ölçme ve değerlendirme yöntemlerinin, öğrencilerden beklenen üst düzey zihinsel becerilerinin ölçülmesi ve değerlendirilmesi sürecinde yetersiz kalması sebebiyle, açık uçlu maddeler ve performans değerlendirmeye dayalı yaklaşımlar önem kazanmıştır. Öğrencilerin cevaplarını kendilerinin oluşturması (Haladyna & Rodriguez, 2013) ve öğrencilerden üst düzey bilişsel cevapların alınması (Popham, 1997) açık uçlu maddelerin ölçme ve değerlendirme sürecindeki yerini önemli hale getirmektedir.

Açık uçlu maddeler, bilişsel süreçlerin değerlendirilmesinde ve bilimsel araştırmalarda veri toplama amacıyla farklı ölçme araçları içinde kullanılabilen; birden fazla cevabın doğru olabileceğini gösteren madde türüdür (Brookhards, 2023; Karakaya, 2022). Matematik dersinde, açık uçlu sorular için, problem kavramı önem kazanmaktadır. Problem, belirli açık sorunlar taşıyan, problemle karşılaşmış olan kişinin ilgisini çeken ve bu sorunu çözecek, yeteri kadar algoritma ve çözüm yöntemi bilgisine sahip olmadığı durumlara denir (Bloom & Niss, 1991). Altun (2014), problemleri, rutin problemler ve rutin olmayan problemler olarak iki farklı biçimde sınıflandırmıştır. Bir problem, günlük yaşamda sık karşılaşılan olaylardan oluşturulmuş; daha önceden çözülmüş ve genelleştirilebilir bir probleme, verilen özel verileri yerleştirerek ya da hiçbir yenilik yapmadan, sadece bilinen bir örneğin adımları takip edilerek, sadece probleme yeni değerleri yerleştirerek çözülebiliyorsa, bu probleme rutin

bir problem denir (Polya, 1973). Rutin olmayan problemler ise, daha önceden öğrenilmiş formül ya da yöntemlerle çözülemeyen (Altun, 2014); çözüm yapabilmek için yöntemin açık ve net bir şekilde gözükmediği; çözümü için, öğrencinin elinde olan verileri dikkatli bir biçimde analiz etmesini, kendine has bir problem çözme girişiminde bulunmasını, bir veya birden çok, farklı stratejileri kullanmasını gerektiren; matematiksel işlem becerilerinden ziyade, verileri düzenleme, sınıflandırma ve ilişkileri görme gibi bazı becerilere sahip olmanın ve birbiriyle ilişkili eylemleri arka arkaya planlı bir şekilde yapmanın gerektiği çözümler gerektiren problemler olarak tanımlanabilir (Altun, 2014; Arıkan,2024; Çelik ve Güler, 2013; Gürbüz ve Güder, 2016; Işık ve Kar, 2012; Şahin, Güler ve Taşdelen, 2016)

Öğrencilerin ögrenmelerine ilişkin yeterli bilgi edinebilmek için portfolyolar, performans görevleri veya projeler gibi farklı performans değerlendirme yöntemlerinden yararlanılması önerilmiş (Kantrov, 2000; MEB, 2013; NCTM,1995); çoktan seçmeli testlerin, öğrencilerin matematik başarıları hakkında değerlendirmeleri için sınırlılıkların bulunduğu (Alharby, 2006; Bağcan vd., 2013; Bal ve Doğanay, 2010; Baştürk, 2008; Berberoğlu, 1988; Stecher, 2010; Turgut ve Baykul, 2012; Woodward vd., 2001), bunun için öğrencilerin çözüme nasıl ulaştıklarının görülebileceği, çözümlerin gerekçelerini açıklayabilecekleri açık uçlu soruların kullanılması üzerinde durulmuştur (Archbald & Grant, 2000). Ancak, açık uçlu etkinlikler, proje ve performans görevleri için objektif puanlanama yapılamaması bir güvenirlik problemidir (Kutlu vd., 2014). Açık uçlu soruların ya da performans değerlendirme sürecinde öğrencilerin ne şekilde puanlanacağına karar vermek ve yapılan ölçmenin güvenirliğini sağlamak yaşanan güçlüklerdendir. Ölçme ve değerlendirme süreçlerinde, puanlayıcı etkisini en aza indirebilmek için puanlama sürecinde birden çok puanlayıcı kullanılması, puanlayıcılara eğitim verilmesi (Ebel,1951; Görgülü Öztürk, 2023; Haladyna, 1997; İlhan ve Çetin, 2014; Kubiszyn & Borich, 2013; Şata,2019; Woehr,1994) ve puanlamalarda puanlama anahtarları kullanılması (Kutlu vd., 2014; Wolf & Stevens, 2007) gibi öneriler getirilmiştir.

Performans değerlendirme sürecinde, yapılacak ölçme ve değerlendirmenin güvenirliğini sağlamak için puanlayıcı etkisinin en aza indirilmesi gerekmektedir. Puanlayıcı etkisi, bireyin performans değerlendirme puanlarına, ölçülen yapıdan bağımsız olarak, puanlayıcılardan kaynaklanan hata, puanlayıcıların sistematik davranışları (Bachman, 2004; Eckes, 2005; Hoyt, 2000) olarak tanımlanabilir. Alan yazında performans değerlendirme sürecinde otuzdan fazla problemli puanlayıcı davranışı olduğu belirtilmekte (Royal & Hecker, 2016); bununla birlikte, problemli puanlayıcı davranışlarından merkeze yönelim, halo etkisi, katılık ve cömertlik, tutarsızlık ve farklılaşan puanlayıcı katılık ve cömertliği diğer davranışlardan daha fazla ortaya çıkan davranışlardır (Myford & Wolfe, 2004).

Puanlayıcı katılığı ve cömertliği, puanlayıcının devamlı bir şekilde, diğer puanlayıcılardan ya da belirlenmiş puanlama ölçütlerinden yüksek ya da düşük puan verme eğiliminde olması şeklinde tanımlanabilir (Şata, 2019). Cronbach (1990), puanlama sürecinde, puanlamaya karışan en önemli puanlayıcı etkisinin, puanlayıcı katılığı ve cömertliği olduğunu ifade etmiştir. Puanlayıcıların katılığı veya cömertliği, kriterler, puanlama zamanı, birey gibi değişkenlere göre değişebilmektedir (Şata, 2019). Bir diğer puanlayıcı etkisi olan Halo etkisini Thorndike (1920), puanlayıcının, performansı puanlanan kişi hakkında edindiği genel izlenimine göre, genel olarak iyi ya da genel olarak kötü şeklinde düşünerek puanlamayı yapması olarak tanımlamıştır. Merkeze yönelim etkisi, puanlayıcının, puanlama ölçeğinin orta kategorisini diğer kategorilerden daha fazla eğilim göstermesi ve daha fazla kullanılması olarak tanımlanmaktadır (Royal & Hecker, 2016; Wolfe & McVay, 2010). Merkeze yönelim etkisi, puanlamanın değişkenliğinin azalmasından dolayı hem geçerlilik hem de güvenirliğin azalmasına neden olur (Anastasi, 1988). Ranj sınırlaması, puanlayıcının, öğrenci performansından bağımsız olarak, puanlama ölçeğinin belirli kategorilerini daha fazla kullanma eğiliminde olması olarak tanımlanmaktadır (Moore,

2009). Öğrencilerin performans değerlendirilmesi sürecinde geçerliği tehdit eden unsurlardandır (Saal vd., 1980). Tutarsızlık, puanlayıcılardan birinin, puanlama ölçeğini diğerlerinden daha farklı şekilde kullanması şeklinde tanımlanabilir (Myforde & Wolfe, 2004). Yanlılık ise, puanlayıcının, değerlendirme süreci içinde bireylerin cinsiyet, yaş ya da diğer kültürel faktörler gibi çeşitli özelliklerine göre, olması gerekenden daha yüksek veya düşük puan verme eğiliminde olması şeklinde tanımlanmaktadır (Kumar, 2005).

Puanlayıcı etkisini en aza indirebilmek için, puanlama sürecinde birden çok puanlayıcı kullanılması, puanlayıcılara eğitim verilmesi (Ebel,1951; Haladyna, 1997; İlhan ve Çetin, 2014; Kubiszyn & Borich, 2013; Woehr, 1994) ve puanlamalarda puanlama anahtarları kullanılması (Kutlu vd., 2014; Wolf & Stevens, 2007) gibi öneriler getirilmiştir. Puanlamalarda kullanılan dereceli puanlama anahtarları, istenen ölçütleri belirlediğinden, puanlayıcılar arasındaki farklılıkları azaltarak, tutarsızlıkların da oluşmasını azaltabilir (Moskal & Leydens, 2000).

Dereceli puanlama anahtarları (DPA), etkili ve doğru dönütler yoluyla, öğrencilerin değerlendirmeden sonraki görevlerinde, kavramları ve becerileri daha iyi anlamasını kolaylaştırdığı için de matematik başarısının değerlendirilmesinde büyük bir öneme sahiptir (Lau vd., 2015). Kritik kararların verildiği ya da öğrenme amaçlı değerlendirme süreçlerinde öğretmene de öğrenciye de ne arandığını ve aslında neyin daha önemli olduğunu anlatır (Jonsson & Svingby, 2007).

Brookhart (2023), dereceli puanlama anahtarını, ölçütlerde yer alan öğrenci performansının niteliklerine ait düzeyleri tanımlayan, tutarlı bir ölçüt seti olarak tanımlamıştır. Dereceli puanlama anahtarları, ölçmede tutarlılığı iyileştirirken değerlendirme için kullanılan başarı standartlarını daha da netleştirdiği için geçerliliği de artırır. Öğrenci cevapları, dereceli puanlama anahtarı ölçütleriyle tutarsız bir şekilde eşleştirildiğinde, puanlama sürecinin güvenirliği de etkilenmektedir (Brookhart & Nitko, 2014). Bu yüzden, açık uçlu maddelerin değerlendirilmesinde ölçmeye karışan puanlayıcı hatalarını en az seviyede tutmak için puanlama anahtarlarının doğru şekilde kullanılması gerekmektedir. Rutin olmayan matematik problemlerinin puanlanmasında, analitik ve bütünsel dereceli puanlama anahtarlarının ikisi de kullanılabilir.

Bütünsel DPA, bireyin performansının tamamının ya da ortaya koyduğu ürünün bir bütün olarak ele alınarak, tek bir puan ile değerlendirildiği; farklı seviyelerdeki performansların kalitesinin ortaya konulduğu puanlama anahtarlarıdır (Kutlu vd., 2014). Bütünsel DPA'nın, daha hızlı olması (Mertler, 2001), puanlayıcılar arası güvenirliğe daha kısa sürede ulaşabilmesi (Reddy, 2010), ekonomik ve daha pratik olması (Weigle, 2002), düzey belirleyici değerlendirme için daha uygun olması (Brookhart, 2013) gibi avantajları vardır. Öğrencilerin hakkında ayrıntılı bilgi edinilmesin imkân sağlamaması (Bargainnier, 2003), farklı eksiklikleri bulunan iki bireyin aynı puanı almasına sebep olabilmesi (Arter & Mctighe, 2001), öğrencilerin güçlü ya da zayıf oldukları yerlere ilişkin bilgi elde edilememesi dolayısıyla öğrencilere geri bildirim verilememesine (Arter & Mctighe, 2001) ve dolayısıyla nasıl bir ek öğretime ihtiyaç olduğuna dair hedef belirlenmesinde yeterince açıklayıcı olmaması (Nelson & VanMeter, 2007) bütünsel DPA'nın dezavantajlarındandır. Seviye belirlemek amacıyla yapılan değerlendirmelerde bütünsel dereceli puanlama anahtarlarının kullanılması önerilmektedir (Mertler, 2001).

Analitik DPA, bireyin performansının farklı ölçütlerdeki düzeylerini, güçlü̈ ve zayıf yönlerini belirten puanlama araçlarıdır (Kutlu vd., 2014). Analitik DPA'da, performansı oluşturan bileşenler ayrı ayrı puanlanarak (Klein vd.,1998), her bir bileşen için puanlar toplanır ve performansın genel puanı elde edilir (Petkov & Petkova, 2006). Öğrencilerin performansının farklı bileşenlerinin daha zayıf veya daha güçlü̈ yönlerini ortaya koyması ve ayrıntılı geri bildirim verilebilmesinden (Nitko, 2004) dolayı tanıma ve biçimlendirmeye yönelik değerlendirmeler için daha fazla tercih edilmektedir (Reddy, 2010) Her

ölçüt için farklı bir düzey belirlenmesi ve her ölçüte farklı puan verilmesinden dolayı, puanlama yapmak daha uzun sürer, fakat daha ayrıntılı geribildirim sağlar (İlhan, 2015). Analitik DPA'nın oluşturulması ve kullanılmasının zaman alması ve maliyetli olması (Wiseman, 2012); her ölçüt için puanların iyi tanımlanmaması, puanlayıcıların öğrenci performanslarına aynı puanı verememesi analitik puanlama anahtarlarının sınırlılıkları (Moskal, 2000) olarak görülmektedir.

Analitik ve bütünsel DPA'lar öğrenci performanslarını değerlendirmede farklı açılardan yaklaşmaktadırlar. Analitik ve bütünsel DPA'ları karşılaştırmak üzerine yapılan çalışmalarda, analitik DPA ile yapılan puanlamaların tutarlılıklarının daha fazla olduğu ve öğrencileri daha iyi tanımladığı görülmüştür (Meier vd., 2006). Bütünsel DPA ile yapılan puanlamalarda puanlayıcıların kişisel görüş ve beklentilerinden daha fazla etkilendiği, bunun da tutarlılığı düşürdüğü görülmüştür. Ayrıca analitik DPA ile yapılan puanlamalarda ayrıntılı geribildirim verilmesinin biçimlendirci değerlendirmeler için önemli olduğu vurgulanmıştır (Badia, 2019; Chukwuere, 2021; Jönsson, vd., 2021).

Uluslararası olarak yapılan öğrenci başarılarını değerlendirme çalışmalarına (PISA, TIMSS, PIRLS) ülkemiz uzun yıllardır katılmasına rağmen, bu çalışmaların sonuçları incelendiğinde, ülkemizin ortalamanın altında kaldığı görülmektedir. Bu sebeple, ülkemizde üst düzey zihinsel becerileri geliştirmek için hazırlanan eğitim programları ve bu programlar içinde uygun ölçme değerlendirme yöntemleri kullanılması bir gerekliliktir. Bununla birlikte, açık uçlu sorular ve rutin olmayan problemlerin ölçme ve değerlendirme süreçlerinde kullanılması, daha derin öğrenmeye teşvik ederek (Karakaya, 2022), öğrenciler için üst düzey düşünme becerilerinin kazandırılmasına ve geliştirilmesine katkı sağlayacaktır. Bu çalışma, açık uçlu soruların doğru bir şekilde puanlanması için, puanlama sürecinde hangi puanlama anahtarının kullanılacağına ilişkin bilgiler verecek olmasından dolayı, eğitim öğretimin değerlendirilmesi süreçlerine ve literatüre katkı sağlayacaktır.

ÖSYM yaptığı açık uçlu maddelerden oluşan sınavlarda standart bir puanlama deseni tercih etmiştir. Bu deseninin her bir alt test için güvenilir sonuçlar verip vermediği, puanlayıcılar arası güvenirliği bu standart puanlama deseninden daha yüksek olan puanlama desenlerinin kullanılıp kullanılamayacağı konusunda ülkemizde yapılan çalışmalar oldukça azdır (İlhan, 2015). Açık uçlu maddelerin puanlanmasında kullanılabilecek analitik ve bütünsel puanlama anahtarlarının karşılaştırılması; bunlara göre, puanlayıcı hatalarının minimum düzeyde olduğu yöntemlerin uygulanması, öğrenci başarısının daha güvenilir bir biçimde belirlenmesine katkı sağlayacaktır. Alan yazında, matematik dersi için açık uçlu maddelerin ve rutin olmayan problemlerin puanlanmasında, bütünsel ve analitik puanlama anahtarlarının kullanılmasının ve puanlayıcı etkilerinin analiz edildiği bir çalışmaya rastlanmadığı dikkate alındığında, uygulamaya yönelik sonuçlarının yanında, bu araştırmanın matematik eğitimi ile ölçme ve değerlendirme alanındaki literatüre de katkısı olacağı düşünülmektedir.

Bu çalışmada, analitik DPA ve bütünsel DPA ile puanlanan rutin olmayan matematik problemlerinden oluşan matematik başarı testinin puanlanmasının, puanlayıcı davranışları üzerine etkilerinin çok yüzeyli Rasch modeli ile incelenmesi amaçlanmıştır.

Buna göre, aşağıda verilen problemlere yanıt aranmıştır.

1. Analitik dereceli puanlama anahtarlarına göre yapılan puanlamalarda; puanlayıcı, birey ve madde yüzeyleri için güvenirlik değerleri ve uygunluk istatistikleri nasıldır?
2. Analitik dereceli puanlama anahtarlarına göre yapılan puanlamalarda; puanlayıcı katılığı-cömertliği ve yanlılık davranışları nasıldır?
3. Bütünsel dereceli puanlama anahtarlarına göre yapılan puanlamalarda; puanlayıcı, birey ve madde yüzeyleri için güvenirlik değerleri ve uygunluk istatistikleri nasıldır?
4. Bütünsel dereceli puanlama anahtarlarına göre yapılan puanlamalarda; puanlayıcı katılığı-cömertliği ve yanlılık davranışları nasıldır?

2. Yöntem

2.1. Araştırmanın Deseni

Bu araştırma betimsel bir araştırma niteliğindedir. Bu araştırmada, betimsel araştırma yöntemlerinden tarama modeli kullanılmıştır. Tarama modeli, zaman içerisinde gerçekleşen değişikliklerle birlikte, belirli bir zamanda ortaya çıkan durum hakkında değerlendirilmelerin yapılması ve incelenen durumun içyüzünün aydınlatılmaya çalışılmasıdır (Aypay, 2015; Karasar, 2020).

2.2. Çalışma Grubu

Bu çalışmanın katılımcıları, rutin olmayan matematik problemlerinden oluşan başarı testinin uygulandığı öğrenciler ve cevaplanan başarı testini değerlendiren puanlayıcılar olmak üzere iki farklı kısımdan oluşmaktadır. Çalışmanın öğrenci grubunu 2023-2024 Eğitim Öğretim Yılı'nda Sakarya İli'nde ortaokul sekizinci sınıfa devam eden 10 kız (%50), 10 erkek (%50) toplam 20 öğrenci oluşturmaktadır. Çalışmanın asıl çalışma grubunu yedi kadın (%44) ve dokuz erkek (%56) olmak üzere toplam 16 matematik öğretmeni oluşturmaktadır. Bu öğretmenler, öğrenciler tarafından cevaplanan başarı testini dereceli puanlama anahtarlarını kullanarak değerlendirmiştir. Ayrıca araştırmada gönüllülük esasına göre görev almıştır.

2.3. Veri Toplama Araçları

Araştırma verilerinin toplanmasında, rutin olmayan açık uçlu maddeler oluşan matematik başarı testi, analitik ve bütünsel DPA'lar ile puanlayıcı formları kullanılmıştır.

2.3.1. Matematik Başarı Testi

Rutin olmayan 20 matematik sorusundan oluşan başarı testi araştırmacı tarafından hazırlanmıştır. Üç matematik alan uzmanı, bir ölçme ve değerlendirme uzmanının görüşleri doğrultusunda sekizinci sınıf düzeyinde olan 17 soru ile 30 öğrenciyle pilot çalışma sonrasında yapılan analizlerin sonuçları ve uzman görüşlerine başvurularak hazırlanan başarı testinde, madde ayırıcılık indeksi 0,30 ve üzeri, madde güçlük düzeyi 0,40 ile 0,60 olan toplam 15 soru seçilmiştir.

2.3.2. Puanlama Anahtarları

Öğrencilerin cevaplamış olduğu matematik problemlerinin puanlanmasında, analitik ve bütünsel dereceli puanlama anahtarları kullanılmıştır. Altun (2020) tarafından matematiksel okuryazarlık için gerekli olduğu vurgulanan matematiksel yeterlilikler bu araştırmada kullanılan dereceli puanlama anahtarlarında ölçüt olarak kullanılmıştır. Temel işlem becerileri, sembolik ve teknik dil ve işlemleri kullanma, modelleme, muhakeme ve argümantasyon, problem çözme için stratejiler oluşturma, temsil ile gösterim ve iletişim olarak verilen matematiksel yeterlikler incelenmiş, yapılan pilot çalışma ve uzman görüşleri de dikkate alınarak hazırlanmıştır. Analitik DPA, 0 ile 4 puan arasında performans düzeyi olacak şekilde, her bir puan altı ölçütle incelenecek genel puanlama anahtarı olarak; bütünsel DPA ise, performans düzeyleri 0 ile 4 puan arasında olacak şekilde matematiksel yeterliklere göre hazırlanmıştır. Araştırmacı tarafından hazırlanmış olan DPA'ların geçerlik ve güvenirliğinin sağlanması için uzman görüşleri alınmış ve dereceli puanlama anahtarları için Lawshe (1975) tarafından geliştirilmiş olan kapsam geçerlik indeksi her bir madde için hesaplanmış; 0,05 anlamlılık düzeyinde her bir madde için eşik değeri olan 0,736'nın üzerinde bulunmuştur. Her iki dereceli puanlama anahtarı da matematik alanında yüksek lisans mezunu iki matematik öğretmeni ile ölçme ve değerlendirme

alanında uzman iki öğretim üyesi olmak üzere toplam dört uzman tarafından incelenmiştir. Pilot uygulama, uzman görüşleri ve kapsam geçerlilik indeksleri hesaplamaları sonucunda gerekli düzenlemeler yapılmıştır. Matematiksel yeterliliklerden biri olan iletişim, puanlama anahtarlarından çıkarılmıştır. Daha sonra cümle ve yapı olarak uygun şekilde dereceli puanlama anahtarlarına son halleri verilmiştir.

2.4. Veri Toplama Süreci

Çalışma için belirlenen okullardaki sekizinci sınıf öğrencileri öncelikle çalışma hakkında bilgilendirilmiş. Gönüllülük esasına göre katılım sağlanmış ve gerekli izinler alınmıştır. Rutin olmayan matematik problemlerinden oluşan matematik başarı testi, beşer sorudan oluşan iki farklı oturum olarak deneme formatında hazırlanmış ve iki farklı günde çalışmaya katılan 90 sekizinci sınıf öğrencisine uygulanmıştır. Uygulamanın sonrasında, cevaplanan kağıtlar içerisinden, tüm soruları cevaplayan 20 öğrenciye ait cevaplar seçilmiştir.

Puanlayıcılar için, puanlama süreci öncesinde çalışma hakkında süreci, puanlama anahtarları, puanlama sürecinde puanlamaya karışan ölçme hataları, analitik ve bütünsel puanlama anahtarlarını ve ölçütlerini tanıtan, araştırmacı tarafından örnek bir puanlamanın yapıldığı bir puanlayıcı eğitimi yapılmıştır.

Öğrencilerden toplanan başarı testleri, her bir öğrenciye kod verilerek numaralandırılmış (Ö1, Ö2, …, Ö20), her soru her bir öğrenci için ayrı ayrı çoğaltılmıştır. Puanlayıcılar için başarı testleri ve cevap anahtarları, puanlama anahtarları, puanlama formları, öğretmen bilgi formları ve bilgilendirme sunumunu içeren bir dosya hazırlanmış ve her bir öğretmene verilmiştir.

2.5. Verilerin Analizi

Bu çalışmada elde edilen veriler öğrencilerin açık uçlu sorulara verdikleri cevapların, puanlayıcılar tarafından analitik ve bütünsel dereceli puanlama anahtarları ile puanlaması ile elde edilmiştir. DPA'larda 0-4 arası derecelendirme kullanılmıştır, ancak analitik DPA'da her bir puan 6 farklı ölçüt ile incelendiği için, bütünsel DPA ile arasında fark olmasından dolayı analiz sırasında puanlarda eşitleme yapılmıştır.

Araştırma kapsamında elde edilen veri çok yüzeyli Rasch ölçme modeli ile analiz edilmiştir. Madde tepki kuramının tek parametreli modellerinden biri olan ve modeli geliştiren George Rasch'ın (1960) ismiyle anılan Rasch modeli, klasik test kuramının sınırlılıklarını gidermek amacıyla geliştirilmiştir. Linacre (1989) ise, ölçme işlemine yüzeyleri veya birçok boyutu dahil etmemize olanak sağlayan çok yüzeyli Rasch ölçme modelini geliştirmiştir. Çok yüzeyli Rasch modelinde, test veya değerlendirme puanlarını etkileme olasılığı olan değişkenlik kaynaklarının her birine yüzey adı verilmektedir (Eckes, 2015). Bu çalışmada da dereceli puanlama anahtarı türünün, puanlayıcı davranışları üzerindeki etkisi çok yüzeyli Rasch modeliyle analiz edilerek incelenmiştir. Araştırma kapsamında birey, madde ve puanlayıcı olmak üzere üç yüzey yer almaktadır. Analizler yapılırken FACET paket programından faydalanılmıştır. Çok yüzeyli Rasch modeli analizine başlamadan önce çok yüzeyli Rasch modelinin tek boyutluluk, yerel bağımsızlık ve model veri uyumu varsayımları incelenmiş ve varsayımların sağlandığı görülmüştür. Çok yüzeyli Rasch modeli varsayımlarının karşılandığı belirlendikten sonra analize geçilmiştir.

2.6. Geçerlik, Güvenirlik ve Etik

Araştırmada veri toplama araçlarının güvenirliğini sağlamak için her bir veri toplama aracı için uzman görüşlerine başvurulmuştur Dereceli puanlama anahtarları için Lawshe(1975) tarafından geliştirilmiş olan kapsam geçerlik indeksi her bir madde için hesaplanmıştır. Bu araştırmanın, Gazi Üniversitesi Etik Komisyonu tarafından 17.10.2023 tarihinde yapılan 18 sayılı toplantısında alınan karar ile etik kurul onayı bulunmaktadır. Araştırmanın planlanması, uygulanması, verilerin toplanması ve verilerin analizi süreçlerinde "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" nde uyulması belirtilen tüm kurallara uyulmuştur.

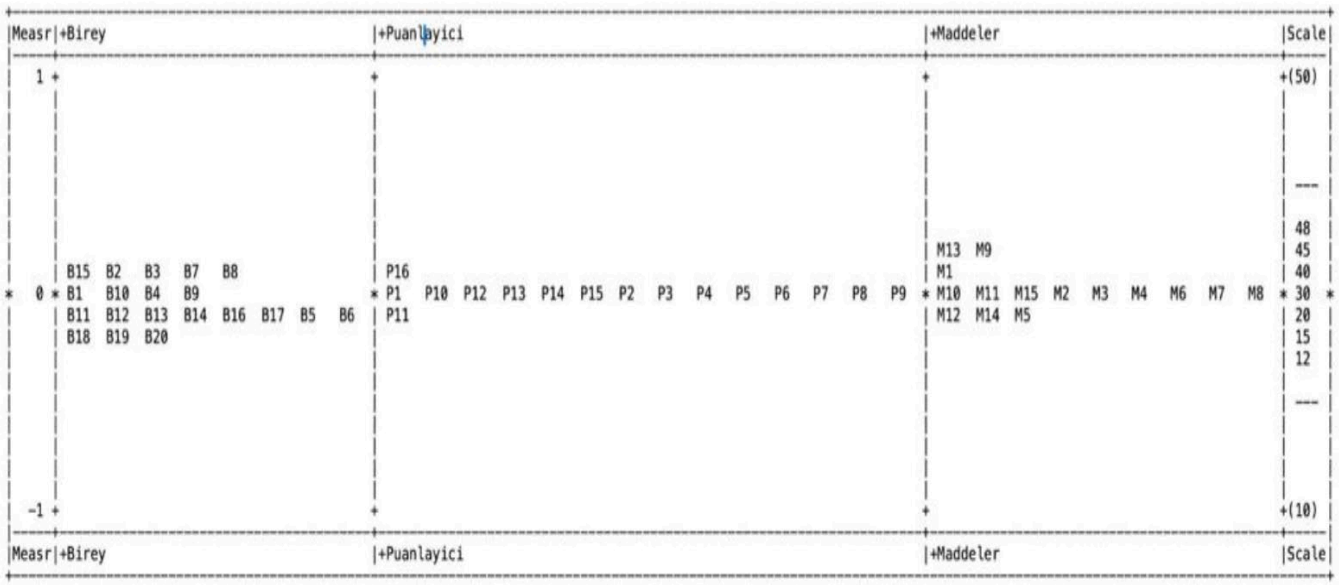
3. Bulgular

Bu bölümde dereceli puanlama anahtarı türüne göre, araştırma sorularına ilişkin bulgulara yer verilmiştir.

3.1. Analitik Dereceli Puanlama Anahtarına Göre Yapılan Puanlamalar

Öncelikle, "Analitik dereceli puanlama anahtarına göre yapılan puanlamada, puanlayıcı, birey ve madde yüzeyleri için hesaplanan güvenirlik değerleri ve uygunluk istatistikleri nasıldır?" araştırma sorusu için yapılan analiz sonuçlarında bulunan veri değişkenlik haritası incelenmiştir. Aynı ölçek üzerinde birey, puanlayıcı ve maddeler arasındaki ilişkiyi öz ve uygun bilgiyi hızlıca edinmemizi veri değişken haritası sağlamaktadır (Nakamura, 2000). Analitik dereceli puanlama anahtarı kullanan puanlayıcılardan elde edilen verilerin analiz sonuçlarına ait değişkenlik haritası Şekil 1'de verilmiştir.

Şekil 1. *Analitik Dereceli Puanlama Anahtarına Ait Değişken Haritası*



Şekil 1'e bakıldığında ilk sütunda logit ölçeği, ikinci sütunda birey yüzeyi, üçüncü sütunda puanlayıcı yüzeyi, dördüncü sütunda ise madde yüzeyine ilişkin ölçümler bulunmaktadır. Değişken haritasındaki puanlayıcılara ilişkin ölçümlere göre, sütunun alt tarafında yer alan ve düşük logit değerine sahip puanlayıcıların daha katı puanlamalar yaptığı; sütunun üst tarafında yer alan ve yüksek logit değerine sahip olan puanlayıcıların ise daha cömert puanlamalar yaptığı söylenebilir. Buna göre, en cömert puanlamaların 16 numaralı puanlayıcı (P16) (0,06 logit birimde), en katı puanlamaların ise 11 numaralı puanlayıcı (P11) (-0,12 logit birimde) tarafından yapıldığı görülmektedir. Birey yüzeyine ait

ölçümler incelendiğinde, sütunun üst ucundan alt ucuna doğru yetenek düzeyleri düştüğü için, analiz sonucu elde edilen değerler de incelendiğinde en yüksek yetenek düzeyindeki öğrencinin üç numaralı öğrenci (B3) (0,11 logit birimde), en düşük yetenek düzeyindeki öğrencilerin ise 18, 19 ve 20 numaralı öğrenciler (B18, B19, B20) (-0,16 logit birimde) olduğu görülmektedir. Madde yüzeyine ait ölçümler incelendiğinde, sütunun alt ucundan üst ucuna doğru maddeler kolaylaştığı için, en kolay maddenin 13.madde (0,17 logit biriminde), en zor maddenin ise beşinci madde olduğu (-0,13 logit biriminde) belirlenmiştir.

Birey, madde ve puanlayıcı yüzeyleri için, logit ölçeğinin pozitif ve negatif uçları arasında uzanan ölçümlerin elde edilmesi; maddelerin güçlük düzeyleri açısından farklılık olduğuna, bireylerin yetenek düzeyleri bakımından ayırt edilebildiğine ve puanlayıcılar arasında katılık ve cömertlikleri açısından fark olduğuna işaret eder, fakat bu konuda daha kesin bir sonuca varabilmek için her bir yüzey için ölçüm raporları ayrı ayrı incelenmelidir. Tablo 1'de analitik DPA'yla yapılan puanlamalara ait puanlayıcı yüzeyine ait ölçüm raporları gösterilmiştir. Puanlayıcılar için logit ölçümlerinin pozitif olması puanlayıcının cömert; logit ölçümleri negatif olması ise puanlayıcının katı puanlamalar yaptığını göstermektedir.

Tablo 1. *Analitik Dereceli Puanlama Anahtarına Ait Puanlayıcı Yüzeyi İçin Elde Edilen Ölçüm Raporları*

Puanlayıcı	*Gözlenen Ortalama*	*Düzeltilmiş Ortalama*	*Logit Ölçüsü*	*Standart Hata*	*Uygunluk İçi*	*Uygunluk Dışı*	*t değeri*
P1	24,29	22,06	-0,02	0,01	0,69	0,63	-2
P2	28,62	28,30	0,04	0,01	0,81	0,84	4
P3	26,43	24,95	0,01	0,01	0,80	0,68	1
P4	23,9	21,47	-0,02	0,01	1,06	1,02	-2
P5	25,72	23,96	0	0,01	0,93	0,84	0
P6	26,18	24,57	0,01	0,01	0,80	0,73	1
P7	26,49	25,07	0,01	0,01	0,77	0,68	1
P8	26,45	25,04	0,01	0,01	1,76	2,74	1
P9	28,92	28,75	0,04	0,01	1,12	1,09	4
P10	29,11	29,09	0,05	0,01	1,29	1,31	5
P11	17,90	15,21	-0,12	0,01	0,78	0,71	-12
P12	23,48	20,94	-0,03	0,01	1,05	1,00	-3
P13	23,41	20,86	-0,03	0,01	1,10	1,05	-3
P14	24,87	22,75	-0,01	0,01	0,88	0,90	-1
P15	25,82	24,04	0	0,01	1,00	0,90	0
P16	30,35	31,02	0,06	0,01	1,36	1,40	6
Ortalama	25,75	24,25	0,00	0,01	1,01	1,03	
Standart Sapma (Evren)	2,86	3,77	0,04	0,00	0,27	0,49	
Standart Sapma (Örneklem)	2,96	3,89	0,04	0,00	0,28	0,51	
Model, Evren: RMSE =0,01 Standart Sapma= 0,04 Ayırma Oranı= 4,60 Ayırma İndeksi=6,46 Güvenirlik= 0,95							
Model, Örneklem: RMSE = 0,01 Standart Sapma= 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							
Puanlayıcılar arası mutlak uyum: % 38,2							

Puanlayıcılar arası beklenen uyum: % 23,5
Puanlayıcılar arası Güvenirliğe İlişkin Kappa İstatistiği: 0,19

Not. RMSE: Hata kareleri ortalamasının karekökü, sd: serbestlik derecesi

Wright ve Linacre (1994), uygunluk istatistiklerine (uygunluk içi ve dışı) dair kabul edilebilir aralığın 0,6 ile 1,4 olduğunu; Myford ve Wolfe (2003) ise, 1,5 ile 2 arasındaki değerlerin, ölçüm için yararlı da zararlı da olmadığını, ikiye kadar olan uygunluk istatistiklerini kabul edilebilir olduğunu belirtmişler, 0,5 ile 2 aralığının dikkate alınması önermişlerdir. Bununla birlikte, Sudweeks, Reeve ve Bradshaw (2004), 2'nin üstündeki uygunluk istatistiklerinin ölçüm için zararlı olduğunu belirtmişlerdir. Puanlayıcı yüzeyi için uygunluk içi ve uygunluk dışı istatistiklerine ait ortalamalara bakıldığında, uygunluk içi ölçüm ortalamasının 1,0; uygunluk dışı ölçüm ortalamasının ise 1,03 olduğu görülmüştür. Buna göre, elde edilen değerlere göre verinin model ile uyumlu olduğu görülmektedir. Uygunluk istatistikleri incelendiğinde, puanlayıcılardan P8 hariç diğerlerinin kabul edilebilir aralık içinde kaldığı görülmüştür. Buna göre, P8'in model ve veri uyumunu olumsuz şekilde etkileyen puanlayıcı olarak yansıdığı belirlenmiştir.

Tablo 1'e göre, güvenirlik indeksi 0,96; ayırma oranı 4,76 ve ayırma indeksi 6,67 olarak bulunmuştur. Puanlayıcı ayırma indeksinin 0,00 ve 0,00'a yakın bir değer olması beklenirken bu araştırma için istenen değerlerin üstünde bir değer (6,67) bulunması, puanlayıcılar arasında puanlama noktasında farklılıklarının olduğunu, puanlayıcıların katılık-cömertlik düzeylerine göre birbirlerinden farklılaştığını ve puanlamalarda cömertlik/katılık hatasının olabileceğini göstermektedir. Puanlayıcı yüzeyine ilişkin güvenirlik indeksinin yüksek bir değer olması (0,96), puanlayıcıların katılık-cömertlik açısından farklılık gösterdiğini belirtse de bu farkın anlamlılığı Ki-Kare değerine göre yorumlanmaktadır. Ki- Kare değeri incelendiğinde, istatistiksel olarak anlamlı [χ2(sd)=299,8 (15) p=0,00] olduğu, dolayısıyla katılık ve cömertlik davranışları açısından puanlayıcılar arasında anlamlı bir farkın olduğu bulunmuştur.

Tablo 1'de verilen puanlayıcılar arası beklenen uyum, mutlak uyum ve güvenirliğe ilişkin Kappa istatistiği değerleri de puanlayıcı güvenirliğine dair verilen istatistiklerdir. Literatür incelendiğinde, 0,40'ın altındaki Kappa istatistiği değerlerinin kötü uyumu yansıttığı belirtilmiştir (Landis ve Knoch, 1977). Buna göre, elde edilen bu değer (0,19) analitik DPA kullanılarak yapılan puanlamalar için, puanlayıcılar arası uyumun iyi olmadığını göstermektedir. Birey yüzeyine ait ölçüm raporu Tablo 2'de sunulmuştur.

Tablo 2. *Analitik Dereceli Puanlama Anahtarına Ait Birey Yüzeyi için Elde Edilen Ölçüm Raporları*

Birey	*Gözlenen Ortalama*	*Düzeltilmiş Ortalama*	*Logit Ölçüsü*	*Standart Hata*	*Uygunluk İçi*	*Uygunluk Dışı*
1	26,62	25,90	-0,03	0,01	0,57	0,56
2	36,93	38,86	0,08	0,01	1,12	1,17
3	39,75	41,77	0,11	0,01	1,38	1,39
4	26,87	26,18	-0,03	0,01	1,48	1,61
5	20,21	18,29	-0,12	0,01	0,81	0,85
6	22,61	20,92	-0,08	0,01	0,87	0,89
7	36,83	38,74	0,08	0,01	1,24	1,19
8	35,69	37,43	0,06	0,01	0,96	0,87
9	30,93	31,54	0,01	0,01	0,73	0,79
10	28,55	28,41	-0,01	0,01	1,19	1,15
11	18,95	17,05	-0,14	0,01	0,68	0,59
12	19,13	17,19	-0,13	0,01	0,95	1,02
13	19,38	17,46	-0,13	0,01	0,85	0,95
14	21,07	19,18	-0,10	0,01	1,15	1,37
15	37,06	38,97	0,08	0,01	1,10	1,06
16	21,57	19,76	-0,10	0,01	1,09	0,94
17	20,25	18,34	-0,11	0,01	0,90	0,99
18	17,62	15,82	-0,16	0,01	0,93	0,98
19	17,43	15,67	-0,16	0,01	1,55	1,59
20	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
Standart Sapma (Evren)	7,59	9,15	0,09	0,00	0,26	0,29
Standart Sapma (Örneklem)	7,78	9,39	0,09	0,00	0,27	0,30
Model, Evren: RMSE = 0,01 Standart Sapma= 0,09 Ayırma Oranı= 9,15 Ayırma İndeksi=12,54 Güvenirlik= 0,99						
Model, Örneklem: RMSE = 0,01 Standart Sapma= 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						

Birey yüzeyine ilişkin analiz sonuçları incelendiğinde, öğrencilerin yetenek düzeylerine ilişkin kestirimlerin 0,11 logit ile -0,16 logit arasında (0,11- (-0,16) = 0,27 logit) değiştiği görülmüştür (Tablo 2). Logit değerlerinin geniş bir aralıkta olması, birey yüzeyi için performansın geniş bir aralıkta dağıldığına işaret eder. Bireylerin yetenek düzeylerinin ortalaması -0,05 ve standart sapması 0,09 logit bulunmuştur. Uygunluk içi istatistiğinin ortalaması 1,01ve uygunluk dışı istatistiğinin ortalaması da 1,03 olarak bulunmuştur. Birey yüzeyi için elde edilen uygunluk içi ve uygunluk dışı istatistiklerinin ortalamasının, model ve veri arasındaki uyumun iyi olduğu görülmüştür.

Tablo 2'ye göre, ayırma oranı 9,39; ayırma indeksi 12,86 ve güvenirlik indeksi 0,99 olarak hesaplanmıştır. Analiz sonucunda elde dilen bu değerlerin yüksek olması, farklı yetenek düzeyindeki bireylerin birbirinden iyi bir şekilde ayırt edildiğini göstermekle birlikte, istatistiksel olarak öğrenci yetenek düzeylerinin iyi ayırt edilip edilmediğine bakmak için Ki-Kare değeri incelenmelidir. Ki-Kare değeri incelendiğinde değerin [χ2(sd)=1702,0 (19), p=0,00] anlamlı olduğu, yetenek düzeyleri

bakımından öğrenciler arasında anlamlı fark bulunduğu görülmüştür. Madde yüzeyine ait ölçüm raporu Tablo 3'te sunulmuştur.

***Tablo 3**. Analitik Dereceli Puanlama Anahtarına Ait Madde Yüzeyi İçin Elde Edilen Ölçüm Raporları*

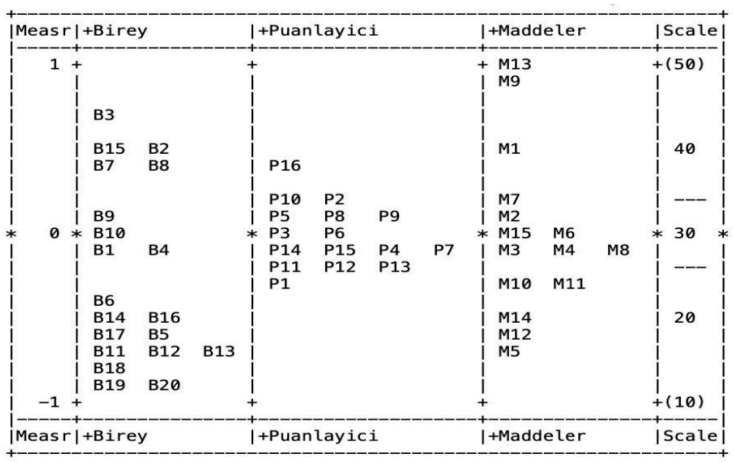
Madde	*Gözlenen Ortalama*	*Düzeltilmiş Ortalama*	*Logit Ölçüsü*	*Standart Hata*	*Uygunluk İçi*	*Uygunluk Dışı*
1	32,67	33,97	0,08	0,01	0,54	0,57
2	26,76	25,76	0,02	0,01	0,74	0,66
3	24,05	22,15	-0,02	0,01	0,83	0,79
4	24,23	22,35	-0,01	0,01	0,60	0,56
5	16,83	14,94	-0,13	0,01	1,20	1,27
6	23,71	21,76	-0,02	0,01	1,02	0,87
7	28,37	28,01	0,04	0,01	0,56	0,53
8	23,62	21,61	-0,02	0,01	1,16	1,23
9	39,12	41,53	0,16	0,01	1,40	1,29
10	21,96	19,68	-0,04	0,01	1,63	1,83
11	22,09	19,83	-0,04	0,01	1,16	1,09
12	18,23	16,06	-0,10	0,01	1,62	1,47
13	39,82	42,24	0,17	0,01	0,89	0,94
14	19,21	16,92	-0,09	0,01	1,64	1,54
15	25,53	24,08	0,00	0,01	0,84	0,82
Ortalama	25,75	24,73	0,00	0,01	1,06	1,03
Standart Sapma (Evren)	6,60	8,16	0,08	0,00	0,37	0,38
Standart Sapma (Örneklem)	6,83	8,45	0,09	0,00	0,39	0,40
Model, Evren: RMSE = 0,01 Standart Sapma= 0,08 Ayırma Oranı= 9,74 Ayırma İndeksi= 13,32 Güvenirlik= 0,99						
Model, Örneklem: RMSE = 0,01 Standart Sapma= 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						

Madde yüzeyine ait analiz sonuçları incelendiğinde, maddelerin güçlük düzeylerinin 0,17 logit ile -0.13 logit arasında değiştiği görülmektedir. Tablo 3'e göre, maddelerin logit değerlerinin ortalaması 0,00 ve standart sapması 0,09 logit olarak bulunmuştur. Madde yüzeyi için uygunluk içi istatistiğinin ortalaması 1,06 ve uygunluk dışı istatistiğinin ortalaması da 1,03 olarak bulunmuştur. Buna göre, model ve veri arasındaki uyumun iyi olduğu ve bu araştırma için oluşturulan maddelerin içinde, model ve veri arasındaki uyumu olumsuz etkileyen bir soru olmadığı söylenebilir. Madde yüzeyine ait ayırma oranı 10,08, ayırma indeksi 13,78 ve güvenirlik indeksi 0,99 olarak hesaplanmıştır. Analiz sonucunda elde edilen değerlerin yüksek olması, maddelerin güçlük düzeylerinin birbirlerinden farklı olduğunu göstermektedir, ancak bu farkın, istatistiksel olarak anlamlılığı Ki-Kare değeri ile yorumlanmaktadır. Ki-Kare değeri incelendiğinde [χ2(sd)=1359,4 (14), p=0,00], maddeler arasında güçlük düzeyleri açısından istatistiksel olarak anlamlı olarak fark olduğu görülmüştür.

3.2. Bütünsel Dereceli Puanlama Anahtarına Göre Yapılan Puanlamalar

"Bütünsel dereceli puanlama anahtarına göre yapılan puanlamada, puanlayıcı, birey ve madde yüzeyleri için hesaplanan güvenirlik değerleri ve uygunluk istatistikleri nasıldır?" araştırma sorusu için yapılan analiz sonuçlarına ilişkin bulgular aşağıda verilmiştir.

Şekil 2. *Bütünsel Dereceli Puanlama Anahtarına Ait Değişken Haritası*



Şekil 2'de verilen bütünsel DPA'ya göre yapılan puanlamalara ait değişken haritası incelendiğinde, ilk sütunda logit ölçeği, ikinci sütunda birey yüzeyine ait, üçüncü sütunda puanlayıcı yüzeyine ve dördüncü sütunda madde yüzeyine ölçümler bulunmaktadır. Buna göre, yetenek düzeyi en yüksek öğrencinin 3 numaralı öğrenci (0,71 logit birimde), yetenek düzeyi en düşük öğrencilerin ise 19 ve 20 numaralı öğrenciler (-0.88 logit birimde) olduğu görülmektedir. Puanlayıcılara ilişkin ölçümler incelendiğinde, en cömert puanlayıcının 16 numaralı puanlayıcı (0,38 logit birimde), en katı puanlayıcının ise 1 numaralı puanlayıcı (-0,26 logit birimde) olduğu görülmektedir. Maddelere ait veriler incelendiğinde, en kolay maddenin 13. madde (0,96 logit biriminde), en zor maddenin ise 5. madde olduğu (-0,67 logit biriminde) belirlenmiştir. Daha kesin bir sonuca ulaşılması için her bir yüzeye ait ölçüm raporları aşağıda incelenmiştir. Puanlayıcı yüzeyine ait ölçüm raporu Tablo 4'te sunulmuştur.

***Tablo 4**. Bütünsel Dereceli Puanlama Anahtarına Ait Puanlayıcı Yüzeyi İçin Elde Edilen Ölçüm Raporları*

Puanlayıcı	*Gözlenen Ortalama*	*Düzeltilmiş Ortalama*	*Logit Ölçüsü*	*Standart Hata*	*Uygunluk İçi*	*Uygunluk Dışı*	*t Değeri*
P1	22,83	20,26	-0,26	0,05	1,10	1,01	-5,2
P2	28,70	28,56	0,18	0,05	0,78	0,79	3,6
P3	26,17	24,77	0,00	0,05	0,75	0,64	0,0
P4	24,67	22,65	-0,11	0,05	0,84	0,74	-2,2
P5	27,97	27,45	0,13	0,05	0,85	0,83	2,6
P6	26,53	25,3	0,02	0,05	0,77	0,74	0,4
P7	25,37	23,62	-0,06	0,05	0,74	0,79	-1,2
P8	27,43	26,64	0,09	0,05	1,78	2,67	1,8
P9	28,00	27,50	0,13	0,05	1,05	0,95	2,6
P10	29,53	29,84	0,24	0,05	1,21	1,14	4,8
P11	24,10	21,88	-0,16	0,05	1,22	1,06	-3,2
P12	23,33	20,88	-0,22	0,05	1,01	0,97	-4,4
P13	23,47	21,05	-0,21	0,05	1,04	1,00	-4,2
P14	24,97	23,06	-0,09	0,05	0,83	0,82	-1,8
P15	25,40	23,67	-0,06	0,05	1,03	0,93	-1,2
P16	31,63	33,00	0,38	0,05	1,28	1,23	7,6
Ortalama	26,26	25,01	0,00	0,05	1,02	1,02	
Standart Sapma (Evren)	2,39	3,47	0,17	0,00	0,26	0,45	
Standart Sapma (Örneklem)	2,47	3,59	0,18	0,00	0,27	0,47	
Model, Evren: RMSE =0,05 Standart Sapma= 0,17 Ayırma Oranı= 3,36 Ayırma İndeksi=4,82 Güvenirlik= 0,92							
Model, Örneklem: RMSE = 0,05 Standart Sapma= 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							
Puanlayıcılar arası mutlak uyum: % 53 Puanlayıcılar arası beklenen uyum: %36,8 Puanlayıcılar arası Güvenirliğe İlişkin Kappa İstatistiği: 0,26							

Tablo 4'te bütünsel DPA'yla yapılan puanlamalara ait puanlayıcı yüzeyine ilişkin ölçüm raporları verilmiştir. Puanlayıcılara ilişkin logit ölçülerinin 0,38 ile -0,26 (0,64 logit) arasında değiştiği görülmektedir. Puanlayıcılar için uygunluk içi ve uygunluk dışı istatistikleri ortalamalarına bakıldığında, her iki ölçüm için de 1,02 olan ölçüm sonucunun 1'e oldukça yakın değerler olduğu ve verinin model ile uyumlu olduğu görülmektedir. Uygunluk istatistikleri incelendiğinde, puanlayıcılardan P8 hariç diğerlerinin kabul edilebilir aralık içinde kaldığı görülmüştür. Buna göre, P8'in model ve veri uyumunu olumsuz şekilde etkileyen puanlayıcı olarak yansıdığı belirlenmiştir.

Tablo 4'e göre güvenirlik indeksi 0,92; ayırma oranı 3,48 ve ayırma indeksi 4,98 olarak bulunmuştur. Puanlayıcı ayırma indeksinin 0,00 ve 0,00'a yakın bir değer olması beklenirken, bu değerlerin üstünde bir değer (4,98) bulunması, puanlayıcılar arası puan verme noktasında farklılıkların olduğunu, puanlayıcıların katılık- cömertlik davranışlarını gösterme düzeylerine göre farklılaştığını ve puanlayıcıların verdikleri puanlarda katılık-cömertlik hatasının olabileceğine işaret etmektedir. Puanlayıcı yüzeyine ait güvenirlik indeksinin 0,92 gibi yüksek bir değer olması, puanlayıcıların katılık-cömertlik davranışları açısından farklılık gösterdiği, bununla birlikte bu farkın anlamlılığını

yorumlamamızı sağlayan Ki-Kare değerine göre istatistiksel olarak anlamlı [χ2(sd)=198,2 (15), p=0,00] olduğu, dolayısıyla katılık ve cömertlik davranışları bakımından puanlayıcılar arasında anlamlı fark olduğu bulunmuştur.

Tablo 4'te puanlayıcı güvenirliğine dair verilen istatistiklerden, puanlayıcılar arası mutlak uyum, beklenen uyum ve güvenirlik için Kappa istatistiği değerleri incelenmiştir. Puanlayıcılar arası mutlak uyum ise %53; beklenen uyum %36,8 olarak hesaplanmıştır. Mutlak uyum yüzdesi için dikkate alınması gereken ölçütün net bir sınırı olmasa da puanlayıcılar arası mutlak uyumun en az %75 olması tavsiye edilmektedir (Graham vd., 2012). Buna göre, bütünsel DPA'ya göre yapılan puanlamalar için puanlayıcılar arası güvenirliğin düşük olduğu görülmektedir. Bütünsel DPA kullanılarak yapılmış puanlamalardan elde edilen Kappa istatistiği değeri (0,26) puanlayıcılar arası uyumun iyi olmadığını göstermektedir. Birey yüzeyine ait ölçüm raporu Tablo 5'te sunulmuştur.

Tablo 5. *Bütünsel Dereceli Puanlama Anahtarına Ait Birey Yüzeyi İçin Elde Edilen Ölçüm Raporları*

Birey	*Gözlenen Ortalama*	*Düzeltilmiş Ortalama*	*Logit Ölçüsü*	*Standart Hata*	*Uygunluk İçi*	*Uygunluk Dışı*
1	27,50	27,24	-0,12	0,05	0,58	0,57
2	38,00	39,90	0,51	0,05	1,08	1,09
3	40,75	42,59	0,71	0,06	1,38	1,35
4	27,92	27,78	-0,10	0,05	1,40	1,45
5	21,00	19,28	-0,56	0,06	0,78	0,83
6	23,33	21,99	-0,39	0,05	0,89	0,90
7	36,62	38,44	0,42	0,05	1,32	1,22
8	36,58	38,39	0,42	0,05	0,97	0,87
9	31,33	32,17	0,10	0,05	0,78	0,82
10	28,83	28,97	-0,04	0,05	1,22	1,23
11	19,67	17,85	-0,67	0,06	0,74	0,69
12	19,58	17,77	-0,68	0,06	0,91	0,88
13	19,42	17,60	-0,69	0,06	0,78	0,83
14	21,46	19,79	-0,53	0,06	1,10	1,25
15	37,17	39,02	0,46	0,05	1,25	1,39
16	22,21	20,65	-0,47	0,06	1,13	1,05
17	20,58	18,82	-0,60	0,06	0,91	0,97
18	18,25	16,45	-0,80	0,06	0,89	0,85
19	17,50	15,75	-0,88	0,07	1,40	1,46
20	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
Standart Sapma (Evren)	7,68	9,18	0,51	0,01	0,25	0,26
Standart Sapma (Örneklem)	7,88	9,42	0,52	0,01	0,25	0,27
Model, Evren: RMSE = 0,06 Standart Sapma= 0,50 Ayırma Oranı= 8,94 Ayırma İndeksi= 12,26 Güvenirlik= 0,99						
Model, Örneklem: RMSE = 0,06 Standart Sapma= 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						

Birey yüzeyine ait analiz sonuçları incelendiğinde, öğrencilerin yetenek düzeylerine ilişkin kestirimlerin 0,71 logit ile -0,88 logit (1,59 logit) arasında değiştiği görülmüştür. Bireylerin yetenek düzeylerinin ortalaması -0,24 ve standart sapması 0,52 logit, uygunluk içi istatistiğinin ortalaması 1,01 ve uygunluk dışı istatistiğinin ortalaması da 1,02 olarak bulunmuştur. Buna göre, model ve veri arasındaki uyumunun sağlandığı görülmüştür. Ayırma oranı 9,18, ayırma indeksi 12,57 ve güvenirlik indeksi 0,99 olarak hesaplanmıştır. Bu değerlerin yüksek olması, farklı yetenek düzeyindeki öğrencilerin birbirinden iyi bir şekilde ayırt edilebildiği gösterir, bununla birlikte öğrenci yetenek düzeyleri arasındaki değişkenliklerin anlamlılığı Ki-Kare değerine göre incelendiğinde, istatistiksel olarak anlamlı [χ2(sd)=1585,6 (19), p=0,00] fark bulunduğu görülmüştür. Madde yüzeyine ait ölçüm raporu Tablo 6'da sunulmuştur.

Tablo 6. *Bütünsel Dereceli Puanlama Anahtarına Ait Madde Yüzeyi İçin Elde Edilen Ölçüm Raporları*

Madde	*Gözlenen Ortalama*	*Düzeltilmiş Ortalama*	*Logit Ölçüsü*	*Standart Hata*	*Uygunluk İçi*	*Uygunluk Dışı*
1	33,5	35,08	0,48	0,04	0,59	0,66
2	27,56	27,09	0,11	0,04	0,85	0,79
3	24,88	23,49	-0,07	0,05	0,96	0,93
4	24,97	23,61	-0,06	0,05	0,73	0,69
5	17,47	15,48	-0,67	0,06	1,06	1,01
6	25,28	24,02	-0,04	0,05	0,98	0,88
7	28,81	28,82	0,19	0,04	0,6	0,57
8	23,75	22,06	-0,15	0,05	1,14	1,21
9	39,94	42,15	0,91	0,05	1,53	1,44
10	22,31	20,33	-0,25	0,05	1,58	1,77
11	22,31	20,33	-0,25	0,05	1,12	1,08
12	17,88	15,83	-0,63	0,06	1,49	1,35
13	40,5	42,66	0,96	0,05	0,82	0,91
14	18,97	16,82	-0,52	0,05	1,47	1,22
15	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
Standart Sapma (Evren)	6,80	8,26	0,47	0,00	0,32	0,32
Standart Sapma (Örneklem)	7,04	8,55	0,49	0,00	0,33	0,33
Model, Evren: RMSE = 0,05 Standart Sapma= 0,47 Ayırma Oranı= 9,64 Ayırma İndeksi=13,19 Güvenirlik= 0,99						
Model, Örneklem: RMSE = 0,05 Standart Sapma= 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						

Madde yüzeyine ait analiz sonuçları incelendiğinde, maddelerin güçlük düzeylerinin 0,96 logit ile -0,67 logit arasında (1,63 logit) değiştiği görülmektedir. Maddelerin logit değerlerine ait ortalamanın 0,00; standart sapmanın ise 0,49 logit olduğu bulunmuştur. Maddelere ilişkin uygunluk içi istatistiğinin ortalaması 1,05 ve uygunluk dışı istatistiğinin ortalaması da 1,02 olarak olduğu ve uygunluk istatistiklerinin ortalamalarının, beklenen değer olan bire çok yakın olduğu; dolayısıyla model ve veri arasındaki uyumun sağlandığı görülmüştür. Uygunluk istatistikleri için analiz sonuçları incelendiğinde model ve veri arasındaki uyumu olumsuz etkileyen herhangi bir madde bulunmadığı görülmüştür.

Analiz sonucunda madde yüzeyine ait ayırma oranı 9,64, ayırma indeksi 13,19 ve güvenirlik indeksi 0,99 olarak bulunmuştur. Elde dilen bu değerlerin yüksek olması, maddelerin güçlük düzeyleri bakımından birbirinden farklı olduğunu göstermektedir. İstatistiksel olarak maddelerin güçlük düzeyleri arasındaki farkın anlamlı olup olmadığına bakmak için Ki-Kare değeri incelendiğinde, bu değerin [χ2(sd)=1297,9 (14), p=0,00] anlamlı olduğu, maddeler arasında güçlük düzeyleri açısından istatistiksel olarak anlamlı fark olduğu söylenebilir.

3.3. Analitik Dereceli Puanlama Anahtarına Göre Yapılan Puanlamalarda Puanlayıcı Davranışları

Analitik dereceli puanlama anahtarlarına göre yapılan puanlamalarda; puanlayıcı katılığı ve cömertliği ve yanlılık davranışlarına ilişkin bulgular aşağıda verilmiştir.

3.3.1. Puanlayıcı Katılığı ve Cömertliği

Katılık ve cömertlik davranışlarını belirlemek için bireysel ve grup düzeyinde istatistiksel göstergeler incelenmiştir. Puanlayıcıların katılık ve cömertlik davranışları için grup düzeyinde bakılması gereken istatistiksel değerler ayırma oranı, ayırma indeksi, güvenirlik ve Ki-Kare değerleridir. Ayırma oranının ve güvenirlik indeksinin yüksek olması, Ki Kare değerinin de istatistiksel olarak anlamlı olmasından faydalanılarak yorumlanabilir. Puanlayıcı yüzeyine ait ölçüm raporları incelendiğinde, ayırma oranı 4,76, ayırma indeksi 6,46 ve güvenirlik indeksi 0,96 olarak bulunmuştur. Yüksek ayırma oranı ve güvenirlik indekisinin yüksek çıkması, puanlayıcıların katılık ve cömertlikleri bakımından aralarında fark bulunduğunu göstermektedir. Ayrıca sabit etkili Ki-Kare istatistiği [χ2(sd)=299,8 (15) p=0,00] incelendiğinde, bu değerler istatistiksel olarak anlamlıdır ve puanlayıcı katılığı ve cömertliği için grup düzeyinde fark vardır denilebilir. Bulunan farkın hangi puanlayıcılardan kaynaklandığı bireysel düzeydeki istatistiksel göstergeler incelenerek belirlenebilir. Puanlayıcıların değişken haritasındaki yerleri ve her puanlayıcı için elde edilen logit değerleri ve t değerlerinin istatistiksel olarak anlamlı olup olmadığı bulunur (Linacre, 2017).

Şekil 1'de verilen analitik DPA ile yapılan puanlamalara ait değişkenlik haritası ve Tablo 1 'de her bir puanlayıcı için hesaplanan t değerleri incelenip kritik t değeri (t_{kritik} = 2,131; sd=15, p=0,00$<$ 0,05) ile karşılaştırılmıştır.

Tablo 7. *Analitik Dereceli Puanlama Anahtarına Ait Katılık ve Cömertlik Davranışları Bakımından Puanlayıcılar Arasında Gözlenen Farkın Anlamlılığının t Testi Sonuçları*

Puanlayıcı	*t değeri*	*Anlamlılık*
P11	-3,2	$\lvert t_{heseplanan} \rvert > t_{kritik}$ olduğu için fark anlamlıdır. Puanlayıcılardan P11, P12 ve P13 değişkenlik haritasının negatif ucunda olduğundan, anlamlı derecede katı puanlamalar yaptıkları görülmüştür.
P12	-4,4	
P13	-4,2	
P1	-2	$\lvert t_{heseplanan} \rvert < t_{kritik}$ olduğu için fark anlamlı değildir.
P3	1	
P4	-2	
P5	0	
P6	1	
P7	1	
P8	1	
P14	-1,8	
P15	-1,2	
P2	4	$\lvert t_{heseplanan} \rvert > t_{kritik}$ olduğu için fark anlamlıdır. Puanlayıcılardan P2, P9, P10 ve P16 değişkenlik haritasının pozitif ucunda olduğundan, anlamlı derecede cömert puanlamalar yaptıkları görülmüştür.
P9	4	
P10	5	
P16	6	

Tablo 7'de analitik DPA kullanılarak yapılan puanlamaların puanlayıcıların katılık ve cömertlik davranışları bakımından puanlayıcılar arasında gözlenen farkın anlamlılığının t testi sonuçları verilmiştir. Buna göre, puanlayıcılardan P11, P12 ve P13'ün daha katı, P2, P9, P10 ve P16'nın ise daha cömert puanlama yaptığı belirlenmiştir.

3.3.2. Yanlılık

Yanlılık; puanlayıcıların, bazı öğrencilere diğerlerinden daha cömert ya da daha katı davranması olarak açıklanabilir. Puanlayıcıların yanlı davranıp davranmadığını belirleyebilmek için analiz çıktılarında birey ve puanlayıcı etkileşimlerine (birey × puanlayıcı); puanlayıcıların maddeler arasında yanlı puanlama yapıp yapmadığını belirlemek için de madde ve puanlayıcı (madde × puanlayıcı) yüzeyleri arasındaki etkileşim değerlerine bakılır. Bu çalışmada analitik DPA kullanılarak yapılan puanlamalarda, puanlayıcıların yanlı davranıp davranmadıklarını belirleyebilmek için analiz çıktılarında birey ve puanlayıcı (birey × puanlayıcı) etkileşimleri incelenmiştir.

Ayrıca puanlayıcıların yanlılık davranışlarının incelenmesinde, öncelikle grup düzeyinde Ki-Kare istatistiğinin anlamlılığı; daha sonra bireysel düzeyde puanlayıcı ve birey yüzeylerinin etkileşimindeki yanlılık için ise t istatistiği incelenir, t değerinin ±2 aralığının dışında kalması, anlamlı etkileşime işaret eder ve puanlayıcıların yanlı davrandıkları ifade edilebilir (Bond ve Fox, 2015; Linacre, 2014). Bu çalışmada grup düzeyinde Ki-Kare istatistiğinin değeri incelendiğinde Ki-Kare testinin istatistiksel olarak anlamlı olmadığı görülmüştür [χ2=286,8; sd=320 p=0,91> 0,05]. Ki- Kare testinin anlamsız çıkması sonucunda, puanlayıcıların analitik DPA ile yaptıkları puanlamalarda, grup düzeyinde yanlılık (farklılaşan puanlayıcı katılığı veya cömertliği) göstermedikleri ifade edilebilir.

Bu çalışmada 20 öğrencinin, 15 tane rutin olmayan matematik problemine verdiği cevaplar 16 puanlayıcı tarafından puanlanmıştır. Buna göre birey × puanlayıcı etkileşiminde 320 olası etkileşim vardır. Aşağıda Tablo 8'de yanlı davranışlar gösteren puanlayıcılara ait analiz sonuçları verilmiştir.

Tablo 8. *Puanlayıcı × Birey Etkileşim Tablosu (Analitik Dereceli Puanlama Anahtarı)*

Puanlayıcı	*Birey*	*Gözlenen Puan*	*Beklenen Puan*	*Yanlılık Büyüklüğü*	*Standart Hata*	*t değeri*
P8	17	513	390,81	0,15	0,03	4,47
P8	19	426	330,13	0,14	0,03	4,11
P8	12	450	355,98	0,13	0,03	3,74
P8	5	444	363,35	0,10	0,03	3,12
P8	18	373	310,78	0,10	0,04	2,73
P13	2	628	559,93	0,09	0,04	2,2
P8	13	403	338,54	0,09	0,03	2,61
P12	2	618	556,69	0,08	0,04	2,00
P14	4	487	426,38	0,07	0,03	2,09
P8	8	433	504,91	-0,08	0,03	-2,43
P8	15	431	516,22	-0,10	0,03	-2,92
P8	2	422	512,35	-0,10	0,03	-3,07
P8	7	415	508,35	-0,11	0,03	-3,15
P8	3	463	553,05	-0,11	0,03	-3,32
P12	13	178	228,18	-0,19	0,1	-2,03

Tablo 8'de sadece yanlı davranan puanlayıcılara ait analiz sonuçları verilmiştir. Veri analizinden elde edilen çıktılara göre, 320 etkileşimin 15 tanesinde puanlayıcı katılığı ve cömertliği davranışı ortaya çıkmıştır. Farklılaşan puanlayıcı katılığı ve farklılaşan puanlayıcı cömertliği davranışları birey × puanlayıcı etkileşim raporunda verilen t değerinin 2'den büyük olması durumunda farklılaşan puanlayıcı cömertliği; t değerinin -2'den küçük olması durumunda ise farklılaşan puanlayıcı katılığı olduğu söylenir.

Buna göre, analitik DPA ile yapılan puanlamalarda dokuz puanlayıcının farklılaşan puanlayıcı cömertliği, altı puanlayıcının ise farklılaşan puanlayıcı katılığı davranışı gösterdiği belirlenmiştir. Tablo 8 incelendiğinde, puanlayıcılardan P12, P13, P14 ve P8'in (11 farklı bireye) yanlılık davranışı gösterdiği belirlenmiştir.

3.4. Bütünsel Dereceli Puanlama Anahtarına Göre Yapılan Puanlamalarda Puanlayıcı Davranışları

Bütünsel DPA'ya göre yapılan puanlamalarda; puanlayıcı katılığı ve cömertliği ve yanlılık davranışlarına ilişkin bulgular aşağıda verilmiştir.

3.4.1. Puanlayıcı Katılığı ve Cömertliği

Bütünsel dereceli puanlama anahtarı ile yapılan puanlamaların katılık ve cömertlik davranışlarını belirlemek için öncelikle grup düzeyinde istatistiksel göstergeler incelenmiştir. Puanlayıcı katılık ve cömertlik davranışları için, grup düzeyinde bakılması gerekli olan istatistiksel göstergeler, ayırma oranı, ayırma indeksi, güvenirlik ve Ki-Kare değerleridir. Puanlayıcı katılığı ve cömertliği davranışları için; analiz sonuçlarından elde edilen puanlayıcı yüzüne ait ölçüm raporları incelendiğinde, ayırma oranı 3,48, ayırma indeksi 4,82 ve güvenirlik indeksi 0,92 olarak bulunmuş; ayırma oranı ve güvenirlik değerlerinin yüksek olması, puanlayıcıların katılık ve cömertlikleri bakımından aralarında fark bulunduğunu göstermektedir. Ayrıca sabit etkili Ki-Kare istatistiği [χ2=198,2; sd=15 p=0,00] incelendiğinde, istatistiksel olarak anlamlı olduğu sonucuna ulaşılmıştır.

Grup düzeyinde puanlayıcıların katılık ve cömertlik davranışları gösterdiği bulunmuştur, bireysel düzeydeki istatistiksel göstergeler incelenerek de bulunan farkın hangi puanlayıcılardan kaynaklandığı belirlenebilir. Puanlayıcıların değişken haritasında bulundaki yerler ve her puanlayıcı için

elde edilen logit ve t değerlerinin istatistiksel olarak anlamlı olup olmadığı bulunur. Aşağıda verilen Tablo 9'da, her bir puanlayıcı için hesaplanan t değerleri incelenip kritik t değeri (t_{kritik} = 2,131) ile karşılaştırılmıştır. Buna göre, puanlayıcının t değeri $|t_{heseplanan}| > t_{kritik}$ olduğunda anlamlıdır denilebilir. Bütünsel Dereceli Puanlama Anahtarına Ait Katılık ve Cömertlik Davranışları Bakımından Puanlayıcılar Arasında Gözlenen Farkın Anlamlılığının t Testi Sonuçları Tablo 9'da verilmiştir.

Tablo 9. *Bütünsel Dereceli Puanlama Anahtarına Ait Katılık ve Cömertlik Davranışları Bakımından Puanlayıcılar Arasında Gözlenen Farkın Anlamlılığının t Testi Sonuçları*

Puanlayıcı	*t değeri*	*Anlamlılık*
P1	-5,2	$\|t_{heseplanan}\| > t_{kritik}$ olduğu için fark anlamlıdır. Puanlayıcılardan P1, P4, P11, P12 ve P13 değişkenlik haritasının negatif ucunda olduğundan, anlamlı derecede katı puanlamalar yaptıkları görülmüştür.
P4	-2,2	
P11	-3,2	
P12	-4,4	
P13	-4,2	
P3	0	$\|t_{heseplanan}\| < t_{kritik}$ olduğu için fark anlamlı değildir.
P6	0,4	
P7	-1,2	
P8	1,8	
P14	-1,8	
P15	-1,2	
P2	3,6	$\|t_{heseplanan}\| > t_{kritik}$ olduğu için fark anlamlıdır. Puanlayıcılardan P2, P5, P9, P10 ve P16 değişkenlik haritasının pozitif ucunda olduğundan, anlamlı derecede cömert puanlamalar yaptıkları görülmüştür.
P5	2,6	
P9	2,6	
P10	4,8	
P16	7,6	

Tablo 9'da bütünsel DPA kullanılarak yapılan puanlamaların katılık ve cömertlik davranışları bakımından puanlayıcılar arasında gözlenen farkın anlamlılığının t- Testi sonuçları verilmiştir. Buna göre, bütünsel DPA ile puanlama yapan puanlayıcılardan P1, P4, P11, P12 ve P13'ün anlamlı derecede katı puanlamalar yaptıkları; P2, P5, P9, P10 ve P16 anlamlı derecede cömert puanlamalar yaptıkları görülmüştür.

3.4.2. Yanlılık

Puanlayıcıların, bütünsel DPA kullanarak yaptıkları puanlamalarda, yanlılık davranışı gösterip göstermediğini incelemek için puanlayıcı ve birey etkileşimleri (puanlayıcı x birey) incelenmiştir. Puanlayıcıların yanlılık davranışlarının incelenmesinde, grup düzeyinde Ki-Kare istatistiğinin anlamlılığına bakılır. Buna göre, yapılan analiz sonucunda, Ki-Kare testinin istatistiksel olarak anlamlı olmadığı görülmüştür [χ2=271,6; sd=320 p=0,98> 0,05]. Ki-Kare testinin anlamsız çıkması sonucunda, puanlayıcıların bütünsel dereceli puanlama anahtarı ile yaptıkları puanlamalarda, grup düzeyinde yanlılık davranışı (farklılaşan puanlayıcı katılığı veya cömertliği) göstermedikleri ifade edilebilir.

Bireysel düzeyde ise puanlayıcı ve birey yüzeylerinin etkileşimindeki yanlılık olup olmadığını belirlemek için t değerlerinin ±2 aralığının dışında kalıp kalmadığı incelenir. t değerinin ±2 aralığının dışında kalması, anlamlı etkileşime işaret eder ve puanlayıcıların yanlı davrandıkları ifade edilebilir (Bond & Fox, 2015; Linacre, 2014).

Bu çalışmada veriler 20 öğrencinin, 15 tane rutin olmayan matematik problemine verdiği cevapların 16 puanlayıcı tarafından puanlanması ile elde edilmiştir. Buna göre birey × puanlayıcı

etkileşiminde 320 olası etkileşim vardır. Aşağıda Tablo 10'da, sadece yanlı davranışlar gösteren puanlayıcılara ait analiz sonuçları verilmiştir.

Tablo 10. *Puanlayıcı × Birey Etkileşim Tablosu (Bütünsel Dereceli Puanlama Anahtarı)*

Puanlayıcı	*Birey*	*Gözlenen Puan*	*Beklenen Puan*	*Yanlılık Büyüklüğü*	*Standart Hata*	*t değeri*
P8	17	530	509,34	0,84	0,2	4,22
P8	19	430	414,34	0,74	0,2	3,76
P8	13	440	426,45	0,59	0,2	3,04
P8	5	460	447,01	0,54	0,19	2,77
P8	18	390	379,56	0,5	0,2	2,49
P8	12	420	408,72	0,5	0,2	2,54
P8	9	380	391,36	-0,44	0,2	-2,16
P8	7	440	453,24	-0,52	0,2	-2,67
P8	8	430	444,18	-0,56	0,2	-2,84
P8	15	430	445,02	-0,6	0,2	-3,03
P8	2	430	446,21	-0,65	0,2	-3,31
P8	3	480	495,04	-0,66	0,19	-3,39

Veri analizinden elde edilen çıktılar incelendiğinde 320 olası etkileşimin 12 (%3,8) tanesinde yanlılık davranışı ortaya çıkmıştır.

Farklılaşan puanlayıcı katılığı ve farklılaşan puanlayıcı cömertliği davranışları birey × puanlayıcı etkileşim raporunda verilen t değerine göre yorumlanmaktadır. t değeri >2 ise farklılaşan puanlayıcı cömertliği; t değeri < -2 ise farklılaşan puanlayıcı katılığı vardır şeklinde yorum yapılır. Buna göre, bütünsel dereceli puanlama anahtarıyla yapılan puanlamalarda, birey × puanlayıcı etkileşim raporundan elde edilen t değerleri incelendiğinde, P8'in 12 farklı yanlılık davranışı gösterdiği belirlenmiştir. Bunlardan altısının farklılaşan puanlayıcı cömertliği, altısının ise farklılaşan puanlayıcı katılığı davranışı gösterdiği belirlenmiştir.

3.5. Analitik ve Bütünsel Dereceli Puanlama Anahtarlarına Göre Yapılan Puanlamaların Karşılaştırılması

Aşağıda Tablo 11'de *dereceli puanlama anahtarı türüne* birey, madde ve puanlayıcı yüzeyleri ve puanlayıcı davranışlarına ait genel bilgiler sunulmuştur. Tablo 11'de, öğrenci, madde ve puanlayıcı yüzeylerine ait sonuçlar incelendiğinde her iki dereceli puanlama anahtarı için de sonuçların birbirine benzer olduğu görülmüştür. Bununla birlikte, bütünsel DPA ile yapılan puanlamalarda, katılık ve cömertlik davranışı gösteren puanlayıcıların daha fazla, analitik DPA ile yapılan puanlamalarda ise, yanlılık davranışı gösteren puanlayıcıların daha fazla olduğu görülmüştür. Ayrıca, katılık ve cömertlik davranışı gösteren puanlayıcıların her iki DPA türünde de benzer olduğu görülmüştür. Her iki DPA ile yapılan puanlamada da puanlayıcılardan P2, P9, P10 ve P16'nin cömert; P11, P12 ve P13'ün de katı davrandığı görülmüştür. Puanlayıcılardan P8'in tüm DPA'larda yanlı davrandığı görülmüştür.

Tablo 11. *Dereceli Puanlama Anahtarları Türüne Göre Öğrenci, Madde, Puanlayıcı Yüzeyleri ve Puanlayıcı Davranışlarına Ait Özet Tablosu*

Yüzey ve Davranışlar		*Analitik DPA*	*Bütünsel DPA*
Öğrenci	En Yüksek Yetenek	3	3
	En Düşük Yetenek	18, 19, 20	19, 20
Puanlayıcı	En Katı	11	11
	En Cömert	16	16
Madde	En Zor	5	5
	En Kolay	13	13
Puanlayıcı Davranışları	Katılık	P11, P12, P13	P1, P4, P11, P12, P13
	Cömertlik	P2, P9, P10, P16	P2, P5, P9, P10, P16
	Farklılaşan Katılık (Yanlılık)	P8, P14, P12	P8
	Farklılaşan Cömertlik (Yanlılık)	P8, P13, P12	P8

4. Sonuç, Tartışma ve Öneriler

Bu çalışma, rutin olmayan matematik problemlerinin puanlanmasında analitik ve bütünsel DPA kullanımının puanlayıcı davranışlarına etkisini incelemek amacıyla yapılmıştır. Puanlayıcı davranışları için yapılan analizler için çok yüzeyli Rasch ölçme modeli kullanılmıştır. Araştırma kapsamında, her iki dereceli puanlama anahtarı türü için sonuçlar verilmiştir.

4.1.Analitik Dereceli Puanlama Anahtarı Kullanılarak Yapılan Puanlamalardan Elde Edilen Verilerin Analizi Sonuçları

Analitik DPA'ya göre yapılan puanlamalarda, puanlayıcı, birey ve madde yüzeylerine ait hesaplanan güvenirlik değerleri ve uygunluk istatistikleri incelenmiştir. Buna göre; birey, madde ve puanlayıcı yüzeylerinin her birinin model veri uyumunun iyi olduğu bulunmuştur. Bireylerin istatistiksel açıdan anlamlı bir şekilde birbirlerinden ayrıldığı, madde güçlükleri arasında istatistiksel açıdan anlamlı fark olduğu ve puanlayıcıların puanlama davranışları açısından birbirlerinden istatistiksel olarak anlamlı bir şekilde farklı davrandıkları belirlenmiştir.

Puanlayıcı katılığı ve cömertliği davranışları incelendiğinde, grup düzeyinde katılık ve cömertlik davranışları gözlenmemiştir. Bireysel düzeyde ise, puanlayıcılardan P11, P12 ve P13'ün daha katı, P2, P9, P10 ve P16'nın ise daha cömert puanlama yaptığı belirlenmiştir.

Puanlayıcıların yanlılık davranışları incelendiğinde, grup düzeyinde yanlılık davranışı gözlenmemiştir. Bireysel düzeyde ise, dört farklı puanlayıcıda (P8, P12, P13 ve P14)yanlılık davranışı gözlenmiştir. Yanlılık davranışının belirlendiği dört puanlayıcının puanlayıcı x birey etkileşimlerinden 15 tanesinin istatistiksel olarak anlamlı olduğu, buna göre puanlayıcılardan P8, P12, P13 ve P14 tarafından yapılan puanlamalardan dokuzunun cömert, puanlayıcılardan P8, P12'nin yaptığı puanlamalardan altısının ise katı davrandığı görülmüştür.

4.2. Bütünsel Dereceli Puanlama Anahtarı Kullanılarak Yapılan Puanlamalardan Elde Edilen Verilerin Analizi Sonuçları

Bütünsel DPA'ya göre yapılan puanlamalarda, puanlayıcı, birey ve madde yüzeylerine ait hesaplanan güvenirlik değerleri ve uygunluk istatistikleri incelenmiştir. Buna göre; birey, madde ve puanlayıcı yüzeylerinin her birinin model veri uyumunun iyi olduğu bulunmuştur. Bireylerin istatistiksel açıdan anlamlı bir şekilde birbirlerinden ayrıldığı, madde güçlükleri arasında istatistiksel açıdan anlamlı fark olduğu ve puanlayıcıların puanlama davranışları açısından birbirlerinden istatistiksel olarak anlamlı bir şekilde ayrıştığı belirlenmiştir.

Puanlayıcı katılığı ve cömertliği davranışları incelendiğinde, grup düzeyinde katılık ve cömertlik davranışları gözlenmemiştir. Bireysel düzeyde ise, puanlayıcılardan P1, P4, P11, P12 ve P13'ün anlamlı derecede katı puanlamalar yaptıkları; P2, P5, P9, P10 ve P16 anlamlı derecede cömert puanlamalar yaptıkları görülmüştür.

Puanlayıcıların yanlılık davranışları incelendiğinde, grup düzeyinde yanlılık davranışı gözlenmemiştir. Bireysel düzeyde ise, yalnızca bir puanlayıcıda (P8) yanlılık davranışı gözlenmiştir. Yanlılık davranışının belirlendiği puanlayıcının puanlayıcı x birey etkileşimlerinden 12 tanesinin istatistiksel olarak anlamlı olduğu, buna göre bu puanlayıcının altı puanlamada cömert davrandığı, altı puanlamada ise katı davrandığı görülmüştür.

Rutin olmayan matematik problemleri çözümlerinin puanlanmasına yönelik hazırlanmış olan analitik DPA ile yapılan puanlamanın bütünsel DPA ile yapılan puanlamaya göre, daha güvenilir sonuçlar verdiği belirlenmiştir. Literatürde, Klasik test kuramında analitik DPA ile yapılan puanlamalardan elde edilen puanların, bütünsel DPA'dan elde edilmiş puanlara göre az da olsa daha yüksek güvenirlikler verdiğini ifade eden birçok çalışma, elde edilen bu sonuçları desteklemektedir (Boring, 2002; Jonsson & Svingby, 2007; Svingby, 2007). Bunun yanında, Kutlu ve arkadaşlarının (2009)'nin, analitik DPA'dan elde edilen sonuçların, bütünsel DPA'dan elde edilen sonuçlara göre güvenirlik düzeylerinin daha yüksek olduğu yönündeki açıklamalarıyla da uyumludur.

Her iki DPA ile yapılan puanlamalar sonucunda elde edilen puanlayıcı yüzeyleri karşılaştırıldığında, analitik DPA kullanılan puanlamalarda puanlayıcı katılık ve cömertlik düzeylerinin, bütünsel DPA kullanılan puanlamalardan daha fazla olduğu belirlenmiştir. Çok yüzeyli Rasch ölçme modeli analizlerine göre, analitik DPA kullanılan puanlamalar arasındaki uyumun, bütünsel DPA kullanılan puanlamalardan daha kötü olduğu sonucuna varılmıştır. Bu durum, analitik DPA'da ölçülecek özellik daha ayrıntılı olarak incelendiği için daha objektif sonuçlar sağlasa da, bütünsel DPA'da ölçülecek özellik bir bütün olarak ele alındığı için puanlayıcıların benzer şekilde puanlama yapmasına sebep olabilir.

Alanyazında, matematik dersi için açık uçlu rutin olmayan maddelerin puanlanmasında, bütünsel ve analitik puanlama anahtarlarının kullanılmasının ve puanlayıcı etkilerinin analiz edildiği bir çalışmaya rastlanmadığı dikkate alındığında, uygulamaya yönelik sonuçlarının yanında, araştırmanın matematik eğitimi alanında ve ölçme ve değerlendirme alanında literatüre de katkısı olacağı düşünülmektedir.

Araştırma sonucu elde edilen veriler incelendiğinde, analitik DPA kullanılarak yapılan puanlamalarda birey, madde ve puanlayıcı yüzeylerinin, bütünsel DPA kullanılarak yapılan puanlamalara göre daha benzer olduğu sonucuna ulaşılmıştır. Analitik DPA kullanımının puanlayıcılar arasındaki farklılıkları yok etmese de puanlamalar arasındaki tutarlılığı artırarak objektiflik düzeyini artırdığına dair elde edilen bu sonuçlar, literatürde yer alan görüşlerle de desteklenmektedir (Bıkmaz Bilgen ve Doğan, 2017; Büyükkıdık, 2012; Şanlı, 2010). Her iki dereceli puanlama anahtarının kullanımında puanlayıcılarda grup düzeyinde olmasa da bireysel düzeyde merkeze eğilim, halo etkisi ve yanlılık davranışları ortaya çıktığı görülmüştür. Bu durum da Esfandiari (2015), Esfandiari (2021), Jones & Bergin (2019), Linlin (2020), Yılmaz (2017)'ın yapmış oldukları çalışmaları desteklemektedir.

Bu araştırmada puanlayıcılara puanlayıcı eğitimi verilmiştir, puanlayıcı eğitimi verilerek ve verilmeden, deney-kontrol grubu oluşturularak benzer çalışmalar yapılabilir. Ayrıca yapılacak yeni çalışmalarda, puanlayıcı olarak öz ve akran değerlendirmeleri de eklenebilir. Bu araştırmada DPA türüne göre bazı puanlayıcı davranışlarının farklılaştığı, bazılarının da değişmediği tespit edilmiştir; bu puanlayıcıların özelliklerinin belirlenmesine yönelik çalışmalar yapılabilir.

Kaynaklar

Akgün, M. (2016). Yüksek öğretimde ideal öğretim elemanı nasıl olmalıdır? *Kaygı Uludağ Üniversitesi Fen-Edebiyat Fakültesi Felsefe Dergisi*, *26*, 197-204. https://doi.org/10.20981/kuufefd.97116

Alharby, E.R. (2006). *A comparison between two scoring methods, holistic vs. analytic using two measurement models, generalizability theory and the many facet Rasch measurement within the context of performance assessment.* [Doctoral Dissertation, Pennsylvania State University].

Alkan, H. (1999). Matematikte ölçme ve değerlendirme, A. Özdaş (Ed.), *Matematik öğretimi* içinde (ss. 93-110). Anadolu Üniversitesi Açık Öğretim Fakültesi.

Altun, M. (2014). *Ortaokullarda matematik öğretimi*. Aktüel.

Altun, M. (2020). *Matematik okuryazarlığı el kitabı-yeni nesil soru yazma ve öğretim düzenleme teknikleri.* Aktüel.

Anastasi, A. (1988). *Psychological testing*. Macmillan.

Archbald, D.A., & Grant, T.J. (2000) What's on the test? An analytical framework and findings from an examination of teachers' math tests. *Educational Assessment*, *6*(4), 221-256. http://dx.doi.org/10.1207/S15326977EA0604_2

Arıkan, S. (2023). *Rutin olmayan problemler nedir ve nasıl hazırlanır?* Kanguru Matematik Derneği Yayınevi.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance.* Corwin Press, Inc.

Aypay, A. (Ed.). (2015). *Araştırma yöntemleri desen ve analiz.* Anı.

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

Badia, G. (2019). Holistic or analytic rubrics? Grading information literacy instruction. *College & Undergraduate Libraries*, *26*(2), 109–116. https://doi.org/10.1080/10691316.2019.1638081

Bağcan Büyükturan, E. & Çıkrıkçı Demirtaşlı, N. (2013). Çoktan seçmeli testler ile yapılandırılmış gridlerin psikometrik özellikleri bakımından karşılaştırılması. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *46*(1), 395-415.

Bargainnier, S. (2003). *Fundamentals of rubrics*.
http://www.webpages.uidaho.edu/ele/scholars/Practices/Evaluating_Projects/Resources/Using_Rubrics.pdf

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. ÖSYM.

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

Bloom, W. & Niss, M. (1991). Applied mathematical problem solving, modelling, applications and links to other subjects. *Educational Studies in Mathematics*, 22, 37-68. https://doi.org/10.1007/BF00302716

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

Brookhart, S. M. (2013). *Dereceli puanlama anahtarlarının hazırlanması ve uygulanması*. (İ. Karakaya, & F.N., Fişne, Çev.). Pegem A Yayıncılık.

Chukwuere, J.E. (2021). The comparisons between the use of analytic and holistic rubrics in information systems discipline. *Academia Letters, Article* 3579. https://doi.org/10.20935/AL3579.

Crocker, L. M., & Algina, L. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

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

Ebel, R.L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*(4), 407- 424. https://psycnet.apa.org/doi/10.1007/BF02288803

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

Gürbüz, R., & Güder, Y. (2016). Matematik öğretmenlerinin problem çözmede kullandıkları stratejiler. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi (KEFAD) 17*(2), 371-386.

Güler, N. (2008). *Klasik test kuramı, genellenebilirlik kuramı ve Rasch modeli üzerine bir araştırma.* (Yayın No. 257569) [Doktora Tezi, Hacettepe Üniversitesi]. YÖK. https://tez.yok.gov.tr

Güler, N. & Taşdelen Teker, G. (2015). Açık uçlu maddelerde farklı yaklaşımlarla elde edilen puanlayıcılar arası güvenirliğin değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(1), 12-24. https://doi.org/10.21031/epod.63041

Gölcür, Z. (2022). *Çok yüzeyli Rasch modeli puanlama desenlerine göre açık uçlu maddelerin puanlayıcılar arası güvenirliklerinin karşılaştırılması*. (Yayın No. 708236) [Yüksek Lisans Tezi, Gazi Üniversitesi]. YÖK. https://tez.yok.gov.tr

Görgülü Öztürk, B. (2023). *Açık uçlu maddelerinin puanlanmasında öğrencilere verilen puanlayıcı eğitiminin puanlayıcı davranışlarına etkisinin çok yüzeyli Rasch ölçme modeli ile incelenmesi.* (Yayın No. 839567) [Yüksek Lisans Tezi, Gazi Üniversitesi]. YÖK. https://tez.yok.gov.tr

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

Haladyna, T. M. & Rodriguez, M. C. (2013) Developing and validating test items. Routledge. https://doi.org/10.4324/9780203850381

Hoyt, W.T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5(*1), 64-86. https://psycnet.apa.org/doi/10.1037/1082-989X.5.1.64

Işık, C. & Kar T. (2012). Sınıf öğretmeni adaylarının problem kurma becerileri. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 23*(1), 190-214.

İlhan, M. (2015). *Standart ve Solo taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi.* (Yayın No. 398394) [Doktora Tezi, Gaziantep Üniversitesi]. YÖK. https://tez.yok.gov.tr

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

Jönsson, 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

Jönsson, A., Balan, A., & Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assessment in Education: Principles, Policy & Practice*, *28*(3), 212–227. https://doi.org/10.1080/0969594X.2021.1884041

Kantrov, I. (2000). Assessing students' mathematics learning. *Issues in Mathematics Education* (pp. 1-11). Educational Development Center.

Karakaya, İ. (Ed.). (2022). *Açık uçlu soruların hazırlanması, uygulanması ve değerlendirilmesi*. Pegem Akademi.

Karakaya, İ. & Fişne, F. N. (Ed.). (2023). *Dereceli puanlama anahtarlarının hazırlanması ve uygulanması*. Pegem A Yayıncılık.

Karasar, N. (2020). *Bilimsel araştırma yöntemi*. Nobel.

Klein, S.P., Stecher, B.M., Shavelson, R.J., McCaffrey, D., Ormseth, T., Bell, R.M., Comfort, K., & Othman, A.R. (1998) Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11*(2), 121- 137. DOI: 10.1207/s15324818ame1102_1

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

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

Kutlu, Ö., Doğan, C. D., & Karakaya, İ. (2014). *Performansa ve portfolyoya dayalı durum belirleme*. Pegem.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563-575.

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

Mertler, C. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation, 7*(25). http://pareonline.net/getvn.asp?v=7&n=25

Milli Eğitim Bakanlığı (2015). *Akademik becerilerin izlenmesi ve değerlendirilmesi (ABİDE) projesi.* http://abide.meb.gov.tr/proje-hakkinda.asp

Moore, B.B. (2009). *Consideration of rater effects and rater design via signal detection theory*. [Unpublished Doctoral Dissertation]. Columbia University.

Moskal, B.M. (2000). Scoring rubrics: What, when, how? *Practical Assessment, Research and Evaluation, 7*(3). http://pareonline.net/getvn.asp?v=7&n=3

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*, 45-85.

Nelson, N.W., & Van Meter, A.M. (2007). Measuring written language ability in narrative samples. *Reading & Writing Quarterly, 23*(3), 287-309. https://psycnet.apa.org/doi/10.1080/10573560701277807

Nitko, A.J. (2004). *Educational assessment of students*. Pearson.

Petkov, D., & Petkova, O. (2006). Development of scoring rubrics for IS projects as an assessment tool. *Issues in Informing Science and Information Technology, 3*, 499-510. DOI:10.28945/910

Polya, G. (1973). *How to solve it: A new aspect of mathematical method* (2nd ed.). Princeton University Press.

Popham, W.J. (1997). What's wrong-and what's right-with rubrics. *Educational Leadership, 55*(2), 72-75.

Reddy, M.Y. (2010). Design and development of rubrics to improve assessment outcomes. A pilot study in a master's level business program in India. *Quality Assurance in Education, 19*(1), 84-104.

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

Saal, F. E., Downey, R. G. & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413-428. https://doi.org/10. 1037/0033-2909.88.2.413

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Eds.), *Handbook of Experimental Psychology* (22). John Wiley and Sons.

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

Şahin, M. G., Güler, N., & Taşdelen Teker, G. (2016). An analysis of peer assessment through many facets of the Rasch model. *Journal of Education and Practice, 7*(32), 172-181.

Şata, M. (2019). *Performans değerlendirme sürecinde puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisinin incelenmesi.* (Yayın No. 626117) [Doktora Tezi, Gazi Üniversitesi]. YÖK. https://tez.yok.gov.tr

Şata, M. & Karakaya, İ. (2020). Investigation of the use of electronic portfolios in the determination of student achievement in higher education using many-facet Rasch measurement model. *Educational Policy Analysis and Strategic Research, 15(1),* 7-21. https://doi.org/10.29329/epasr.2020.236.1

Thorndike, R. M. & Thorndike-Christ, T. (2017). *Psikolojide ve eğitimde ölçme ve değerlendirme* (M. Otrar, Çev. Ed.). Nobel.

Tobaş, C. (2020). *Performansın değerlendirilmesinde farklılaşan puanlayıcı davranışlarının çok yüzeyli Rasch ölçme modeli ile incelenmesi.* (Yayın No. 629923) [Yüksek Lisans Tezi, Gazi Üniversitesi]. YÖK. https://tez.yok.gov.tr

Turgut, M.F. & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme*. Pegem Akademi.

Uluman, M. (2015). *Çok değişkenlik kaynaklı Rasch ölçme modeli ve hiyerarşik puanlayıcı modeli ile kestirilen parametrelerin karşılaştırılması.* (Yayın No. 396151) [Doktora Tezi, Ankara Üniversitesi]. YÖK. https://tez.yok.gov.tr

Weigle, S.C. (2002). *Assessing writing*. Cambridge University.

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

Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *The Journal of Effective Teaching, 7*(1), 3-14.

Wolfe, E.W., & McVay, A. (2010). *Rater effects as a function of rater training context*. http://www.pearsonassessments.com/NR/rdonlyres/6435A0AF-0C12-46F7-812E-908CBB7ADDFF/0/RaterEffects_101510.pdf

Woodward, J., Monroe, K., & Baxter, J. (2001). Enhancing student achievement on performance assessments in mathematics. *Learning Disability Quarterly, 24*(1), 33-46. https://doi.org/10.2307/1511294

Extended Abstract

Introduction

In general, developments at social and academic levels occur through teaching in education. Therefore, it is important to plan and evaluate the teaching (Akgün, 2016). Measurement and evaluation are used in the process of determining the preliminary knowledge of individuals, determining and measuring the effectiveness of the methods and techniques used in the training process, finding and eliminating the deficiencies in learning and their causes during the implementation of the training program, and checking whether or not the training program has achieved its goals or to what extent. In order to make an accurate assessment of the measured quality, appropriate criteria must be used, and the measurement results must be valid and reliable. Since classical measurement and evaluation methods are inadequate in measuring and evaluating high-level mental skills, open-ended items and approaches based on performance evaluation have gained importance.

In mathematics class, the concept of a problem becomes important for open-ended questions. A problem is a situation with some open issues that attract the attention of the person who has encountered the problem and does not have sufficient knowledge of algorithms and solution methods to solve this problem (Bloom & Niss, 1991). Altun (2014) divided the problems into routine and non-routine problems. Non-routine problems can not be solved with previously learned formulas or methods. It can be defined as problems that require solutions where the method is not visible and where it is necessary to have some skills such as organizing data, classifying and seeing relationships rather than mathematical operation skills, and performing interrelated actions in a planned manner, one after the other (Altun, 2014).

The rater effect can be defined as the error caused by raters and the systematic behavior of raters on an individual's performance evaluation scores, regardless of the measured structure (Bachman, 2004; Eckes, 2005; Hoyt, 2000). Rater rigidity and leniency can be defined as the rater's tendency to consistently give higher or lower scores than other raters or established scoring criteria. Bias is defined as the tendency of the rater to give higher or lower scores than they should during the evaluation process, depending on various characteristics of individuals, such as gender, age or other cultural factors (Kumar, 2005).

Rubrics are of great importance in evaluating mathematics achievement as they facilitate students' better understanding of concepts and skills in their post-evaluation tasks through effective and accurate feedback (Lau, Kuehl & Sofronas, 2015). Brookhart (2023, Trans. Karakaya & Fişne) defined rubrics as a consistent set of criteria that determine the levels of student performance qualities included in the criteria. Holistic rubrics are where the individual's performance or the product is evaluated as a whole and evaluated with a single score, where the quality of performances at different levels is revealed. Analytical rubrics, on the other hand, are scoring tools that indicate the levels, strengths and weaknesses of the individual's performance in different criteria (Kutlu et al., 2014).

Method

This research utilized a survey model, which is a descriptive research method. The participants of this study comprised two different groups: 20 students who were administered an achievement test consisting of non-routine mathematical problems and 16 raters who evaluated the completed achievement tests.

Research data was collected using a mathematics achievement test containing non-routine items, analytical and holistic rubrics and rater forms. These rubrics were employed in scoring the students' responses to mathematical problems based on mathematical competencies outlined by Altun (2020) as necessary for mathematical literacy. A mathematics test comprising non-routine problems was administered to 90 students across two sessions, each with five questions, on two different days. Then, 20 students were selected who answered all the questions. Before the scoring process, raters underwent training covering the study process, scoring keys, measurement errors relevant to scoring, both during and involved in the process, and analytical and holistic scoring criteria. Students were assigned unique codes (S1, S2, ..., S20) for each question for every student. The obtained data were analyzed according to the many-facet Rasch model.

Results

Scoring With The Analytical Rubric

Sudweeks, Reeve, and Bradshaw (2004) stated that fit statistics above two are detrimental to measurement. Accordingly, it is seen that the in and out-of-fit statistics for the individual, item and rater surfaces are very close to 1, and according to the obtained values, the data is compatible with the model. It was observed that no items, individuals or raters negatively affected the model and data compatibility.

The statistics related to individual, item, and rater surfaces, separation rate, separation index, reliability index, and chi-square values, which show the significance of these statistics, were examined. Accordingly, there was a difference in terms of difficulty levels of the items for the item surface (separation ratio 10.08, separation 13.78 and reliability 0.99 and χ2(df)=1359.4 (14), p=0,00). Rater surface (reliability 0.96; separation rate 4.76 and separation 6.67 and χ2(df)=299.8 (15) p=0.00) when examined, it was seen that there were differences in the scoring point between the raters. It was found that the statistical values of the individual surface are high (separation rate 9.39, discrimination index 12.86 and the reliability index 0.99 and χ2(df)=1702.0, (19), p=0.00)) indicating that different ability levels can be well separated from each other.

Scoring With The Holistic Rubric

The fit and non-fit statistics for individual, item, and rater surfaces are very close to 1, indicating compatibility with the model according to the obtained values. No substance or individual negatively affects model and data compatibility. However, R8 was reflected as negatively impacting the model and data fit among raters.

The statistics related to individual, item and rater surfaces, separation rate, separation index reliability index and chi-square values showing the significance of these statistics were examined. Accordingly, there was a difference in terms of difficulty levels of the items for the item surface (separation ratio 9.64, separation index 13.19 and reliability index 0.99 and χ2(df)=1297.9 (14), p=0.00). Rater surface (reliability index 0.92, separation rate 3.48 and separation index 4.98 and χ2(df)=198.2 (15) p=0.00) when examined, it was seen that there were differences between the raters in the scoring point. It was found that the statistical values of the individual surface are high (separation rate 9.18, separation index 12.57 and reliability index 0.99 and χ2(sd)=1585,6 (19), p=0,00) indicating that different ability levels can be well separated from each other.

Raters' Behaviors

When examining student, item, and rater facets, consistent results were seen across both rubrics. Regarding individual surface examination, student S3 demonstrated the highest talent, while students S18, S19, and S20 displayed the lowest. Rater R11 was identified as the strictest, while R16 was deemed the most generous. When the analysis results for the item surface were examined, the most difficult item was 5, and the easiest item was 13.

In the scoring according to analytical rubrics, findings regarding rater severity, leniency, and bias behaviors are given below. Statistical indicators at the individual and group levels were examined to determine severity and leniency behaviors. Group-level examination for the severity and leniency behaviors of the raters included the separation ratio, separation index, reliability and Chi-Square values, and the high separation rate and reliability index can be interpreted by taking advantage of the statistically significant Chi-Square value. When the analyses of the rater surface were examined, the discrimination rate was found to be 4.76, the discrimination index was 6.46 and the reliability index was 0.96. The high separation rate and reliability index show a difference between the raters in terms of their severity and leniency. In addition, fixed effect Chi-Square statistics [χ2(df)=299.8 (15) p=0.00] when examined, these values are statistically significant and it can be said that there is a difference at the group level for rater severity and leniency. The difference can be determined from which raters the difference is caused by examining statistical indicators at the individual level. The t values calculated for each rater were analysed and compared with the critical t value (t_(critical) = 2.131; df=15, p=0.00<0.05). Accordingly, it was determined that raters R11, R12 and R13 scored more strictly, while R2, R9, R10 and R16 scored more generously.

Bias can be explained as raters treating some students more generously or harshly than others. In order to determine whether the raters were biased, the interactions of the individual and the raters (individual × rater) were examined in the analysis outputs. In examining rater bias, firstly the group-level significance of the Chi-Square statistic. Then, t statistics are examined for bias in the interaction of rater and individual faces at the individual level. If the t value is outside the ±2 range, it indicates a significant interaction, and it can be stated that the raters are biased (Bond & Fox, 2015; Linacre, 2014). Differentiating rater severity and varying rater generosity behaviors; if the t value given in the individual × rater interaction report is greater than 2, different rater generosity; If the t value is less than -2, it is said to have varying rater rigidity. When the Chi-Square value was examined at the group level, it was seen that the value was not statistically significant and the raters did not show bias at the group level [χ2(df)=286.8 (320) p=0.91> 0.05].

In this study, 20 students were scored by 16 raters, resulting in 320 possible individual × rater interactions. Data analysis revealed rater severity and leniency behaviors in 15 of these interactions. It was determined that raters R12, R13, R14 and R8 showed biased behavior, including nine raters with varying rater leniency and six with varying rater severity.

In the ratings made according to holistic DPA, findings regarding rater severity, leniency, and bias behaviors are given below. Statistical indicators at both individual and group levels were examined to determine severity and leniency behaviors. Rater rigor and generosity for their behavior; When the measurement reports of the rater's face obtained from the analysis results were examined, it was seen that the separation rate and reliability values were high, there was a difference between the raters in terms of their severity and leniency, and the Chi-Square statistic [χ2(df)=198.2 (15), p=0.00]. Individual-level statistical indicators and t values calculated for each rater were examined and compared with the critical t value. Accordingly, among the raters who scored with holistic rubrics, R1, R4, R11, R12 and P13 made significantly stricter ratings; R2, R5, R9, R10 and R16 were found to give

significantly generous ratings. The Chi-Square statistic[χ2(df)=271.6 (320), p=0.98> 0.05] for rater bias behavior was not significant at the group level was observed, indicating no bias among raters at this group level. When the t values obtained from the individual × rater interaction report were examined, it was determined that R8 showed 12 different bias behaviors.

When rater behaviors are examined, in the ratings made with holistic rubric, R1, R4, R11, R12, R13 are strictness, R2, R5, R9, R10, R16 are generosity, R8 is both Differential Strictness and Differentiated Generosity). In the scores made with analytical rubric, R11, R12, R13 are strictness and R2, R9, R10, R16 are generosity and R8, R14, R12 are differentiated strictness, R8, R13, R12 are differentiated generosity has been observed to exhibit this behavior.

Conclusion, Discussion and Recommendations

This study was conducted to examine the effect of using analytical and holistic rubrics on rater behavior in scoring non-routine mathematics problems. The many-facet Rasch measurement model was used for the analysis of rater behavior. Within the scope of the research, upon examining the results for both types of rubrics, it was determined that scoring with analytical rubric gave more reliable results than scoring with holistic rubric. Numerous studies in the literature support these results, stating that the scores obtained from the analytical rubric in classical test theory give slightly higher reliability than the scores obtained from the holistic rubric (Boring, 2002; Jonsson & Svingby, 2007). In addition, it is compatible with Kutlu et al.'s (2009) statements that the reliability levels of the results obtained from analytic rubric are higher than the results obtained from holistic rubric.

Yayın Etiği Beyanı

Bu araştırmanın, Gazi Üniversitesi Etik Komisyonu tarafından 17.10.2023 tarihinde yapılan 18 sayılı toplantısında alınan karar ile etik kurul onayı bulunmaktadır. Araştırmanın planlanması, uygulanması, verilerin toplanması ve verilerin analizi süreçlerinde "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" nde uyulması belirtilen tüm kurallara uyulmuştur. Yönergenin ikinci bölümü olan "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbiri gerçekleştirilmemiştir. Bu araştırmanın yazım sürecinde bilimsel, etik ve alıntı kurallarına uyulmuş; toplanan veriler üzerinde herhangi bir tahrifat yapılmamıştır. Bu çalışma herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiştir.

Araştırmacıların Katkı Oranı Beyanı

Birinci yazar %50 (araştırmanın tasarlanması, araştırmanın uygulanması, geçerlik ve güvenirlik çalışmaları, veri analizi, raporlaştırma) ve ikinci yazar %50 (araştırmanın tasarlanması, yöntemin belirlenmesi, danışmanlık) oranında katkı sağlamıştır.

Çatışma Beyanı

Araştırmanın yazarları olarak makalenin herhangi bir aşamasında maddi veya manevi çıkar sağlamadığımızı ifade ederiz