
GAZİ ÜNİVERSİTESİ
Gazi
1926

ÇOK YÜZEYLİ RASCH MODELİ PUANLAMA DESENLERİNE GÖRE AÇIK UÇLU MADDELERİN PUANLAYICILAR ARASI GÜVENİRLİKLERİNİN KARŞILAŞTIRILMASI

Zeynep Gölcür

YÜKSEK LİSANS TEZİ

EĞİTİM BİLİMLERİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

OCAK, 2022

TELİF HAKKI ve TEZ FOTOKOPİ İZİN FORMU

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

YAZARIN

Adı: Zeynep

Soyadı: GÖLCÜR

Bölümü: Eğitimde Ölçme ve Değerlendirme

İmza:

Teslim Tarihi:

TEZİN

Türkçe Adı: Çok Yüzeyli Rasch Modeli Puanlama Desenlerine Göre Açık Uçlu Maddelerin Puanlayıcılar Arası Güvenirliklerinin Karşılaştırılması

İngilizce Adı: Comparison of the Inter-rater Reliability of Open-Ended Items According to the Many Facet Rasch Model Rating Designs

ETİK İLKELERE UYGUNLUK BEYANI

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

Yazar Adı Soyadı: Zeynep GÖLCÜR

İmza: ……..

JÜRİ ONAY SAYFASI

Zeynep GÖLCÜR tarafından hazırlanan "Çok Yüzeyli Rasch Modeli Puanlama Desenlerine Göre Açık Uçlu Maddelerin Puanlayıcılar Arası Güvenirliklerinin Karşılaştırılması" adlı tez çalışması aşağıdaki jüri tarafından oy birliği/oy çokluğu ile Gazi Üniversitesi Eğitim Bilimleri Ana Bilim Dalı'nda Yüksek Lisans tezi olarak kabul edilmiştir.

Danışman: Prof. Dr. İsmail KARAKAYA
(Eğitim Bilimleri Ana Bilim Dalı, Gazi Üniversitesi) ……..…..….

Başkan: Prof. Dr. Selahattin GELBAL
(Eğitim Bilimleri Ana Bilim Dalı, Hacettepe Üniversitesi) ……....

Üye: Doç. Dr. Dilara BAKAN KALAYCIOĞLU
(Eğitim Bilimleri Ana Bilim Dalı, Gazi Üniversitesi) ……..

Tez Savunma Tarihi: 07/01/2022

Bu tezin Eğitim Bilimleri Ana Bilim Dalı'nda Yüksek Lisans tezi olması için şartları yerine getirdiğini onaylıyorum.

Prof. Dr. Yücel GELİŞLİ
Eğitim Bilimleri Enstitüsü Müdürü ……

Aileme

TEŞEKKÜR

Ders döneminden başlayarak ölçme ve değerlendirme alanında farklı kaynaklar önererek bilgi birikimimin artmasını sağlayan ve yaşadığım sorunları çözmede yardımını esirgemeyen danışman hocam Sayın Prof. Dr. İsmail KARAKAYA'ya çok teşekkür ederim.

Savunma jüri üyesi olmalarından onur duyduğum, yaptıkları önerilerle tez çalışmama katkıda bulunan Sayın Prof. Dr. Selahattin GELBAL ve Sayın Doç. Dr. Dilara BAKAN KALAYCIOĞLU'na çok teşekkür ederim.

Yüksek lisans eğitimim boyunca derslerini dinlemekten gurur duyduğum Eğitimde Ölçme ve Değerlendirme Bilim Dalı hocalarıma çok teşekkür ederim.

Yüksek Lisans tez çalışmam kapsamında Milli Eğitim Bakanlığı Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü'nden talep ettiğim veri, bilgi ve belge istekleri karşılandığı için Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü Veri Analizi, İzleme ve Değerlendirme Daire Başkanlığına teşekkürlerimi sunarım.

Hep daha iyisini yapabileceğime inanan, karşılaştığım her zorlukta kurdukları destekleyici cümleleriyle beni motive eden, yazdığım bu tezi okumasını çok istediğim merhum babam Rıdvan SEİS'e ve annem Fikriye SEİS'e çok teşekkür ederim.

Yardımlarıyla hayatımı kolaylaştıran, desteğini, gücünü yanımda ve arkamda hissettiğim eşim Mehmet GÖLCÜR'e, birlikte geçireceğimiz zamanlardan çalarak yazdığım bu tezi okuduğunda adının burada yazılı olmasından mutluluk duyacağını bildiğim oğlum Eymen Alp GÖLCÜR'e çok teşekkür ederim.

ÇOK YÜZEYLİ RASCH MODELİ PUANLAMA DESENLERİNE GÖRE AÇIK UÇLU MADDELERİN PUANLAYICILAR ARASI GÜVENİRLİKLERİNİN KARŞILAŞTIRILMASI

(Yüksek Lisans Tezi)

Zeynep Gölcür

GAZİ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Ocak 2022

ÖZ

Bu çalışma matematik başarı testindeki açık uçlu maddeleri özel bütünsel dereceli puanlama anahtarı ile puanlayan puanlayıcılar arası güvenirliği çok yüzeyli Rasch modeli puanlama desenlerine göre karşılaştırarak güvenilir ve kullanışlı bir puanlama deseni ortaya koymayı amaçlamıştır. Ankara'daki devlet okullarının sekizinci sınıfında öğrenim gören 137 öğrenciye uygulama yapılmıştır. Başarı testi, ikisi PISA 2012 matematik testi pilot uygulama sorusu olmak üzere 11 maddeden oluşmuştur. Araştırmacı tarafından hazırlanan dokuz açık uçlu maddenin ve maddelere ilişkin özel bütünsel puanlama anahtarlarının kapsam geçerliğini kontrol etmek için uzman değerlendirme formu oluşturularak bir ölçme değerlendirme ve beş matematik alan uzmanından görüş alınmıştır. Belirtke tablosunda belirtilen kazanımlara ve bilişsel düzeylere uygun iyi maddeler hazırlandığı konusunda görüş birliğine varılmıştır. Açık uçlu maddelerin ve dereceli puanlama anahtarının her madde için hesaplanan kapsam geçerlik indeksi değerinin kapsam geçerlik ölçütü değeri 0,80 (N=6 ve α =0,05)'den büyük olduğu, kapsam geçerliklerinin sağlandığı görülmüştür. Matematik başarı testinin uygulanmasından elde edilen verilerle açımlayıcı faktör analizi yapılmıştır. Analiz sonucunda maddeler iki faktörde kümelenmiş, birinci faktörün açıkladığı varyans oranının %43,52 bulunması matematik başarı testinin tek faktörlü bir yapı

oluşturabileceğini göstermiştir. Güvenirlik katsayısı ω değeri 0,89 bulunmasıyla açık uçlu maddelerin, dereceli puanlama anahtarının puanlama güvenirliğinin sağlandığı görülmüştür. Öğrenci cevapları dört matematik öğretmeni tarafından birbirlerinden bağımsız olarak puanlanmış, veriler analiz edilmiştir. Eckes (2015)'in tanımladığı çok yüzeyli Rasch modeli puanlama desenlerinden tamamen çaprazlanmış puanlama deseni (Desen A), tamamlanmamış-bağlantılı puanlama deseni (Desen B ve C), tamamlanmamış-bağlantılı olmayan puanlama deseni (Desen D), tamamlanmamış-bağlantılı sarmal puanlama deseni (Desen E), Tharbai ve Young (2001)'ın kullandığı sarmal puanlama deseninin (Desen F) tümünde puanlayıcıların katı puanlama yaptığı, katılık-cömertlik davranışları arasında istatistiksel olarak anlamlı fark olduğu, puanlamaya puanlayıcıdan kaynaklanan hataların karıştığı ve Madde 6'nın güçlüğünün en yüksek olduğu görülmüştür. Tüm desenlerde madde güçlüğü yüksek ve düşük madde gruplarının, logit değerlerinin farklılaşabildiği görülmüştür. Rasch-Cohen'in kappa istatistiği değerleri Desen A, B, C, D, E, F için sırasıyla; 0,65; 0,66; 0,66; 0,93; 0,43; 0,45 bulunmuştur. Puanlayıcıların puanlama anahtarı ölçütleriyle tutarlı puanlama yaptıkları ve puanlayıcılar arası güvenirliğin (uyumun) yüksek olduğu anlaşılmıştır. Açık uçlu maddeleri cevaplandırmada en yüksek ve en düşük beceriye sahip öğrenciler Desen F dışındaki desenlerde aynı öğrencilerden oluşmaktadır. Desen B'nin Desen A ile puanlayıcı sıralamalarının aynı, logit değerlerinin öğrenci, açık uçlu madde yüzeyleri ile dereceli puanlama anahtarı kategorilerine ilişkin ölçüm raporlarının, puanlayıcılar arası güvenirliklerinin oldukça benzer olduğu görülmüştür. Desen F'ye ilişkin ölçüm raporları diğer desenlerden oldukça farklıdır. Desen A ile B'nin benzer Rasch-Cohen'in kappa değerlerinin de benzerlik gösterdiği tüm puanlayıcıların tüm öğrenci cevaplarını puanladığı Desen A yerine daha kullanışlı ve daha az zaman gerektiren dört puanlayıcıdan herhangi üçünün bir öğrenciyi puanladığı Desen B'nin alternatif olarak kullanılabileceği sonucuna ulaşılmıştır.

Anahtar Kelimeler: Açık Uçlu Maddeler, Çok Yüzeyli Rasch Modeli Puanlama Desenleri, Özel Bütünsel Dereceli Puanlama Anahtarı, Puanlayıcılar Arası Güvenirlik.

Sayfa Adedi: xx+167

Danışman: Prof. Dr. İsmail KARAKAYA

COMPARISON OF THE INTER-RATER RELIABILITY OF OPEN-ENDED ITEMS ACCORDING TO THE MANY FACET RASCH MODEL RATING DESIGNS

(M.S Thesis)

Zeynep Gölcür

GAZI UNIVERSITY

GRADUATE SCHOOL OF EDUCATIONAL SCIENCES

January 2022

ABSTRACT

This study aimed to reveal a reliable and useful rating design by comparing the reliability of the open-ended items in the mathematics achievement test with the multi-faceted Rasch model rating designs, which scored with a task-specific holistic rubric. An application was made to 137 students studying in the eighth grade of public schools in Ankara. In order to check the content validity of the nine open-ended items and the task specific holistic rubrics for the items, an expert evaluation form was created and opinions were received from an assessment and evaluation and five mathematics domain experts. It was agreed that good items were prepared in accordance with the acquisitions and cognitive levels specified in the specification table. It was observed that the content validity index value calculated for each item of the open-ended items and rubric was greater than the content validity criterion value of 0,80 (N=6 and α=0,05), and the content validity was achieved. Exploratory factor analysis was performed with the data obtained from the application of the mathematics achievement test. As a result of the analysis, the items were clustered in two factors, and the variance rate explained by the first factor was 43,52%, which showed that the mathematics achievement test could create a single factor structure. With the reliability coefficient ω value of 0,89 it was seen that the rating reliability of the open-ended items and the rubric was ensured.

Student answers were rated independently by four mathematics teachers, and the data were analyzed. Among the many facet Rasch model rating designs defined by Eckes (2015), fully crossed rating design (Design A), incomplete-connected rating design (Design B), incomplete-unconnected rating design (Design C and D), incomplete-connected spiral rating design (Design E), the spiral rating design (Design F) used by Tharbai and Young (2001) where raters rate severe, there is a statistically significant difference between severity and generosity behaviors, mistakes caused by raters are involved in rating, and the difficulty of item 6 was found to be the highest. It was observed that the item groups with high and low item difficulty and logit values could differ in all designs. Rasch-Cohen's kappa statistic values for Design A, B, C, D, E, F respectively; 0.65; 0.66; 0.66; 0.93; 0.43; 0.45 was found. It was understood that the raters made consistent rating with the rubric criteria and the inter-rater reliability was high. The students with the highest and lowest skills in answering open-ended items consist of the same students in designs other than Design F. It was seen that the order of the raters of Design B was the same with Design A, and the inter-rater reliability of the measurement reports regarding the logit values of the student, open-ended item facets and rubric categories were quite similar. The measurement reports for Design F are quite different from other designs. It has been concluded that Design B, in which three of the four raters score a student, which is more useful and requires less time, can be used as an alternative to Design A, in which all raters score all student answers, with similar Rasch-Cohen's kappa values in Design A and B.

Key Words: Open Ended Items, Many Facet Rasch Model Rating Designs, Task-Specific Holistic Rubric, Inter-Rater Reliability.

Page Number: xx+167

Supervisor: Prof. Dr. İsmail KARAKAYA

SİMGELER VE KISALTMALAR LİSTESİ

MEB	Milli Eğitim Bakanlığı
LGS	Liselere Geçiş Sistemi
ÖSYM	Öğrenci Seçme ve Yerleştirme Merkezi
TYT	Temel Yeterlik Testi
AYT	Alan Yeterlik Testi
YDT	Yabancı Dil Testi
KPSS	Kamu Personeli Seçme Sınavı
ABİDE	Akademik Becerilerin İzlenmesi ve Değerlendirilmesi
KTK	Klasik Test Kuramı
GK	Genellenebilirlik Kuramı
MTK	Madde Tepki Kuramı
ÇYRM	Çok Yüzeyli Rasch Modeli
DPA	Dereceli Puanlama Anahtarı
AFA	Açımlayıcı Faktör Analizi
DFA	Doğrulayıcı Faktör Analizi

İÇİNDEKİLER

TELİF HAKKI ve TEZ FOTOKOPİ İZİN FORMU i

ETİK İLKELERE UYGUNLUK BEYANI ii

JÜRİ ONAY SAYFASI iii

TEŞEKKÜR v

ÖZ vi

ABSTRACT viii

SİMGELER VE KISALTMALAR LİSTESİ x

TABLOLAR LİSTESİ xvii

ŞEKİLLER LİSTESİ xx

BÖLÜM I 1

GİRİŞ 1

1.1. Problem Durumu 1

1.2. Araştırmanın Amacı 5

1.3. Araştırmanın Önemi 7

1.4. Araştırmanın Sınırlılıkları 9

1.5. Araştırmanın Sayıltıları 9

BÖLÜM II 10

KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR 10

2.1. Kuramsal Çerçeve 10

2.1.1. Alternatif Durum Belirleme 10

2.1.2. Dereceli Puanlama Anahtarları 11

2.1.2.1. Bütünsel (Holistik) Dereceli Puanlama Anahtarları 14

2.1.2.2. Analitik Dereceli Puanlama Anahtarları 15

2.1.2.3. Genel Dereceli Puanlama Anahtarı 16

2.1.2.4. Göreve Özel Dereceli Puanlama Anahtarı 16

2.1.3. Açık Uçlu Maddeler 17

2.1.4. Güvenirlik 19

2.1.4.1. Puanlayıcı Güvenirliği 19

2.1.4.1.1. Puanlayıcılar Arası Güvenirlik 21

2.1.4.1.2. Puanlayıcı İçi Güvenirlik 21

2.1.5. Çok Yüzeyli Rasch Modeli 22

2.1.5.1. Çok Yüzeyli Rasch Modeli Puanlama Desenleri 24

2.1.5.1.1. Tamamen (Tam) Çaprazlanmış Puanlama Deseni 26

2.1.5.1.2. Tamamlanmamış-Bağlantılı Puanlama Deseni 27

2.1.5.1.3. Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni. 29

2.1.5.1.4. Tamamlanmamış - Bağlantılı Sarmal Puanlama Deseni. 30

2.3. Çok yüzeyli Rasch Modeli ve Puanlayıcılar Arası Güvenirlikle İlgili Çalışmalar 34

BÖLÜM III 40

YÖNTEM 40

3.1. Araştırmanın Modeli 40

3.2. Çalışma grubu 40

3.3. Veri Toplama Aracı 41

3.3.1. Açık Uçlu Maddelerden Oluşan Matematik Başarı Testinin Hazırlanması 41

3.3.2. Açık Uçlu Maddeler için Özel Bütünsel Dereceli Puanlama Anahtarının Hazırlanması 42

3.4. Veri Toplama Süreci 47

3.5. Verilerin Analizi 48

BÖLÜM IV 50

BULGULAR VE YORUMLAR 50

4.1. Tamamen Çaprazlanmış Puanlama Desenine (Desen A) İlişkin Bulgular 50

4.1.1. Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Veri Kalibrasyon Haritasına İlişkin Bulgular 51

4.1.2. Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Puanlayıcı Yüzeyine İlişkin Bulgular 53

4.1.3. Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Öğrenci Yüzeyine İlişkin Bulgular 55

4.1.4. Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Açık Uçlu Madde Yüzeyine İlişkin Bulgular 57

4.1.5. Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Özel Bütünsel Dereceli Puanlama Anahtarı Ölçütlerine İlişkin Bulgular 59

4.2. Tamamlanmamış-Bağlantılı Puanlama Desenine (Desen B) İlişkin Bulgular 61

4.2.1. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen B) Veri Kalibrasyon Haritasına İlişkin Bulgular 61

4.2.2. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen B) Puanlayıcı Yüzeyine İlişkin Bulgular 63

4.2.3. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen B) Öğrenci Yüzeyine İlişkin Bulgular 64

4.2.4. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen B) Açık Uçlu Madde Yüzeyine İlişkin Bulgular 66

4.2.5. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen B) Özel Bütünsel Dereceli Puanlama Anahtarı Ölçütlerine İlişkin Bulgular 68

4.3. Tamamlanmamış - Bağlantılı Puanlama Desenine (Desen C) İlişkin Bulgular.... 70

4.3.1. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Veri Kalibrasyon Haritasına İlişkin Bulgular 70

4.3.2. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Puanlayıcı Yüzeyine İlişkin Bulgular 71

4.3.3. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Öğrenci Yüzeyine İlişkin Bulgular 72

4.3.4. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Açık Uçlu Madde Yüzeyine İlişkin Bulgular..... 74

4.3.5. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Özel Bütünsel Dereceli Puanlama Anahtarı Ölçütlerine İlişkin Bulgular..... 76

4.4. Tamamlanmamış - Bağlantılı Olmayan Puanlama Desenine (Desen D) İlişkin Bulgular 78

4.4.1. Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Veri Kalibrasyon Haritasına İlişkin Bulgular 78

4.4.2. Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Puanlayıcı Yüzeyine İlişkin Bulgular..... 79

4.4.3. Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Öğrenci Yüzeyine İlişkin Bulgular..... 80

4.4.4. Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Açık Uçlu Madde Yüzeyine İlişkin Bulgular..... 82

4.4.5. Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Özel Bütünsel Dereceli Puanlama Anahtarı Ölçütlerine İlişkin Bulgular 84

4.5. Tamamlanmamış-Bağlantılı Sarmal Puanlama Desenine (Desen E) İlişkin Bulgular 86

4.5.1. Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Veri Kalibrasyon Haritasına İlişkin Bulgular 86

4.5.2. Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Puanlayıcı Yüzeyine İlişkin Bulgular..... 87

4.5.3. Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Öğrenci Yüzeyine İlişkin Bulgular 88

4.5.4. Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Açık Uçlu Madde Yüzeyine İlişkin Bulgular 90

4.5.5. Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Özel Bütünsel Dereceli Puanlama Anahtarı Ölçütlerine İlişkin Bulgular 92

4.6. Sarmal Puanlama Desenine (Desen F) İlişkin Bulgular 94

4.6.1. Sarmal Puanlama Deseni (Desen F) Veri Kalibrasyon Haritasına İlişkin Bulgular..... 94

4.6.2. Sarmal Puanlama Deseni (Desen F) Puanlayıcı Yüzeyine İlişkin Bulgular.....95

4.6.3. Sarmal Puanlama Deseni (Desen F) Öğrenci Yüzeyine İlişkin Bulgular 96

4.6.4. Sarmal Puanlama Deseni (Desen F) Açık Uçlu Madde Yüzeyine İlişkin Bulgular..... 98

4.6.5. Sarmal Puanlama Deseni (Desen F) Özel Bütünsel Dereceli Puanlama Anahtarı Ölçütlerine İlişkin Bulgular..... 100

BÖLÜM V 104

SONUÇLAR VE ÖNERİLER 104

5.1. Sonuçlar 104

5.1.1. Veri Kalibrasyon Haritasına İlişkin Sonuçlar..... 104

5.1.2. Puanlayıcı Yüzeyine İlişkin Sonuçlar 105

5.1.3. Öğrenci Yüzeyine İlişkin Sonuçlar 106

5.1.4. Açık Uçlu Madde Yüzeyine İlişkin Sonuçlar..... 106

5.1.5. Özel Bütünsel Dereceli Puanlama Anahtarı Kategorilerine İlişkin Sonuçlar 107

5.1.6. Puanlama Desenlerine İlişkin Sonuçlar..... 107

5.2. Öneriler 108

5.2.1. Araştırmacılara Yönelik Öneriler 108

5.2.2. Uygulamaya Yönelik Öneriler 109

KAYNAKLAR 111

EKLER 122

EK 1. Açık Uçlu Maddelere İlişkin Belirtke Tablosu 123

EK 2. Lawshe(1975)'nin Kapsam Geçerlik Oranı için Kritik Değerler Tablosu..... 124

EK 3. Matematik Başarı Testi Yönergesi 125

EK 4: Matematik Başarı Testi 126

EK 5. Uzman Değerlendirme Formu 139

EK 6. Özel Bütünsel Dereceli Puanlama Anahtarları 143

EK 7. Öğretmen Değerlendirme Formu 165

EK 8. PISA 2012 Pilot Uygulama Soruları Kullanım İzni 166

EK 9. Etik Kurul Onayı 167

TABLOLAR LİSTESİ

Tablo 1. *Puanlayıcı Aracılığıyla Performans Değerlendirmenin Üç Yüzey İçin Tamamen Çaprazlanmış Puanlama Deseni Gösterimi (Desen A)* *26*
Tablo 2. *Çapraz Desen Örneği* 27
Tablo 3. *Puanlayıcı Aracılığıyla Performans Değerlendirmenin Üç Yüzey İçin Tamamlanmamış- Bağlantılı Puanlama Deseni Gösterimi (Desen B)* 27
Tablo 4. *Puanlayıcı Çifti Desen Örneği* 28
Tablo 5. *Puanlayıcı Aracılığıyla Performans Değerlendirmenin Üç Yüzey İçin Tamamlanmamış-Bağlantılı Puanlama Deseni Gösterimi (Desen C)* 28
Tablo 6. *Karma Desen Örneği* 29
Tablo 7. *Puanlayıcı Aracılığıyla Performans Değerlendirmenin Üç Yüzey İçin Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni Gösterimi (Desen D)* 30
Tablo 8. *Puanlayıcı Aracılığıyla Performans Değerlendirmenin Üç Yüzey İçin Tamamlanmamış-Bağlantılı Sarmal Puanlama Desen Gösterimi (Desen E)* 30
Tablo 9. *İç İçe Geçmiş (Yuvalanmış) Desen Örneği* 31
Tablo 10. *Sarmal Puanlama Deseni Örneği* 31
Tablo 11. *Açık Uçlu Maddelere İlişkin Kapsam Geçerlik İndeksleri (KGİ)* 44
Tablo 12. *Özel Bütünsel Dereceli Puanlama Anahtarına İlişkin Kapsam Geçerlik İndeksleri (KGİ)* 45
Tablo 13. *KMO Değerlerinin Faktör Analizi İçin Uygunluğunun Yorumlanmasına İlişkin Değer Aralıkları ve Açıklama* 46
Tablo 14. *Model Veri Uyumu Kontrolüne İlişkin Aralıklara Göre Standartlaştırılmış Artık Değer Sayıları ve Yüzdeleri* 49
Tablo 15. *Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu* 54

Tablo 16. *Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Öğrenci Yüzeyine İlişkin Ölçüm Raporu* 56
Tablo 17. *Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Açık Uçlu Madde Yüzeyine İlişkin Ölçüm Raporu* 58
Tablo 18. *Özel Bütünsel Dereceli Puanlama Anahtarı Puanlama Kategorilerine Ait Ölçüm Raporu* 59
Tablo 19. *Tamamlanmamış-Bağlantılı Puanlama Deseni Gösterimi (Desen B) Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu* 63
Tablo 20. *Tamamlanmamış-Bağlantılı Puanlama Deseni Gösterimi (Desen B) Öğrenci Yüzeyine İlişkin Ölçüm Raporu* 65
Tablo 21. *Tamamlanmamış-Bağlantılı Puanlama Deseni Gösterimi (Desen B) Açık Uçlu Madde Yüzeyine İlişkin Ölçüm Raporu* 67
Tablo 22. *Özel Bütünsel Dereceli Puanlama Anahtarı Puanlama Kategorilerine Ait Ölçüm Raporu* 68
Tablo 23. *Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu* 71
Tablo 24. *Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Öğrenci Yüzeyine İlişkin Ölçüm Raporu* 73
Tablo 25. *Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Açık Uçlu Madde Yüzeyine İlişkin Ölçüm Raporu* 75
Tablo 26. *Özel Bütünsel Dereceli Puanlama Anahtarı Puanlama Kategorilerine Ait Ölçüm Raporu* 76
Tablo 27. *Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu* 79
Tablo 28. *Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Öğrenci Yüzeyine İlişkin Ölçüm Raporu* 81
Tablo 29. *Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Açık Uçlu Madde Yüzeyine İlişkin Ölçüm Raporu* 83
Tablo 30. *Özel Bütünsel Dereceli Puanlama Anahtarı Puanlama Kategorilerine Ait Ölçüm Raporu* 84
Tablo 31. *Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu* 87

Tablo 32. *Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Öğrenci Yüzeyine İlişkin Ölçüm Raporu*..... 89
Tablo 33. *Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Açık Uçlu Madde Yüzeyine İlişkin Ölçüm Raporu*..... 91
Tablo 34. *Özel Bütünsel Dereceli Puanlama Anahtarı Puanlama Kategorilerine Ait Ölçüm Raporu*..... 92
Tablo 35. *Sarmal Puanlama Deseni (Desen F) Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu* 95
Tablo 36. *Sarmal Puanlama Deseni (Desen F) Öğrenci Yüzeyine İlişkin Ölçüm Raporu*.....97
Tablo 37. *Sarmal Puanlama Deseni (Desen F) Açık Uçlu Madde Yüzeyine İlişkin Ölçüm Raporu*..... 99
Tablo 38. *Özel Bütünsel Dereceli Puanlama Anahtarı Puanlama Kategorilerine Ait Ölçüm Raporu*..... 100
Tablo 39. *Çok Yüzeyli Rasch Modeli Puanlama Desenlerine İlişkin Bulguların Karşılaştırılması* 101

ŞEKİLLER LİSTESİ

Şekil 1. Tamamen çaprazlanmış puanlama deseni (Desen A) veri kalibrasyon haritası 52
Şekil 2. Tamamen çaprazlanmış puanlama desenine (Desen A) göre dereceli puanlama anahtarı ölçütlerine ait olasılık eğrileri..... 61
Şekil 3. Tamamlanmamış-bağlantılı puanlama deseni (Desen B) veri kalibrasyon haritası 62
Şekil 4. Tamamlanmamış-bağlantılı puanlama deseni gösterimi (Desen B) dereceli puanlama anahtarı ölçütlerine ait olasılık eğrileri..... 69
Şekil 5. Tamamlanmamış-bağlantılı puanlama deseni (Desen C) veri kalibrasyon haritası 70
Şekil 6 . Tamamlanmamış-bağlantılı puanlama deseni (Desen C) dereceli puanlama anahtarı ölçütlerine ait olasılık eğrileri 77
Şekil 7. Tamamlanmamış-bağlantılı olmayan puanlama deseni (Desen D) veri kalibrasyon haritası 78
Şekil 8. Tamamlanmamış-bağlantılı olmayan puanlama deseni (Desen D) dereceli puanlama anahtarı ölçütlerine ait olasılık eğrileri..... 85
Şekil 9. Tamamlanmamış-bağlantılı sarmal puanlama deseni (Desen E) veri kalibrasyon haritası 86
Şekil 10. Tamamlanmamış-bağlantılı sarmal puanlama desenine (Desen E) göre dereceli puanlama anahtarı ölçütlerine ait olasılık eğrileri..... 93
Şekil 11. Sarmal puanlama deseni (Desen F) veri kalibrasyon haritası 94
Şekil 12. Sarmal puanlama desenine (Desen F) göre dereceli puanlama anahtarı ölçütlerine ait olasılık eğrileri 101

BÖLÜM I

GİRİŞ

Bu bölümde problem durumu, kuramsal çerçeve, araştırmanın amacı, araştırmanın önemi hakkında bilgi verilmiştir.

1.1. Problem Durumu

Eğitim ve öğretim iç içedir. Bireylerin sosyal ve akademik başarılarının gelişmesi eğitim kurumlarında yapılan öğretimlerle gerçekleşmektedir. Bu nedenle yapılan öğretimin planlanması ve değerlendirilmesi önemlidir (Akgün, 2016).

Küçük yaşlardan itibaren eğitim alan bireylerin ön bilgilerinin belirlenmesi, eğitim sürecinde kullanılan yöntem ve tekniklerin belirlenerek etkililiğinin ölçülmesi, eğitim programının uygulanması sırasında öğrenmede eksiklikler varsa bu eksikliklerin nedenlerinin bulunması ve eksikliklerin giderilmesi, öğretimin kazanımlar doğrultusunda ne kadar etkili olduğunun, eğitim programının hedeflerine ulaşıp ulaşamadığının ya da ne kadarına ulaşıldığının kontrol edilmesi, eğitim paydaşlarına bireylerin akademik başarıları hakkında geri bildirimler vererek yapılması gereken yeniliklere karar verilmesi aşamasında ölçme ve değerlendirmeden faydalanılmaktadır (Demirel, 2004). Ölçme, belirli bir kural dizisine göre niteliklere sayılar veya semboller atama işlemidir (Stevens, 1951). Başka bir ifadeyle, nesnelerin herhangi bir özelliğe sahip olma durumunun ya da derecesinin gözlemlenerek sonuçlarının sayı, sembol veya sıfatlarla ifade edilmesidir. Değerlendirme ise ölçme sonuçlarını bir ölçütle karşılaştırıp yorumlama ve ölçülen nitelikle ilgili karara varma sürecidir (Turgut & Baykul, 2010). Ölçülen nitelikle ilgili doğru değerlendirmenin yapılması uygun ölçütün kullanılmasına ve güvenilir, geçerli ölçme sonuçlarına bağlıdır. Ölçme

işlemlerinde yüksek güvenirlik ve geçerliğe sahip ölçme araçları kullanılarak yerinde değerlendirmeler yapma olasılığı artırılabilir (Güler & Gelbal, 2010).

Eğitimin her kademesinde bireylerin bilgi ve becerilerinin geliştirilmesine önem verilmekte derslerde eğitim programına göre bireylere bilişsel, duyuşsal ve psikomotor alan davranışlarının kazandırılması sağlanmaktadır. Ülkemizde bireylerin bir üst eğitim kurumuna geçişleri bu alan davranışlarının ölçülüp belli bir ölçüte göre değerlendirilmesiyle mümkündür. Bireylerin gidecekleri bir üst eğitim kurumu belirlenirken sadece bilişsel alan davranışlarını ölçen maddeler kullanılarak oluşturulan geniş ölçekli sınavlar uygulanmaktadır.

Milli Eğitim Bakanlığı (MEB)'nın bireyleri liseye yerleştirmek amacıyla uyguladığı Liselere Geçiş Sistemi (LGS) sınavında, Ölçme, Seçme ve Yerleştirme Merkezi (ÖSYM)'nin bireyleri yükseköğrenime yerleştirmek amacıyla uyguladığı Temel Yeterlik Testi (TYT), Alan Yeterlik Testi (AYT), Yabancı Dil Testi (YDT) ile Kamu Personeli Seçme Sınavı (KPSS) gibi geniş ölçekli sınavlarda çoktan seçmeli maddelere yer verilmektedir. Çoktan seçmeli maddeler bilişsel alanın bilgi, kavrama ve uygulama basamağındaki davranışların ölçülmesi için uygundur. Öğrencilerin verilen seçenekler içerisinden doğru cevaba ya da en doğru olanına ulaşarak işaretleme yapması nedeniyle bu maddeler üst bilişsel davranış olarak kabul edilen analiz, sentez ve değerlendirme basamaklarındaki özellikleri ölçmede yetersizdir (Tan, 2008, s. 272). Bireylerin bilişsel alanın tüm basamaklarındaki davranışları gerçekleştirecek düzeyde yetiştirilmesi eğitim programının hedeflerindendir. Bu hedeflere ulaşılıp ulaşılamadığının kontrolü ve üst bilişsel davranışların ölçülmesi için açık uçlu maddelerden oluşan testlerin kullanılması daha uygundur. Çoktan seçmeli maddelerden oluşan testler, her soruda farklı kazanımların sorgulanabilmesi ve soru sayısının fazla olması sebebiyle kapsam geçerliği yüksek, madde güçlük düzeylerinin kolay ayarlanabilir, değerlendirilmesinin objektif ve kullanışlı olması nedeniyle puanlama güvenirliği yüksek kabul edildiğinden bu maddelerin geniş ölçekli sınavlarda kullanılması avantajlıdır. Maddenin seçenek sayısını artırarak şans başarısı azaltılmaya çalışılsa da şansla doğru cevaba ulaşılabilmesi nedeniyle konuyu hiç bilmeyen birey ile kısmen bilen bireyi ayırt etmenin mümkün olmaması ya da konuyu kısmen bilen bireyin iki seçenek arasında kaldığında yanlış cevabı işaretlemesinin bireyin bilgi seviyesinin tam göstergesi olmadığından güvenirlik ve geçerliği azaltmakta, bu testleri dezavantajlı hale getirmektedir (Yaman, 2016). Bireylerin düşüncelerini özgürce yazabildikleri açık uçlu maddelerden oluşan testlerin değerlendirilmesinin subjektif ve soru sayısının az olması sonucu farklı

kazanımlarının ölçülebilmesinin sınırlılığı güvenirlik ve geçerliği düşürürken, şans başarısının düşük olması güvenirliği artırmaktadır (Temizkan & Sallabaş, 2011).

MEB ve ÖSYM, birey başarısının daha güvenilir bir şekilde ölçülmesi amacıyla uyguladıkları sınavlarda açık uçlu maddelere yer verilmesi gerektiğini belirtmiştir. Milli Eğitim Bakanlığı (MEB, 2015) 8. sınıf öğrencilerine içerisinde açık uçlu maddelerin de yer aldığı farklı madde formatlarından oluşan Akademik Becerilerin İzlenmesi ve Değerlendirilmesi (ABİDE) projesinin pilot uygulamasını yapmıştır. Bu projenin amacı, üst bilişsel düşünme becerilerini ölçmek için farklı madde türlerinin kullanıldığı beceri testleri geliştirmek ve bireylerin düşünme beceri seviyelerini belirlemektir (Milli Eğitim Bakanlığı [MEB], 2015).

Resmi Gazete'de yayımlanan yönetmelik değişikliği ile açıköğretim sınavlarında çoktan seçmeli maddelerin yanında açık uçlu, kısa cevaplı, doğru-yanlış, eşleştirme gibi farklı madde tiplerinin kullanılabilmesi sağlanmıştır (Anadolu Üniversitesi Açıköğretim, İktisat ve İşletme Fakülteleri Eğitim-Öğretim ve Sınav Yönetmeliğinde Değişiklik Yapılmasına Dair Yönetmelik, 2017). Bazı derslerin ara sınavlarında çoktan seçmeli soruların yanında kısa cevaplı ve açık uçlu soruların da kullanılacağını belirtmiştir (Gündoğan, 2017). Bu çalışmaların yanında İçişleri Bakanlığı Kaymakam Adaylığı Giriş Sınavında da açık uçlu maddelere yer verilmesi düşünülmüştür (Ölçme, Seçme ve Yerleştirme Merkezi [ÖSYM], 2015). Bireylerin akademik başarılarının ve üst bilişsel becerilerinin ölçülmesine yönelik yapılmaya çalışılan bu yenilikler, ölçme ve değerlendirme alanında çalışan araştırmacıları açık uçlu maddelerden oluşan testleri daha iyi nasıl değerlendirebilecekleri konusunda çalışmaya yönlendirmiştir. Açık uçlu maddelerin puanlanmasında puanlayıcıdan, maddeden ya da ölçme aracından kaynaklanan hatalar meydana gelebilmektedir (Tekin, 2014). Açık uçlu maddelerin puanlanmasının subjektif olması nedeniyle puanlamaya puanlayıcıdan kaynaklanan hatalar karışabilmekte ve bu maddelerin puanlama güvenirliğini düşürebilmektedir. Daha güvenilir sonuçlar elde etmek amacıyla bu maddelerin puanlanmasında holistik (bütünsel) ya da analitik dereceli puanlama anahtarları kullanılarak puanlayıcı içi tutarlılığı sağlamak amacıyla bir puanlayıcıya belli aralıklarla puanlama yaptırılmakta ya da puanlayıcılar arası tutarlılığı sağlamak amacıyla birden çok puanlayıcıdan bu maddeleri puanlandırmaları istenmektedir.

Puanlayıcıların açık uçlu maddeleri benzer ya da farklı puanlar ile değerlendirmesi sonucunda, puanlayıcıdan kaynaklanan ögelerin ölçme sonuçlarını etkileme durumunun yani puanlayıcı güvenirliğinin belirlenmesi gerekmektedir (Atılgan, 2004; Kan, 2013). Puanlayıcı

güvenirliğinin belirlenmesinde kullanılabilecek yöntemler; klasik test kuramına (KTK) dayanan yöntemler, genellenebilirlik kuramına (GK) dayanan yöntemler ve madde tepki kuramına (MTK) dayanan yöntemler olmak üzere üç başlıkta incelenmektedir (MacMillan, 2000). KTK'da madde güçlüğüne bakılmadan doğru cevap verildiğinde bireyin tam puan aldığı, yanlış cevap verildiğinde puan alamadığı, farklı maddeleri doğru cevaplayarak aynı puanı alan iki bireyin bilgi ve becerilerinin eşit seviyede olduğu kabul edilmektedir (Kaptan, 1994; Bahar, 2001). Alanda uzun süredir kullanılan KTK'nın varsayımları kolay sağlanabilirdir ancak madde ve yetenek parametrelerinin gruba bağlı olarak elde edilmesi, tüm bireyler için tek bir hata kestirimi yapması gibi sınırlılıkları bulunmaktadır (Köse, Usta ve Yandı, 2016). Son yıllarda KTK'nın bu sınırlılıklarını güçlü varsayımlarıyla ortadan kaldıran ölçme kuramlarından biri de Madde Tepki Kuramı'dır. MTK, bireylerin maddelere verdikleri cevapları kullanarak onların ölçülen özellikteki yerini matematiksel modeller yardımıyla ortaya koyan ölçme kuramıdır (Bobcock, 2009). Madde tepki kuramı modelleri doğrusallığına, boyut sayısına, puanlama şekline (1-0, çoklu puanlama) göre sınıflara ayrılmaktadır (Özkan, 2012). Doğrusal ilişkinin sağlanmasında, kayıp verilerle baş etmede, yapılan ölçümün duyarlılığını tahmin etmede ve standart değerlere uymayan sonuçları değerlendirmede Madde Tepki Kuramı'na dayalı modellerden faydalanılabilmektedir (Elhan & Atakurt, 2005).

Temelinde Madde Tepki Kuramı olan, madde güçlüğü parametresini örneklemden bağımsız olarak tahmin eden ve bir parametreli Rasch modeline (Rasch, 1960) puanlayıcıların katılığı/cömertliği yüzeyinin eklenmesiyle Çok Yüzeyli Rasch Modeli (ÇYRM) geliştirilmiştir (Linacre, 1993).

Lynch ve McNamara (1998) ÇYRM'yi bireylerin yetenek düzeyleri ve maddelerin güçlük düzeylerinin yanında puanlayıcılar, puanlama ölçütleri, puanlama anahtarı gibi test puanlarını etkileme olasılığına sahip diğer değişkenlik kaynaklarının da dikkate alınmasını sağlayan bir model olarak tanımlamaktadır. Bu nedenle sayısal alanda açık uçlu maddeler ile oluşturulan bir testin puanlayıcılar arası güvenirliğini belirlemek için birçok değişkenlik kaynağı arasındaki ilişkiyi aynı anda değerlendirmeye olanak sağlayan çok yüzeyli Rasch modeli puanlama desenlerinden hangisinin birey başarısının değerlendirilmesinde güvenilir sonuçlar verdiğini ortaya çıkarmak gerekmektedir.

1.2. Araştırmanın Amacı

Bu çalışmanın amacı, matematik başarı testi için hazırlanan açık uçlu maddeleri özel bütünsel dereceli puanlama anahtarı kullanarak puanlayan puanlayıcılar arasındaki güvenirliği, uyumu çok yüzeyli Rasch modeli puanlama desenlerine göre karşılaştırarak bu testlerin puanlanmasında kullanılabilecek güvenilir ve kullanışlı puanlama desenini ortaya koymaktır. Bu amaç doğrultusunda aşağıdaki araştırma sorularına cevap aranmıştır:

1) Tamamen (tam) çaprazlanmış puanlama deseninde (Desen A);
    a. Veri kalibrasyon haritasına göre puanlayıcıların katılık/cömertlik, maddelerin güçlük ve öğrencilerin başarı düzeyleri,
    b. Puanlayıcı yüzeyine ilişkin ölçüm raporları,
    c. Öğrenci (birey) yüzeyine ilişkin ölçüm raporları,
    d. Açık uçlu madde yüzeyine ilişkin ölçüm raporları,
    e. Dereceli puanlama anahtarı ölçütlerine ilişkin ölçüm raporları ve olasılık eğrileri nasıldır?
2) Tamamlanmamış bağlantılı puanlama deseninde (Desen B);
    a. Veri kalibrasyon haritasına göre puanlayıcıların katılık/cömertlik, maddelerin güçlük ve öğrencilerin başarı düzeyleri,
    b. Puanlayıcı yüzeyine ilişkin ölçüm raporları,
    c. Öğrenci (birey) yüzeyine ilişkin ölçüm raporları,
    d. Açık uçlu madde yüzeyine ilişkin ölçüm raporları,
    e. Dereceli puanlama anahtarı ölçütlerine ilişkin ölçüm raporları ve olasılık eğrileri nasıldır?
3) Tamamlanmamış bağlantılı puanlama deseninde (Desen C);
    a. Veri kalibrasyon haritasına göre puanlayıcıların katılık/cömertlik, maddelerin güçlük ve öğrencilerin başarı düzeyleri,
    b. Puanlayıcı yüzeyine ilişkin ölçüm raporları,
    c. Öğrenci (birey) yüzeyine ilişkin ölçüm raporları,
    d. Açık uçlu madde yüzeyine ilişkin ölçüm raporları,

e. Dereceli puanlama anahtarı ölçütlerine ilişkin ölçüm raporları ve olasılık eğrileri nasıldır?

4) Tamamlanmamış bağlantılı olmayan puanlama deseninde (Desen D);

a. Veri kalibrasyon haritasına göre puanlayıcıların katılık/cömertlik, maddelerin güçlük ve öğrencilerin başarı düzeyleri,

b. Puanlayıcı yüzeyine ilişkin ölçüm raporları,

c. Öğrenci (birey) yüzeyine ilişkin ölçüm raporları,

d. Açık uçlu madde yüzeyine ilişkin ölçüm raporları,

e. Dereceli puanlama anahtarı ölçütlerine ilişkin ölçüm raporları ve olasılık eğrileri nasıldır?

5) Tamamlanmamış bağlantılı sarmal puanlama deseninde (Desen E);

a. Veri kalibrasyon haritasına göre puanlayıcıların katılık/cömertlik, maddelerin güçlük ve öğrencilerin başarı düzeyleri,

b. Puanlayıcı yüzeyine ilişkin ölçüm raporları,

c. Öğrenci (birey) yüzeyine ilişkin ölçüm raporları,

d. Açık uçlu madde yüzeyine ilişkin ölçüm raporları,

e. Dereceli puanlama anahtarı ölçütlerine ilişkin ölçüm raporları ve olasılık eğrileri nasıldır?

6) Sarmal puanlama deseninde (Desen F);

a. Veri kalibrasyon haritasına göre puanlayıcıların katılık/cömertlik, maddelerin güçlük ve öğrencilerin başarı düzeyleri,

b. Puanlayıcı yüzeyine ilişkin ölçüm raporları,

c. Öğrenci (birey) yüzeyine ilişkin ölçüm raporları,

d. Açık uçlu madde yüzeyine ilişkin ölçüm raporları,

e. Dereceli puanlama anahtarı ölçütlerine ilişkin ölçüm raporları ve olasılık eğrileri nasıldır?

1.3. Araştırmanın Önemi

Bireylerin seçeneklerden yola çıkarak doğru cevaba ulaşmalarını sağlayarak onları düşünmekten uzaklaştıran, özgün ve yaratıcı fikirler üretmelerini sınırlayan, tesadüfi olarak verdikleri cevaplar ile başarılı kabul edilmelerinin mümkün olduğu çoktan seçmeli maddelerle yapılan özellikle geniş ölçekli sınavların eğitim sistemimize olumsuz etkileri, eğitim sistemimizin tüm paydaşları tarafından eleştirilen bir durumdur. Son yıllarda birey başarılarının daha iyi nasıl ölçüleceği ve değerlendirileceği konusunda çalışmalar yapılmış, geçerliği ve güvenirliği yüksek sınavların oluşturulması için MEB ve ÖSYM tarafından bazı değişiklikler yapılması düşünülmüştür. Üst bilişsel davranışların ölçülebilmesi için bireylere bilgilerini organize ederek cevaplarını kendilerinin oluşturacakları ve bu bilgileri kendi kelime ya da cümleleriyle ifade edecekleri şans başarısını engelleyecek madde türlerinin kullanılması uygun bulunmuştur. ÖSYM seçme, yerleştirme ve kamu kurumlarına atama yapma amacıyla uyguladığı sınavlarda birey başarısını belirlemede çoktan seçmeli maddelerin yetersiz olduğunu, farklı madde türleri kullanılarak üst düzey bilişsel davranış ve becerilerin ölçülmesi gerektiğini belirtmiştir. Öğrencilerin üst düzey düşünme becerilerini ölçmeyi ülkelerin eğitim politikalarını yönlendirmeyi hedefleyen aralarında Türkiye'nin de bulunduğu birçok ülkenin katıldığı PISA gibi uluslararası sınavlarda başarı sıralamamızın istenilen düzeyde olmaması ulusal olarak uygulanan bir sınava ihtiyaç duyulduğunu göstermiş, ABİDE projesi planlanmıştır. 8. sınıf düzeyinde öğrencilerin Türkçe, matematik, fen bilimleri ve sosyal bilimler alanlarında üst düzey düşünme becerilerine sahip olma durumlarını belirlemek amacıyla Ankara'da rastgele seçilen 26 ortaokulda yaklaşık 5000 öğrenciyle gerçekleştirilen pilot uygulamada açık uçlu maddelerin ulusal sınavlarda kullanılabileceği öngörülmüştür (MEB, 2015). Buna ek olarak ÖSYM (2015) yaptığı açıklamada kendileri tarafından hazırlanacak ve uygulanacak sınavlarda açık uçlu maddelerin kullanılacağı; sınavlarda kısa cevaplı, cevabı sınırlandırılmış ve cevabı sınırlandırılmamış olmak üzere üç farklı madde türünün yer alacağı belirtilmiştir. Ayrıca 2016-2017 eğitim öğretim yılı bahar döneminden itibaren açıköğretim fakülte sınavlarında ve İçişleri Bakanlığı Kaymakam Adaylığı Giriş Sınavında da kısa cevaplı, açık uçlu maddelere yer verilmeye başlanmış ancak bu uygulamadan daha sonra vazgeçilmiştir (ÖSYM, 2015).

Cevabı bir kelime, bir cümle ya da bir sayı olan yazılı sınavlarda kullanılan kısa cevaplı, bireylerin verilen metinde belirtilen sınırlamalara uygun bir şekilde cevap oluşturdukları üst bilişsel becerileri ölçen cevabı sınırlandırılmış, cevapları kişiden kişiye değişen ve belirli bir

sınırı olmayan genellikle kompozisyon şeklinde yazılan eleştirel düşünme özgün ve yaratıcı fikir üretme, problem çözme gibi üst bilişsel becerilerin ölçülmesinde etkili ölçümü sağlayan cevabı sınırlandırılmamış (serbest yanıtlı) maddelerin (Tekin, 2014) birey başarıları hakkında güvenilir ölçümler vermesinin yanında bu tip maddelerden elde edilen ölçümlerin değerlendirilmesi de önemlidir.

ÖSYM (2015), uzman akademisyenler tarafından hazırlanan açık uçlu maddelere verilen cevapların doğru, kısmen doğru, yanlış şeklinde puanlanması için dereceli puanlama anahtarlarının oluşturulacağını, puanlamanın bu anahtara göre bilgisayar ortamında alan uzmanları tarafından yapılacağını belirtmiştir. Adayların kimliği belli olmayacak şekilde elektronik ortamda her bir maddeye aynı puanı vermeleri beklenen iki puanlayıcı tarafından puanlanacağı, puanlayıcılar farklı puanlama yaptığında ise üçüncü bir puanlayıcının puanlamaya katılacağı böylece objektif bir başarı değerlendirmesi yapılabileceği açıklanmıştır (ÖSYM, 2015). Birey başarısını değerlendiren puanlayıcılar arasındaki güvenirliğin yüksek olması gerekmektedir. Çünkü puanlayıcılarının yanılabilirliği, sınav yapılan bireylere verilen puanların psikometrik kalitesiyle ilgili ciddi kaygılar uyandırmaktadır. Bu kaygılar özellikle de bireylerin kariyerini veya çalışma planlarını büyük ölçüde etkileyen geniş ölçekli sınavlarda dikkatle ele alınmalıdır. Ölçülmekte olan yapı ile ilişkili olanlar dışındaki pek çok değişkenlik kaynağı, değerlendirme sonuçları üzerinde göz ardı edilemez bir etkiye sahip olabilir. Bu nedenle, bireylerin yeteneklerinin güvenilir, geçerli yapıda ve adil ölçümlerle belirlenmesi, çok yüzeyli değerlendirme durumlarını etkileyen değişkenlik kaynaklarıyla başa çıkmak için iyi tasarlanmış yöntemlerin uygulanmasına bağlıdır (Eckes, 2015).

ÖSYM tarafından açıklanan puanlama yönergesinde tüm alt testler için standart bir puanlama deseni tercih edilmiştir. Bu standart puanlama deseninin her bir alt test için güvenilir sonuçlar verip vermediği, sayısal ya da sözel alandaki alt testlerin puanlanmasında bu standart puanlama deseninden daha yüksek puanlayıcılar arası güvenirliğe sahip olan puanlama desenlerinin kullanılıp kullanılamayacağı konusunda ülkemizde yapılan çalışmalar oldukça azdır (İlhan, 2016). Çok yüzeyli Rasch modeli, birden fazla değişkenlik kaynağı arasındaki ilişkileri aynı anda değerlendirmeye ve katılık/cömertlik yüzeyine bakarak puanlayıcılar arasındaki tutarlılığı keşfetmeye olanak sağlamaktadır. Bu nedenle birey, madde ve puanlayıcı değişkenlerinin birlikte karşılaştırılabildiği çok yüzeyli Rasch modeli puanlama desenlerinin hangisinden matematik alt testi için güvenilir sonuçlar elde

edilebileceğinin belirlenmesiyle açık uçlu maddelerden oluşan geniş ölçekli sınavların değerlendirilmesine katkıda bulunulacağı düşünülmektedir.

1.4. Araştırmanın Sınırlılıkları

Çalışma;

1. Ankara ilinde sekizinci sınıfta öğrenim gören 137 öğrenci ve dört puanlayıcı ile sınırlıdır.
2. Çok yüzeyli Rasch modeli puanlama desenlerinden hangisinin puanlama sürecinde en güvenilir sonuç verdiğinin incelenmesi matematik branşı ile sınırlıdır.
3. Açık uçlu maddelerden oluşan matematik başarı testinin puanlanması sadece puanlayıcılar tarafından yapılacaktır; öz ya da akran değerlendirme yapılmayacaktır.

1.5. Araştırmanın Sayıltıları

Çalışmaya katılan;

1) Puanlayıcıların öğrenci cevaplarını hassasiyetle ve birbirlerinden bağımsız bir şekilde puanladıkları varsayılmıştır.

2) Öğrencilerin gönüllülük esasına dayanarak çalışmaya katıldıkları maddeleri içtenlikle yanıtladıkları varsayılmıştır.

3) Öğrencilerin matematik başarı testindeki açık uçlu maddeleri cevaplandıracak hazırbulunuşluğa sahip oldukları varsayılmıştır.

BÖLÜM II

KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

Bu bölüm araştırmanın amacına uygun kavramsal çerçeve ve ilgili araştırmalar olmak üzere iki başlıktan oluşmaktadır.

2.1. Kuramsal Çerçeve

Bu başlıkta alternatif durum belirleme, dereceli puanlama anahtarları ve çeşitleri, açık uçlu maddeler, güvenirlik, puanlayıcı güvenirliği ve çeşitleri, çok yüzeyli Rasch modeli, çok yüzeyli Rasch modeli puanlama desenleri ile ilgili bilgilere yer verilmiştir.

2.1.1. Alternatif Durum Belirleme

Öğrenci başarısının bir göstergesi olan kağıt kalem testleri gibi geleneksel değerlendirmelerden duyulan memnuniyetsizlik ile kendilerine verilen görevler anlamlı olduğunda öğrenci en iyi öğrenmesini gerçekleştirir düşüncesinin 1980'lerde ve 1990'larda birleştiği ve bunun sonucunda da öğretmenlerin projeler, özgün ve performansa dayalı değerlendirmeler, örnekler yazma ve portfolyolar ile öğrenci becerilerini değerlendirebilmesi tavsiyesi ortaya çıkmıştır (Campbell, 2013).

Alternatif durum belirleme, geleneksel durum belirleme yöntemlerinin algılanan yetersizlikleri ve özellikle de özel ihtiyaçları olan öğrencilere uygulanmasındaki eksikliklerine yanıt olarak ortaya çıkmıştır. Bu alandaki gelişmelerin öncülüğünü Amerika Birleşik Devletleri'nin 1990 yılında yürürlüğe giren ve önümüzdeki yıllarda yeniden yetkilendirilen Engelli Bireyler Eğitim Yasası ile engelli öğrenciler için uygun alternatif

durum belirleme yöntemlerinin öğrencilerin durum değerlendirmelerine dahil edilmesinin ve öğrencilerin zeka çeşitleri ile tercih edilen öğrenme stillerini bir araya getirmenin faydalı olduğu görülmüştür (Reardon, 2017). Bu fayda, alternatif durum belirlemenin çeşitli görevleri içeren ve hepsi de öğrencilerin gerçek hayatta veya gerçek durumlarda üst düzey düşünme becerilerini kullanmalarını gerektiren değerlendirme biçimlerini kapsamasıyla ilişkilidir (Al Ruqeishi, 2015).

Performansa dayalı durum belirleme, sergilenen deneyimlerin, bireyler tarafından yapılandırılan cevapların ya da davranışların ne seviyede gerçekleştirildiğinin değerlendirilmesi için bir alternatiftir. Bu değerlendirme, seçenekler içerisinden doğru olanın seçilmesine karşın cevapların yapılandırılmasını ya da görev sergilemeyi gerekli kılmaktadır (Arsan, 2012).

Russell ve Ariasian (2012), performansa dayalı durum belirlemenin, alternatif ya da özgün durum belirleme olarak isimlendirilebileceğini belirtmişlerdir. Performansa dayalı durum belirlemeyi tanımlarken "alternatif " kelimesinin kullanılması, çoktan seçmeli veya kısa cevaplı testlere bir alternatif oluşturacağını; "özgün" kelimesinin kullanılması ise bazı performansa dayalı durum belirlemelerin öğrencilerin gerçek durumlarda neler yapabileceklerini göstermeye fırsat vereceğini göstermektedir (Wiggins, 1992).

Alternatif durum belirleme, geleneksel durum belirlemelerde kullanılmayan ölçme araçları ile gerçekleştirilir (Çepni, 2011). Öğrencilerin kendi bilgi, beceri ve seçimleri ile cevabı özgün bir şekilde yapılandırdıkları açık uçlu bazı madde tipleri de "alternatif durum belirleme" yöntemleri içerisinde yer almaktadır (Kutlu, Doğan ve Karakaya, 2014).

2.1.2. Dereceli Puanlama Anahtarları

Dereceli puanlama anahtarı (DPA), birey çalışmalarının hangi kriterlere göre değerlendirildiğini ve performanslarının hangi düzey puanına karşılık geldiğini belirten puanlama araçlarıdır (Kutlu vd., 2014). Khitab, Zaman ve Ghaffar (2014), dereceli puanlama anahtarının öğrenci performansını derecelendirmek için bir kılavuz olduğunu ve bu kılavuzların performansın üstün, mükemmel, iyi ve zayıf gibi çeşitli seviyelerde neye benzediğini belirttiğini ifade etmişlerdir. Ayrıca dereceli puanlama anahtarını alternatif durum belirlemenin uygulanabilir bir formu olan bir dizi ölçüt listesi ve puanlayıcının değerlendirme veya derecelendirmeyi nasıl yapacağını gösteren açık bir kılavuz olarak tanımlamışlardır. Goodrich (1997)'e göre ise dereceli puanlama anahtarı, bir eserin

ölçütlerini ve her bir ölçüt için kaliteyi mükemmelden kötüye kadar derecelendiren bir puanlama aracıdır.

Öğretmenlerin öğrencilerden beklentilerini netleştiren ve öğrencilere bu beklentileri nasıl karşılayacaklarını gösteren dereceli puanlama anahtarları ile öğrenci performansı artırılabilir ve izlenebilir. Dereceli puanlama anahtarlarının kullanımı hakkındaki en yaygın görüş, öğrencinin çalışması ve öğrenmesindeki kalitenin tanımlanmasına yardımcı olmalarıdır (Marcus, 1995). Brookhart (1999) da kaliteye vurgu yaparak DPA'ları, elde edilen cevapların kalitesi hakkında değerlendirme yapılırken puanlama için yararlı olan betimsel puanlama ölçekleri olarak tanımlamış ve DPA'ların, çalışma kalitesinin kötüden mükemmele doğru sıralamada iyi olduklarını belirtmiştir.

Jönssen ve Panadero (2016)'ya göre öğrenci performansındaki nitel farklılıkları belirlemeye ve değerlendirmeye yardımcı olmak için tasarlanmış değerlendirme araçları olan dereceli puanlama anahtarları iki amaca hizmet etmektedir: (1) Puanlayıcılara performans görevlerini puanlarken daha yüksek tutarlılık düzeylerine ulaşmada yardımcı olmak ve değerlendirme beklentilerini açık hale getirmek ve (2) geri bildirim sürecine yardımcı olarak öğrenmeyi teşvik etmek ve / veya öğretimi iyileştirmek.

DPA'nın öğrencilerin cevap ve performanslarına performans ölçütünün uygulanmasını sağlayan yönergeler dizisi olduğunu belirten Miller, Linn ve Gronlund (2009) DPA'nın genellikle performansın sözel tanımlarından veya öğrenci performanslarını ileri düzey, yeterli, kısmen yeterli ve başlangıç seviyesi olarak ayırt etmeyi sağlayan öğrenci cevaplarından oluştuğunu belirtmişlerdir.

Genellikle bir kalite kararı gerektiğinde ve çok çeşitli konuları, faaliyetleri değerlendirmek için kullanılabilen dereceli puanlama anahtarları ile puanlama yapılması, örneklerin değerlendirilmesine rehberlik etmektedir. Verilen bir yazının kalitesine ilişkin puanlamalar, puanlayıcının belirlediği ölçütlere bağlı olarak değişebilir. Bir puanlayıcı, değerlendirme sürecini dil yapısı üzerine yoğunlaşırken diğer bir puanlayıcı doğru bilgi kullanılıp kullanılmadığına daha çok ilgi gösterebilir. Bu nedenle güvenilir bir değerlendirme süreci için önceden tanımlanmış bir plan geliştirerek değerlendirmede yer alan öznellik daha objektif hale gelmelidir (Moskal, 2000).

Kutlu vd. (2014) dereceli puanlama anahtarında bulunan temel elemanları şöyle sıralamıştır:

- “Öğrencinin belli bir göreve ya da çalışmaya bağlı olarak gösterdiği başarı hakkındaki durumunu belirlemede kullanılacak olan bir ya da daha çok sayıdaki ölçütler.”
- “Her bir ölçütün ne anlama geldiğini açıklayan betimsel anlatımlar.”
- “Her bir ölçüt düzeyinde öğrencinin başarısını gösterecek olan dereceler (düzey).”
- “Olanaklı ise her bir düzeyin ne anlama geldiğini tanımlayan betimsel açıklamalar.”

Bu özelliklere sahip DPA’lar kullanıldığında öğrenci başarısının objektif değerlendirmesinin yapılması ve öğrenci performans düzeyinin doğru bir şekilde tanımlanması mümkün olacaktır.

Moskal ve Leydens (2000)’e göre, puan kategorileri iyi tanımlanmış bir dereceli puanlama anahtarı, puanlamanın ne zaman tamamlandığına veya kim tarafından puanlandığına bakılmaksızın tutarlı puanlama yapılmasına yardımcı olmalıdır. İyi tanımlanmış bir dereceli puanlama anahtarının (1) puan kategorilerinin iyi tanımlanmış olması, (2) puan kategorileri arasındaki farklılıkların açık bir şekilde ortaya konması, (3) iki bağımsız puanlayıcının verilen cevaba karşılık aynı puanı verebilmesine olanak sağlaması gerekmektedir. Eğer bu özelliklerden bir tanesinde dahi sorun varsa o dereceli puanlama anahtarı gözden geçirilerek düzeltilmelidir.

Dereceli puanlama anahtarı geliştirme sürecini McTighe (2016) şu şekilde sıralamıştır:

1. Kavranması istenen bilgi veya beceriyi gösteren öğrenci çalışmalarından mümkün olduğu kadar fazla ve çeşitli örnekler seçilerek örnekler toplanır.
2. Öğrenci çalışmaları (güçlü, orta, zayıf vb.) farklı gruplara ayrılır ve gruplara ayırma nedenleri yazılır. Örnekleri gruplara yerleştirme nedenleri için tanımladığınız nitelikler ölçütleri oluşturur.
3. Çalışmanın özelliklerine veya önemli boyutlarına göre nedenler bir araya getirilir.
4. Her özelliğin, değerden bağımsız olarak iyi performansın neye benzediğiyle ilgilenmeyerek özelliğin ne hakkında olduğunu açıklayan bir tanımı yazılır. Özellikle ilgili iyi performans açıklamaları "yüksek" derecelendirmelere yazılmalıdır.
5. Ölçütler ve dereceli puanlama anahtarları kullanımla birlikte gelişir. Puanlama anahtarının iyi çalışmayan kısımlarındaki açıklamaları değiştirerek ya da ekleme yaparak özellikleri gözden geçirerek özetle sürekli iyileştirerek bireylere uygulamaya hazır hale getirilmelidir.

6. Bireylere pilot uygulama yapılarak gereken değişikliklerle dereceli puanlama anahtarına son şekli verilir.

Dereceli puanlama anahtarları, bütünsel (holistik) ve analitik dereceli puanlama anahtarları olmak üzere iki farklı yapıda oluşturulabilmektedir. Bütünsel ve analitik puanlama anahtarlarının her biri genel ya da göreve özel olarak hazırlanabilmektedir.

2.1.2.1. Bütünsel (Holistik) Dereceli Puanlama Anahtarları

Birey performansının tamamının tek bir puan ile değerlendirildiği ve farklı seviyelerdeki performanslarının kalitesinin ortaya koyulduğu puanlama anahtarıdır (Kutlu vd., 2014). Elde edilen ürünü ya da süreci alt ölçütlere ayırmakla beceri ve ölçütleri bağımsız bir şekilde değerlendirmekle ilgilenmeyip ürün ya da süreci bir bütün olarak ele alır (Çepni, 2011). Bu tür puanlama anahtarları, öğrenme ürünlerinin toplam puan olarak değerlendirilmesi istendiğinde ve düzey belirlemeye dayalı değerlendirme yapılırken kullanılmaktadır (Kutlu vd., 2014; Çepni, 2011).

Bütünsel dereceli puanlama anahtarları, bir görevin genel performansını değerlendirmek için uygundur ve tek bir puan verildiğinden daha kolay puanlama yapmaya fırsat vermektedir. Performans düzeyleri bir bütün olarak ele alındığından her ölçüt için öğrencinin performansı hakkında ayrıntılı bilgi vermezler ve öğrencilerin öğrenme eksikliklerini belirlemede kullanılmazlar. Akıcı okuma veya kompozisyon sorulara cevap yazmak gibi basit görevleri ve performansları değerlendirmek, genel kalite veya başarıya ilişkin hızlı bir şekilde sonuç almak ve bir ürünün performansın etkililiğini değerlendirmek için kullanılabilir (Arter & McTighe, 2001, s. 21).

Bütünsel puanlama anahtarları, bir betimsel puanlama ölçeği üzerinde tüm ölçütler aynı anda değerlendirildiğinde ve cevapların en iyi şekilde tanımlandığı seviyede puanlanılması istendiğinde kullanılır. Puanlamanın analitik dereceli puanlama anahtarına göre daha hızlı olması, kısa sürede puanlayıcılar arası güvenirliğe ulaşılabilmesi nedeniyle avantajlıdır ve düzey belirleyici (summative) değerlendirme için oldukça uygundur (Brookhart, 1999, 2013).

Bütünsel puanlama anahtarlarının bazı sınırlılıkları vardır. Bu sınırlılıkların ilki, bütün olarak değerlendirilen birey performansının alt ölçütlere ayrılmaması sonucunda iki bireyin farklı eksiklikleri bulunsa da aynı puanı alabilmeleridir (Arter & McTighe, 2001). İkincisi

ise her ölçüt farklı betimleme derecelerine ayrılmadığı ve ayrıntılı olmadığı için puanlayıcıların bir ölçüte aynı puanı verecekken birbirlerinden farklı bir puan verme eğiliminde olabilmeleridir. Bu durum puanlama anahtarının güvenirliğini düşürmektedir (Haladyna, 1997).

2.1.2.2. Analitik Dereceli Puanlama Anahtarları

Birey performansının farklı ölçütlerdeki başarı seviyelerini, güçlü ve zayıf yönlerini gösteren puanlama araçları olan analitik dereceli puanlama anahtarları (Kutlu vd., 2014) diğer dereceli puanlama anahtarlarına göre daha ayrıntılı ve kapsamlı betimlemeler yapılmasına olanak sağlar (Gronlund, 1998).

Bu puanlama anahtarının her ölçütünün ayrı bir derecelendirmesinin olduğunu ve toplam puan için toplanmış ya da ortalaması alınmış puanlamalarla performansın her bir ölçüt üzerinde ayrı ayrı değerlendirildiği durumlarda kullanılır. Bireylere biçimlendirici geribildirim vermesi, yönergelerle bağlantı kurulmasının bütünsel dereceli puanlama anahtarına göre kolay olması nedeniyle avantajlıdır. Biçimlendirici (formative) değerlendirme için kullanışlıdır ancak bireylerin puanlarına göre sıralanması istendiğinde puanları birleştirerek düzey belirleyici değerlendirme için uygun hale getirilebilir (Brookhart, 2013).

Ürün veya süreci bölümlere ayırmak, her beceriyi veya ölçütü bağımsız olarak değerlendirmek ve sonra da puanları toplayarak toplam puanın hesaplanmasını gerektirir (Çepni, 2011). Analitik puanlama anahtarlarında her ölçüt farklı betimleme derecelendirmeleri kullanarak ayrı ayrı değerlendirilir ve her ölçüte ayrı bir puan verilir. Puanlama yapmak daha fazla zaman alır ancak daha ayrıntılı geribildirim sağlar. Bu puanlama anahtarları, önemli ölçütler içererek karmaşık performansları değerlendirmede ve öğrencilere daha belirgin bilgi veya geri bildirim sağlamada kullaılabilir (Arter & McTighe, 2001, s. 22). Oluşturulması ve kullanılmasının zaman alması ve her ölçüt için her bir puan iyi tanımlanmadığında puanlayıcıların birey performanslarına aynı puanı verememesi bu dereceli puanlama anahtarının sınırlılıklarıdır (Moskal, 2000).

2.1.2.3. Genel Dereceli Puanlama Anahtarı

Farklı görevlerle birlikte kullanılabilen ya da farklı görevlere genelleştirilebilen, aynı öğrenme çıktısına örnek olan bütün görevler için geçerli performans tanımlarının ve ölçütlerinin kullanıldığı dereceli puanlama anahtarıdır (Brookhart, 2013). Bireyler tamamen aynı görevi gerçekleştirmeyecekse, öğrencilerin belirli bir beceri veya ürünle ilgili yetkinliklerini göstermek için hangi bulguların seçileceğine dair bir seçim yapıldığında ya da eğitmenlerin tutarlı bir şekilde karar vermeye çalıştığı farklı kurs bölümlerinde de kullanılabilir (Arter & McTighe, 2001).

Kendi çalışmalarını planlamalarına ve izlemelerine yardımcı olmak için bir görevin başında öğrencilerle paylaşılabilmesi, öğrencileri zaman içinde geliştirdikleri bilgi ve becerilere odaklayarak birçok farklı görevde kullanılabilmesi, öğrenci performansını başarıya giden birçok farklı yol ile tanımlayabilmesi, öğretmenleri öğrencilerin görevi tamamlamaları yerine becerilerini öğrenmelerine ve geliştirmelerine yönlendirmesi nedeniyle avantajlıdır. Göreve özel dereceli puanlama anahtarına göre güvenirliğinin düşük olması ve iyi uygulanması için önceden pratik yapılmasının gerekli olması dezavantajlarıdır (Brookhart & Nitko, 2008).

2.1.2.4. Göreve Özel Dereceli Puanlama Anahtarı

Ölçütlerin bireylerin görev bileşenlerine dayalı olarak oluşturulduğu, bileşenlerin nasıl anlaşıldığını ve bu bileşenler ile bireylerin kendi öğrenmeleri arasındaki ilişkiyi açıklayarak belirli bir görevin belirli içeriğine ait iş tanımlamaları yapan puanlama anahtarıdır (Rosenow, 2014; Brookhart & Nitko, 2008). Gerçeklerin, denklemlerin, yöntemlerin veya kavramların hatırlandığını anlamak ya da bilip bilmediğini kontrol etmek gibi bir bilgi birikiminin hatırlanmasını ve anlaşılmasını değerlendirmeyi amaçlayan birey çalışmalarını notlandırırken kullanmak yararlı olacaktır (Brookhart, 2013; Arter & McTighe, 2001).

Bireylerin belirli bir göreve verdiği cevapta aranacak unsurları detaylandırdıkları için öğrencilerin cevaplarını göreve özel puanlama anahtarı ile puanlamak bireylerin cevaplarını genel puanlama anahtarlarıyla puanlamaktan daha gerçekçi sonuçlar elde etmemizi sağlar. Bu nedenle, büyük ölçekli değerlendirmelerde göreve özel puanlama anahtarı kullanan puanlayıcılar kabul edilebilir puanlama güvenirliği düzeylerine ulaşmak için daha hızlı eğitilmektedir. Ayrıca bu puanlama anahtarının kullanılmasının puanlamayı kolaylaştırdığı da belirtilmektedir (Brookhart, 2013).

2.1.3. Açık Uçlu Maddeler

Açık uçlu maddeler, tek bir cevap yerine birden fazla cevabın doğru olabileceğini gösteren madde türüdür. Bu özellik açık uçlu maddeleri çoktan seçmeli, kısa cevaplı, doğru-yanlış, eşleştirmeli madde türlerinden ayırır. Açık uçlu maddelerin en karakteristik özelliği, bireylerin çözüm yöntemlerine küçük bir sınırlama getirmekle birlikte onların problemlere farklı bakış açılarıyla bakmalarını sağlamasıdır. Bu madde türünün bireylerin düşüncelerini ortaya çıkarmada ve bireylerin konuyu kavrayışları hakkında öğreticilerin daha güçlü bilgiler elde etmesini sağlamada daha iyi olduğu düşünülmektedir. Ancak bu maddelerin hazırlanması, uygulanması ve puanlanması çok zaman gerektirmektedir (Hancock, 1995).

Açık uçlu maddelerde bireyler sadece "ne"yi değil, aynı zamanda "ne zaman, niçin ve nerede"yi açıklayarak daha fazla yargı, neden ve gerekçe üretmek ve düşüncelerini açıklamak zorundadır (Dochy, 2001). Bireylerin problemi hangi yöntem ya da strateji ile çözecekleri ve probleme nasıl bir cevap oluşturacakları konusunda özgür olmaları (Hadenfeldt, Bernholt, Liu, Neumann, & Parchmann, 2013), bireylerin bilgi ve potansiyellerini gösterebilecekleri bilişsel alanın analiz, sentez ve değerlendirme basamaklarındaki üst taksonomik bilişsel düzey davranışları ölçmek için açık uçlu maddelerin kullanılmasına olanak sağlamıştır. Ayrıca sayısal alanda matematiksel dil ve ifade yeteneği ile problemin çözüm yollarını gösterme, kavram haritaları oluşturma, bilgilerini probleme uyarlama beceri seviyelerinin belirlenmesinin yanında sözel alanda bireylerin bir konu hakkındaki görüş ve düşüncelerini yazılı bir şekilde anlatma, yazma, yaratıcı ve özgün düşünme becerilerinin ölçülmesinde bu maddeler kullanılmaktadır. Bu nedenle üst bilişsel davranışların ölçülmesi için oluşturulacak sayısal ve sözel alandaki testlerde açık uçlu maddelerin kullanılması uygundur (Lee, Liu, & Linn, 2011).

Açık uçlu maddelerin kullanılmasının bazı avantajları ve dezavantajları vardır. Açık uçlu maddelerin avantajları aşağıda sıralanmıştır:

- Doğru cevapları şansla seçme durumunu ortadan kaldırırlar, böylece diğer yöntemlerle ilgili ölçme hatalarını azaltırlar. Şans başarısını ve ölçme hatalarını ortadan kaldırdığından ölçme aracının güvenirlik ve geçerliği artmaktadır (Güler, 2014).
- Büyük ölçekli değerlendirmelerde derin ve karmaşık muhakeme yapmaktan yoksun; (Delandshere & Petrosky, 1998; Kennedy, 1999; Lane, 2004) öğretmenleri, öğrencileri izole edilmiş gerçekler, formüller üzerine alıştırma yapmaya teşvik eden

(Frederiksen, 1984; Shepard, 2000) çoktan seçmeli maddeler yerine bireylerin yazma, çizme ve kavram haritalama gibi yollarla cevaplarını yapılandırdığı, bu şekilde aynı zamanda bir üretim gerçekleştirdiği açık uçlu maddelerin kullanılması amaca uygun ölçme yapılmasına yardımcı olacaktır (Rodriguez, 1993).

- Çoktan seçmeli maddelerde olduğu gibi seçeneklerdeki değerleri eleme ya da tek tek deneme yoluyla amaçlanandan daha farklı bir yaklaşımla doğru cevaba ulaşmaya fırsat vermemektedir (Güler, 2014).
- Analiz, sentez ve değerlendirme gibi üst bilişsel davranışların geçerli ölçümlerinin elde edilmesini sağlamaktadır. Ayrıca bireylerin eğitim ve iş hayatında karşılaşabileceği durumlara benzer problemlere açık uçlu maddelerde yer verildiğinde bireylerin problem çözme becerilerinin gelişmesine katkı sağlamaktadır (Birenbaum & Tatsuoka, 1987; Ward, Frederiksen, & Carlson, 1980).
- Verilen maddenin çözümü sırasında başarısız bir girişimde bulunulduğunda doğru cevabı bulmak için bireylere düzeltici geribildirim vermemektedirler. Bu durum çözümde yapılan hatanın birey tarafından fark edilmemesine neden olmaktadır. Çoktan seçmeli maddelerde verilen cevabın yanlış olduğu fark ettiklerinde bireyler yeni bir strateji geliştirerek cevabı bulmaya çalışabilmektedirler (Bridgeman, 1992).

Açık uçlu maddelerin dezavantajları ise şöyle sıralanmıştır:

- Testin cevaplanmasına ayrılan tüm sürenin cevabı düşünmek yerine cevabı yazmak için kullanılması açık uçlu maddelerden oluşan testlerde daha az maddeye yer verilmesine neden olmakta ve bu durum ölçme aracının duyarlılığını, kapsam geçerliliğini düşürmektedir (Özçelik, 1998).
- Maddelere verilen cevapların puanlanmasının objektif olmaması, maddelerin puanlayıcılar tarafından sadece doğru-yanlış (0-1) olarak puanlanması, aynı cevapların farklı zamanlarda puanlanması, puanlayıcıdan puanlayıcıya puanlamanın değişmesi puanlama güvenirliğini düşürmektedir (Güler, 2014).
- Açık uçlu maddelerin uygulama ve puanlama maliyetlerinin fazla olması puanlayıcılar arasındaki tutarsızlıklar nedeniyle eleştirilmektedir (Wainer & Thissen, 1993).

2.1.4. Güvenirlik

Güvenirlik, değerlendirme puanlarının tutarlılığını ifade eder. Örneğin, güvenilir bir testte öğrenci, değerlendirmeyi ne zaman tamamladığını, cevapların ne zaman ve kim tarafından puanlandığını dikkate almaksızın aynı puana ulaşmayı bekler. Güvenilir olmayan bir sınavda bir öğrencinin puanı değerlendirmenin amacı ile ilgili olmayan faktörlere bağlı olarak değişebilir (Moskal & Leydens, 2000).

Linn, Miller ve Gronlund (2009) güvenirliği, bir ölçmeden diğer ölçmeye test puanlarının ve durum belirleme sonuçlarının nasıl tutarlı olduğunu gösteren ölçme tutarlılığı olarak tanımlamıştır. Ölçme sonuçları farklı durumlarda, farklı puanlayıcılarda ve aynı performans alanının farklı örneklerinde makul ölçüde tutarlı olmadıkça sonuçların güvenilir olduğunun söylenemeyeceğini belirtmişlerdir. Güvenirlik, bir ölçümün değişmemiş bir kişiye veya sürece tekrar tekrar uygulandığında aynı değerleri üretme derecesi ya da ölçümlerin tekrarlanabilirliğidir. Ölçme işlemine rastgele hatalar karışmadığında ya da çok az karıştığında güvenilir ölçümler elde edilir (Shrout & Lane, 2012). Güvenirliğin sağlanması için ölçme sonuçlarının duyarlı, tutarlı ve kararlı olması gerekir.

Sınavlarda ve durum belirlemede uygulanan güvenirliğin anlamı şu genel ifadelerle açıklanabilir (Linn vd., 2009):

1. Güvenirlik, durum belirleme dokümanlarına değil, bu dokümanlardan elde edilen sonuçlara işaret eder.
2. Güvenirlik tahmini, her zaman belirli bir tutarlılık çeşidine işaret eder.
3. Güvenirlik, geçerlik için zorunludur ancak yeterli değildir.
4. Güvenirlik öncelikle istatistiksel indislerle değerlendirir.

2.1.4.1. Puanlayıcı Güvenirliği

Karar verme, birçok durum belirleme çeşidinde öğrenci cevaplarını puanlamayı gerektirir. Yazılı sınavlarda bu puanlamaya göre karar verileceği açıktır, bunun yanında açık uçlu matematik problemlerinde ve fen alanındaki laboratuvar uygulamalarında da cevapları puanlayarak karar verilmesi gerekir (Linn vd., 2009). Öğrencilerin performans durumunun puanlayıcı puanlamasına göre belirlenerek öğrencilerin başarılı olup olmadığına karar verilmesi, yapılan puanlamanın güvenilirliğinin kontrol edilmesini gerekli kılmıştır.

Bir puanlayıcının ölçme aracını güvenilir ve tutarlı bir şekilde değerlendirebilmesi için dış etkilerden ve yanlılıktan uzak olması gerekir. Bir puanlayıcının gerçekten güvenilir olup olmaması değerlendirmenin geçerliğini de belirlemektedir (Craig, 2001). Süreçte puanlamanın yapıldığı ortam ile eğitimli ve kalibre edilmiş, benzer geçmişlere sahip puanlayıcılar kullanıldığında değerlendirmenin güvenirlik ve geçerliği kabul edilebilir bir düzeye yükseltilebilmektedir (Cooper, 1997).

Puanlayıcıların puanlamalarındaki tutarlılık, puanlayıcıların güvenirliğinin bir göstergesi olarak kabul edilmektedir. Öğrenci performansına puanlayıcı tarafından bir puanın atanıp aradan belirli bir süre geçtikten sonra yine aynı puanın atanması ya da birden çok puanlayıcının öğrenci performansını aynı puan ile değerlendirmesi puanlayıcıların tutarlı aynı zamanda güvenilir puanlama yaptıklarını göstermektedir. Linn vd. (2009)'e göre, seçilen öğrenci örnekleminden elde edilen performanslar iki veya daha fazla puanlayıcı tarafından puanlanmalıdır. Tutarlılık, aynı niteliğe sahip bir puanlayıcı ile diğer puanlayıcı tarafından atanan puanlarla ilişkilendirilerek ya da öğrenci performanslarının puanlayıcı çiftinden aynı puanı alma oranları hesaplanarak değerlendirilebilir.

Test / tekrar test güvenirliği, eşdeğer form güvenirliği, iki yarı güvenirliği ve rasyonel eşdeğerlik güvenirliği terimlerinden her biri, belirli bir testte veya birden fazla testte öğrenci performanslarının tutarlılığını sağlamak için kullanılan istatistiksel yöntemleri ifade eder. Bu güvenirlik türleri, sınıf içi değerlendirmeden çok, standartlaştırılmış veya bireyler için önemli sonuçları olan testlerde daha endişe vericidir. Bir sınıfta, öğrencilerin bilgisi tekrar tekrar değerlendirilir ve öğretmen, yeni kavrayışlar kazanılırken buna uyum sağlanmasına izin verir. Bu öğrencilerin ölçme sonuçlarının güvenirliğinin belirlenmesi ile puanlayıcıların yaptıkları puanlamaların güvenirliğinin belirlenmesi birbirine benzer durumlardır (Moskal & Leydens, 2000).

Sınıf değerlendirmesinde ve dereceli puanlama anahtarı geliştirmede dikkate alınan iki güvenirlik çeşidi, puanlayıcı güvenirliğini içerir. Puanlayıcı güvenirliği, genellikle, iki bağımsız puanlayıcı tarafından atanan puanlar arasındaki tutarlılık "puanlayıcılar arası güvenirlik"; farklı zamanlarda aynı puanlayıcı tarafından atanan puanlar arasındaki tutarlılık ise "puanlayıcılar içi güvenirlik" olarak tanımlanır (Moskal & Leydens, 2000).

2.1.4.1.1. Puanlayıcılar Arası Güvenirlik

Yapılan ölçmede ölçülmek istenen özelliğe iki ya da daha fazla puanlayıcının birbirinden bağımsız olarak verdikleri puanlar arasındaki uyum ve tutarlılık derecesine puanlayıcılar arası güvenirlik denir (Crocker & Algina, 1986; Güler, 2008; Güler & Gelbal, 2010).

Puanlayıcılar arası güvenirlik başka bir deyişle birey puanının puanlayıcıya göre değişebilme endişesidir. Bir yazılı sınavı analiz etmenin yolu, öğrencilerin cevaplarını okumak ve öğrencilerin yazılı ürünlerinin kalitesine ilişkin yargılarda bulunmaktır. Puanlama sürecine rehberlik etmek için belirlenmiş ölçütler olmaksızın iki bağımsız puanlayıcı belirli bir yanıta aynı puanı veremez. Bu durumda bireyler puanlayıcıların kendi değerlendirme ölçütlerine göre yaptıkları öznel değerlendirmelerine dayanan sınavlardan aldıkları puanlar konusunda endişe duyarlar. Dereceli puanlama anahtarları, her puan seviyesinde aranan ölçütleri belirleyerek endişe verici durumu ortadan kaldırır ve puanlayıcılar arasındaki farklılıkları tamamen ortadan kaldırmasa da tutarsızlıkların oluşumunu azaltabilir (Moskal & Leydens, 2000).

Tinsley ve Weiss (1975) puanlayıcılar arası güvenirliği, birden fazla puanlayıcının yaptığı puanlamalardan elde edilen puan ortalamalarının sapma derecesi olarak tanımlamışlardır. Bu bilgiye dayanarak çok sayıda puanlayıcıya puanlama yaptırılması, her puanlayıcı için puan ortalamalarının hesaplanması ile ölçmeye karışan tesadüfi hataların azalacağı ve puanlayıcı güvenirliğinin artacağı düşünülmektedir (Deliceoğlu, 2009).

2.1.4.1.2. Puanlayıcı İçi Güvenirlik

Anastasi ve Urbina (1997) puanlayıcı içi güvenirliği, bir puanlayıcının farklı zamanlarda aynı bireylerin kağıtlarını birden fazla kez puanlamasından elde edilen verilerin tutarlılığı; Jöreskog (1971) ise aynı puanlayıcının iki farklı zamanda yaptığı puanlamalar arasındaki korelasyon, kararlılık olarak tanımlamışlardır.

Puanlama sürecindeki tutarsızlıklar, birey performanslarındaki gerçek farklılıklar yerine puanlayıcının puanlama sürecinde yorulması, yapılan analizlere zamanla daha az özen göstermesi, o günkü ruh hali ya da puanladığı bireyin kim olduğunu bilmesi sonucunda başarısız bir bireyin doğru yanıtını, başarılı olduğu bilinen bir bireyden alınan özdeş bir cevaptan daha eleştirel olarak analiz etmesi gibi değerlendirme amacının dışındaki faktörler bir puanlayıcının birey performanslarını puanlama biçimini etkileyebilir. Bu durum birey

performansının puanlamasının önceki puanlamadan farklı olmasına neden olabilmektedir. Moskal ve Leydens (2000)'e göre puanlayıcı içi güvenirlik, puanlayıcının puanlama sürecinin zaman içerisinde değiştiği durumların her biridir. Birey performansının güvenilir bir şekilde puanlanması için puanlayıcıların kendilerinden kaynaklanan etkilerinin en aza indirgenmesi gerekmektedir. Bu olumsuz durumun puanlama ölçütlerinin açıklamalarının daha önceden oluşturulduğu iyi tasarlanmış dereceli puanlama anahtarları ile giderilebileceği ve puanlama sürecinde puanlayıcı tutarlılığının devam etmesini sağlamak için puanlama ölçütlerinin tekrar gözden geçirilmesinin faydalı olacağı düşünülmektedir.

2.1.5. Çok Yüzeyli Rasch Modeli

Bireylerin yetenek düzeyleri ile madde güçlük düzeyleri yüzeyleri üzerine geliştirilen Rasch modeline, Linacre (1993) tarafından puanlayıcıların katılığı/cömertliği yüzeyi eklenerek çok yüzeyli Rasch modeli (ÇYRM) oluşturulmuştur. Madde Tepki Kuramı'nın özel bir hali olan bir parametreli Rasch modelinin (Rasch, 1980) bir uzantısıdır ve bir parametreli lojistik model olarak da adlandırılır. Test sonuçlarının analizinde, ölçme işlemine yüzeyleri veya birçok boyutu dahil etmemize olanak sağlar. ÇYRM, bir gözlem oluşturmak için üç yüzeyin (puanlayıcı, puanlanan ve özellikler) etkileşimini ortaya koymaktadır (Linacre, 2002). Genel olarak üç yüzey içeren bu modelin olasılık denklemi logaritmik olarak aşağıdaki gibidir (Linacre, 1993, s. 6):

$$\log\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \beta_n - \delta_i - \lambda_j - \tau_k$$

P_{nijk} : n. bireyin i. maddede j. puanlayıcı tarafından k olarak puanlandırılma olasılığı

P_{nijk-1} : n. bireyin i. maddede j. puanlayıcı tarafından k-1 olarak puanlandırılma olasılığı

β_n : n. bireyin yeteneği

δ_i : i. maddenin güçlük düzeyi

λ_j : j. puanlayıcının katılığı

τ_k : k olarak puanlamadan k-1 olarak puanlamaya geçişin güçlüğü

Ölçme yüzeyi, test geliştiricilerin test puanlarını etkileyebileceklerini düşündüğü ve bu nedenle test geliştirme kısmında araştırma ihtiyacı duydukları ölçme işleminin bir boyutudur. Yüzeylere görevler, puanlayıcı katılığı, puanlama koşulu, madde güçlüğü örnek

olarak verilebilir (Haiyang, 2010). Sudweeks, Reeve ve Bradshaw (2005) da yüzeyi test puanlarını etkileme potansiyeline sahip olan değişkenlik kaynaklarından her biri olarak tanımlamışlardır. Eckes (2015) ise daha genel anlamda yüzeyin test veya değerlendirme puanlarını sistematik bir şekilde etkilediği varsayılan, herhangi bir faktör, değişken veya ölçüm durumunun bir bileşeni olarak tanımlanabileceğini belirtmiştir.

ÇYRM, puanlayıcılar bireyleri test maddeleri üzerinde puanlarken oluşan çeşitliliği kontrol eder. Ayrıca her birey için ham puandan daha adil olan, bireylerin karşılaştığı maddeler ve belli puanlayıcılar için tanımlanan bir ölçü (logit değeri) hesaplar (Linacre, 1993). Bu ölçüler, logit cetvel olarak adlandırılan genelde dikey olan bir cetvel üzerinde yer alırlar. Logit cetvelde tüm yüzeyler pozitif tanımlanmışsa, yer alan yüksek pozitif değerlere sahip ölçmeler, bireyler için yüksek yetenek, puanlayıcılar için cömert puanlama ve maddeler içinse kolaylığı ifade eder; yüksek negatif değerler ise düşük yetenekli bireyleri, katı puanlayıcıları ve güç maddeleri ifade etmektedir (Güler, 2008).

ÇYRM 'nin karşılanması gereken üç özel gerekliliği vardır (Farrokhi, Esfandiari, & Dalili, 2011):

1. Puanlar ve yetenek arasında pozitif bir ilişki olmalıdır. Daha yüksek beceriye sahip öğrencilerin, daha az yetenekli öğrencilere göre daha fazla maddeden daha yüksek puan alma olasılığı daha yüksektir.

2. Yerel bağımsızlık: Öğrencilerin performans veya cevapları, önceki maddelere verdikleri cevaba bağlı olmamalıdır.

3. Tek boyutluluk: Bir testteki tüm maddelerin tek bir değişkeni ya da yapıyı ölçmesi gerektiği fikridir. Bir seferde yalnızca bir gizil veya altta yatan yetenek ölçülür.

ÇYRM bireylerin, belirli puanlayıcıların, görevlerin ve maddelerin ayrıntılarını ortadan kaldırarak her birey için bir ölçü elde etmektedir. Her birey için elde edilen ölçümü, aynı analizde yer alan diğer bireylerin dağılım ayrıntılarından istatistiksel olarak ayrı tutarak her bir bireyi bireysel değerlendirir. Bu modelin amacı, her ham puanı belirli puanlayıcılar, görevler veya bireylerin karşılaştığı maddeler için düzelterek bu puanı lineer olmayan formundan lineer bir ölçüye dönüştürmektir. ÇYRM analizinin sonucu, standart hata ve kalite kontrol uyum istatistikleri ile nitelendirilen her bireyin performans seviyesinin doğrusal(lineer) bir ölçüsüdür. Ayrıca ÇYRM her bireye ait ölçümler için yerel hata varyansını tahmin etmekte ve güvenirlik katsayılarını hesaplamaktadır. Tek boyutluluk varsayımını sağlamasının bir gereği olarak ÇYRM ölçümleri, bir matematik testine daha zor

madde eklenmesi ya da puanlamayı daha cömert puanlayıcıların yapması gibi nitelik açısından benzer nicelik açısından farklı durumlara genellenebilmektedir (Andrich, 1988, s. 9). Stemler (2004)'e göre ÇYRM, puanlayıcı katılığını aynı ölçeği (logit ölçeği) kullanarak kişi yeteneği ve madde güçlüğü olarak türetilmesini sağlar. Maddeleri puanlayan tüm puanlayıcıların katılıklarının yanında her maddenin güçlüğü de doğrudan karşılaştırılabilir.

ÇYRM puanlayıcılar arası güvenirliği tahmin etmede de kullanılır (Linacre, 1994). Yüzeyler yaklaşımı (ÇYRM), her puanlayıcının diğer puanlayıcılarla ilişkili olarak puanlama katılıklarının değerlendirilmesine yardım eden bilgilerin sağlanmasına ek olarak bireysel puanlayıcıların her birinin iç tutarlılığı yüksek olan bir dereceli puanlama anahtarı kullanıp kullanmadığını değerlendirmesine yani puanlayıcılar arası güvenirlik tahmini yapmasına olanak sağlar. Diğer bir deyişle, puanlayıcıların kendi ölçeğini nasıl kullandıklarına dair kendi tanımlamaları farklı olsa da uyum istatistiklerinin, belirli bir puanlayıcının madde ve bireyler üzerindeki ölçek kategorileri tanımlamalarına uygun olup olmadığını belirleyecektir (Stemler, 2004). ÇYRM, puanlayıcı aracılığıyla yapılan değerlendirmeler için bir derecelendirme ölçeği genelleştirmek için tasarlanmıştır (Engelhard, 2013) ve bu model derecelendirme ölçeği kullanılan ölçmelerdeki puanlayıcı hatalarını tanımlamak için de kullanılabilir (Rahman, Ahmad, Yasin, & Hanafi, 2017).

2.1.5.1. Çok Yüzeyli Rasch Modeli Puanlama Desenleri

Puanlayıcılar birtakım görevlerdeki birey performansları için puanlar verdiğinde; tüm puanlayıcıların tüm sınav sonuçlarını değerlendirmelerinin gerekip gerekmediği, puanlayıcılardan alınan bir alt kümedeki her puanlayıcının bireylerin belirli bir alt kümesini puanlamasının yeterli olup olmayacağı, her bir bireyin kaç puanlayıcı tarafından puanlanmasının uygun olacağı, her puanlayıcının kaç tane bireyi puanlaması gerektiği ya da her görev için birey performansının her puanlayıcı tarafından puanlanıp puanlanmayacağı gibi durumlar oluşmaktadır. Ayrıca bireylerin bir alt kümesini puanlayan birkaç puanlayıcıyla birlikte yüzeylerin yani puanlayıcı, birey ve görevlerin tüm ögelerinin aynı referans çerçevesinde temsil edilebilmesini sağlamak için bireylere puanlayıcıların nasıl atanması gerektiği sorusu da ortaya çıkmaktadır. Bu nedenle, puanlayıcı aracılığıyla değerlendirmede, puanlama verisinin toplandığı desene önem verilmelidir (Eckes, 2015).

Linacre ve Wright (2002), ÇYRM modellemesinin puanlama desenlerinin uygulanmasında karşılaşılan hatalara karşı dayanıklı olduğunu ve değerlendirmeden sorumlu kişilerin veri

toplamaya başlar başlamaz ÇYRM analizini başlatabileceklerini belirtmişlerdir. Bu şekilde, puanlama deseninin uygulanmasındaki hatalar veya puanlayıcıların sorunlu davranışları, puanlama işlemi tamamlanmadan tanımlanabilir ve düzeltilebilir (Eckes, 2015).

Belirli bir puanlama deseninin seçiminin, ölçme ve uygulama kavramlarının bir karışımı olduğunu belirten Eckes (2015), aşağıdaki üç duruma vurgu yapmıştır:

- Diğer değişkenler eşit olduğunda, daha fazla veri toplanırsa, model parametrelerinin ölçme hassasiyeti o kadar yüksek olur. Örneğin, sınav yapan kişi başına düşen puanlayıcı sayısı ne kadar fazla olursa sınav sonuçlarının yeterliliği ve görev güçlüğü tahminleri o kadar kesin olur.
- Puanlama deseni, yüzeyler arasında yeterli bağlantı sağlamazsa bireyleri puanlayan puanlayıcılar, ciddi ölçme problemlerine karşı koruma sağlamaz. Bu önemli desen boyutu, ortaya çıkan veri setinin bağlantılılığı ile ilgilidir. Bağlantılı bir veri seti, bir gözlemde yer alan her öğenin aynı değerlendirmenin her bir öğesine doğrudan veya dolaylı olarak bağlandığı bir bağlantı ağının bulunduğu veri kümesidir (Engelhard, 1997; Linacre & Wright, 2002; Wright & Stone, 1979). Belli bir yüzeyin ögeleri arasında bağlantının olmaması, o yüzeyin tüm ögelerini aynı ölçekte kalibre etmeyi imkansız hale getirecektir; yani bu öğeler için elde edilen ölçümler (örneğin puanlayıcı katılığı ölçümleri) aynı ölçekte doğrudan karşılaştırılamayacaktır.
- Birçok değerlendirme aşamasında, özellikle geniş ölçekli değerlendirmelerde, zaman kısıtlamaları, puanlayıcı iş yükü ve bütçe sorunları gibi uygulama hususları puanlama deseninin seçimini büyük ölçüde sınırlandırmaktadır.

"Bir puanlama oturumunda görev alan çok sayıda puanlayıcı olduğunda ve puanlayıcılar arasındaki örtüşme çok az olduğunda, puanlayıcılar arasındaki bağlantıları güçlendirmek için ne yapılabilir?" sorusuna karşılık olarak Myford, Marr ve Linacre (1996), bir puanlama oturumunda görev alan tüm puanlayıcıların birey çalışmalarının küçük bir setini puanlamalarının yanında birey çalışmalarının tümünü standart iş yüküyle puanlamalarını gerektiren bir puanlama deseninin oluşturulmasını önermektedir.

Schumacker (1999), Linacre (2010) ve Eckes (2015) puanlama desenlerini farklı başlıklar kullanarak oluşturmuşlardır. Kullanılan yüzeyler aynı olmakla birlikte Schumacker (1999) çalışmasında desenlerini çapraz desen, yuvalanmış desen ve karma desen olarak isimlendirmiştir. Buna ek olarak Linacre (2010) puanlayıcı çifti desenini oluşturmuştur. Eckes (2015) de tamamen (tam) çaprazlanmış desen, tamamlanmamış-bağlantılı desen,

tamamlanmamış-bağlantılı olmayan desen ve tamamlanmamış-bağlantılı sarmal desen isimlerini kullanmıştır.

Bu çalışmada açık uçlu maddelerden oluşan matematik başarı testinin puanlanmasında Eckes (2015)'in tanımladığı çok yüzeyli Rasch modeli puanlama desenlerinden hangisinin yüksek puanlayıcılar arası güvenirliğe sahip olduğu incelenmiştir. Puanlayıcılar arası güvenirliği belirlemede aşağıda belirtilen puanlama desenleri karşılaştırılmıştır.

2.1.5.1.1. Tamamen (Tam) Çaprazlanmış Puanlama Deseni

Tüm puanlayıcıların tüm bireyleri tüm görevlerde puanladıkları desendir. Tamamen çaprazlanmış puanlama deseni, mümkün olan model parametre tahminlerinin en yüksek hassasiyetine ve tek bir kayıp halka içermeyen bir veri setine yol açtığı için ölçme bakış açısına göre en uygun desendir. Bununla birlikte bu desen gerçek değerlendirme durumlarında nadiren kullanılmaktadır. Aşağıda verilen tabloda her (✓) işareti parametre tahmini için kullanılabilen bir gözlem, yanıt veya puan anlamına gelir (Eckes, 2015).

Tablo 1

Puanlayıcı Aracılığıyla Performans Değerlendirmenin Üç Yüzey İçin Tamamen Çaprazlanmış Puanlama Deseni Gösterimi (Desen A)

		Bireyler									
Puanlayıcı	Görev	1	2	3	4	5	6	7	8	9	10
A. Tamamen (Tam) Çaprazlanmış Desen											
1	1, 2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	1, 2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	1, 2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	1, 2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

"*Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*" Eckes, T., 2015, Frankfurt am Main: Peter Lang kaynağından uyarlanmıştır.

Tamamen çaprazlanmış puanlama deseninden alınan puan blokları, bir başka yalın veri yapısına yerleştirilerek tüm puanlayıcıların bu ortak birey çalışmalarının diğer puanlamalarıyla bağlantılandırılması sağlanır (Myford & Wolfe, 2000).

Çaprazlanmış puanlama deseninin mümkün olan en üst düzeyde bilgi sağladığı ve ölçüm hatalarını azalttığı düşünülürken çoğu test durumunda özellikle geniş ölçekli değerlendirmelerde kullanışlı değildir.

Tablo 2

Çapraz Desen Örneği

	Puanlayıcılar								
	P1			P2			P3		
Görevler	A	B	C	A	B	C	A	B	C
Bireyler									
B1	3	4	5	4	2	5	4	3	5
B2	2	5	4	4	3	5	4	4	3
B3	3	4	5	2	3	4	5	4	3

"Many-facet Rasch analysis with crossed, nested and mixed designs" Schumacker, R.E., 1999, *Journal of Outcome Measurement, 3*(4), 323-338 kaynağından uyarlanmıştır.

Schumacker (1999)'in çalışmasında tüm puanlayıcıların tüm bireyleri tüm görevlerde puanlaması olarak tanımladığı çapraz desen bu desen ile örtüşmektedir.

2.1.5.1.2. Tamamlanmamış-Bağlantılı Puanlama Deseni

Eckes (2015)'in tanımına göre tamamlanmamış – bağlantılı puanlama deseni, bağlantılı bir veri setinin üretilmesinin ölçüm kısıtlamalarını aynı anda gideren tamamlanmamış bir desendir ve bu desen iki farklı şekilde tasarlanabilir. Birinci tasarımda (Desen B), her puanlayıcı yalnızca bazı bireyleri puanlar ve her bir birey dört puanlayıcıdan yalnızca üçü tarafından puanlanır ancak üç yüzeyin tüm ögeleri birbirine bağlıdır. Tablo 3'e bakıldığında 1. puanlayıcı, 2., 4., 6., 8. ve 10. bireylerin ortak puanlanmasıyla 3. puanlayıcı ile bağlantılıdır. Her birey diğer bireylerle (Birey 1 ile Birey 2 veya Birey 3 ile Birey 10) en az iki ortak puanlayıcı tarafından yapılan puanlama ile bağlantılıdır.

Tablo 3

Puanlayıcı Aracılığıyla Performans Değerlendirmenin Üç Yüzey İçin Tamamlanmamış-Bağlantılı Puanlama Deseni Gösterimi (Desen B)

		Bireyler									
Puanlayıcı	Görev	1	2	3	4	5	6	7	8	9	10
B. Tamamlanmamış – Bağlantılı Desen											
1	1, 2	✓	✓		✓	✓	✓		✓	✓	✓
2	1, 2	✓	✓	✓		✓	✓	✓		✓	✓
3	1, 2		✓	✓	✓		✓	✓	✓		✓
4	1, 2	✓		✓	✓	✓		✓	✓	✓	

"Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments" Eckes, T., 2015, Frankfurt am Main: Peter Lang kaynağından uyarlanmıştır.

Eckes (2015)'in oluşturduğu bu desen tasarımı (Desen B) ile Linacre (2010)'nin puanlayıcı çifti puanlama deseni örtüşmektedir. Katılımcı sayısının fazla olduğu sınavlarda puanlayıcı

ya da puanlayıcıların tüm bireylerin tüm performanslarını puanlamasını gerektiren çapraz ve karma desenin kullanılması zor ve zaman alıcıdır.

Tablo 4

Puanlayıcı Çifti Desen Örneği

	Puanlayıcılar								
	P1			P2			P3		
Görevler	A	B	C	A	B	C	A	B	C
Bireyler									
B1	3	4	-	4	-	5	-	3	5
B2	-	2	4	5	3	-	4	-	4
B3	4	-	5	-	2	4	5	4	-

"Many-facet Rasch analysis with crossed, nested and mixed designs" Schumacker, R.E., 1999, *Journal of Outcome Measurement, 3*(4), 323-338 kaynağından uyarlanmıştır.

Tablo 4'e bakıldığında bireylerin her bir görevdeki performansının iki puanlayıcı tarafından puanlandığı görülmektedir. Bireyin bir görevdeki performansını birden fazla puanlayıcının puanladığı puanlayıcı çifti deseni, zaman ve iş yükü açısından çapraz ve karma desene (Desen C ile benzer) alternatif olarak kullanılabilir.

Tablo 5

Puanlayıcı Aracılığıyla Performans Değerlendirmenin Üç Yüzey İçin Tamamlanmamış-Bağlantılı Puanlama Deseni Gösterimi (Desen C)

		Bireyler									
Puanlayıcı	Görev	1	2	3	4	5	6	7	8	9	10
C. Tamamlanmamış – Bağlantılı Desen											
1	1, 2	✓			✓			✓			✓
2	1, 2					✓			✓		✓
3	1, 2			✓			✓				✓
4	1, 2		✓							✓	✓

"Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments" Eckes, T., 2015, Frankfurt am Main: Peter Lang kaynağından uyarlanmıştır.

Tablo 5'e bakıldığında, Birey 10 dışındaki bireyler sadece bir puanlayıcı tarafından puanlanmıştır. Birey 10, dört puanlayıcı tarafından puanlandığı için bağlantılılık durumu korunur. Desen C ile her bir puanlayıcının iş yükünde azalma sağlanır. Her puanlayıcı üç veya dört bireyi puanlamak zorundadır. Tamamen çaprazlanmış desene kıyasla bu desendeki gözlem sayısı % 34 oranında azaltılır (Eckes, 2015). Bu durum desenin kullanışlılığını artırmaktadır.

Tablo 6

Karma Desen Örneği

	Puanlayıcılar											
	P1			P2			P3			P4		
Görevler	A	B	C	A	B	C	A	B	C	A	B	C
Bireyler												
B1	3	-	-	-	2	-	-	-	5	2	4	5
B2	2	-	-	-	3	-	-	-	3	2	3	5
B3	3	-	-	-	3	-	-	-	3	3	4	4

"Many-facet Rasch analysis with crossed, nested and mixed designs" Schumacker, R.E., 1999, *Journal of Outcome Measurement,* 3(4), 323-338 kaynağından uyarlanmıştır.

Schumacker (1999)'in karma desen olarak isimlendirdiği desen Desen C ile örtüşmektedir. Çalışmasında ortak bir logit cetvelin oluşturulması için bağlantı kurmanın yolu, en az bir puanlayıcının bir yüzeyin tüm ögeleriyle çaprazlanması gerektiğini belirtmiş ve iç içe geçmiş desene dördüncü bir puanlayıcı eklendiğinde karma desenin elde edileceğini söylemiştir. Ancak dördüncü puanlayıcı, tüm bireyleri üç görevde de puanlamak zorundadır.

Karma puanlama deseni, yüzeylerin bileşenlerinin tek bir kalibrasyon haritası üzerinde gösterilmesine ve birbirleriyle karşılaştırılmasına yardımcı olsa da puanlayıcılardan birinin tüm öğrencilerin bütün maddelere verdikleri cevapları puanlamasını gerektirdiğinden katılımcı sayısının fazla olduğu sınavlarda kullanışlı görünmemektedir (İlhan, 2016).

2.1.5.1.3. Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni

Tamamlanmamış-bağlantılı olmayan desenin, Desen C'den daha fazla gözleme sahip olmasına rağmen bireylere atanan puanlayıcılar iki ayrık veya bağlantısız alt kümeye ayrılır. Tablo 7'ye bakıldığında Puanlayıcı 1 ile Puanlayıcı 2, birinci alt kümeyi; Puanlayıcı 3 ve Puanlayıcı 4 ikinci alt kümeyi oluşturur. Böyle bir durumda sadece aynı alt kümedeki ölçümler doğrudan karşılaştırılabilir. Puanlayıcı 1 ile Puanlayıcı 3'ün ölçümlerinin katılığını karşılaştırmak yanıltıcı olacaktır. Ayrıca Puanlayıcı 1, Puanlayıcı 3'ten daha düşük puanlar verdiğinde, bunun Puanlayıcı 1'in katılığından veya Puanlayıcı 1 tarafından puanlanan bireylerin ortalama yeterliliklerinin düşük olmasından kaynaklanıp kaynaklanmadığı belirsiz olacaktır. Puanlayıcı katılığı ve birey yeterliliğini incelemek karıştırılacaktır (Eckes, 2015).

Tablo 7

Puanlayıcı Aracılığıyla Performans Değerlendirmenin Üç Yüzey İçin Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni Gösterimi (Desen D)

		Bireyler									
Puanlayıcı	Görev	1	2	3	4	5	6	7	8	9	10
D. Tamamlanmamış - Bağlantılı Olmayan Desen											
1	1, 2	✓	✓	✓	✓	✓	✓				
2	1, 2	✓	✓	✓	✓	✓	✓				
3	1, 2							✓	✓	✓	✓
4	1, 2							✓	✓	✓	✓

"Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments" Eckes, T., 2015, Frankfurt am Main: Peter Lang kaynağından uyarlanmıştır.

2.1.5.1.4. Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni

Bu desen, görevlerin bir alt kümesinde birey performanslarını puanlayan puanlayıcıların iş yükünü azaltan tamamlanmamış-bağlantılı olan puanlama deseninin özel bir şeklidir. Her bir görevdeki performans farklı bir puanlayıcı grubu tarafından puanlandığından (Görev 1, Puanlayıcı 1 ve Puanlayıcı 3 ile; Görev 2, Puanlayıcı 2 ve Puanlayıcı 4 ile puanlanır) bu desene "iç içe geçmiş" ya da "yuvalanmış desen" de denir. Diğer bir deyişle, puanlayıcılar görevler içine yerleştirilir (Eckes, 2015).

Tablo 8

Puanlayıcı Aracılığıyla Performans Değerlendirmenin Üç Yüzey İçin Tamamlanmamış-Bağlantılı Sarmal Puanlama Desen Gösterimi (Desen E)

		Bireyler									
Puanlayıcı	Görev	1	2	3	4	5	6	7	8	9	10
E. Tamamlanmamış –Bağlantılı Sarmal Desen											
1	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

"Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments" Eckes, T., 2015, Frankfurt am Main: Peter Lang kaynağından uyarlanmıştır.

Performans değerlendirme durumlarında yaygın olarak zamana bağlı kısıtlamalar, puanlayıcının dayanıklılığı ve maliyet, puanlayıcıların sadece belirli bir alt birey grubunu puanlamak için görevlendirilmeleri gerektiğini ortaya koymaktadır. Bu durum puanlama desenlerinden iç içe geçmiş (yuvalanmış) desene işaret etmektedir (Brennan, 1992). Aşağıdaki tabloda yuvalanmış desenin farklı bir tasarımı gösterilmiştir.

Tablo 9

İç İçe Geçmiş (Yuvalanmış) Desen Örneği

	Puanlayıcılar								
	P1			P2			P3		
Görevler	A	B	C	A	B	C	A	B	C
Bireyler									
B1	3	-	-	-	2	-	-	-	5
B2	2	-	-	-	3	-	-	-	3
B3	3	-	-	-	3	-	-	-	3

"Many-facet Rasch analysis with crossed, nested and mixed designs" Schumacker, R.E., 1999, *Journal of Outcome Measurement, 3*(4), 323-338 kaynağından uyarlanmıştır.

Tablo 9'a bakıldığında her bir puanlayıcı bireyleri sadece bir görevde puanlamıştır. Puanlamalar her bir görevin içerisinde yuvalanmıştır (Schumacker, 1999). Bu desen puanlayıcı üzerindeki iş yükünü azaltır, maliyeti düşüktür ancak bireylerin yeteneklerini doğru bir şekilde tahmin etmeyebilir (Hombo, Donoghue, & Thayer, 2001).

Taherbai ve Young (2001)'ın çalışmalarında kullandıkları sarmal puanlama deseni tasarımı, Eckes (2015)'in sarmal puanlama deseni tasarımından farklı bir şekilde oluşturulmuştur.

Tablo 10

Sarmal Puanlama Deseni Örneği

	Puanlayıcılar								
	P1 (cömert)			P2 (ılımlı)			P3 (katı)		
Görevler	A	B	C	A	B	C	A	B	C
Bireyler									
B1	3	-	-	-	2	-	-	-	5
B2	-	-	4	4	-	-	-	4	-
B3	-	4	-	-	-	4	5	-	-
B4	3	-	-	-	2	-	-	-	5
B5	-	-	4	4	-	-	-	4	-

"The impact of rater effects on weighted composite scores under nested and spiraled scoring designs using the multifaceted Rasch model" Taherbhai, H. M. & Young, M. J., 2001, *Journal of Outcome Measurement, 5*(1), 819-838 kaynağından uyarlanmıştır.

Her bir birey, her bir görevde farklı bir puanlayıcı tarafından puanlanmaktadır. Üç birey, üç görev ve üç puanlayıcının bulunduğu bir çalışmada, ilk üç bireyin puanlanması için oluşturulan desen tasarımı her üç bireyde tekrarlanarak sarmal bir şekilde puanlama yapılmıştır.

Hombo, Donoghue ve Thayer (2001)'e göre bir puanlayıcının, bireylerin bir alt kümesi için birey performansını sarmal puanlama deseni ile puanlamasıyla elde edilen birey yeteneklerinin tahminlerindeki yanlılık diğer desenlere göre daha azdır. İç içe ve sarmal

desenlerin, puanlayıcı parametrelerini tahmin etmede uygun olmayacaklarını belirtmişlerdir. Çünkü puanlayıcıları ortak bir cetvele yerleştirmek için desenin bir yerinde puanlayıcılar ile puanlayıcılar tarafından puanlanan bireyler örtüşmelidir (Hombo vd., 2001, s. 6).

2.2. Çok Yüzeyli Rasch Modeli Puanlama Desenleri İle İlgili Araştırmalar

Wind ve Jones (2019) farklı türde tamamlanmamış puanlama desenleri kullanıldığında puanlayıcı etkilerinin örtük özellik modeli indekslerinin duyarlılığını araştırmıştır. Çalışmada özellikle üç tür puanlayıcı etkisiyle (puanlayıcı katılığı, puanlayıcı merkezi eğilimi, puanlayıcı cömertliği) puanlayıcı aracılı değerlendirmeler için kullanılabilecek üç tür tamamlanmamış puanlama desenine (puanlayıcıların sistematik bağlantılar kullanarak bağlandığı desenler, puanlayıcıların istasyon tekniğinin uygulandığı performanslar gibi bir grup birey performansı aracılığıyla bağlandığı desenler, puanlayıcıların sarmal bir puanlama ile bağlandığı desenler) odaklanılmıştır. Puanlamalar sistematik bağlantılara, istasyon performanslarına veya sarmal desene dayalı tamamlanmamış bir desen kullanılarak toplandığında puanlayıcı etkilerinin örtük özellik modeli indekslerinin hassasiyetinin, puanlayıcılar farklı oranlarda puanlayıcı etkileri gösterdiğinde bu indekslerin duyarlılığının, farklı puanlama desenleri kullanıldığında bu indekslerin değişiminin, puanlayıcı ve bireylerin örneklem büyüklükleri farklı olduğunda ne ölçüde duyarlı olduğu incelenmiştir. Üç desenin de katı, merkezi ve cömert puanlayıcıları doğru bir şekilde tanımladığı görülmüştür. Puanlayıcı etkilerine duyarlılıkları açısından puanlama desenleri arasında önemli bir fark görülmemiştir. Ranj daralmasına duyarlılığı en yüksek olan desen birey performansı aracılığıyla puanlayıcıların bağlandığı desendir. Herhangi bir göstergenin tamamlanmamış puanlama deseniyle birlikte puanlayıcı etkilerinin belirlenmesi için yeterli bilgi sağlamadığı görülmüştür.

McEwen (2018) yaptığı çalışmada sekiz puanlayıcı tarafından puanlanan 24 makaleden toplanan, önceden var olan tamamen çaprazlanmış puanlama veri kümesinden seksen farklı veri alt kümesi oluşturulmuştur. Bu alt kümeler 20 farklı tamamlanmamış puanlama desenini ve makaleleri puanlamak amacıyla belirlenen dört puanlayıcıyı temsil etmiştir. Tamamlanmamış puanlama desenlerinin çok yüzeyli Rasch modeli sonuçları üzerindeki etkilerini araştırmak için puanlama verilerinin alt kümeleri FACETS yazılımında analiz edilmiştir. Tamamlanmamış tasarımlarda bağlantı çeşitliliği, makale başına düşen puanlayıcı sayısına, alt desen modelinin bir tekrarında puanlanan makale sayısına, tamamlanmamış

desenin bağlantı ağı yapısına ve puanlayıcıların sırasına göre değişmektedir. 80 tane çok yüzeyli Rasch modeli analizinden puanlayıcılar ve makaleler için gözlenen ve düzeltilmiş ortalamalar dahil olmak üzere çeşitli ölçümler, tamamen çaprazlanmış desen için düzeltilmiş ortalamalarla karşılaştırılmıştır. Sonuçlar, desen özelliklerinin ve puanlayıcı sıralarının farklı kombinasyonlarına bağlı olarak büyük ölçüde değişiklik göstermiştir. Makale başına düşen puanlayıcı sayısının en büyük etkiye sahip olduğu ve puanlayıcı düzeninin daha seyrek tasarımlar için daha geniş değer aralıkları ürettiği görülmüştür. Puanlayıcılar ve makaleler için gözlenen ortalamaların birçoğu düzeltilmiş ortalamalara kıyasla özellikle daha seyrek bağlantılı desenlerde, tamamen çaprazlanmış desenden elde edilen sonuçlara daha çok yaklaşmıştır.

İlhan (2016)'nın yaptığı çalışmada çapraz ve puanlayıcı çifti desenlerin çok yüzeyli Rasch modeli analizlerinin sonuçlarını karşılaştırmak amacıyla 168 sekizinci sınıf öğrencisi ve beş puanlayıcı görev almıştır. Çalışma verileri, açık uçlu sorular içeren bir başarı testi ve yanıtları puanlamak için kullanılan bütünsel dereceli puanlama anahtarı ile toplanmıştır. Veri toplama sürecinde ilk olarak öğrencilere başarı testi uygulanmış, öğrenci yanıtları beş puanlayıcı tarafından puanlanmıştır. Puanlamalar çapraz puanlama deseni için veri setini sağlamıştır. Puanlayıcı çifti desende, öğrencilerin test maddelerine verdikleri yanıtlar beş puanlayıcıdan herhangi ikisi tarafından puanlanmıştır. Çok yüzeyli Rasch analizi gerçekleştirilmiş ve iki puanlama deseninin analiz çıktıları karşılaştırılmıştır. Analize dahil edilen yüzey bileşenleri arasındaki farkın çapraz desende puanlayıcı çifti desenine göre daha etkili olduğu belirlenmiştir. Bununla birlikte çalışmanın, puanlayıcı çifti desende yüzey bileşenleri arasında önemli ayrımlar yaptığı görülmüştür. Çapraz ve puanlayıcı çifti desenlerine göre hesaplanan yetenek tahminleri arasındaki göreceli uyuşmanın çok yüksek olduğu ve iki desen tarafından hesaplanan yetenek tahminleri arasında önemli bir fark olmadığı bulunmuştur. Bu nedenle, daha az maliyetli olacağı için, çok sayıda katılımcının bulunduğu sınavlarda puanlayıcı çifti deseninin seçilmesinin daha uygun olacağı anlaşılmıştır.

Taherbai ve Young (2001) yaptığı çalışmada öğrencilerin yetenek tahminlerini, farklı puanlama desenleri ve farklı ağırlıklı bileşik puanlardaki gerçek yetenekleri ile karşılaştırmayı amaçlamıştır. Üç açık uçlu görev ve yirmi çoktan seçmeli maddeden oluşan Yeni Standartlar İngiliz Dili Sanatları Sınavı'nın (ELA) Okuma: Temel Anlayış bölümünden elde edilen verilerini, iç içe geçmiş ve sarmal puanlama desenlerini kullanarak puanlama yapan puanlayıcı etkilerinin bileşik puanlar ile etkileşimini incelemek için kullanmışlardır.

Açık uçlu maddeler sıfırdan dörde kadar derecelendirilen dereceli puanlama anahtarı ile puanlanmıştır. Analizler, çok yüzeyli Rasch modeli kullanılarak FACETS programı ile yapılmıştır. Sonuçlar, puanlayıcı etkilerinin ağırlıklı bileşik puanlarla etkileşiminin, öğrenci yetenek tahminlerini önemli ölçüde değiştirebileceğini göstermiştir. İç içe geçmiş desende yetenek dağılımının sonundaki bireylerle çok katı ya da cömert puanlayıcılar eşleştirildiğinde, sınavın yetenek tahminlerinin sistematik olarak bozulabileceği, ağırlıklandırılan madde ve görevlerin gerçek yetenek tahminlerinin iyileştirilmesini daha da karmaşık hale getirebileceği görülmüştür. Bununla birlikte, açık uçlu görevlerin çoktan seçmeli maddelere kıyasla çok daha fazla olan bileşik puanları kullanıldığında sarmal desenin yetenek tahminlerinin yanlılığını azalttığı sonucuna ulaşılmıştır.

2.3. Çok yüzeyli Rasch Modeli ve Puanlayıcılar Arası Güvenirlikle İlgili Çalışmalar

Çetin (2019) tarafından yapılan çalışmada açık uçlu biyoistatistik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi amaçlanmıştır. Araştırmanın çalışma grubunu Tokat Gaziosmanpaşa Üniversitesi'nde öğrenim gören Biyoistatistik dersini alan 100 lisans öğrencisi oluşturmuştur. Veri toplama aracı olarak Biyoistatistik eğitimi alanına ilişkin literatürdeki açık uçlu çalışmalar incelenerek 12 adet açık uçlu sorudan oluşan sınav hazırlanmıştır. Açık uçlu soruların daha önceden belirlenen puanlama kriterlerine göre 1-5 arasında puanlayan (1: çözüm veya seçilen strateji tamamıyla yanlış, 2: çözüm veya seçilen strateji yanlış, 3: çözüm ve seçilen strateji kısmen doğru, 4: çözümün büyük bir kısmı doğru ve seçilen strateji uygun, 5: çözüm tamamıyla doğru ve seçilen strateji uygun) bağımsız iki puanlayıcı tarafından değerlendirilmesiyle veriler elde edilmiştir. Öğrencilerin problem çözme sınavından aldıkları puanlar iki farklı puanlayıcının değerlendirmeleri sonucunda çok yüzeyli Rasch analizi modeli ile incelenmiştir. Puanlamadan önce puanlama kriterleri belirlenmiş olmasına rağmen puanlayıcılar arasında puanlama sürecinde farklılıklar oluşmuştur. Sınavda yer alan bir soru haricinde 1. puanlayıcının tüm sorularda 2. puanlayıcıya göre daha hoşgörülü davrandığı tespit edilmiştir. Ayrıca puanlayıcılara göre problem çözme testinde yer alan soruların en kolay ve en zor olanlarının değişkenlik gösterdiği belirlenmiştir.

Doğru (2019) tarafından yapılan çalışmada, karma testlerin psikometrik özelliklerini belirlemede Klasik Test Teorisi ve Rasch modelinin benzerlik ve farklılıklarını karşılaştırarak hangisinin daha ayrıntılı istatistiksel bilgiler sağladığını göstermek istemiştir.

Çalışmada ABİDE uygulama sınavına katılan 11193 öğrenci cevabı veri seti olarak kullanılmış; klasik test teorisi ve Rasch modeline göre madde ve test parametreleri Microsoft Excel ve Winsteps ile hesaplanmıştır. ABİDE sınavının matematik ve fen alt testlerinin her biri, dokuz çoktan seçmeli ve dokuz açık uçlu soru içeren on sekiz maddeden oluşmaktadır. Analizler sonucunda madde güçlüklerinin tespitinde iki kuram benzer sonuçlar göstermiştir. Güvenirlik ve geçerlik açısından iki kuram da fen alt testinde aynı problemli maddeleri işaret etmiştir. Rasch modelinin matematik alt testinde KTK'ya ek olarak bir maddeyi daha problemli olarak işaret ettiği görülmüştür. İki kuramda da test güvenirliğinin aynı olduğu; Rasch modelinin KTK'ya göre birey ve madde yüzeyi ile ilgili daha ayrıntılı istatistiksel bilgi sağlamıştır. Bu durum geniş ölçekli değerlendirmelerden istatistiksel bilgi elde edilirken Rasch modelinin klasik test kuramına iyi bir alternatif olabileceği anlaşılmıştır.

Fernandez ve Siddiqui (2017) Pakistan'ın güneyindeki Punjab eyaletinde lise diploması almak isteyenler için 12. sınıfta ulusal düzeyde yapılan sınavda İngilizce makaleleri değerlendiren puanlayıcıların değerlendirme ölçütlerini araştırmıştır. Çalışmaya üç sınav kurulunun her birinden beş puanlayıcının seçilmesiyle oluşan 15 puanlayıcı ve 30 öğrenci katılmıştır. Puanlayıcıların her biri üç makaleyi ayrı ayrı puanlamıştır. Çalışmanın nicel verileri, 15 puanlayıcı tarafından değerlendirilen üç makale setinin puanlarından; nitel veriler ise yarı yapılandırılmış görüşmelerden ve denemelerdeki öğrenci puanlarını rasyonelleştirmek için puanlayıcıların not ettiği kısa yazılı yorumlardan elde edilmiştir. Puanlayıcıların performansını (puanlayıcı içi ve puanlayıcılar arası güvenilirlik açısından) ve makale zorluğunu tahmin etmek için FACETS (Linacre & Wright, 1993) kullanılarak çok yüzeyli Rasch modeli analizleri gerçekleştirilmiştir. Bu analizler sonucunda puanlayıcıların puanlama tutarlılığı ve gösterdikleri katılıklar arasında farklılıklar olduğu saptanmış; değerlendirme ölçütlerinde puanlayıcıların verdikleri puanları doğrulayan görüşmelerin ve yorumların analizi, dil bilgisi, hatalara karşı tutum, el yazısı, uzunluk, yaratıcılık, organizasyon ve cümleleri bağlayan kelimelerin kullanımı açısından büyük ölçüde değişkenlik olduğu görülmüştür.

Rahman, Ahmad, Yasin ve Hanafi (2017), Malezya Meslek Yüksekokulu'ndaki Elektronik Devre Tasarımı bölümü öğrencilerinin yeterliliklerini analitik dereceli puanlama anahtarı kullanarak değerlendiren öğretmenlerin merkezi eğilim etkisine odaklanması sonucu oluşan değerlendirici hatasını araştırmıştır. Elektronik Teknoloji: Elektronik Devrelerin Uygulanması dersi bulunan üç meslek yüksekokulunun her birinden üç öğretmen puanlayıcı olarak seçilmiştir. Çalışmanın örneklemini dokuz öğretmen ve 68 öğrenci oluşturmaktadır.

Uygulamaya başlamadan önce dereceli puanlama anahtarının dört dereceden (1: zayıf, 2: orta, 3: iyi, 4: mükemmel) oluştuğu ve bu puanlama anahtarı ile 14 görev tanımının puanlanacağı puanlayıcılara kısaca tanıtılmıştır. Toplanan veriler, çok yüzeyli Rasch modeli analizlerini gerçekleştiren FACETS programında yapılmıştır. Analiz sonuçları bireysel ve grup düzeyinde elde edilmiştir. Grup düzeyinde, ölçek kategori istatistiklerinden elde edilen bilgiler merkezi eğilim etkisini göstermiştir; ancak, ayırma istatistiklerinin hiçbiri böyle bir etki göstermemiştir. Bireysel düzeyde ise iki puanlayıcını merkeze yöneldiği görülmüştür.

Salmaner (2015) yaptığı çalışmada zaman içerisindeki değişimlere dikkat ederek öğrencilerin öz ve akran puanlamalarını öğretmen ya da uzman puanlamalarıyla karşılaştırılarak incelenmesini amaçlamıştır. 60 öğrenci ve üç puanlayıcı ile öğrencilerin yazma becerilerinin değerlendirilmesi amacıyla beşinci sınıf Türkçe dersinde beş haftalık bir değerlendirme programı uygulanmıştır. Araştırmacı tarafından görsel okuma ve yaratıcı yazma konularında yedi tane sınırlandırılmış yazma görevi hazırlanarak öğrencilerin yazma becerileri ölçülmek istenmiştir. Yazma görevlerinin öz, akran ve öğretmen değerlendirmelerinde kullanılmaları için analitik ve göreve özel dereceli puanlama anahtarları oluşturulmuştur. Öğrenciler ilk hafta yazacakları yazılar, dereceli puanlama anahtarları ile öz ve akran değerlendirmeler hakkında bilgilendirilmiş, örneklerle anlaşılmayan noktalar tamamlanmıştır. İkinci hafta ve sonrasında önceki haftanın tekrarı yapılarak konuyla ilgili beyin jimnastikleri yapılmıştır. Daha sonra öğrencilere görev ve yönergelere uygun bir şekilde konuyla alakalı yazılar yazmaları için 15-20 dakika süre verilmiştir. Süre bittiğinde öğrencilerden öz değerlendirme yapmaları için dereceli puanlama anahtarındaki kategorileri ve kategorilerin derecelerine ait tanımları anlamaları, 10 dk içerisinde kendi yazılarını okuyarak derecelendirmelerden uygun olanları işaretlemeleri istenmiştir. Akran değerlendirme için yanlılığı önlemek amacıyla sadece okul numaraları yazılı olan kağıtlar karıştırılarak sınıfa dağıtılmış ve 15 dakika süre verilmiştir. Son haftada öğrencilerden uygulamanın başlangıcından sonuna kadar geçen süreçte öğrencilerin yazma ve değerlendirme konularında kendilerinde gerçekleşen değişimleri kompozisyon olarak yazmaları istenmiştir. Yazılan kompozisyonlardan bazıları okunduktan sonra öğrencilerden puanlama anahtarına uygun olarak kompozisyonları değerlendirmeleri sağlanmıştır. Yapılan etkinlikler üç puanlayıcı tarafından puanlanmıştır. Analizler sonucunda birinci uygulamada öğretmenlerin en katı değerlendirmeyi yaptığı son uygulamaya kadar cömertleşerek orta seviyeye ulaştığı görülmüştür. Öz değerlendirmeler, birinci ve ikinci uygulamada katılık düzeyinin anlamlı bir düzeyde değişmemiş, üçüncü uygulamada ise daha katılaşmıştır.

Akran değerlendirmeler, birinci uygulamada en cömert, ikinci uygulamada en katı, üçüncü uygulamada ise orta düzeye yaklaşmıştır. Tüm değerlendirme türleri birbirlerine yakınlaşarak orta düzey katılığa ulaşmıştır.

İlhan (2016) yaptığı çalışmada, açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılmasını amaçlamıştır. Araştırma, sekizinci sınıfta öğrenim gören 100 öğrenci ve çalışmada puanlayıcı olarak görev alan ilköğretim matematik öğretmenliği alanından mezun dört matematik öğretmeninin katılımıyla gerçekleşmiştir. KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimleri öğrencilerin açık uçlu matematik sorularındaki performansları esas alınarak elde edilmiştir. Çalışmanın verileri açık uçlu altı sorudan oluşan bir matematik başarı testi ve bu testteki soruların puanlanmasında kullanılan bütüncül bir rubrik (dereceli puanlama anahtarı) yardımıyla toplanmıştır. Araştırmada kullanılan rubrikte; *Yetersiz* (1), *Geliştirilmesi Gerek* (2), *İyi* (3) ve *Çok İyi* (4) şeklinde dörtlü bir derecelendirme kullanılmıştır. Puanlama işlemi öncesinde, puanlayıcılara kullanılacak rubrik tanıtılmış; puanlama sırasında dikkat edilmesi gerekenler açıklanarak kısa bir eğitim verilmiştir. Öğrencilerin birinci maddeye verdikleri cevapları değerlendirirken, tüm kâğıtları tek oturuşta puanlamaları, bütün öğrencilerin birinci maddedeki performanslarını puanladıktan sonra, ikinci maddeye ilişkin puanlamalara hemen başlamayıp bir süre ara vermeleri önerilmiştir. Veriler KTK'da; puanlayıcı güvenirliğinin belirlenmesinde puanlayıcılar arası korelasyon katsayısı ve tekrarlı ölçümler için varyans analizinden yararlanılmış; öğrencilerin matematik performanslarına ait yetenek kestirimlerini elde etmek amacıyla dört puanlayıcı tarafından verilen puanlamaların ortalaması alınmıştır. Daha sonra ÇYRM'nin varsayımları olan tek boyutluluk, yerel bağımsızlık ve model-veri uyumu test edilmiştir. Yetenek kestirimleri arasındaki göreli uyumu belirlemek için Pearson Momentler Çarpımı korelasyon katsayısından, mutlak uyumu saptamak amacıyla ilişkili örneklem *t*-testinden yararlanılmıştır. Hesaplanan yetenek kestirimlerinin öğrencilerin matematik karne notları ve TEOG sistemi kapsamında uygulanan merkezi matematik sınavındaki doğru sayıları ile arasındaki korelasyonlara bakılarak, iki kurama göre hesaplanan yetenek kestirimlerinin ölçüt geçerliği karşılaştırılmıştır. KTK'ya ilişkin analizlerin gerçekleştirilmesinde, iki farklı kurama göre elde edilen yetenek kestirimlerinin göreli ve mutlak uyumu ile ölçüt geçerliğinin karşılaştırılmasında SPSS paket programı, Çok yüzeyli Rasch analizi için FACETS paket programı kullanılmıştır. Araştırma sonucunda; KTK ve ÇYRM'ye göre hesaplanan yetenek kestirimleri arasındaki göreli uyumun oldukça yüksek olduğu

belirlenmiştir. İki kurama göre hesaplanan yetenek kestirimlerine ait ortalamalar arasında anlamlı fark bulunduğu ve mutlak bir uyumun olmadığı saptanmıştır. ÇYRM'de rapor edilen yetenek kestirimlerinin ölçüt geçerliğinin KTK'dan elde edilen yetenek kestirimlerine oranla daha yüksek olduğu tespit edilmiştir.

Farrokhi, Esfandiari ve Dalili (2011), öz değerlendirme, akran değerlendirme ve öğretmen değerlendirmesinde merkezi eğilim etkisini saptamak için yeni geliştirilen bir ölçüm modeli olan çok yüzeyli Rasch modeli (ÇYRM) kullanmayı amaçlayan bir çalışma yapmıştır. Yazma becerilerini değerlendirmede yaygın olarak kullanılan altılı likert tipi analitik dereceli puanlama anahtarı ile İran'daki üniversitelerde İngilizce anabilim dalları tarafından yazılan 188 denemeyi değerlendirmek için İngilizce anabilim dalı öğrencilerinden 188'i öz ve akran değerlendirici ve altısı öğretmen değerlendirici olmak üzere 194 değerlendiriciden faydalanılmıştır. Analitik dereceli puanlama anahtarı, beş paragraflık denemenin sahip olması gereken temel özelliklerinin yer aldığı 15 maddeden oluşturulmuştur. Çalışma, farklı bir İngilizce Yabancı Dili (EFL) ortamında İran'da yürütülmüştür. Önceki araştırmalardan farklı olarak, oldukça tecrübesiz değerlendiriciler kullanılmış ve daha önceki çalışmalarda yer almayan tamamen çaprazlanmış desene göre puanlama yaptırılmıştır. ÇYRM analizinin sonuçları, öz, akran ve öğretmen değerlendirmesinin bireysel ve grup düzeyinde merkezi eğilim etkisi göstermediği görülmüştür. Öz ve akran değerlendirmenin öğretmen değerlendirmesi yerine kullanılabileceği sonucuna ulaşılmıştır.

Haiyang'ın (2010) yaptığı çalışma, bilgi teknolojileri, ekonomi ve mühendislik anabilim dallarından yaşları 23-30 arasında değişen İngiliz olmayan, ulusal sınavlara girmiş ve bu sınavlardaki İngilizce puanları 100 üzerinden 55'in üzerinde olan 56 lisansüstü öğrencisine uygulanmıştır. Bu çalışma KTK ve ÇYRM 'ye dayanılarak İngilizce anabilim dalı mezunu olmayanlar için uygun nitelikteki standart olmayan bir İngilizce testinin güvenirliğini analiz etmek; KTK ve ÇYRM 'nin test puan analizindeki tamamlayıcı rollerini göstermek amacıyla yürütülmüştür. Dinleme, anlama, kelime bilgisi ve okuma anlama çoktan seçmeli sorular ile İngilizce-Çince ve Çince-İngilizce çeviri ve yazma iki puanlayıcı tarafından onlu likert tipi bütünsel dereceli puanlama anahtarı ile değerlendirilmiştir. KTK için puanlayıcılar arası güvenirlik, verilerin aralıksız olması ve Cohen'in kappa veya Pearson korelasyonunun kullanılmasına uygun olmaması nedeniyle iki puanlayıcının her bir görevdeki Spearman rho (ρ) değeri hesaplanarak; iç tutarlılık güvenirliği ise görevlerin Cronbach alfa (α) değeri hesaplanarak değerlendirilmiştir. Bu çalışmada, bireyler, puanlayıcılar ve görevler yüzeyleri arasındaki tüm iki yönlü etkileşimler için yanlılık analizleri gerçekleştirilmiştir. Analizler

sonucunda, geleneksel KTK güvenilirlik çalışmasının ve ÇYRM 'nin test güvenirliği analizinde oynadığı tamamlayıcı rolleri ortaya koyulmuştur. KTK güvenirlik çalışması, testin ve taraflar arasındaki görev düzeylerinin tutarlılığını göstermekte; ÇYRM ise tutarsızlığın veya uyuşmazlığın nerede olduğu ve neden kaynaklandığı hakkında ayrıntılı bilgi vermektedir. Bu nedenle her bir analizin bilgilerinin diğerini tamamlamak için kullanılabilineceği sonucuna ulaşılmıştır.

Myford ve Wolfe (2004), bireysel seviyelerde merkezi eğilim etkisini belirlemek için uyum ve uyumsuzluk ortalama karelerine, grup seviyesinde ise sabit ki kare testi, ayrılma oranı, ayrılma indeksi, ayrılma indeksinin güvenirliği ve maddeler için uyum indislerine bakılması gerektiğini belirtmiştir. Bu bilgiye dayanarak yapılan FACETS analizinin sonuçları, öz değerlendirmenin, akran değerlendirmenin ve öğretmen değerlendirmesinin grup düzeyinde ya da bireysel düzeyde herhangi bir merkeziyet belirtisi göstermediğini göstermiştir. Öğretmenlerin zorlandıkları ve öğrencilerin ödevlerini puanlamak için yeterli zamana sahip olmadığı sınıflarda, öğretmenlerle birlikte hem öz değerlendiriciler hem de akran değerlendiricilerden puanlama yapmak için yararlanılabilineceği ortaya çıkmıştır.

BÖLÜM III

YÖNTEM

Bu bölümde araştırmanın türü, çalışma grubu, veri toplama araçları, verilerin toplanması ve analizi hakkında bilgi verilmiştir.

3.1. Araştırmanın Modeli

Çok Yüzeyli Rasch Modeli puanlama desenlerinden hangisinin matematik alanında hazırlanan başarı testinin puanlanmasında daha güvenilir sonuçlar verdiği ortaya koymaya çalışıldığından bu çalışma betimsel bir araştırma niteliğindedir. Betimsel araştırma yöntemlerinden tarama modeli kullanılarak mevcut durum hakkında değerlendirmeler yapılmaya ve incelenen durum aydınlatılmaya çalışılmıştır (Karasar, 2020).

3.2. Çalışma grubu

Çalışma grubunda üç alt grup bulunmaktadır. Birinci grup 2017-2018 eğitim öğretim yılında Ankara ili Yenimahalle ilçesinde bulunan devlet okullarının 8. sınıfında öğrenim gören 15 öğrenciden oluşan pilot uygulama grubu, ikinci grup 76'sı kız 46'sı erkek toplam 122 öğrenciden oluşan uygulamanın yapıldığı grup ve üçüncü grup ilköğretim matematik öğretmenliği mezunu biri erkek ve üçü kadın olmak üzere MEB eğitim kurumlarında çalışan dört matematik öğretmeninin yer aldığı puanlayıcı grubudur. Açık uçlu maddelerin puanlanmasında en az iki en fazla beş puanlayıcının görev alması (Turgut & Baykul, 2012), Rasch analizinin parametre tahminleri için de katılımcı sayısının 100 ile 200 arasında olması yeterli bulunmaktadır (DeMars, 2010, s. 34).

3.3. Veri Toplama Aracı

Araştırmacı tarafından hazırlanan 8. sınıf kazanımlarını içeren dokuz açık uçlu madde ile gerekli kullanım izinleri alınarak PISA 2012 Pilot Uygulaması matematik testinde yer alan iki açık uçlu maddenin de eklenmesiyle 11 madde ve iki bölümden oluşan matematik başarı testi uygulanmıştır. Uygulama sonrası özel bütünsel dereceli puanlama anahtarı kullanılarak 122 öğrencinin 11 maddeye verdikleri tüm cevaplar dört puanlayıcı tarafından puanlanmıştır. Özel bütünsel dereceli puanlama anahtarı her bir maddeye özel tanımlamaların yer aldığı beş düzeyden oluşmuştur. Puanlayıcılara güvenilir puanlama yapmaları için açık uçlu maddeler ile özel bütünsel dereceli puanlama anahtarının içeriği hakkında bilgilendirme yapılmıştır.

3.3.1. Açık Uçlu Maddelerden Oluşan Matematik Başarı Testinin Hazırlanması

Matematik başarı testinde yer alacak açık uçlu maddelere ilişkin belirtke tablosu oluşturulmuştur. Belirtke tablosuna uygun olarak hazırlanan açık uçlu maddelerin kazanımları ölçüp ölçmediğinin, sınıf düzeyine uygunluğunun ve anlaşılabilirliğinin incelenmesi amacıyla uzmanlardan bu maddeleri "Uzman Değerlendirme Formu"nu doldurarak değerlendirmeleri ve düzeltilmesi gereken yerler varsa bu kısımların nasıl düzeltilmesi gerektiğini ifade eden önerilerini belirtmeleri istenmiştir. Uzman görüşleri sonucunda açık uçlu maddelerin sınıf seviyelerine uygun olduğu ve belirtke tablosunda belirtilen kazanımları ölçmede iyi hazırlanmış maddeler oldukları konusunda görüş birliğine varılması ile başarı testinin kapsam geçerliğinin sağlandığı görülmüştür. Ayrıca açık uçlu maddelerin öğrenciler tarafından anlaşılabilirliğini, kitapçık üzerinde bulunan test yönergesinin yeterli bilgi sağlayıp sağlamadığının, maddelerin çözümüne ayrılan alan ve süre yeterliliğini kontrol etmek amacıyla 8. sınıfta öğrenim gören 15 öğrenciye pilot uygulama yapılmıştır. Maddelerin kitapçığa öğrencilerin rahatça çözüm yapabilmeleri için yeterli alanlar oluşturacak şekilde yerleştirilmesine dikkat edilmiştir. Başlangıçta açık uçlu maddelerin tamamının çözümü için verilen süre 40 dakika olarak düşünülmüş ön uygulama sonrasında ise maddelerin tamamının çözümü için verilen sürenin yeterli olmadığı görülmüştür. Bunun sonucunda matematik başarı testi ilk beş madde birinci bölüm, sonraki altı madde ise ikinci bölüm olacak şekilde iki oturum halinde ve her oturum için verilen süre 40'ar dakika toplamda 80 dakika olarak belirlenmiştir. Değerlendirmeler ve pilot uygulama

sonucunda maddelerde bulunan eksikler düzeltilerek başarı testindeki açık uçlu maddelerin uygulamaya hazır hale gelmesi sağlanmış, başarı testinin uygulanışı sırasında ortaya çıkabilecek sorunlar giderilmiştir.

3.3.2. Açık Uçlu Maddeler için Özel Bütünsel Dereceli Puanlama Anahtarının Hazırlanması

Bütünsel dereceli puanlama anahtarları düzey belirlemeye dayalı değerlendirme amaçlı kullanıldığından bireyler hakkında çok ayrıntılı bilgi vermezler ve bireylerin eksik yönlerinin belirlenmesinde kullanılmazlar. Birey performansının hangi duyarlılıkta belirleneceğine ve bireyin bulunduğu sınıf düzeyine bağlı olarak 3, 4 veya 5 düzeyden oluşan DPA'lar kullanılabilir (Kutlu vd., 2014). Bu çalışmada açık uçlu maddelerin puanlanması için beş düzeyden oluşan özel bütünsel dereceli puanlama anahtarı hazırlanmış ve kullanılmıştır.

Bütünsel dereceli puanlama anahtarı oluştururken Kutlu vd. (2014) tarafından önerilen aşağıdaki basamaklar dikkate alınmıştır:

1) Bireylerin gösterdikleri performansın yüksekten en düşüğe doğru sıralanıp her düzeye bir puan verilmelidir. Düzeyler mükemmel, iyi, orta, düşük gibi sözel ifadelerle de nitelendirilebilir.
2) Birey performansı her düzeyde tanımlanmalıdır. Birey çalışması bir bütün olarak düşünülmeli ve performans tanımlamaları bireyden beklenen davranışlar düşünülerek yetkinden basite doğru yapılmalıdır. Her bir düzey, farklı bir başarıyı ifade edecek şekilde yazılmalı ve puanlayıcı çalışmanın hangi başarı düzeyinde yer aldığına karar vermekte zorlanmamalıdır.
3) Performans tanımlamaları yapıldıktan sonra dereceli puanlama anahtarı kullanılabilir. Puanlayıcıların puanlama anahtarını kullanırken karşılaştıkları sorunlar sonucunda güvenilir, geçerli ve kullanışlı olmaları için bu anahtarlar geliştirilmelidir.

Araştırmacı tarafından hazırlanan özel bütünsel dereceli puanlama anahtarında derecelendirme düzeyleri şu şekilde ifade edilmiştir:

"4" : Tam ve doğru cevap verilmesi.

"3" : Yapılan işlemlerin ve sonucun doğru olup sonuca ulaştıran işlemlerde yeterli açıklamanın yapılmaması.

"2" : Çözüm yolu doğru olup temel işlemlerde hata yapılması nedeniyle doğru sonuca ulaşamama ya da uygun çözümle başlayıp doğru işlemler yapıp küçük hatalar dolayı doğru sonuca ulaşamama.

"1" : Herhangi bir işlem yapmadan doğru cevabın yazılması.

"0" : Boş ya da tamamen yanlış cevap verilmesi.

Matematik başarı testinde yer alan açık uçlu maddelerdeki öğrenci başarısının durumunu güvenilir bir şekilde belirleyebilmek amacıyla her maddeye özel yukarıdaki derecelendirme düzeylerini kullanarak taslak bir bütünsel dereceli puanlama anahtarı oluşturulmuştur. Taslak dereceli puanlama anahtarının eksik ve değiştirilmesi gereken yönlerini öğrenmek amacıyla altı uzmandan "Uzman Değerlendirme Formu" ile puanlama anahtarını değerlendirmeleri istenmiştir. Bunun yanında pilot uygulamada 15 öğrencinin açık uçlu maddelere vermiş olduğu cevaplar doğrultusunda bazı maddelere ait diğer çözüm yolları eklenerek özel bütünsel dereceli puanlama anahtarı güncellenmiştir.

3.3.3. Ölçme Araçlarına İlişkin Geçerlik Güvenirlik Çalışması

Açık uçlu maddelerin ve bu maddelere ait özel bütünsel dereceli puanlama anahtarının kapsam ve yapı geçerliğinin kontrolü için Lawshe (1975) tekniğine uygun olarak (3) *uygun,* (2) *düzeltilmeli,* (1) *tamamen düzeltilmeli* şeklinde 3'lü likert tipi derecelendirmeden oluşan açık uçlu maddeler için sekiz, dereceli puanlama anahtarı için yedi ölçütün yer aldığı "Uzman Değerlendirme Formu" kullanılmıştır. Bu teknikte en az 5 en fazla 40 uzmandan görüş alınması önerilmektedir. Biri ölçme değerlendirme ve matematik alan uzmanı araştırma görevlisi, beşi matematik öğretmeni olarak görev yapan altı uzmandan açık uçlu maddelere ve dereceli puanlama anahtarlarına ilişkin görüş ve öneri alınmıştır.

Kapsam geçerliği: Uzmanlardan alınan görüşler doğrultusunda açık uçlu maddeler ve özel bütünsel dereceli puanlama anahtarına ait kapsam geçerlik indeksi (KGİ) hesaplanmıştır. Birçok boyutun bir araya gelmesiyle oluşan ölçülen özelliğin kapsam geçerliği kontrol edilirken her bir boyut için KGİ elde edilmelidir (Yurdugül, 2005). Hesaplamada Lawshe (1975) tekniğine ait formül kullanılmıştır.

$$KGO = \frac{N_U - \frac{N}{2}}{\frac{N}{2}} \qquad KG\dot{I} = \frac{\sum_{i}^{m} KGO}{m}$$

Kapsam geçerlik oranı (KGO), belirtilen ölçütün ya da maddenin kapsam geçerlik oranını; N_U, belirtilen ölçüt ya da maddeye "uygun" diyen uzman sayısı; N, toplam uzman sayısı; KGİ, belirtilen ölçüt ya da maddenin kapsam geçerlik indeksi; m, ölçme aracında bulunan ölçüt ya da madde sayısı olmak üzere yukarıda verilen formüller kullanılarak Tablo 11 ve Tablo 12'deki değerler elde edilmiştir.

Tablo 11

Açık Uçlu Maddelere İlişkin Kapsam Geçerlik İndeksleri (KGİ)

Maddeler	KGİ
1	0,92
2	0,92
3	0,92
4	1
5	1
6	0,92
7	1
8	0,92
9	1
10	0,83
11	1

KGİ ≥ KGÖ= 0,80

KGO geliştirilen ölçekteki maddelere ait uzman görüşlerinin niceliklendirilerek kapsam geçerliğinin sağlanmasında kullanılır. KGİ değeri, KGO değerlerinin ortalamalarının alınması ile bulunur (Yurdugül & Bayrak, 2012). Wilson, Pan ve Schumsky (2012) tarafından açıklanan tabloda kapsam geçerlik ölçüt (KGÖ) değeri, görüş bildiren uzman sayısı altı (N=6) olduğunda $\alpha = 0{,}05$ anlamlılık düzeyinde 0,80 olarak belirlenmiştir (EK 2). Başarı testinde yer alacak her bir açık uçlu maddenin KGİ değerinin ölçüt değeri olan 0,80'den büyük olduğu görülmüş ve testin kapsam geçerliğinin sağlandığı belirlenmiştir.

Tablo 12

Özel Bütünsel Dereceli Puanlama Anahtarına İlişkin Kapsam Geçerlik İndeksleri (KGİ)

Maddeler	KGİ
1	0,91
2	0,91
3	1
4	1
5	0,86
6	0,91
7	1
8	0,91
9	0,91
10	0,86
11	1

KGİ ≥ KGÖ= 0,80

Açık uçlu maddeleri güvenilir bir şekilde puanlamak amacıyla hazırlanan özel bütünsel dereceli puanlama anahtarının her bir maddeye ait KGİ değerinin kapsam geçerlik ölçüt değeri olan 0,80'den büyük olduğu görülmüş ve kapsam geçerliğinin sağlandığı sonucuna varılmıştır.

Yapı Geçerliği: Özel bütünsel dereceli puanlama anahtarının kapsam geçerliği sağlandıktan sonra 8. sınıfta öğrenim gören 122 öğrenciye 11 maddeden oluşan matematik başarı testi 40'ar dakikalık iki oturum şeklinde uygulanmıştır. Uygulama sonrasında matematik öğretmeni olan bir puanlayıcıdan öğrencilerin başarı testindeki cevaplarını hazırlanan dereceli puanlama anahtarını kullanarak puanlaması istenmiştir. Puanlamadan elde edilen verilerle yapı geçerliğinin kontrolünü sağlamak amacıyla SPSS 21.0 programı kullanılarak açımlayıcı faktör analizi (AFA) yapılmıştır. Faktör analizi için örneklem büyüklüğünün madde sayısının 10 katı büyüklükte olması önerilmektedir (Kline, 1994). Açımlayıcı faktör analizi sayıltılarının karşılanıp karşılanmadığı kontrol edilmiştir. Veri setinin faktörleşme yapısına uygunluğunun belirlenmesi için Kaiser-Meyer-Olkin örneklem yeterlik ölçütü (KMO) ve Barlett'in küresellik testi sonuçlarına bakılmıştır.

Tablo 13

KMO Değerlerinin Faktör Analizi İçin Uygunluğunun Yorumlanmasına İlişkin Değer Aralıkları ve Açıklama

Değer aralığı	Açıklama
0,90 ≤KMO≤1,00	Mükemmel
0,80 ≤KMO< 0,90	Çok iyi
0,70 ≤KMO< 0,80	İyi
0,60 ≤KMO< 0,70	Orta
0,50 ≤KMO< 0,60	Zayıf
KMO< 0,50	Kabul Edilemez

"Applied Multivariate Techniques" Sharma, S., 1996, New York: John Wiley and Sons kaynağından uyarlanmıştır.

Analiz sonucunda KMO değeri 0,808 olarak bulunmuştur. Sharma (1996)'ya göre bu değer, faktör analizinin gerçekleştirilebileceğini göstermiştir. Barlett'in küresellik testi $p<0,05$ olduğunda istatistiksel olarak anlamlıdır (Tabachnick & Fidell, 2013). Barlett'in küresellik testinin 0,05 düzeyinde anlamlı olduğu görülmüştür (χ^2 =243,136, s.d=55, p=0,00). Daha sonra temel bileşenler analizi uygulandığında maddeler iki faktörde kümelenmiştir. Açık uçlu maddelere ilişkin faktör yük değerleri; Madde 1: 0,701, Madde 2: 0,793, Madde 3: 0,548, Madde 4: 0,497, Madde 5: 0,576, Madde 6: 0,724, Madde 7: 0,800, Madde 8: 0,689, Madde 9: 0,821, Madde 10: 0,846, Madde 11: 0,828 olarak bulunmuştur. Faktör özdeğerleri incelendiğinde ise birinci faktörün yük değerine (4,791) ait açıklanan varyans oranı % 43,55; ikinci faktörün yük değerine (1,391) ait açıklanan varyans oranı % 12,65 bulunmuş ve bu da dereceli puanlama anahtarının tek boyutlu bir yapı oluşturabileceğini göstermiştir. Açıklanan varyans oranının %30 ve üzerinde olması ölçme aracının tek faktörlü yapıya sahip olduğuna işaret etmektedir (Büyüköztürk, 2008; Bayram, 2009). Bu durum açık uçlu maddelerin matematik başarı testinin genel amacına hizmet ettiği şeklinde yorumlanabilir. Başarı testinin tek boyutluluğu dereceli puanlama anahtarı ile verilen puanlar kullanılarak elde edildiğinden özel bütünsel dereceli puanlama anahtarının da tek boyutlu ve iyi bir ölçme aracı olduğu düşünülmüştür.

Güvenirlik: Yapılan faktör analizinden elde edilen faktör yükleri eşit olmadığında klasik güvenirlik katsayıları (Guttmann ve Cronbach α, Armor'un θ Katsayısı, Heise ve Bohrnstedt'ın Ω Katsayısı) gerçek güvenirlikten daha düşük değer almaktadır. McDonald (1999) tarafından geliştirilen omega (ω) güvenirlik katsayısı, doğrulayıcı faktör analizini kullanarak α'ya eşit ya da daha büyük değer üretmektedir (Yurdugül, 2005). Doğrulayıcı faktör analizini kullanarak güvenirlik katsayısı üretmesi nedeniyle "yapısal güvenirlik"

olarak (Nunnally & Bernstein, 1994) da isimlendirilen McDonald'ın ω katsayısı nokta ya da güven aralığı tahmini yaparak iç tutarlılığın hesaplanmasını sağlamaktadır (Dunn, Baguley, & Brunsden, 2014). ω katsayısının değeri 0,87 ve %95 güven aralığı değerleri 0,76–0,92 bulunmuştur. Bu bilgilerden faydalanarak matematik başarı testinin dolayısıyla özel bütünsel dereceli puanlama anahtarının iç tutarlığının sağlandığı, güvenilir ve geçerli olduğu düşünülmüştür.

3.3.3. Açık Uçlu Maddelere İlişkin Öğretmen (Puanlayıcı) Değerlendirme Formu

Puanlayıcıların açık uçlu maddelere özel bütünsel dereceli puanlama anahtarını kullanarak verdiği puanlar bu forma yazılmıştır. Öğrenci ve açık uçlu madde numaralarının bulunduğu formda ayrıca puanlayıcılara ait demografik bilgilere de yer verilmiştir.

3.4. Veri Toplama Süreci

Bireylere ilk olarak açık uçlu maddelerden oluşan 11 maddelik matematik başarı testi 40'ar dakikalık iki oturum şeklinde Ankara'da bulunan iki ortaokulda toplam 122 öğrenciye uygulanmıştır. Testi uygulamadan önce bireylere sınavın not ile değerlendirilmeyeceği sonuçların sadece bilimsel bir araştırmaya katkı sağlamak için kullanılacağı bildirilmiştir. Ayrıca bireylerden kendilerini gerçek bir sınavı cevaplandırıyormuş gibi düşünmeleri ve maddelere ciddiyetle cevap vermeleri istenmiştir. Güvenilir ve geçerli sonuçlar elde etmek için bu konuya dikkat edilmesi gerektiği, sınav süresinin 80 dakika olduğu ve maddelere cevap verirken yaptıkları çözümün her bir adımını açıkça yazmaları, sonuca ulaşamasalar dahi akıllarına gelen çözümün araştırmacı tarafından değerli olduğu söylenerek cevaplarını kağıt üzerine yazmaktan çekinmemeleri öğrencilere hatırlatılmıştır. Uygulama sonrasında 122 öğrencinin cevap kağıtları fotokopi ile çoğaltılarak belirlenen matematik alan uzmanı ya da ilköğretim matematik öğretmenliği mezunu olan dört puanlayıcıdan özel bütünsel dereceli puanlama anahtarı kullanarak her bir puanlama deseni için her bir kağıdı puanlamaları istenmiştir. Birbirlerinden bağımsız olarak yaptıkları puanlamalarını "Matematik Başarı Testindeki Açık Uçlu Maddelere İlişkin Öğretmen Değerlendirme Formu"na yazmaları istenmiştir. Turgut ve Baykul (2012) açık uçlu maddeleri değerlendiren puanlayıcı sayısının beşten fazla olmasının puanlama güvenirliğinde önemli bir artış sağlamayacağını belirtmiştir.

3.5. Verilerin Analizi

Eckes (2015)'in tanımladığı çok yüzeyli Rasch modeli puanlama desenlerine göre açık uçlu maddelerin puanlayıcılar arası güvenirlikleri karşılaştırılırken Linacre (1993) tarafından geliştirilen FACETS programı kullanılmıştır. FACETS, özellikle performans değerlendirmesi ve eşleştirilmiş karşılaştırmalar alanlarında Rasch modelinin pek çok uygulamasına yardımcı olan Windows tabanlı bir yazılımdır. Her bir öğe için bir ölçü (doğrusal miktar), standart hata (hassasiyet) ve beş uyum istatistiği (istatistiksel geçerlik) sağlar. Uyum istatistikleri, anormal gözlemlerin ve kendine has ögelerin teşhisine olanak tanır. FACETS programı her bir sıralı gözlemi birey, madde ve puanlayıcı gibi ögeler arasındaki bir etkileşimin sonucu olarak kavramsallaştırmaktadır. Her bir puanlayıcıyı "puanlama makinesi" olarak değil, belirli bir miktarda katılık veya cömertlik gösteren bağımsız bir uzman olarak değerlendirerek etkileşimli ögelerin bağımsız olarak çalıştırılmasını ve bu ögelerden elde edilen ölçümlerin gizil değişken üzerinde birleştirilmesini modellemektedir (Linacre, 2010, s. 12). FACETS, doğru/yanlış, her zaman/bazen/asla, kesinlikle katılıyorum/katılıyorum /katılmıyorum gibi nicel puanları, karşılaştırmaları ve gözlemleri analiz eder. Her bir gözlem yüzeyler arasındaki ilişkinin bir sonucudur. FACETS, yazılı performansın puanlanması, portfolyo değerlendirme ve puanlayıcıların puanladığı diğer performanslar için uygundur (Köse vd., 2016). Kullanımı eğitimsel ve psikolojik testlerle sınırlı değildir. Ürün geliştirme, spor bilimi, kirlilik denetimi, halka hitap etme ve sanat gibi araştırma ve uygulamanın birçok alanında niteliksel gözlemleri, doğrusal ölçümlere dönüştürmek için de kullanılır (Linacre, 2010, s.12).

Çok yüzeyli Rasch modelinin tek boyutluluk, yerel bağımsızlık ve model veri uyumu sayıltılarının kontrolü yapılmıştır. Açımlayıcı faktör analizi SPSS ve doğrulayıcı faktör analizi ise LISREL programları ile gerçekleştirilmiş, analizler sonucunda matematik başarı testinin ve testin değerlendirilmesinde kullanılan özel bütünsel dereceli puanlama anahtarının tek boyutlu bir yapıya sahip olduğu ve sayıltının sağlandığı görülmüştür.

Yerel bağımsızlık, birey yeteneğinin ve maddelerin etkisinin, bireylerin maddelere verdiği cevaplar ile istatistiksel olarak hiçbir ilgisinin olmadığının kabul edilmesi anlamına gelmektedir. Linacre (2009) yerel bağımsızlık varsayımının tek boyutluluk varsayımını kapsadığını belirtmiştir. Tek boyutluluğun sağlanması nedeniyle yerel bağımsızlık sayıltısının da sağlandığı kabul edilmiştir (Hambleton, Swaminathan, & Rogers, 1991).

Model veri uyumu sağlandığında standartlaştırılmış artık değerlerin (StRes) ± 2 aralığı dışında kalan değerleri tüm gözlem sayısının (puanlayıcı sayısı x madde sayısı x birey sayısı) yaklaşık % 5'i, ± 3 aralığı dışında kalan değerleri ise tüm gözlem sayısının yaklaşık %1'i kadardır (Linacre, 2010). Tablo 15'teki değerlere bakıldığında model veri uyumu sayıltısının sağlandığı görülmüştür.

Tablo 14

Model Veri Uyumu Kontrolüne İlişkin Aralıklara Göre Standartlaştırılmış Artık Değer Sayıları ve Yüzdeleri

Desen Adı	Toplam gözlem sayısı	Standartlaştırılmış artık değer (± 2) sayısı-yüzdesi	Standartlaştırılmış artık değer (± 3) sayısı-yüzdesi
Desen A	5368	100 - % 1,86	56 - % 1,04
Desen B	4026	100 - % 2,48	41 - % 1,01
Desen C	1738	71 - % 4,08	17 - % 0,97
Desen D	2684	100 - % 3,73	25 - % 0,93
Desen E	2684	100 - % 3,73	27 - % 1,01
Desen F	1342	56 - % 1,04	14 - %1,04

Çalışmada puanlayıcı, görev (açık uçlu maddeler), birey (öğrenci), ölçüt yüzeyleri arasındaki etkileşimler, veri kalibrasyon haritaları, güvenirlik katsayısı, ayırma oranı, ayırma indeksi gibi değerler ile puanlayıcılar arası güvenirliği belirlemek için Rasch-Cohen'in kappa istatistiği tamamen (tam) çaprazlanmış desen, tamamlanmamış-bağlantılı desen, tamamlanmamış-bağlantılı olmayan desen ve tamamlanmamış-bağlantılı sarmal desenin her biri için ayrı incelenmiştir.

BÖLÜM IV

BULGULAR VE YORUMLAR

Bu bölümde Eckes (2015)'in tanımladığı tamamen çaprazlanmış desen (Desen A), tamamlanmamış-bağlantılı desen (Desen B ve Desen C), tamamlanmamış-bağlantılı olmayan desen (Desen D), tamamlanmamış-bağlantılı sarmal desen (Desen E) ve Taherbai ve Young (2001)'ın tanımladığı sarmal puanlama deseninin (Desen F) her biri için puanlayıcı, öğrenci, açık uçlu madde ve puanlama ölçütleri yüzeylerine ilişkin ölçüm raporları, olasılık eğrileri ile veri kalibrasyon haritaları sunulmuştur. Çok yüzeyli Rasch modeli puanlama desenlerine göre puanlayıcılar arası güvenirliği belirlemek amacıyla Rasch-Cohen'in kappa istatistiği kullanılmış ve yorumlanmıştır.

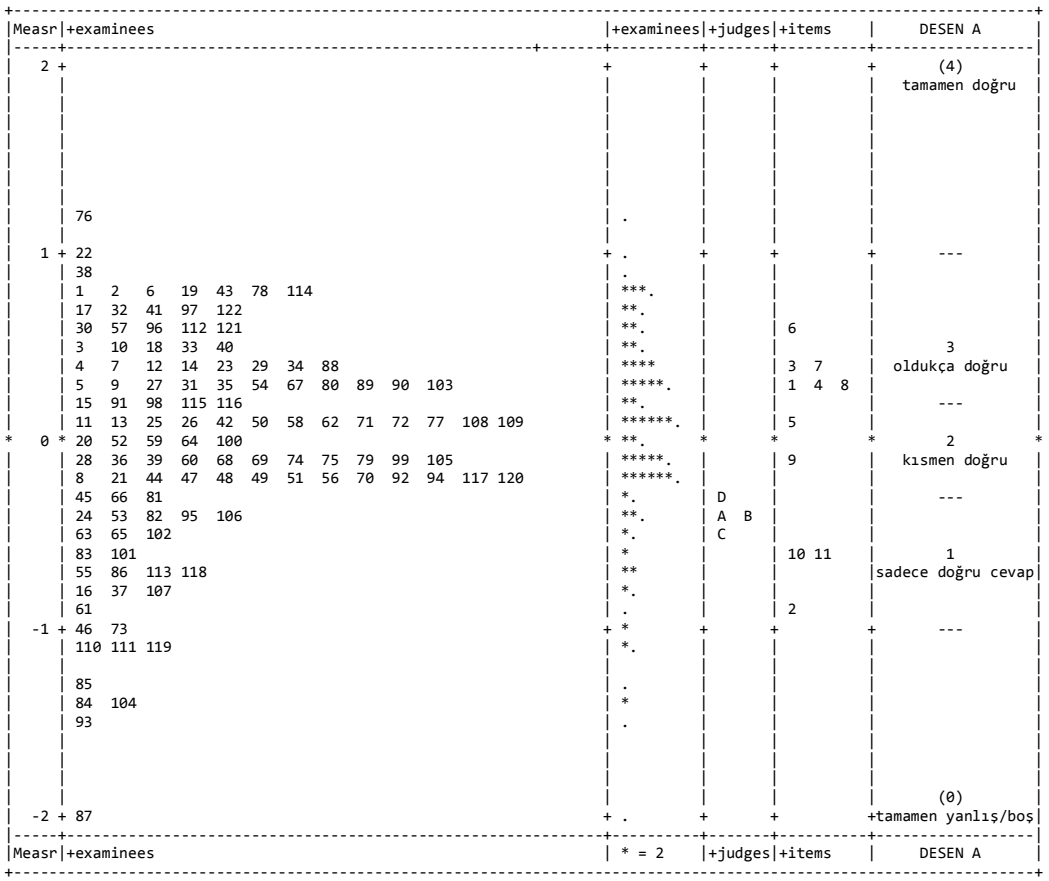
4.1. Tamamen Çaprazlanmış Puanlama Desenine (Desen A) İlişkin Bulgular

Tüm puanlayıcıların tüm öğrenci performanslarını puanladığı bu desene göre öğrenci performanslarının puanlanması sürecinde puanlayıcıların katılık-cömertlik durumları, öğrenci başarıları, açık uçlu madde güçlükleri, dereceli puanlama anahtarları puanlama kategorilerinin puanlamada kullanılma durumları bireysel ve grup düzeyinde ölçüm raporlarından elde edilen istatistiksel değerlerle incelenmiştir. FACETS programıyla elde edilen sabit etkili ve normal ki kare istatistikleri, ayırma oranı, ayırma indeksi, ayırma indeksinin güvenirliği değerleri, puanlayıcılar arası güvenirliği belirlemek amacıyla hesaplanan Rasch-Cohen'in kappa istatistiği ile grup düzeyinde; iç uyum, dış uyum uygunluk ölçütleri ve logit değerleri ile bireysel düzeyde bulgular elde edilmiştir.

4.1.1. Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Veri Kalibrasyon Haritasına İlişkin Bulgular

Çok Yüzeyli Rasch modeli, her değişkenlik kaynağı için doğrusal bir yüzeyler arası bağlantı sağlayarak yüzeylerin değerleri için ortak bir logit ölçeği kullanır (Sarıtaş Akyol ve Karakaya, 2021). Logit ölçekte (kalibrasyon haritası ya da değişken haritası) her bir sütun, analiz edilen yüzeyleri göstermektedir. Tüm yüzeylerin tüm ögeleri, logit ölçeğinde dikey olarak konumlandırılır (Eckes, 2019, s. 14). Logit ölçekte her bir yüzey negatif sonsuz ile pozitif sonsuz arasında değer alabilmektedir.

Kalibrasyon haritasında öğrenci (examinees) yüzeyindeki pozitif değerler açık uçlu maddeleri cevaplandırma becerisi yüksek öğrencileri, negatif değerler ise açık uçlu maddeleri cevaplandırma becerisi düşük öğrencileri göstermektedir. Puanlayıcı (judges) yüzeyindeki pozitif değerler cömert puanlama yapan puanlayıcıları, negatif değerler ise katı puanlama yapan puanlayıcıları temsil etmektedir. Açık uçlu madde (items) yüzeyindeki pozitif değerler maddenin kolay, negatif değerler ise maddenin zor olduğunu ifade etmektedir. Son sütunda kullanılan özel bütünsel dereceli puanlama anahtarı en yüksek değerden en düşüğe doğru derecelendirilmiştir. Şekil 1'de tamamen çaprazlanmış puanlama desenine ilişkin veri kalibrasyon haritası sunulmuştur.



Şekil 1. Tamamen çaprazlanmış puanlama deseni (Desen A) veri kalibrasyon haritası

Kalibrasyon haritası incelendiğinde 76, 22 ve 38 nolu öğrencilerin matematik başarı testindeki açık uçlu maddeleri cevaplandırmada yüksek beceriye, 93 ve 87 nolu öğrencilerin ise düşük beceriye sahip oldukları görülmüştür. Logit ölçekte pozitif logit değerine sahip öğrenci sayısı ile negatif logit değere sahip öğrenci sayısının birbirine yakın olması nedeniyle öğrenci grubunun başarı açısından heterojen olduğu söylenebilir. Haritada 1. Puanlayıcı A, 2. Puanlayıcı B, 3. Puanlayıcı C, 4. Puanlayıcı D harfi ile temsil edilmiştir. Puanlayıcı D'nin en cömert puanlayıcı, puanlayıcı C'nin en katı puanlayıcı olduğu ve puanlayıcı A ile puanlayıcı B'nin birbirine benzer puanlama yaptığı görülmüştür. Açık uçlu maddeler içerisinde Madde 6'nın öğrenciler tarafından kolay; Madde 2'nin ise zor cevaplandığı görülmüştür. Öğrenciler matematik başarı testindeki Madde 1, Madde 3, Madde 4, Madde 5, Madde 6, Madde 7 ve Madde 8'i kolay cevaplandırırken Madde 2, Madde 9, Madde 10, Madde 11'i cevaplandırmada zorlanmışlardır. Puanlayıcıların öğrenci performanslarını puanlarken özel bütünsel dereceli puanlama anahtarının derecelendirmelerini geniş aralıkta ve etkin bir şekilde kullandıkları anlaşılmıştır.

4.1.2. Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Puanlayıcı Yüzeyine İlişkin Bulgular

Matematik başarı testindeki açık uçlu maddelerin puanlanmasında puanlayıcıların katılık ve cömertlik durumlarını grup düzeyinde belirlemek için öncelikle ayırma indeksi ve ayırma indeksi güvenirliği değerlerine bakılmaktadır. Bu değerler puanlayıcıların katılık ve cömertlikleri açısından farklı davranıp davranmadığını göstermektedir. Puanlayıcılar arasında farklılık oluştuğunda bu farklılığın istatistiksel olarak anlamlılığı sabit etkili ki-kare testi ile belirlenmektedir (Karakaya Özyer, 2018). Çok yüzeyli Rasch modeli, bir yüzeyin tüm öğelerinin istatistiksel olarak aynı ölçüme sahip olup olmadığı hipotezini test etmek amacıyla sabit etkili ki-kare testi değerini; bir yüzeyin tüm öğelerinin normal dağılmış bir evrenden rastgele seçilmiş bir örneklem olup olmadığı hipotezini test etmek amacıyla da rastgele ki-kare testi değerini hesaplamaktadır (Linacre, 2010). Sabit etkili ki-kare testi ile ortaya çıkan anlamlı farklılığın hangi puanlayıcıdan kaynaklandığını belirlemek için iç uyum (infit) ve dış uyum (outfit) indeksleri kontrol edilmektedir. Uyum istatistikleri 0,50 ile 1,50 arasında olduğunda tutarlı sonuçlar elde edildiği anlaşılmaktadır (Eckes, 2009). Myford ve Wolfe (2003), çok yüzeyli Rasch modeli analiz sonuçlarının kullanım amacına göre özellikle geniş ölçekli sınavlar gibi yüksek riskli karar verme süreçlerinde uyum indekslerinin alt ve üst sınır değerleri arasındaki farkın azaltılarak 0,80 ile 1,20 olarak belirlenebileceğini belirtmişlerdir. Dış uyum indeksinin uç değerlere karşı duyarlılığı daha yüksek olduğundan puanlayıcıların puanlama tutarsızlığını belirlemede dış uyum indeksi yerine iç uyum indeksinin incelenmesi daha doğru sonuç vermektedir (Engelhard, 1994). Puanlayıcılar arası güvenirlik, gözlenen ve beklenen uyum değerlerinin farkına bağlı olarak hesaplanan Rasch-Cohen'in kappa istatistiği ile kontrol edilmiştir. Logit değerinin pozitif olması puanlayıcıların cömert, negatif değer alması ise katı puanlama yaptıklarını göstermektedir. Tablo 15'de tamamen çaprazlanmış puanlama deseni puanlayıcı yüzeyine ilişkin ölçüm raporu sunulmuştur.

Tablo 15

Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu

	Model		Uygunluk Ölçütleri		Uyum Ölçütleri	
Puanlayıcı	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)	Gözlenen uyum	Beklenen uyum
A	-0,41	0,02	1,02	1,05	82,3	36,7
B	-0,41	0,02	1,05	1,08	82,0	36,6
C	-0,52	0,02	0,92	1,08	0,0	0,0
D	-0,33	0,02	0,93	0,97	68,4	36,4
Ortalama	-0,42	0,02	0,98	1,04		
SS(evren)	0,07	0,00	0,05	0,04		
SS(örneklem	0,08	0,00	0,06	0,05		
Model (evren):	RMSE= 0,02	SS: 0,06	Ayırma oranı= 2,73		Ayırma indeksi= 3,98	
Güvenirlik (puanlayıcılar arası olmayan) =0,88						
Model (örneklem):	RMSE= 0,02	SS: 0,08	Ayırma oranı= 3,21		Ayırma indeksi= 4,61	
Güvenirlik (puanlayıcılar arası olmayan) = 0,91						
Model, sabit etkili ki-kare: 33,7	s.d: 3	p= 0,00				
Model, rastgele ki-kare: 2,8	s.d: 2	p= 0,25				
Puanlayıcılar arası uyumun gözlenen değeri= %77,6						
Puanlayıcılar arası uymun beklenen değeri= %36,6						
Puanlayıcılar arası güvenirlik(Rasch-Cohen'in kappa istatistiği)= 0,65						
RMSE: hata kareleri ortalamasının karekökü		*S.S:standart sapma*			*s.d:serbestlik derecesi*	

Grup düzeyinde elde edilen istatistikler Tablo 15 incelendiğinde ayırma oranı= 3,21, ayırma indeksi= 4,61 ve ayırma indeksi güvenirliği=0,91 olduğu görülmüştür. Puanlayıcı yüzeyinde ayırma indeksinin 2,00'den büyük ve ayırma indeksi güvenirliğinin 0,50'den büyük olması, istenmeyen bir durum olan puanlamaya puanlayıcıdan kaynaklanan hataların karıştığını, puanlayıcıların katılık-cömertlik davranışlarının farklı olduğunu göstermektedir (Karakaya, 2018). RMSE değeri 0,02 bulunmuş ve bu değerin düşük olması, puanlayıcıların katılık davranışının daha kesin bir şekilde tahmin edildiğini göstermiştir. Puanlayıcılar arasındaki farklılığın anlamlı olup olmadığı sabit etkili ki-kare (χ^2) değeri ile belirlenmiştir. Sabit etkili ki-kare değeri $\chi^2(3)= 33,7$, $p=0,00<0,05$ bulunmuş ve puanlayıcılar arasındaki katılık-cömertlik davranışları arasında istatistiksel olarak anlamlı fark olduğu görülmüştür. "Gözlenen uyum yüzdesi" ile şans, puanlayıcı cömertliği ve dereceli puanlama anahtarı yapısına dayalı olarak elde edilen "beklenen uyum yüzdesi" arasındaki farkı temel alan Rasch-Cohen'in kappa istatistiği, puanlayıcıların ne ölçüde birbirlerinden bağımsız puanlama yaptıklarını veya mükemmel puanlayıcılar arası güvenirlik için ne kadar çalıştıklarını belirlemektedir (Eckes, 2015, s. 178).

Rasch modeli şartlarında (Gözlenen uyum% – Beklenen uyum%) / (100 – Beklenen uyum %) formülüyle hesaplanan Rasch-Cohen'in kappa değerinin 0,00'a yaklaşması beklenmektedir. Rasch-Cohen'in kappa değeri –0,20 ile +0,20 arasında ise modelden beklenen seviyede uyumun; -0,20 ile -0,40 ya da +0,20 ile +0,40 arasında ise modellenmiş olandan biraz daha fazla uyumun; -0,50'e eşit ya da küçük, +0,50'e eşit ya da büyük ise yüksek uyumun bulunduğu belirtilmiştir (Taghvafard, t.y). Tamamen çaprazlanmış puanlama deseni için puanlayıcılar arası uyumun gözlenen değeri %77,6, puanlayıcılar arası uyumun beklenen değeri %36,6 bulunmuş ve Rasch-Cohen'in kappa istatistiği değeri 0,65 olarak hesaplanmıştır. Bu değerin +0,50'den büyük olmasıyla farklı puanlayıcıların öğrenci performanslarını değerlendirirken birbirleriyle oldukça tutarlı puanlamalar yaptıkları, puanlayıcılar arası uyumun ve güvenirliğin yüksek olduğu sonucuna ulaşılmıştır. Gözlenen ve beklenen uyum yüzdeleri eşit olduğunda öğrenci performansını puanlamada puanlayıcıdan kaynaklanan hiçbir etkinin bulunmadığı ve o puanlayıcının mükemmel olduğu anlamına gelmektedir. Puanlayıcı C'nin beklenen ve gözlenen uyum yüzdeleri eşit olduğundan mükemmel puanlayıcıdır.

Bireysel düzeyde elde edilen logit değerlerine bakıldığında puanlayıcıların negatif değer alması, tüm puanlayıcıların katı puanlama yaptığını göstermektedir. Veri kalibrasyon haritasından da anlaşıldığı üzere puanlayıcı D en yüksek logit değerine (-0,33) sahip olduğundan en cömert, puanlayıcı C en düşük logit değerine (-0,52) sahip olduğundan en katı, puanlayıcı A ve puanlayıcı B aynı logit değerine (-0,41) sahip olduğundan benzer puanlama yapmıştır. Puanlayıcıların iç uyum indeksleri 0,92 ile 1,05 arasında, dış uyum indeksleri 0,97 ile 1,08 arasında değişmektedir ve 0,50 ile 1,50 arasında olduklarından kabul edilebilir düzeydedir. Bu değerler model ile veri uyumunu bozan puanlayıcının bulunmadığını ve birbirlerinden farklı puanlama yapmalarına rağmen puanlayıcıların dereceli puanlama anahtarı ölçütlerinin tamamını göz önünde bulundurarak tutarlı bir şekilde puanlama yaptıklarını göstermektedir.

4.1.3. Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Öğrenci Yüzeyine İlişkin Bulgular

Her bir öğrencinin her bir maddedeki cevabının tüm puanlayıcılar tarafından puanlandığı *(122x11x4=5368 gözlem)* Desen A için öğrenci yüzeyine ait ölçüm raporu Tablo 16'da verilmiştir. Tabloda öğrencilerin tamamına ilişkin değerlerin verilmesi mümkün olmadığı

için logit değerine göre sıralaması en yüksek ve en düşük 8 öğrenci örnek olarak sunulmuştur.

Tablo 16

Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Öğrenci Yüzeyine İlişkin Ölçüm Raporu

	Ortalamalar		Model		Uygunluk ölçütleri	
Öğrenci sıra no	Gözlenen ortalama	Düzeltilmiş ortalama	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)
76	3,09	3,26	1,16	0,15	0,73	0,87
22	2,89	3,05	0,98	0,14	1,25	0,98
38	2,82	2,98	0,93	0,13	1,22	1,08
1	2,64	2,76	0,79	0,13	0,56	0,57
19	2,64	2,76	0,79	0,13	0,71	0,66
78	2,64	2,76	0,79	0,13	1,37	1,23
114	2,64	2,76	0,79	0,13	1,29	1,35
2	2,61	2,74	0,77	0,13	0,64	0,65
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
110	0,32	0,26	-1,06	0,22	0,83	0,65
119	0,32	0,26	-1,06	0,22	0,74	1,63
111	0,30	0,24	-1,11	0,23	1,13	0,98
85	0,23	0,19	-1,30	0,27	0,85	1,04
84	0,20	0,17	-1,37	0,29	0,75	1,17
104	0,20	0,17	-1,37	0,29	0,75	0,87
93	0,18	0,15	-1,46	0,31	1,08	1,56
87	0,00	0,01	-4,45	1,81	minimum	
Ortalama	1,48	1,40	- 0,04	0,15	0,99	1,04
SS(evren)	0,73	0,79	0,68	0,16	0,36	0,59
SS(örneklem)	0,73	0,80	0,68	0,16	0,36	0,59
Model(evren) Güvenirlik: 0,90	RMSE: 0,22	SS: 0,64	Ayırma oranı= 2,96		Ayırma indeksi= 4,28	
Model(örneklem) Güvenirlik: 0,90	RMSE: 0,22	SS: 0,65	Ayırma oranı= 2,98		Ayırma indeksi= 4,30	
Model, sabit etkili ki-kare: 1377,0			s.d: 121		p= 0,00	
Model, rastgele ki-kare: 78,5			s.d: 120		p= 1,00	
RMSE: hata kareleri ortalamalarının karekökü			*SS: standart sapma*		*s.d: serbestlik derecesi*	

Grup düzeyinde istatistiksel göstergeler incelendiğinde ayırma oranı=2,98, ayırma indeksi= 4,30 ve ayırma indeksi güvenirliği= 0,90 olduğu görülmüştür. Bu değerler yetenek düzeyleri birbirlerinden farklı olan öğrencilerin güvenilir bir şekilde ayırt edilebildiklerini göstermiştir. Öğrenci başarısının yüksek/düşük olma durumunu gösteren RMSE değeri 0,22 olarak bulunmuş ve yüksek beceriye sahip öğrencilerin daha kesin tahmin edildiği anlaşılmıştır. Her değişkenlik kaynağı için yüzey içerisindeki öge düzeyi sayısı; ayırma indeksi (G) olmak üzere (4G+1)/3 formülü ile belirlenmektedir. Desen A için

(*[4.(4,30)+1]/3=6,06*) öğrenci performanslarının altı seviyeye ayrıldığı söylenebilir. Sabit etkili ki-kare değerinin $\chi^2(121)=1377,0$, $p=0,00<0,05$ olması en az iki öğrencinin yetenek düzeyleri arasında anlamlı fark bulunduğunu ve öğrenci grubunun heterojen yapıda olduğunu göstermiştir. Logit değerlerine bakıldığında 76 numaralı öğrenci (1,16) en yüksek yetenek düzeyine, 87 numaralı öğrenci (-4,45) en düşük yetenek düzeyine sahiptir. Bireysel düzeyde iç uyum indeksleri minimum ile 1,90 arasında, dış uyum indeksleri minimum ile 4,66 arasında değişmiştir. Ölçüm üretkenliğini gösteren ağırlıklı ortalama kare uyum istatistiği (iç uyum istatistiği) öğrenci, puanlayıcı ve ölçüt konumlarının ölçüm ölçeğinde birbirine yakın olduğu "içteki" beklenmedik derecelendirmelere; ağırlıklı olmayan ortalama kare uyum istatistiği (dış uyum istatistiği) öğrenci, puanlayıcı ve ölçüt konumlarının ölçüm tablosunda birbirinden çok uzak olduğu "dışta yatan" beklenmedik derecelendirmelere daha duyarlıdır (Eckes, 2019; Linacre, 2013). Uç değerlerden daha çok etkilenen dış uyum indeksi yerine iç uyum indekslerine bakılması ile tutarsızlıkların daha iyi bulunacağı belirtilmiştir (Engelhard, 1994). Bu nedenle bazı öğrencilerin dış uyum indeksleri 2,00'den büyük olsa da iç uyum indekslerinin 2,00'den büyük olmaması model veri uyumunun bozulmadığını göstermiştir (Linacre, 2002). Tüm öğrencilerin % 90,2'sinin iç ve dış uyum indeksleri açısından kabul edilebilir aralıkta (0,50-1,50) olduğu görülmüştür. Analiz sonucunda iç uyum aralığı dışında kalan on iki öğrenci için her bir açık uçlu maddeyi cevaplandırma performanslarının beklenen düzeyde olmadığı; dış uyum aralığı dışında kalan on iki öğrencinin ise uygulamada gösterdikleri performansın beklenen düzeyde olmadığı anlaşılmıştır. İç uyum indeks değeri 1,00 (beklenen değer) ve dış uyum indeks değeri 0,94 olan 32 numaralı öğrencinin her bir açık uçlu maddeyi cevaplandırma performansı ile matematik başarı testindeki genel performans uyumu en yüksek öğrenci olduğu gözlemlenmiştir.

4.1.4. Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Açık Uçlu Madde Yüzeyine İlişkin Bulgular

Matematik başarı testinde yer alan açık uçlu maddelerle ilgili madde güçlük düzeylerine ilişkin istatistikler ve uyum indeksleri Tablo 17'de sunulmuştur.

Tablo 17

Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Açık Uçlu Madde Yüzeyine İlişkin Ölçüm Raporu

	Ortalamalar		Model		Uygunluk ölçütleri	
Madde no	Gözlenen ortalama	Düzeltilmiş ortalama	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)
1	1,91	1,85	0,32	0,04	1,06	1,08
2	0,43	0,32	-0,91	0,06	1,00	0,95
3	2,02	1,99	0,38	0,04	0,78	0,75
4	1,83	1,74	0,27	0,04	0,68	0,66
5	1,64	1,51	0,15	0,04	1,15	1,14
6	2,28	2,33	0,55	0,04	1,21	1,21
7	1,98	1,94	0,36	0,04	1,16	1,12
8	1,85	1,77	0,28	0,04	0,86	0,81
9	1,24	1,04	-0,12	0,04	0,70	0,71
10	0,64	0,48	-0,63	0,05	1,38	1,76
11	0,63	0,47	-0,64	0,05	0,97	1,29
Ortalama	1,50	1,40	0,00	0,04	1,00	1,04
SS(evren)	0,62	0,67	0,48	0,01	0,21	0,31
SS(örneklem)	0,65	0,71	0,50	0,01	0,22	0,32
Model(evren): Güvenirlik: 0,99	RMSE: 0,04	SS: 0,48	Ayırma oranı= 11,58		Ayırma indeksi= 15,77	
Model(örneklem): Güvenirlik: 0,99	RMSE: 0,04	SS: 0,50	Ayırma oranı= 12,15		Ayırma indeksi= 16,53	
Model, sabit etkili ki-kare: 1153,6			s.d: 10		p= 0,00	
Model, rastgele ki-kare: 9,9			s.d: 9		p= 0,36	
RMSE: hata kareleri ortalamalarının karekökü			*SS: standart sapma*		*s.d: serbestlik derecesi*	

Açık uçlu madde yüzeyine ilişkin ayırma oranı=12,15, ayırma indeksi=16,33, ayırma indeksi güvenirliği=0,99 şeklindedir. Ayırma oranı, açık uçlu maddelerin güçlük düzeyleri arasındaki ölçme hatasının karşılayamayacağı kadar fazla olan farklılığın belirlenmesini sağlamaktadır. Ayırma oranı, ayırma indeksi gibi yorumlanmaktadır. Ayırma oranının yüksek olması, açık uçlu madde ölçümlerinin geniş bir alana yayıldığını dolayısıyla madde güçlük düzeylerinin farklı olduğunu göstermektedir (Eckes, 2015; Myford & Wolfe, 2003). Ayırma indeksinin 2,00'den ve ayırma indeksi güvenirliğinin 0,50'den büyük olması, istenen bir durum olan farklı güçlük düzeylerine sahip maddelerin yüksek güvenirlikle birbirlerinden ayırt edildiğini göstermiştir. Açık uçlu maddelerin güçlük düzeyleri arasındaki farklılığın istatistiksel olarak anlamlı olup olmadığı sabit etkili ki-kare testi ile belirlenmiştir. Sabit etkili ki-kare değeri $\chi^2(10)= 1153,6$, $p=0,00<0,05$ bulunmuş ve madde güçlük düzeyleri arasındaki farklılığın istatistiksel olarak anlamlı olduğu anlaşılmıştır.

Açık uçlu maddelerin güçlük düzeylerini gösteren logit değerleri 0,55 ile -0,91 arasında değişmektedir. Veri kalibrasyon haritasından da anlaşıldığı gibi Tablo 18'de Madde 6 (logit=

0,55)'nın güçlük düzeyinin düşük (kolay cevaplanabilen) ve Madde 2 (logit=-0,91)'nin güçlük düzeyinin yüksek (zor cevaplanabilen) olduğu görülmektedir. Logit değerlerine bakılarak güçlük düzeyleri yüksek maddeler Madde 2, Madde 11, Madde 10, Madde 9; güçlük düzeyleri düşük maddeler Madde 6, Madde 3, Madde 7, Madde 1, Madde 8, Madde 4, Madde 5 şeklinde sıralanmıştır. Maddelerin güçlük/kolaylık durumunu gösteren RMSE değeri 0,04 bulunmuştur. Standart hatanın düşük değerler alması madde güçlüğünün daha kesin bir şekilde tahmin edildiğini göstermektedir (Eckes, 2019). Standart hata değeri sırasıyla 0,06; 0,05; 0,05 olan Madde 2, Madde 10 ve Madde 11 dışındaki maddelerin madde güçlükleri daha kesin tahmin edilmiştir. İç uyum indeksleri 0,68 ile 1,38 arasında, dış uyum indeksleri 0,66 ile 1,76 arasında değişmektedir. Tablo 18'de yer alan iç ve dış uyum değerlerinin 0,50 ile 1,50 aralığında ve kabul edilebilir düzeyde olduğu görülmüştür. Bu durum istenilen kazanımı yoklamada maddelerin iyi işlediğini göstermektedir. Sadece Madde 10'da iç uyum değeri (1,38) kabul edilebilir düzeyde olsa da dış uyum değeri (1,76) bulunmuştur. İç uyum indeksinin kabul edilebilir düzeyde olması maddenin istenilen kazanımı ölçmede iyi olduğunu, dış uyum indeksinin kabul edilebilir düzey dışında kalması öğrencilerin matematik başarısını ölçmede yetersiz bir madde olduğunu göstermektedir. Linacre (2002) 1,50 ile 2,00 arasındaki değerlerin ölçümün oluşturulması için verimsiz olduklarını ancak çok kötü değerler olarak nitelendirilmediklerini belirtmiştir. Bu nedenle başarı testinde yer alan maddelerin öğrenci matematik başarısını ölçmede iyi işlediği söylenebilir.

4.1.5. Tamamen Çaprazlanmış Puanlama Deseni (Desen A) Özel Bütünsel Dereceli Puanlama Anahtarı Ölçütlerine İlişkin Bulgular

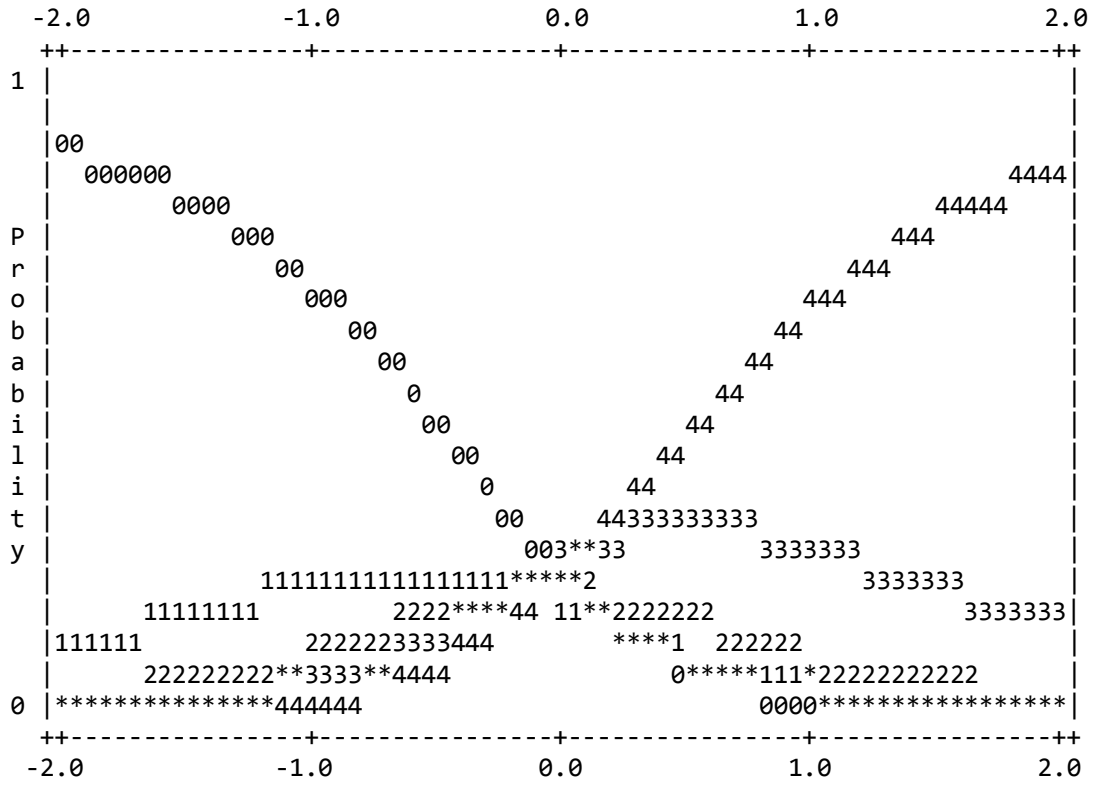
Desen A için dereceli puanlama anahtarı kategorilerine ilişkin değerler Tablo 18'de sunulmuştur.

Tablo 18

Özel Bütünsel Dereceli Puanlama Anahtarı Puanlama Kategorilerine Ait Ölçüm Raporu

Puanlama Kategorileri	Frekans	Yüzde (%)	Birikimli yüzde (%)	Ortalama logit değeri	Beklenen logit değeri	Outfit (dış uyum)
0	2161	41	41	-0,88	-0,90	1,00
1	858	16	57	-0,59	-0,50	1,10
2	662	12	69	-0,19	-0,19	1,00
3	793	15	84	0,11	0,06	0,90
4	850	16	100	0,25	0,26	1,20

Tablo 18'deki frekans ve yüzde değerleri incelendiğinde puanlayıcıların en çok 0 numaralı kategoriyi tercih ettikleri ve 1, 2, 3 ve 4 numaralı kategorileri birbirine yakın sayıda kullanıldığı görülmüştür. Uç değer olan 0 kategorisinin daha fazla kullanılmasının sebebi merkezi eğilim etkisi ya da öğrencinin matematik başarı düzeyinin düşük olması olabilmektedir. Grup düzeyinde bilgi sağlayan kategori istatistikleri merkezi eğilim etkisinin belirlenmesinde yeterli olmayabilir. Bu nedenle öğrenci yüzeyine ilişkin ölçüm raporuna bakılarak yorum yapılabilmektedir. Öğrenci yüzeyi ölçüm raporunda öğrencilerin matematik başarı düzeylerinin yüksek güvenirlikle ayırt edildiği görülmüştür. Bu durum merkezi eğilim etkisinin bulunmadığını ve öğrenci başarı düzeyinden kaynaklandığını göstermiştir. Logit değerlerine bakıldığında 3 ve 4 numaralı kategorilerin tam ve oldukça doğru tanımlarına uygun olarak pozitif değer aldığı, 0, 1, 2 numaralı kategorilerin kısmen yanlış, tamamen yanlış ve boş tanımlarına uygun olarak negatif değerler aldığı anlaşılmıştır. Tablodaki ortalama ve beklenen logit değerleri birbirine yakındır. Ortalama ile beklenen logit değerleri birbirine yakınsa verilerin Rasch modeli beklentilerine uyduğu düşünülmektedir (Eckes, 2019). Dış uyum indeksleri 0,90 ile 1,20 arasında değişmektedir ve 0,50-1,50 aralığında yer aldığından kabul edilebilir düzeydedir. Bu değerler özel bütünsel dereceli puanlama anahtarı kategorilerinin matematik başarı testindeki açık uçlu maddelerin değerlendirilmesinden elde edilen puanlarla uyumlu ve puanlama anahtarının güvenilir bir ölçme aracı olduğunu göstermektedir. 0 ve 2 numaralı kategorilerin 1,00 dış uyum indeks değerine sahip olması mükemmel uyumun göstergesidir.



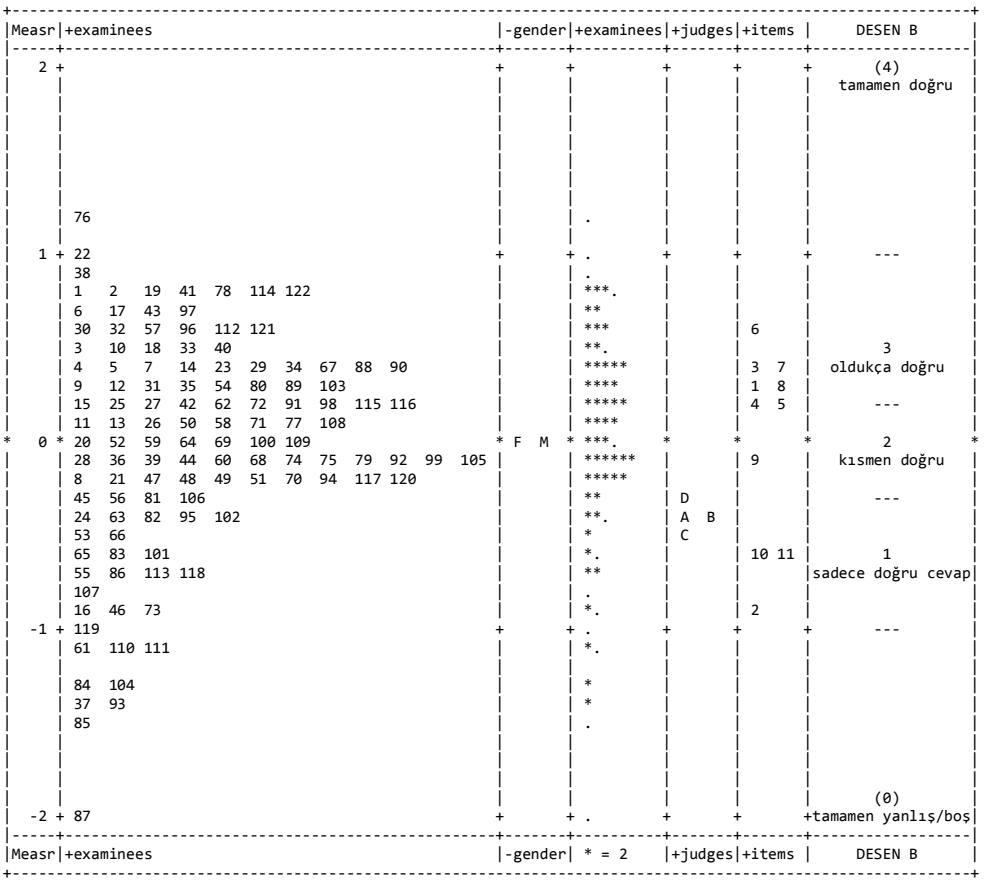
Şekil 2. Tamamen çaprazlanmış puanlama desenine (Desen A) göre dereceli puanlama anahtarı ölçütlerine ait olasılık eğrileri

Şekil 2 incelendiğinde yüksek beceri düzeyindeki öğrencilerin 3 ve 4 numaralı kategoriler gibi üst kategorilerde yer alma olasılığı yüksek; 0, 1, 2 numaralı kategoriler gibi alt kategorilerde yer alma olasılığı ise düşüktür. Düşük beceri düzeyindeki öğrencilerin ise alt kategorilerde yer alma olasılığı yüksek; üst kategorilerde yer alma olasılığı düşüktür. Bu durum özel bütünsel dereceli puanlama anahtarı kategorilerinin açık uçlu maddeleri cevaplandırmadaki öğrenci başarılarını ayırt etmede iyi olduğunu göstermiştir.

4.2. Tamamlanmamış-Bağlantılı Puanlama Desenine (Desen B) İlişkin Bulgular

4.2.1. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen B) Veri Kalibrasyon Haritasına İlişkin Bulgular

Her bir öğrencinin her bir açık uçlu maddedeki cevabının dört puanlayıcının üçü tarafından puanlandığı *(122x11x3=4026 gözlem)* Desen B'ye ilişkin veri kalibrasyon haritası Şekil 3'te sunulmuştur.



Şekil 3. Tamamlanmamış-bağlantılı puanlama deseni (Desen B) veri kalibrasyon haritası

Veri kalibrasyon haritası incelendiğinde matematik başarı testindeki açık uçlu maddeleri cevaplandırmada 76, 22 ve 38 nolu öğrencilerin yüksek beceriye, 85 ve 87 nolu öğrencilerin ise düşük beceriye sahip oldukları görülmüştür. Pozitif logit değerine sahip öğrenci sayısı ile negatif logit değere sahip öğrenci sayısının logit ölçekte eşit bir şekilde dağılması başarı açısından öğrenci grubunun heterojen olduğunu göstermektedir. Puanlayıcı yüzeyine ait sütuna bakıldığında puanlayıcı C'nin en katı, puanlayıcı D'nin en cömert puanlayıcı olduğu, puanlayıcı A ile puanlayıcı B'nin birbirine benzer puanlama yaptığı anlaşılmıştır. Açık uçlu madde yüzeyine bakıldığında madde güçlüğü en yüksek Madde 6 ve madde güçlüğü en düşük Madde 2 olarak belirlenmiştir. Öğrenciler matematik başarı testindeki Madde 1, Madde 3, Madde 4, Madde 5, Madde 6, Madde 7 ve Madde 8'i kolay cevaplandırmış; Madde 2, Madde 9, Madde 10, Madde 11'i cevaplandırmada ise zorlanmışlardır. Öğrenci performanslarını puanlamada özel bütünsel dereceli puanlama anahtarı derecelendirmelerinin puanlayıcılar tarafından geniş aralıkta ve etkin bir şekilde kullanıldığı görülmüştür.

4.2.2. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen B) Puanlayıcı Yüzeyine İlişkin Bulgular

Desen B için kalibrasyon haritasından elde edilen verilerle aynı doğrultuda ve ayrıntılı bilgi edinmek amacıyla puanlayıcı yüzeyine ilişkin ölçüm raporu Tablo 19'da sunulmuştur.

Tablo 19

Tamamlanmamış-Bağlantılı Puanlama Deseni Gösterimi (Desen B) Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu

	Model		Uygunluk ölçütleri		Uyum yüzdeleri	
Puanlayıcı	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)	Gözlenen uyum	Beklenen uyum
A	-0,42	0,03	1,04	1,08	82,9	37,9
B	-0,40	0,03	1,03	1,07	83,5	37,9
C	-0,53	0,03	0,95	1,07	0,0	0,0
D	-0,32	0,03	0,91	0,93	69,4	37,5
Ortalama	-0,42	0,03	0,98	1,04		
SS(evren)	0,07	0,00	0,05	0,06		
SS(örneklem	0,08	0,00	0,06	0,07		
Model (evren):	RMSE= 0,03	SS: 0,07	Ayırma oranı= 2,50		Ayırma indeksi= 3,67	
Güvenirlik (puanlayıcılar arası olmayan) = 0,86						
Model (örneklem):	RMSE=0,03	SS: 0,08	Ayırma oranı= 2,94		Ayırma indeksi= 4,26	
Güvenirlik (puanlayıcılar arası olmayan) = 0,90						
Model, sabit etkili ki-kare: 29,0		s.d: 3	p= 0,00			
Model, rastgele ki-kare: 2,7		s.d: 2	p= 0,25			
Puanlayılar arası uyumun gözlenen değeri= %78,6						
Puanlayıcılar arası uymun beklenen değeri= %37,8						
Puanlayıcılar arası güvenirlik (Rasch-Cohen'in kappa istatistiği)= 0,66						
RMSE: hata kareleri ortalamasının karekökü		*S.S:standart sapma*		*s.d:serbestlik derecesi*		

Tablo 19'da grup düzeyinde elde edilen istatistikler olan ayırma oranı=2,94, ayırma indeksi= 4,26 ve ayırma indeksi güvenirliği= 0,90 olarak bulunmuştur. Puanlayıcı yüzeyinde ayırma indeksinin 2,00'den büyük ve ayırma indeksi güvenirliğinin 0,50'den büyük olması nedeniyle puanlayıcıların katılık-cömertlik davranışlarının farklı olduğu ve puanlamaya puanlayıcıdan kaynaklanan hataların karıştığı düşünülmüştür. RMSE değeri 0,03 bulunmuş ve bu değerin düşük olması puanlayıcıların katılık davranışının daha kesin bir şekilde tahmin edildiğini göstermiştir. Sabit etkili ki-kare değeri $\chi^2(3)=29,0$, $p=0,00<0,05$ bulunmuş ve puanlayıcılar arasındaki katılık-cömertlik davranışları arasında istatistiksel olarak anlamlı fark olduğu görülmüştür.

Desen B için puanlayıcılar arası uyumun gözlenen değeri %78,6 puanlayıcılar arası uyumun beklenen değeri %37,8 bulunmuş ve Rasch-Cohen'in kappa istatistiği değeri 0,66 olarak

hesaplanmıştır. Bu değerin +0,50'den büyük olması, farklı puanlayıcıların öğrencilerin açık uçlu maddeleri cevaplandırma performanslarını değerlendirirken birbirleriyle tutarlı puanlamalar yaptıklarını, puanlayıcılar arası uyumun dolayısıyla güvenirliğin yüksek olduğunu göstermektedir.

Bireysel düzeydeki istatistiklerden logit değerlerine bakıldığında tüm puanlayıcıların negatif logit değerleri aldığı ve katı puanlama yaptığı görülmüştür. Puanlayıcı D en yüksek logit değerine (-0,32) sahip olduğundan en cömert, puanlayıcı C en düşük logit değerine (-0,53) sahip olduğundan en katı puanlayıcıdır. Logit değeri sırasıyla -0,40 ,-0,42 olan puanlayıcı A ve puanlayıcı B'nin benzer puanlama yaptığı anlaşılmıştır. Puanlayıcıların iç uyum indeksleri 0,91 ile 1,04 arasında, dış uyum indeksleri 0,93 ile 1,08 arasında değişmektedir ve 0,50 ile 1,50 arasında olduklarından kabul edilebilir düzeydedir. Bu değerler puanlayıcıların model ile veri uyumunu sağladığını ve birbirlerinden farklı puanlama yapmalarına rağmen puanlayıcıların dereceli puanlama anahtarı ölçütlerinin tamamını göz önünde bulundurarak tutarlı puanlama yaptıklarını göstermektedir.

4.2.3. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen B) Öğrenci Yüzeyine İlişkin Bulgular

Desen B için kalibrasyon haritasından elde edilen bilgilerin ayrıntılı bir şekilde ele alındığı öğrenci yüzeyine ilişkin ölçüm raporu Tablo 20'de sunulmuştur. Öğrencilerin tamamına ait verilerin verilmesi mümkün olmadığından logit değeri sıralaması en yüksek ve en düşük 8 öğrencinin ölçümlerine tabloda yer verilmiştir.

Tablo 20

Tamamlanmamış-Bağlantılı Puanlama Deseni Gösterimi (Desen B) Öğrenci Yüzeyine İlişkin Ölçüm Raporu

	Ortalamalar		Model		Uygunluk ölçütleri	
Öğrenci sıra no	Gözlenen ortalama	Düzeltilmiş ortalama	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)
76	3,15	3,32	1,23	0,17	0,64	0,76
22	2,85	3,05	0,98	0,16	1,21	0,98
38	2,79	2,99	0,93	0,15	1,08	0,91
78	2,67	2,85	0,84	0,15	0,78	0,72
41	2,73	2,82	0,82	0,15	1,23	1,11
1	2,70	2,78	0,79	0,15	0,37	0,38
19	2,64	2,76	0,79	0,15	0,59	0,60
2	2,58	2,74	0,77	0,15	1,29	1,36
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
61	0,30	0,24	-1,12	0,26	0,52	0,42
110	0,27	0,23	-1,13	0,28	0,86	0,68
84	0,21	0,18	-1,34	0,32	0,78	1,16
104	0,21	0,18	-1,34	0,32	0,80	0,99
37	0,21	0,17	-1,38	0,32	0,67	0,48
93	0,21	0,17	-1,38	0,32	1,08	1,48
85	0,18	0,15	-1,49	0,36	0,56	0,50
87	0,00	0,01	-4,17	1,81	minimum	
Ortalama	1,49	1,41	- 0,03	0,17	0,98	1,04
SS(evren)	0,73	0,80	0,68	0,16	0,37	0,61
SS(örneklem)	0,74	0,81	0,68	0,16	0,37	0,62
Model(evren) Güvenirlik: 0,88	RMSE: 0,23	SS: 0,63	Ayırma oranı= 2,73		Ayırma indeksi= 3,97	
Model(örneklem) Güvenirlik: 0,88	RMSE: 0,23	SS: 0,64	Ayırma oranı= 2,74		Ayırma indeksi= 3,98	
Model, sabit etkili ki-kare: 1050,1			s.d: 121		p= 0,00	
Model, rastgele ki-kare: 80,3			s.d: 120		p= 1,00	
RMSE: hata kareleri ortalamalarının karekökü			*SS: standart sapma*		*s.d: serbestlik derecesi*	

Desen B için grup düzeyinde elde edilen istatistiksel göstergeler incelendiğinde ayırma oranı=2,74, ayırma indeksi=3,98 ve ayırma indeksi güvenirliği=0,88 olduğu görülmüştür. Bu değerler açık uçlu maddeleri cevaplandırma becerisi birbirlerinden farklı olan öğrencilerin güvenilir bir şekilde ayırt edilebildiklerini göstermiştir. RMSE değeri 0,23 olarak bulunmuş ve yüksek beceriye sahip öğrencilerin daha kesin bir şekilde tahmin edildiği anlaşılmıştır. Ayırma indeksi G olmak üzere (4G+1)/3 formülü yüzey içerisindeki öge seviye sayısını vermektedir. Desen B için *([4.(3,98)+1)]/3=5,64)* öğrenci performansları altı seviyeye ayrılmıştır. Sabit etkili ki-kare değerinin $\chi^2(121)= 1050,1$, $p=0,00<0,05$ olması en az iki öğrencinin yetenek düzeyleri arasında anlamlı fark bulunduğunu ve öğrenci grubunun başarı açısından heterojen yapıda olduğunu göstermiştir. Logit değerlerine bakıldığında 76

numaralı öğrenci (1,23) en yüksek yetenek düzeyine, 87 numaralı öğrenci (-4,17) en düşük yetenek düzeyine sahiptir. Bireysel düzeydeki istatistiksel göstergeler olan iç uyum indeksleri minimum ile 1,94 arasında, dış uyum indeksleri minimum ile 3,49 arasında değişmiştir. Bu nedenle bazı öğrencilerin dış uyum indeksleri 2,00'den büyük olsa da iç uyum indekslerinin 2,00'den büyük olmaması model veri uyumunun bozulmadığını göstermiştir (Linacre, 2002). Öğrencilerin % 88,5'inin iç ve dış uyum indeksleri açısından kabul edilebilir aralıkta (0,50 – 1,50) olduğu görülmüştür. Analiz sonucunda iç uyum aralığı dışında kalan on dört öğrenci için her bir açık uçlu maddeyi cevaplandırma performanslarının beklenen düzeyde olmadığı; dış uyum aralığı dışında kalan on dört öğrencinin ise matematik başarı testi uygulamasında gösterdikleri genel performansın beklenen düzeyde olmadığı anlaşılmıştır. İç uyum indeks değeri 1,00 (beklenen değer) ve dış uyum indeks değeri 0,98 olan 52 numaralı öğrencinin her bir açık uçlu maddeyi cevaplandırma performansı ile matematik başarı testindeki genel performans uyumu en yüksek öğrenci olduğu gözlemlenmiştir.

4.2.4. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen B) Açık Uçlu Madde Yüzeyine İlişkin Bulgular

Desen B için açık uçlu madde yüzeyine ait madde güçlük düzey istatistikleri ile uyum indeksleri Tablo 21'de sunulmuştur.

Tablo 21

Tamamlanmamış-Bağlantılı Puanlama Deseni Gösterimi (Desen B) Açık Uçlu Madde Yüzeyine İlişkin Ölçüm Raporu

	Ortalamalar		Model		Uygunluk ölçütleri	
Madde no	Gözlenen ortalama	Düzeltilmiş ortalama	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)
1	1,90	1,83	0,31	0,04	1,05	1,08
2	0,43	0,31	-0,92	0,07	1,06	0,99
3	2,02	1,99	0,38	0,04	0,79	0,76
4	1,81	1,71	0,25	0,04	0,65	0,63
5	1,66	1,53	0,16	0,04	1,12	1,09
6	2,29	2,35	0,56	0,04	1,23	1,23
7	2,01	1,97	0,37	0,04	1,19	1,14
8	1,85	1,77	0,28	0,04	0,85	0,80
9	1,25	1,04	-0,12	0,04	0,69	0,68
10	0,64	0,47	-0,64	0,06	1,37	1,72
11	0,64	0,47	-0,64	0,06	0,97	1,30
Ortalama	1,50	1,40	0,00	0,05	1,00	1,04
SS(evren)	0,62	0,68	0,48	0,01	0,22	0,30
SS(örneklem)	0,65	0,71	0,50	0,01	0,23	0,32
Model(evren) Güvenirlik: 0,99	RMSE: 0,05	SS: 0,48	Ayırma oranı= 10,03		Ayırma indeksi= 13,71	
Model(örneklem) Güvenirlik: 0,99	RMSE: 0,05	SS: 0,50	Ayırma oranı= 10,53		Ayırma indeksi= 14,37	
Model, sabit etkili ki-kare: 869,5			s.d: 10		p= 0,00	
Model, rastgele ki-kare: 9,9			s.d: 9		p= 0,36	

RMSE: hata kareleri ortalamalarının karekökü *SS: standart sapma* *s.d: serbestlik derecesi*

Açık uçlu madde yüzeyine ilişkin ayırma oranı=10,53, ayırma indeksi=14,37, ayırma indeksi güvenirliği=0,99 şeklindedir. Açık uçlu maddelerin güçlük düzeyleri arasındaki ölçme hatası ile açıklanamayacak kadar fazla olan farklılığın belirlenmesini sağlayan ayırma oranı, ayırma indeksi gibi yorumlanmaktadır. Ayırma oranının yüksek olması, madde güçlük düzeylerinin farklı olduğunu göstermiştir. Ayırma indeksinin 2,00'den ve ayırma indeksi güvenirliğinin 0,50'den büyük olduğu ve farklı güçlük düzeylerine sahip maddelerin yüksek güvenirlikle birbirlerinden ayırt edildiği görülmüştür. Sabit etkili ki-kare testi ile ($\chi^2(10)$= 869,5, p=0,00<0,05) madde güçlük düzeyleri arasındaki farklılığın istatistiksel olarak anlamlı olduğu anlaşılmıştır.

Açık uçlu maddelerin güçlük düzeylerini gösteren logit değerleri 0,56 ile -0,92 arasında değişmektedir. Veri kalibrasyon haritasında da görüldüğü gibi Tablo 22'de Madde 6 (logit=0,56)'nın madde güçlüğü en yüksek (kolay cevaplanabilen) ve Madde 2 (logit= -0,92)'nin madde güçlüğü en düşük (zor cevaplanabilen) olduğu görülmektedir. Logit değerlerine bakılarak madde güçlüğü düşük maddeler, Madde 2, Madde 11, Madde 10,

Madde 9; madde güçlüğü yüksek maddeler, Madde 6, Madde 3, Madde 7, Madde 1, Madde 8, Madde 4, Madde 5 şeklinde sıralanmıştır. Maddelerin güçlük/kolaylık durumunu gösteren RMSE değeri 0,04 bulunmuştur. Standart hatanın düşük değerler alması ile madde güçlüğünün daha kesin bir şekilde tahmin edildiği anlaşılmıştır. Standart hata değeri sırasıyla 0,07; 0,06; 0,06 olan Madde 2, Madde 10 ve Madde 11 dışındaki maddelerin madde güçlükleri daha kesin tahmin edilmiştir. İç uyum indeksleri 0,65 ile 1,37 arasında, dış uyum indeksleri 0,63 ile 1,72 arasında değişmektedir. Tablo 22'de yer alan iç ve dış uyum değerlerinin genel olarak 0,50 ile 1,50 aralığında ve kabul edilebilir düzeyde olduğu görülmüştür. Sadece Madde 10'da iç uyum değeri (1,37) kabul edilebilir düzeyde olsa da dış uyum değeri (1,72) bulunmuştur. Uyum indeksleri 2,00'den büyük değil ise ölçüm oluşturmada bir sorun çıkmayacağı Linacre (2002) tarafından belirtilmiş ve bu madde de kabul edilebilir düzeyde olarak belirlenmiştir. İç uyum indeksinin kabul edilebilir düzeyde olması maddenin istenilen kazanımı ölçmede iyi olduğunu, dış uyum indeksinin kabul edilebilir düzey dışında kalması öğrencilerin matematik başarısını ölçmede yetersiz bir madde olduğunu göstermektedir. İç uyum indekslerine bakılarak ilgili kazanımı ölçmede her bir maddenin iyi işlediği ve dış uyum indekslerine bakılarak da başarı testinde yer alan maddelerin öğrenci matematik başarısını ölçmede iyi olduğu görülmüştür.

4.2.5. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen B) Özel Bütünsel Dereceli Puanlama Anahtarı Ölçütlerine İlişkin Bulgular

Desen B için dereceli puanlama anahtarı puanlama kategorilerine ilişkin değerler Tablo 22'de sunulmuştur.

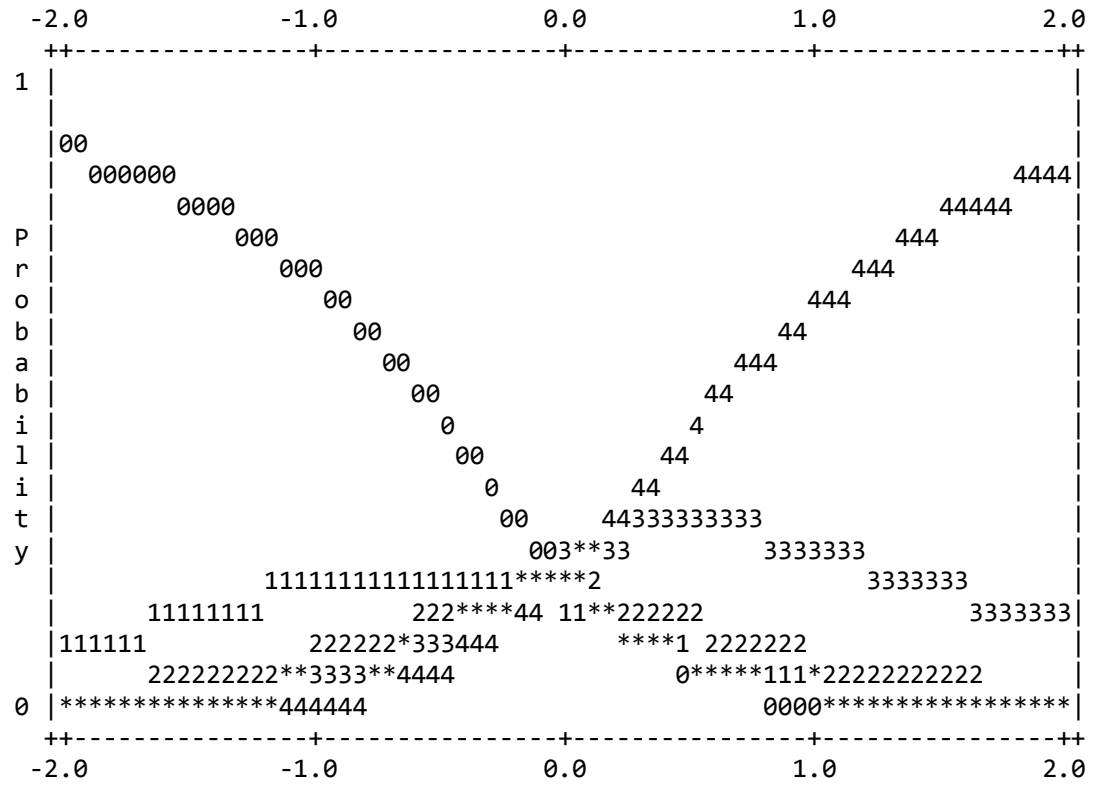
Tablo 22

Özel Bütünsel Dereceli Puanlama Anahtarı Puanlama Kategorilerine Ait Ölçüm Raporu

Puanlama Kategorileri	Frekans	Yüzde (%)	Birikimli yüzde (%)	Ortalama logit değeri	Beklenen logit değeri	Outfit (dış uyum)
0	1622	41	41	-0,88	-0,90	1,00
1	640	16	57	-0,59	-0,50	1,10
2	490	12	69	-0,19	-0,18	0,90
3	594	15	84	0,12	0,06	0,80
4	647	16	100	0,25	0,27	1,20

Tablo 22'deki frekans ve yüzde değerlerine bakıldığında; puanlayıcıların puanlama yaparken en çok 0 numaralı kategoriyi tercih ettikleri ve 1, 2, 3 ve 4 numaralı kategorileri birbirine yakın sayıda kullandıkları görülmüştür. 0 kategorisinin daha fazla kullanılmasının

sebebi öğrenci yüzeyi ölçüm raporunda öğrencilerin matematik başarı düzeylerinin yüksek güvenirlikle ayırt edildiğinden merkezi eğilim etkisininden kaynaklanmamaktadır. Logit değerleri incelendiğinde 3 ve 4 numaralı kategorilerin tam ve oldukça doğru tanımlarına uygun bir şekilde pozitif değer aldığı; 0, 1, 2 numaralı kategorilerin boş, tamamen yanlış ve kısmen yanlış tanımlarına uygun olarak negatif değerler aldığı anlaşılmıştır. Tablodaki ortalama ve beklenen logit değerleri birbirine yakın olması verilerin Rasch modeli beklentilerine uyduğunu göstermiştir. Dış uyum indeksleri 0,80 ile 1,20 arasında değişmektedir ve 0,50-1,50 aralığında yer aldığından kabul edilebilir düzeydedir. Bu durum da özel bütünsel dereceli puanlama anahtarı kategorilerinin matematik başarı testindeki açık uçlu maddelerin değerlendirilmesinden elde edilen puanlarla uyumlu ve puanlama anahtarının güvenilir bir ölçme aracı olduğunu göstermektedir. 0 numaralı kategorinin 1,0 dış uyum indeks değerine sahip olması mükemmel uyumun göstergesidir.



Şekil 4. Tamamlanmamış-bağlantılı puanlama deseni gösterimi (Desen B) dereceli puanlama anahtarı ölçütlerine ait olasılık eğrileri

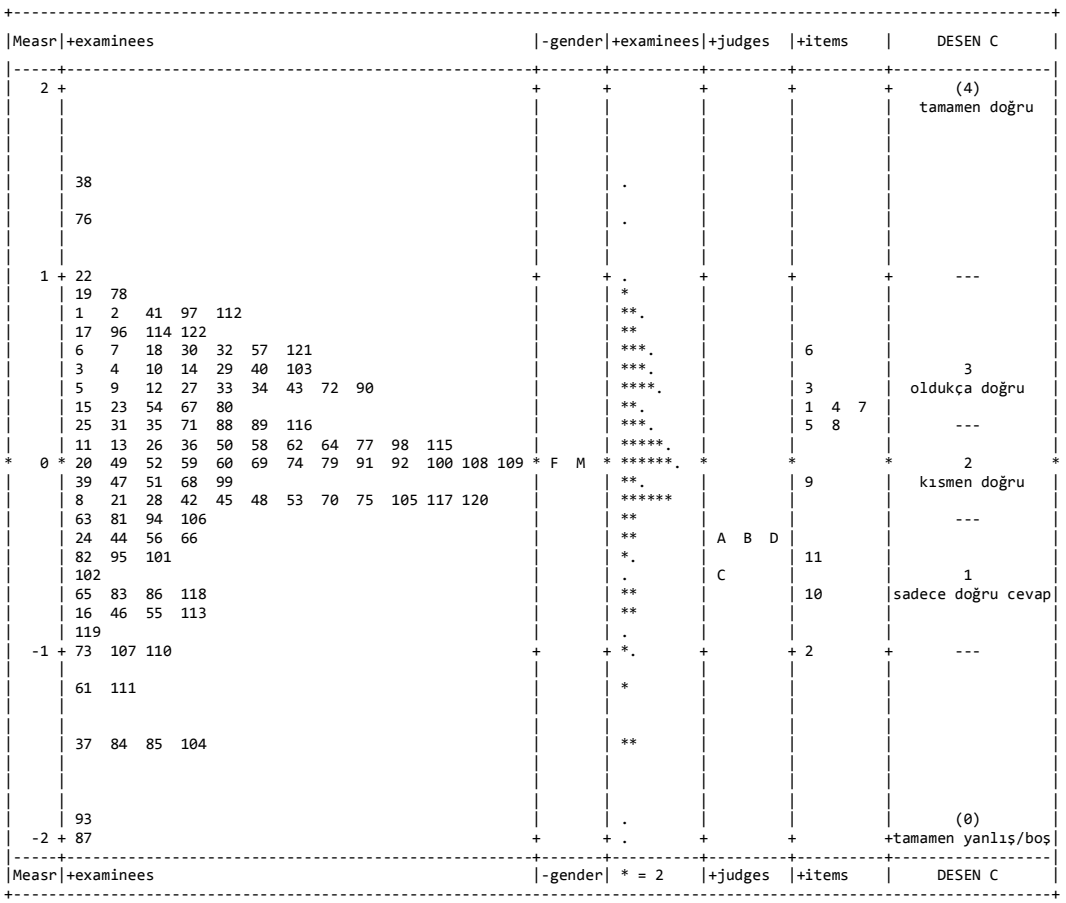
Şekil 4 incelendiğinde düşük beceri düzeyindeki öğrencilerin 0, 1, 2 numaralı kategoriler gibi alt kategorilerde yer alma olasılığı yüksek; 3 ve 4 numaralı kategoriler gibi üst kategorilerde yer alma olasılığı düşüktür. Benzer şekilde yüksek beceri düzeyindeki öğrencilerin üst kategorilerde yer alma olasılığı yüksek; alt kategorilerde yer alma olasılığı

ise düşüktür. Bu durum açık uçlu maddeleri cevaplandırmadaki öğrenci başarılarını ayırt etmede özel bütünsel dereceli puanlama anahtarı kategorilerinin iyi olduğunu göstermiştir.

4.3. Tamamlanmamış-Bağlantılı Puanlama Desenine (Desen C) İlişkin Bulgular

4.3.1. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Veri Kalibrasyon Haritasına İlişkin Bulgular

Her bir öğrencinin öncelikle tüm açık uçlu maddelerde bir puanlayıcı tarafından puanlandığı, sıralama numarası 10 ve 10'un katı olan öğrencilerin performanslarının dört puanlayıcı tarafından puanlandığı [*(110x1x11) +(12x11x4)=1738 gözlem)* Desen C'ye ilişkin kalibrasyon haritası Şekil 5'te sunulmuştur.



Şekil 5. Tamamlanmamış-bağlantılı puanlama deseni (Desen C) veri kalibrasyon haritası

Kalibrasyon haritası incelendiğinde matematik başarı testindeki açık uçlu maddeleri cevaplandırmada 38, 76 ve 22 nolu öğrencilerin yüksek beceriye, 93 ve 87 nolu öğrencilerin

ise düşük beceriye sahip oldukları görülmüştür. Pozitif logit değerine sahip öğrenci sayısı ile negatif logit değere sahip öğrenci sayısının logit ölçekte birbirine yakın olması başarı açısından öğrenci grubunun heterojen olduğunu göstermektedir. Puanlayıcı sütununa bakıldığında puanlayıcı C'nin en katı, puanlayıcı A, B ve D'nin daha cömert ve birbirine yakın puanlama yaptığı anlaşılmıştır. Açık uçlu madde yüzeyine bakıldığında madde güçlüğü en yüksek Madde 6 ve madde güçlüğü en düşük Madde 2 olarak belirlenmiştir. Öğrenciler matematik başarı testindeki Madde 1, Madde 3, Madde 4, Madde 5, Madde 6, Madde 7 ve Madde 8'i kolay cevaplandırmış; Madde 2, Madde 9, Madde 10, Madde 11'i cevaplandırmada ise zorlanmışlardır. Puanlayıcılar tarafından öğrenci performanslarını puanlamada özel bütünsel dereceli puanlama anahtarının derecelendirmeleri etkin bir şekilde geniş aralıkta kullanılmıştır.

4.3.2. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Puanlayıcı Yüzeyine İlişkin Bulgular

Kalibrasyon haritasında puanlayıcı yüzeyi için elde edilen yorumları ayrıntılandırıldığı Desen C'ye ait puanlayıcı ölçüm raporu Tablo 23'te sunulmuştur.

Tablo 23

Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu

	Model		Uygunluk ölçütleri		Uyum yüzdeleri	
Puanlayıcı	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)	Gözlenen uyum	Beklenen uyum
A	-0,41	0,04	1,02	0,97	82,6	33,7
B	-0,42	0,04	1,07	1,15	81,8	33,7
C	-0,56	0,04	0,88	0,97	0,0	0,0
D	-0,37	0,04	0,95	0,98	68,2	33,6
Ortalama	-0,44	0,04	0,98	1,02		
SS(evren)	0,07	0,00	0,07	0,08		
SS(örneklem	0,09	0,00	0,08	0,09		
Model (evren):	RMSE= 0,04	SS: 0,06	Ayırma oranı=1,46		Ayırma indeksi= 2,28	
Güvenirlik (puanlayıcılar arası olmayan) = 0,68						
Model (örneklem):	RMSE= 0,04	SS: 0,07	Ayırma oranı=1,78		Ayırma indeksi= 2,71	
Güvenirlik (puanlayıcılar arası olmayan) = 0,76						
Model, sabit etkili ki-kare: 11,4	s.d: 3	p= 0,01				
Model, rastgele ki-kare: 2,4	s.d: 2	p= 0,30				
Puanlayılar arası uyumun gözlenen değeri= %77,5						
Puanlayıcılar arası uymun beklenen değeri= %33,7						
Puanlayıcılar arası güvenirlik (Rasch-Cohen'in kappa istatistiği)= 0,66						
RMSE: hata kareleri ortalamasının karekökü	*S.S:standart sapma*	*s.d:serbestlik derecesi*				

Tablo 23'te grup düzeyinde elde edilen istatistikler; ayırma oranı=1,78, ayırma indeksi= 2,71 ve ayırma indeksi güvenirliği=0,76 olarak bulunmuştur. Puanlayıcı yüzeyinde ayırma indeksinin 2,00'den büyük ve ayırma indeksi güvenirliğinin 0,50'den büyük olması nedeniyle puanlayıcıların katılık-cömertlik davranışlarının farklı olduğu ve puanlamaya puanlayıcıdan kaynaklanan hataların karıştığı düşünülmüştür. Ayırma oranının 2,00'den küçük çıkması, puanlayıcı yüzeyinde ölçme hatası ile açıklanabilecek bir farklılığın oluştuğuna işaret etmiştir. RMSE değeri 0,02 bulunmuş ve bu değerin düşük olması, puanlayıcıların katılık davranışının daha kesin bir şekilde tahmin edildiğini göstermiştir. Sabit etkili ki-kare değeri $\chi^2(3)=11,4$, $p=0,01<0,05$ bulunmuş ve en az iki puanlayıcının katılık-cömertlik davranışları arasında istatistiksel olarak anlamlı fark olduğu görülmüştür.

Tamamlanmamış bağlantılı puanlama deseni için puanlayıcılar arası uyumun gözlenen değeri %77,5, puanlayıcılar arası uyumun beklenen değeri %33,7 bulunmuş ve Rasch-Cohen'in kappa istatistiği değeri 0,66 olarak hesaplanmıştır. Bu değerin +0,50'den büyük olması, farklı puanlayıcıların öğrencilerin açık uçlu maddeleri cevaplandırma performanslarını değerlendirirken birbirleriyle tutarlı puanlamalar yaptıklarını, puanlayıcılar arası uyumun dolayısıyla güvenirliğin yüksek olduğunu göstermektedir.

Bireysel düzeydeki istatistiklerden logit değerlerine bakıldığında tüm puanlayıcıların negatif logit değerleri aldığı ve katı puanlama yaptığı görülmüştür. Puanlayıcı D en yüksek logit değerine (-0,37) sahip olduğundan en cömert, puanlayıcı C en düşük logit değerine (-0,56) sahip olduğundan en katı puanlayıcıdır. Logit değeri sırasıyla -0,41, -0,42 olan puanlayıcı A ve puanlayıcı B ile puanlayıcı D'nin birbirlerine yakın puanlama yaptığı anlaşılmıştır. Puanlayıcıların iç uyum indeksleri 0,88 ile 1,07 arasında, dış uyum indeksleri 0,97 ile 1,15 arasında değişmektedir ve 0,50 ile 1,50 arasında olduklarından kabul edilebilir düzeydedir. Bu değerler puanlayıcıların model ile veri uyumunu sağladığını ve birbirlerinden farklı puanlama yapmalarına rağmen puanlayıcıların dereceli puanlama anahtarı ölçütlerinin tamamını göz önünde bulundurarak tutarlı puanlama yaptıklarını göstermektedir.

4.3.3. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Öğrenci Yüzeyine İlişkin Bulgular

Öğrencilerin tamamına ait verilerin paylaşılması mümkün olmadığından logit değeri en yüksek ve en düşük 8 öğrencinin verileri Tablo 24'te sunulmuştur.

Tablo 24

Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Öğrenci Yüzeyine İlişkin Ölçüm Raporu

	Ortalamalar		Model		Uygunluk ölçütleri	
Öğrenci sıra no	Gözlenen ortalama	Düzeltilmiş ortalama	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)
38	3,36	3,48	1,46	0,34	0,76	0,91
76	3,09	3,37	1,30	0,30	0,27	0,35
22	3,00	3,08	1,02	0,29	1,50	1,04
19	2,91	2,97	0,94	0,28	0,73	0,77
78	2,82	2,82	0,92	0,27	0,94	0,76
1	2,73	2,83	0,84	0,26	0,40	0,40
2	2,73	2,75	0,79	0,26	1,36	1,41
112	2,73	2,75	0,79	0,26	1,51	1,60
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
61	0,27	0,22	-1,17	0,48	0,41	0,36
111	0,27	0,22	-1,17	0,48	0,98	0,85
85	0,18	0,15	-1,45	0,61	0,52	0,43
37	0,18	0,15	-1,46	0,61	0,49	0,36
84	0,18	0,15	-1,46	0,61	0,60	1,19
104	0,18	0,15	-1,46	0,61	0,56	0,44
93	0,09	0,09	-1,86	0,92	0,86	1,81
87	0,00	0,02	-3,10	1,78	minimum	
Ortalama	1,48	1,40	- ,03	,28	,93	,97
SS(evren)	,77	,83	,66	,17	,39	,64
SS(örneklem)	,78	,83	,67	,18	,39	,65
Model(evren): Güvenirlik: 0,75	RMSE: 0,33	SS: 0,57	Ayırma oranı= 1.71		Ayırma indeksi= 2.62	
Model(örneklem): Güvenirlik: 0,75	RMSE: 0,33	SS: 0,58	Ayırma oranı= 1.72		Ayırma indeksi= 2.63	
Model, sabit etkili ki-kare: 425,1		s.d: 121		p= 0,00		
Model, rastgele ki-kare: 76,1		s.d: 120		p= 1,00		

RMSE: hata kareleri ortalamalarının karekökü SS: standart sapma s.d: serbestlik derecesi

Tablo 24 incelendiğinde ayırma oranı=1,72, ayırma indeksi=2,63 ve ayırma indeksi güvenirliği=0,75 olduğu görülmüştür. Bu değerlerin açık uçlu maddeleri cevaplandırma becerisi birbirlerinden farklı olan öğrencilerin güvenilir bir şekilde ayırt edilebildiklerini göstermiştir. Ayırma oranının 2,00'den küçük olması oluşan farklılığın ölçme hatası ile açıklanabileceğini göstermiştir. Öğrencilerin başarı ölçümlerinin RMSE değeri 0,33 olarak bulunmuş ve yüksek beceriye sahip öğrencilerin daha kesin bir şekilde tahmin edildiği anlaşılmıştır. Ayırma indeksi G olmak üzere (4G+1)/3 formülü yüzey içerisindeki öge seviye sayısını vermektedir. Desen C için *([4.(2,63)+1)]/3=3,84)* öğrenci performanslarının dört seviyeye ayrıldığı söylenebilir. Sabit etkili ki-kare değerinin $\chi^2(121)=425,1$, $p=0,00<0,05$

olması en az iki öğrencinin yetenek düzeyleri arasında anlamlı fark bulunduğunu ve öğrenci grubunun başarı açısından heterojen yapıda olduğunu göstermiştir. Logit değeri (1,46) olan 38 numaralı öğrenci en yüksek yetenek düzeyine, logit değeri (-3,10) olan 87 numaralı öğrenci en düşük yetenek düzeyine sahiptir. Bireysel düzeydeki istatistiksel göstergeler olan iç uyum indeksleri minimum ile 2,07 arasında, dış uyum indeksleri minimum ile 6,26 arasında değişmiştir. Bazı öğrencilerin dış uyum indeksleri 2,00'den büyük olsa da iç uyum indekslerinin 2,00'den büyük olmaması model veri uyumunun bozulmadığını göstermiştir (Linacre, 2002). Öğrencilerin % 85,25'inin iç ve dış uyum indeksleri açısından kabul edilebilir aralıkta (0,50-1,50) olduğu görülmüştür. Analiz sonucunda iç uyum aralığı dışında kalan on sekiz öğrenci için her bir açık uçlu maddeyi cevaplandırma performanslarının beklenen düzeyde olmadığı; dış uyum aralığı dışında da kalan on sekiz öğrencinin matematik başarı testi uygulamasında gösterdikleri genel performansın beklenen düzeyde olmadığı anlaşılmıştır. İç uyum indeks değeri 1,00 (beklenen değer) ve dış uyum indeks değeri 0,97 olan 27 numaralı öğrencinin her bir açık uçlu maddeyi cevaplandırma performansı ile matematik başarı testindeki genel performans uyumu en yüksek öğrenci olduğu gözlemlenmiştir.

4.3.4. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Açık Uçlu Madde Yüzeyine İlişkin Bulgular

Kalibrasyon haritasından elde edilen verilerin ayrıntılandırıldığı, madde güçlük düzeyleri ve uyum indekslerinin yer aldığı Tablo 25 aşağıda sunulmuştur.

Tablo 25

Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Açık Uçlu Madde Yüzeyine İlişkin Ölçüm Raporu

	Ortalamalar		Model		Uygunluk ölçütleri	
Madde no	Gözlenen ortalama	Düzeltilmiş ortalama	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)
1	1,85	1,72	0,27	0,06	1,08	1,15
2	0,40	0,28	-0,97	0,11	1,06	0,99
3	2,03	1,95	0,38	0,06	0,73	0,70
4	1,88	1,76	0,29	0,06	0,75	0,74
5	1,66	1,48	0,15	0,06	1,22	1,18
6	2,39	2,43	0,61	0,07	1,25	1,23
7	1,98	1,89	0,35	0,06	1,15	1,06
8	1,78	1,62	0,22	0,06	0,81	0,75
9	1,25	1,01	-0,12	0,07	0,63	0,62
10	0,62	0,44	-0,66	0,09	1,43	1,71
11	0,75	0,55	-0,52	0,08	1,13	1,34
Ortalama	1,51	1,38	0,00	0,07	1,00	1,01
SS(evren)	0,62	0,67	0,48	0,01	0,25	0,33
SS(örneklem)	0,65	0,70	0,50	0,01	0,26	0,35
Model(evren): Güvenirlik: 0,98	RMSE: 0,07	SS: 0,47	Ayırma oranı= 6,57		Ayırma indeksi= 9,09	
Model(örneklem): Güvenirlik: 0,98	RMSE: 0,07	SS: 0,50	Ayırma oranı= 6,90		Ayırma indeksi= 9,53	
Model, sabit etkili ki-kare: 371,2			s.d: 10		p= 0,00	
Model, rastgele ki-kare: 9,7			s.d: 9		p= 0,37	
RMSE: hata kareleri ortalamalarının karekökü			*SS: standart sapma*		*s.d: serbestlik derecesi*	

Açık uçlu madde yüzeyine ilişkin ayırma oranı= 6,57, ayırma indeksi= 9,09, ayırma indeksi güvenirliği=0,98 şeklindedir. Ayırma oranı madde güçlük düzeylerinin ölçme hatası ile açıklanamayacak kadar birbirlerinden farklı olduğunu göstermiştir. Ayırma indeksinin 2,00' den ve ayırma indeksi güvenirliğinin 0,50'den büyük olduğu ve farklı güçlük düzeylerine sahip maddelerin yüksek güvenirlikle birbirlerinden ayırt edildiği görülmüştür. Sabit etkili ki-kare testi ile ($\chi^2(10)= 371,2$, $p=0,00<0,05$) madde güçlük düzeyleri arasındaki farklılığın istatistiksel olarak anlamlı olduğu anlaşılmıştır.

Açık uçlu maddelerin güçlük düzeylerinin logit değerleri 0,61 ile -0,97 arasında değişmektedir. Tablo 25'te Madde 6 (logit=0,61)'nın madde güçlüğü en yüksek (kolay cevaplanabilen) ve Madde 2 (logit= -0,97)'nin madde güçlüğü en düşük (zor cevaplanabilen) olduğu görülmektedir. Logit değerlerine bakıldığında madde güçlüğü düşük maddeler, Madde 2, Madde 10, Madde 11, Madde 9; madde güçlüğü yüksek maddeler Madde 6, Madde 3, Madde 7, Madde 4, Madde 1, Madde8, Madde 5 şeklinde sıralanmıştır. Maddelerin güçlük/kolaylık durumunu gösteren RMSE değeri 0,07 bulunmuş ve madde güçlüğünün

daha kesin bir şekilde tahmin edildiği anlaşılmıştır. Standart hata değeri 0,06 olan Madde 1, Madde 3, Madde 4, Madde 5, Madde 7 ve Madde 8'in madde güçlükleri daha kesin tahmin edilmiştir. İç uyum indeksleri 0,63 ile 1,43 arasında, dış uyum indeksleri 0,62 ile 1,71 arasında değişmektedir. Tablo 25'te yer alan iç ve dış uyum değerlerinin genel olarak 0,50 ile 1,50 aralığında ve kabul edilebilir düzeyde olduğu görülmüştür. Sadece Madde 10'da iç uyum değeri (1,43) kabul edilebilir düzeyde olsa da dış uyum değeri (1,71) bulunmuştur. Uyum indeksleri 2,00'den büyük değil ise ölçüm oluşturmada bir sorun çıkmayacağı Linacre (2002) tarafından belirtilmiş ve bu madde de kabul edilebilir düzeyde olarak belirlenmiştir. İç uyum indekslerine bakılarak ilgili kazanımı ölçmede her bir maddenin iyi işlediği ve dış uyum indekslerine bakılarak da başarı testinde yer alan maddelerin öğrenci matematik başarısını ölçmede iyi olduğu görülmüştür.

4.3.5. Tamamlanmamış-Bağlantılı Puanlama Deseni (Desen C) Özel Bütünsel Dereceli Puanlama Anahtarı Ölçütlerine İlişkin Bulgular

Desen C için dereceli puanlama anahtarı kategorilerine ilişkin veriler Tablo 26'da sunulmuştur.

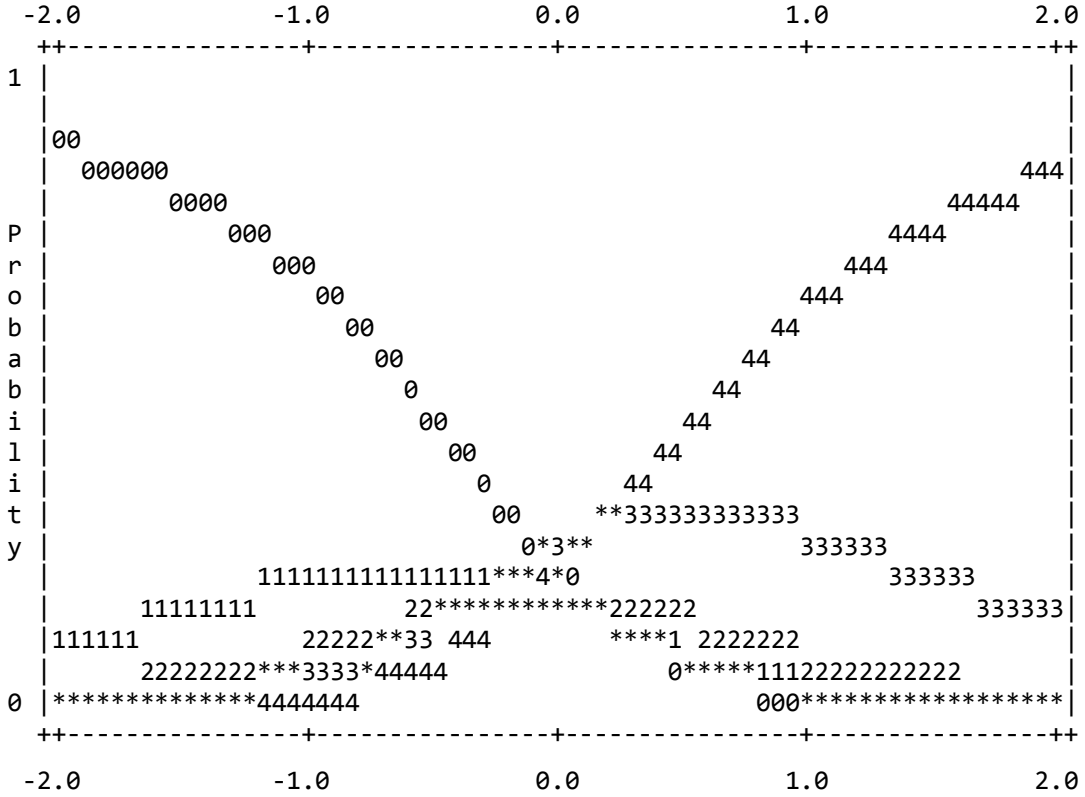
Tablo 26

Özel Bütünsel Dereceli Puanlama Anahtarı Puanlama Kategorilerine Ait Ölçüm Raporu

Puanlama Kategorileri	Frekans	Yüzde (%)	Birikimli yüzde (%)	Ortalama logit değeri	Beklenen logit değeri	Outfit (dış uyum)
0	699	40	40	-0,89	-0,92	1,00
1	274	16	56	-0,57	-0,50	1,10
2	206	12	68	-0,21	-0,19	0,90
3	273	16	84	0,08	0,06	0,90
4	275	16	100	0,27	0,26	1,10

Tablo 26'daki frekans ve yüzde değerlerine bakıldığında; puanlayıcıların puanlama yaparken en çok 0 numaralı kategoriyi tercih ettikleri ve 1, 2, 3 ve 4 numaralı kategorileri birbirine yakın sayıda kullandıkları görülmüştür. Uç değer olan 0 kategorisinin daha fazla kullanılmasının sebebi öğrenci yüzeyi ölçüm raporunda öğrencilerin matematik başarı düzeylerinin yüksek güvenirlikle ayırt edildiğinden merkezi eğilim etkisininden kaynaklanmamaktadır. Logit değerleri incelendiğinde 3 ve 4 numaralı kategorilerin tam ve oldukça doğru tanımlarına uygun bir şekilde pozitif değer aldığı; 0, 1, 2 numaralı kategorilerin boş, tamamen yanlış ve kısmen yanlış tanımlarına uygun olarak negatif

değerler aldığı anlaşılmıştır. Ortalama ve beklenen logit değerlerinin birbirine yakın olması verilerin Rasch modeli beklentilerine uyduğunu göstermiştir. Dış uyum indeksleri 0,90 ile 1,10 arasında değişmektedir ve 0,50-1,50 aralığında yer aldığından kabul edilebilir düzeydedir. Özel bütünsel dereceli puanlama anahtarı kategorilerinin matematik başarı testindeki açık uçlu maddelerin değerlendirilmesinden elde edilen puanlarla uyumlu ve puanlama anahtarının güvenilir bir ölçme aracı olduğu görülmüştür. 1,00 dış uyum indeks değerine sahip olan 0 numaralı kategori, matematik başarı testindeki öğrenci performanslarının değerlendirilme puanları ile mükemmel uyum içerisindedir.



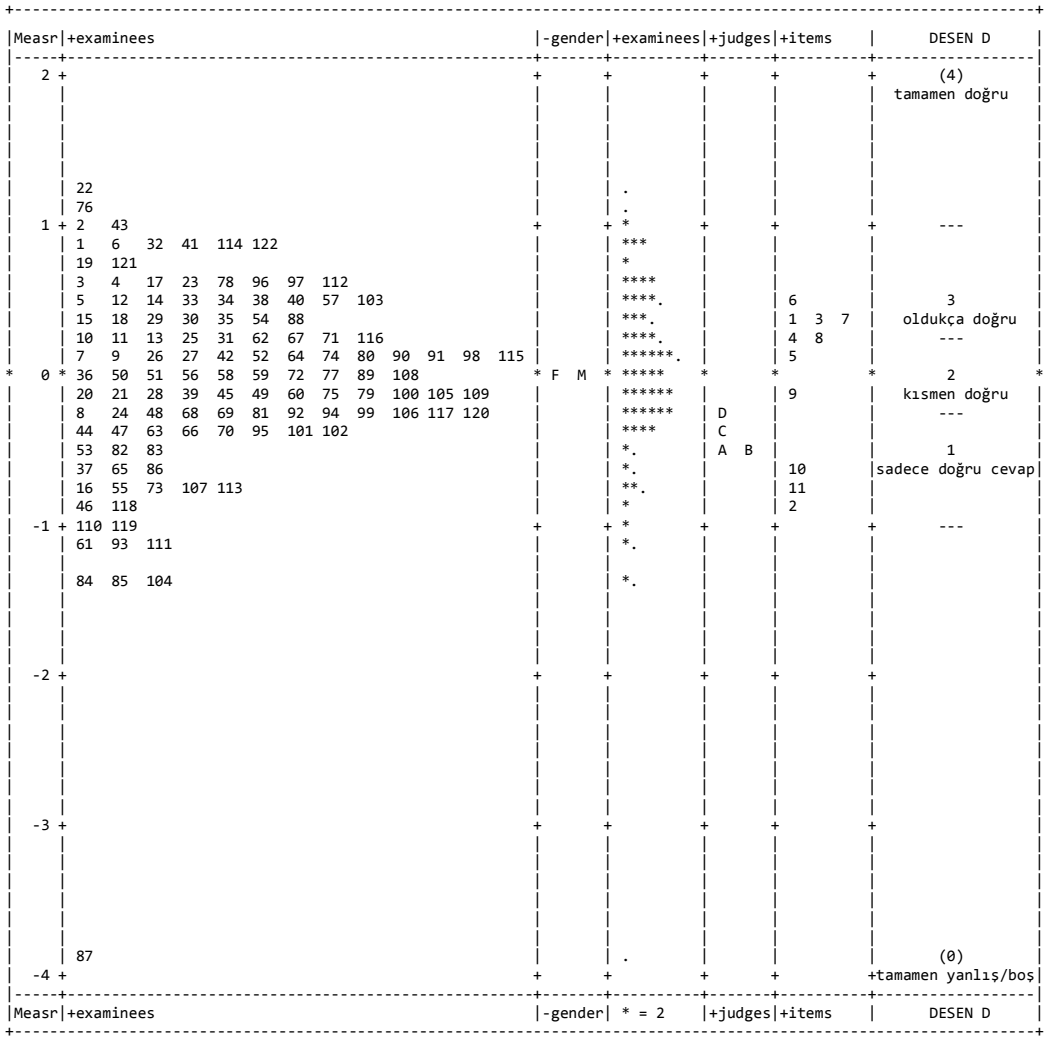
Şekil 6. Tamamlanmamış-bağlantılı puanlama deseni (Desen C) dereceli puanlama anahtarı ölçütlerine ait olasılık eğrileri

Olasılık eğrileri incelendiğinde düşük beceri düzeyindeki öğrencilerin 0, 1, 2 numaralı kategoriler gibi alt kategorilerde bulunma olasılığı yüksek; 3 ve 4 numaralı kategoriler gibi üst kategorilerde bulunma olasılığı düşüktür. Aynı zamanda yüksek beceri düzeyindeki öğrencilerin üst kategorilerde bulunma olasılığı yüksek; alt kategorilerde bulunma olasılığı ise düşüktür. Bunun sonucunda açık uçlu maddeleri cevaplandırmadaki öğrenci başarılarını ayırt etmede özel bütünsel dereceli puanlama anahtarı kategorilerinin iyi olduğu görülmüştür.

4.4. Tamamlanmamış-Bağlantılı Olmayan Puanlama Desenine (Desen D) İlişkin Bulgular

4.4.1. Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Veri Kalibrasyon Haritasına İlişkin Bulgular

Puanlayıcı A ve B'nin sıralama numarasının birler basamağı 1, 2, 3, 4, 5 ve 6 olan öğrencilerin her bir açık uçlu maddedeki performanslarını puanladığı; Puanlayıcı C ve D'nin sıralama numarasının birler basamağı 7, 8, 9 ve 0 olan öğrencilerin her bir açık uçlu maddedeki performanslarını puanladığı *[(2x11x74) + (2x11x48)=2684 gözlem]* Desen D'ye ilişkin kalibrasyon haritası Şekil 7'de sunulmuştur.



Şekil 7. Tamamlanmamış-bağlantılı olmayan puanlama deseni (Desen D) veri kalibrasyon haritası

Kalibrasyon haritası incelendiğinde matematik başarı testindeki açık uçlu maddeleri cevaplandırmada 22 ve 76 nolu öğrencilerin yüksek beceriye, 87 nolu öğrencilerin ise düşük beceriye sahip oldukları görülmüştür. Pozitif logit değerine sahip öğrenci sayısı ile negatif logit değere sahip öğrenci sayısı logit ölçekte birbirine yakın dağıldığı, başarı açısından öğrenci grubunun heterojen olduğu görülmüştür. Puanlayıcı yüzeyine bakıldığında puanlayıcı A ve B'nin birbirine benzer puanlama yaptığı ve en katı, puanlayıcı D'nin en cömert puanlayıcı olduğu anlaşılmıştır. Açık uçlu madde yüzeyine bakıldığında madde güçlüğü en yüksek Madde 6 ve madde güçlüğü en düşük Madde 2 olarak belirlenmiştir. Öğrenciler matematik başarı testindeki Madde 1, Madde 3, Madde 4, Madde 5, Madde 6, Madde 7 ve Madde 8'i kolay cevaplandırmış; Madde 2, Madde 9, Madde 10, Madde 11'i cevaplandırmada ise zorlanmışlardır. Puanlayıcılar tarafından öğrenci performanslarını puanlamada özel bütünsel dereceli puanlama anahtarı derecelendirmeleri geniş aralıkta ve etkin bir şekilde kullanılmıştır.

4.4.2. Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Puanlayıcı Yüzeyine İlişkin Bulgular

Desen D'ye ilişkin uygunluk ölçütleri ve uyum yüzdeleri Tablo 27'de sunulmuştur.

Tablo 27

Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu

	Model		Uygunluk ölçütleri		Uyum yüzdeleri	
Puanlayıcı	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)	Gözlenen uyum	Beklenen uyum
A	-0,48	0,03	1,02	1,02	95,9	39,8
B	-0,47	0,03	1,06	1,05	95,9	39,8
C	-0,42	0,04	0,90	1,01	0,00	0,00
D	-0,23	0,04	0,93	1,02	0,00	0,00
Ortalama	-0,40	0,03	0,98	1,02		
SS (evren)	0,10	0,00	0,06	0,02		
SS (örneklem)	0,12	0,00	0,07	0,02		
Model (evren)	RMSE= 0,03	SS: 0,10	Ayırma oranı=2,83		Ayırma indeksi= 4,10	
Güvenirlik (puanlayıcılar arası olmayan) = 0,89						
Model (örneklem)	RMSE= 0,03	SS: 0,11	Ayırma oranı=3,32		Ayırma indeksi= 4,75	
Güvenirlik (puanlayıcılar arası olmayan) = 0,92						
Model, sabit etkili ki-kare: 34,3	s.d: 3	p= 0,00				
Model, rastgele ki-kare: 2,8	s.d: 2	p= 0,25				
Puanlayılar arası uyumun gözlenen değeri= %95,9						
Puanlayıcılar arası uyumun beklenen değeri= %39,8						
Puanlayıcılar arası güvenirlik (Rasch-Cohen'in kappa istatistiği)= 0,93						

Grup düzeyindeki istatistikler olan ayırma oranı=3,32, ayırma indeksi=4,75 ve ayırma indeksi güvenirliği=0,92 olarak bulunmuştur. Puanlayıcı yüzeyinde ayırma indeksi 2,00'den ve ayırma indeksi güvenirliği 0,50'den büyük olduğundan puanlayıcıların katılık-cömertlik davranışlarının farklılaştığı ve puanlamaya puanlayıcıdan kaynaklanan hataların karıştığı düşünülmüştür. Ayırma oranı da puanlayıcı yüzeyinde ölçme hatası ile açıklanamayacak bir farklılığın oluştuğuna işaret etmiştir. Puanlayıcıların katılık ve cömertliğini gösteren RMSE değeri 0,03 bulunmuş ve bu değerin düşük olması, puanlayıcıların katılık davranışının daha kesin bir şekilde tahmin edildiğini göstermiştir. Sabit etkili ki-kare değeri $\chi^2(3)=34,3$, $p=0,00<0,05$ bulunmuş ve en az iki puanlayıcının katılık-cömertlik davranışları arasında istatistiksel olarak anlamlı fark olduğu görülmüştür.

Tamamlanmamış bağlantılı olmayan puanlama deseni için puanlayıcılar arası uyumun gözlenen değeri %95,9, puanlayıcılar arası uyumun beklenen değeri %39,8 bulunmuş ve Rasch-Cohen'in kappa istatistiği değeri 0,93 olarak hesaplanmıştır. Bu değerin +0,50'den büyük olmasıyla öğrencilerin açık uçlu maddeleri cevaplandırma performanslarını değerlendirirken farklı puanlayıcıların birbirleriyle tutarlı puanlamalar yaptıkları, puanlayıcılar arası uyumun ve güvenirliğin yüksek olduğu görülmüştür.

Logit değerlerine bakıldığında tüm puanlayıcıların negatif logit değerleri aldığı ve katı puanlama yaptığı görülmüştür. Puanlayıcı D en yüksek logit değerine (-0,23) sahip olduğundan en cömert, logit değeri sırasıyla -0,48 ,-0,47 olan puanlayıcı A ve puanlayıcı B'nin birbirlerine yakın puanlama yaptığı ve en katı puanlayıcılar olduğu belirlenmiştir. Puanlayıcıların iç uyum indeksleri 0,90 ile 1,06 arasında, dış uyum indeksleri 1,01 ile 1,05 arasında değişmektedir ve kabul edilebilir düzeydedir. Bu değerler puanlayıcıların model ile veri uyumunu sağladığını ve puanlayıcıların birbirlerinden farklı puanlama yapmalarına rağmen dereceli puanlama anahtarı ölçütlerinin tamamını göz önünde bulundurarak tutarlı puanlama yaptıklarını göstermiştir.

4.4.3. Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Öğrenci Yüzeyine İlişkin Bulgular

Öğrenci verilerinin tamamının gösterimi mümkün olmadığından logit değeri en yüksek ve en düşük 8 öğrenciye ilişkin değerler Tablo 28'de sunulmuştur.

Tablo 28

Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Öğrenci Yüzeyine İlişkin Ölçüm Raporu

	Ortalamalar		Model		Uygunluk ölçütleri	
Öğrenci sıra no	Gözlenen ortalama	Düzeltilmiş ortalama	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)
22	3,14	3,39	1,27	0,21	1,42	0,87
76	2,91	3,19	1,06	0,19	0,99	1,05
2	2,82	3,10	0,99	0,19	1,28	1,32
43	2,91	3,10	0,99	0,19	1,13	0,91
1	2,82	3,00	0,92	0,19	0,36	0,37
6	2,68	2,95	0,88	0,19	0,57	0,55
122	2,68	2,90	0,85	0,18	1,08	0,93
32	2,59	2,95	0,82	0,18	1,10	1,03
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
119	0,36	0,26	-1,06	0,29	0,78	1,59
61	0,27	0,25	-1,09	0,34	0,41	0,35
93	0,27	0,25	-1,09	0,34	1,03	1,31
111	0,27	0,25	-1,09	0,34	0,98	0,85
84	0,18	0,17	-1,39	0,44	0,63	1,24
85	0,18	0,17	-1,39	0,44	0,54	0,45
104	0,18	0,17	-1,39	0,44	0,57	0,44
87	0,00	0,01	-3,87	1,80	0,02	0,01
Ortalama	1,48	1,41	-0,03	0,21	0,98	1,02
SS (evren)	0,74	0,82	0,65	0,15	0,41	0,64
SS (örneklem)	0,74	0,83	0,66	0,15	0,41	0,64
Model (evren) Güvenirlik: 0,84	RMSE: 0,26	SS: 0,60	Ayırma oranı= 2,31		Ayırma indeksi= 3,41	
Model (örneklem) Güvenirlik: 0,84	RMSE: 0,26	SS: 0,60	Ayırma oranı= 2,32		Ayırma indeksi= 3,43	
Model, sabit etkili ki-kare: 726,7			s.d: 121		p= 0,00	
Model, rastgele ki-kare:79,2			s.d: 120		p= 1,00	

RMSE: hata kareleri ortalamalarının karekökü *SS: standart sapma* *s.d: serbestlik derecesi*

Tablo 28 incelendiğinde ayırma oranı=2,32, ayırma indeksi=3,43 ve ayırma indeksi güvenirliği=0,84 olduğu görülmüştür. Bu değerler açık uçlu maddeleri cevaplandırma becerisi birbirlerinden farklı olan öğrencilerin güvenilir bir şekilde ayırt edilebildiklerini göstermiştir. Ayırma oranı öğrenci performansları arasında ölçme hatası ile açıklanamayacak kadar farklılık oluştuğunu göstermiştir. Öğrencilerin başarı ölçümlerinin RMSE değeri 0,26 olarak bulunmuş ve yüksek beceriye sahip öğrencilerin daha kesin bir şekilde tahmin edildiği anlaşılmıştır. Desen D için *([4.(3,43)+1)]/3=4,91)* öğrenci performanslarının beş seviyeye ayrıldığı söylenebilir. Sabit etkili ki-kare değerinin $\chi^2(121)=$ 726,7, $p=0,00<0,05$ olması en az iki öğrencinin yetenek düzeyleri arasında anlamlı fark

bulunduğunu ve öğrenci grubunun başarı açısından heterojen yapıda olduğunu göstermiştir. Logit değeri (1,27) olan 22 numaralı öğrenci en yüksek yetenek düzeyine, logit değeri (-3,87) olan 87 numaralı öğrenci en düşük yetenek düzeyine sahiptir. Bireysel düzeydeki istatistiksel göstergeler olan iç uyum indeksleri 0,02 ile 2,13 arasında, dış uyum indeksleri 0,01 ile 5,02 arasında değişmiştir. Bazı öğrencilerin dış uyum indeksleri 2,00'den büyük olsa da iç uyum indekslerinin 2,00'den büyük olmaması model veri uyumunun bozulmadığını göstermiştir (Linacre, 2002). Öğrencilerin % 84,43'ünün iç ve dış uyum indeksleri açısından kabul edilebilir aralıkta (0,50–1,50) olduğu görülmüştür. Analiz sonucunda iç uyum aralığı dışında kalan on dokuz öğrenci için her bir açık uçlu maddeyi cevaplandırma performanslarının beklenen düzeyde olmadığı; dış uyum aralığı dışında da kalan on dokuz öğrencinin matematik başarı testi uygulamasında gösterdikleri genel performansın beklenen düzeyde olmadığı anlaşılmıştır. İç uyum indeks değeri 1,02 (beklenen değer) ve dış uyum indeks değeri 1,00 olan 77 numaralı öğrencinin her bir açık uçlu maddeyi cevaplandırma performansı ile matematik başarı testindeki genel performans uyumu en yüksek öğrenci olduğu gözlemlenmiştir.

4.4.4. Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Açık Uçlu Madde Yüzeyine İlişkin Bulgular

Desen D'ye ilişkin matematik başarı testindeki açık uçlu maddelerin madde güçlük ve uyum indeksleri Tablo 29'da sunulmuştur.

Tablo 29

Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Açık Uçlu Madde Yüzeyine İlişkin Ölçüm Raporu

	Ortalamalar		Model		Uygunluk ölçütleri	
Madde no	Gözlenen ortalama	Düzeltilmiş ortalama	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)
1	1,95	1,89	0,35	0,05	1,12	1,08
2	0,45	0,32	-0,87	0,08	0,92	0,77
3	2,04	2,02	0,41	0,05	0,73	0,69
4	1,82	1,72	0,27	0,05	0,67	0,68
5	1,63	1,48	0,15	0,05	1,14	1,09
6	2,26	2,32	0,56	0,05	1,25	1,22
7	1,98	1,94	0,37	0,05	1,12	1,07
8	1,81	1,71	0,27	0,05	0,84	0,79
9	1,19	0,97	-0,14	0,05	0,81	0,76
10	0,59	0,43	-0,68	0,07	1,39	1,71
11	0,57	0,42	-0,70	0,07	0,99	1,39
Ortalama	1,48	1,38	0,00	0,06	1,00	1,02
SS (evren)	0,63	0,69	0,49	0,01	0,21	0,31
SS (örneklem)	0,66	0,72	0,51	0,01	0,22	0,33
Model (evren): Güvenirlik: 0,99	RMSE: 0,06	SS: 0,49	Ayırma oranı= 8,34		Ayırma indeksi=11,45	
Model (örneklem): Güvenirlik: 0,99	RMSE: 0,06	SS: 0,51	Ayırma oranı= 8,75		Ayırma indeksi=12,00	
Model, sabit etkili ki-kare: 608,8			s.d: 10		p= 0,00	
Model, rastgele ki-kare: 9,8			s.d: 9		p= 0,36	

RMSE: hata kareleri ortalamalarının karekökü *SS: standart sapma* *s.d: serbestlik derecesi*

Açık uçlu madde yüzeyine ilişkin ayırma oranı=8,75, ayırma indeksi=12,00, ayırma indeksi güvenirliği=0,99 şeklindedir. Bu değerler farklı güçlük düzeylerine sahip maddelerin yüksek güvenirlikle birbirlerinden ayırt edildiği göstermiştir. Ayırma oranı da madde güçlük düzeylerinin ölçme hatası ile açıklanamayacak kadar birbirlerinden farklı olduğunu göstermiştir. Sabit etkili ki-kare testi ile ($\chi^2(10)=608,8$, $p=0,00<0,05$) madde güçlük düzeyleri arasındaki farklılığın istatistiksel olarak anlamlı olduğu anlaşılmıştır.

Açık uçlu maddelerin güçlük düzeylerinin logit değerleri 0,56 ile -0,87 arasında değişmektedir. Tablo 29'da Madde 6 (logit=0,56)'nın madde güçlüğü en yüksek (kolay cevaplanabilen) ve Madde 2 (logit= -0,87)'nin madde güçlüğü en düşük (zor cevaplanabilen) olduğu görülmektedir. Logit değerlerine bakıldığında madde güçlüğü düşük maddeler, Madde 2, Madde 11, Madde 10, Madde 9; madde güçlüğü yüksek maddeler, Madde 6, Madde 3, Madde 7, Madde 1, Madde 4 ve Madde 8, Madde 5 şeklinde sıralanmıştır. Maddelerin güçlük/kolaylık durumunu gösteren RMSE değeri 0,06 bulunmuş ve madde güçlüğünün daha kesin bir şekilde tahmin edildiği anlaşılmıştır. Standart hata değeri sırasıyla

0,08; 0,07; 0,07 olan Madde 2, Madde 11, Madde 10 dışındaki madde güçlükleri daha kesin tahmin edilmiştir. İç uyum indeksleri 0,67 ile 1,39 arasında, dış uyum indeksleri 0,68 ile 1,71 arasında değişmektedir. Tablo 29'da yer alan iç ve dış uyum değerlerinin genel olarak 0,50 ile 1,50 aralığında ve kabul edilebilir düzeyde olduğu görülmüştür. Sadece Madde 10'da iç uyum değeri (1,39) kabul edilebilir düzeyde olsa da dış uyum değeri (1,71) bulunmuştur. Dış uyum indeksi 2,00'den büyük olmadığından bu madde de kabul edilebilir düzeyde olarak belirlenmiştir. İç uyum indeks değerleri ilgili kazanımı ölçmede her bir maddenin iyi işlediğini ve dış uyum indeks değerleri başarı testinde yer alan maddelerin öğrenci matematik başarısını ölçmede iyi olduğunu göstermiştir.

4.4.5. Tamamlanmamış-Bağlantılı Olmayan Puanlama Deseni (Desen D) Özel Bütünsel Dereceli Puanlama Anahtarı Ölçütlerine İlişkin Bulgular

Desen D için dereceli puanlama anahtarı kategorilerine ilişkin değerler Tablo 30'da sunulmuştur.

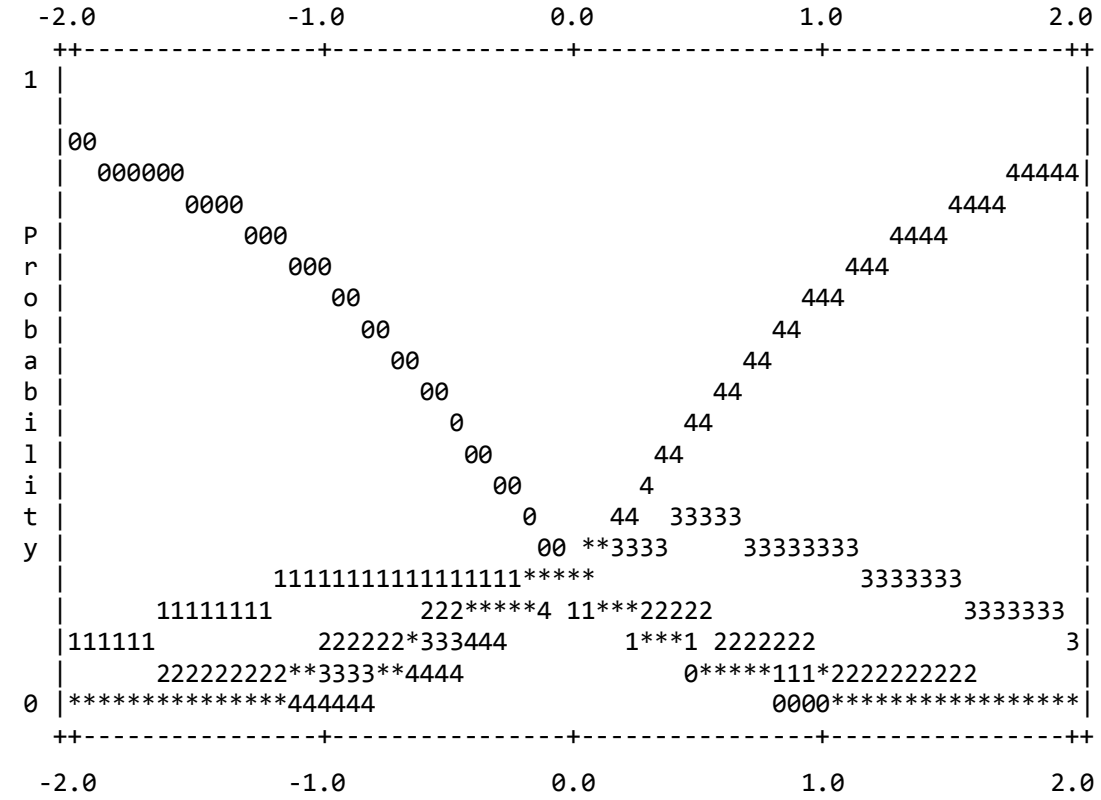
Tablo 30

Özel Bütünsel Dereceli Puanlama Anahtarı Puanlama Kategorilerine Ait Ölçüm Raporu

Puanlama Kategorileri	Frekans	Yüzde (%)	Birikimli yüzde (%)	Ortalama logit değeri	Beklenen logit değeri	Outfit (dış uyum)
0	1115	42	42	-0,95	-0,97	1,00
1	427	16	57	-0,58	-0,51	1,00
2	320	12	69	-0,20	-0,18	0,90
3	381	14	84	0,12	0,07	1,00
4	441	16	100	0,28	0,28	1,10

Frekans ve yüzde değerlerine bakıldığında; puanlayıcıların puanlama yaparken en çok 0 numaralı kategoriyi tercih ettikleri ve 1, 2, 3 ve 4 numaralı kategorileri birbirine yakın sayıda kullandıkları görülmüştür. 0 kategorisinin daha fazla kullanılmasının sebebi öğrenci yüzeyi ölçüm raporunda öğrencilerin matematik başarı düzeylerinin yüksek güvenirlikle ayırt edildiğinden merkezi eğilim etkisininden kaynaklanmamaktadır. Logit değerleri incelendiğinde 3 ve 4 numaralı kategorilerin tam ve oldukça doğru tanımlarına uygun bir şekilde pozitif değer aldığı; 0, 1, 2 numaralı kategorilerin boş, tamamen yanlış ve kısmen yanlış tanımlarına uygun olarak negatif değerler aldığı anlaşılmıştır. Ortalama ve beklenen logit değerlerinin birbirine yakın olduğu ve verilerin Rasch modeli beklentilerine uyduğu görülmüştür. Dış uyum indeksleri 0,90 ile 1,10 arasında değişmektedir ve kabul edilebilir düzeydedir. Matematik başarı testindeki açık uçlu maddelerin değerlendirilmesinden elde

edilen puanlarla özel bütünsel dereceli puanlama anahtarı kategorilerinin uyumlu ve dereceli puanlama anahtarının güvenilir bir ölçme aracı olduğu görülmüştür. 1,00 dış uyum indeks değerine sahip olan 0, 1 ve 3 numaralı kategoriler, matematik başarı testindeki öğrenci performanslarının değerlendirilme puanları ile mükemmel uyum içerisindedir.



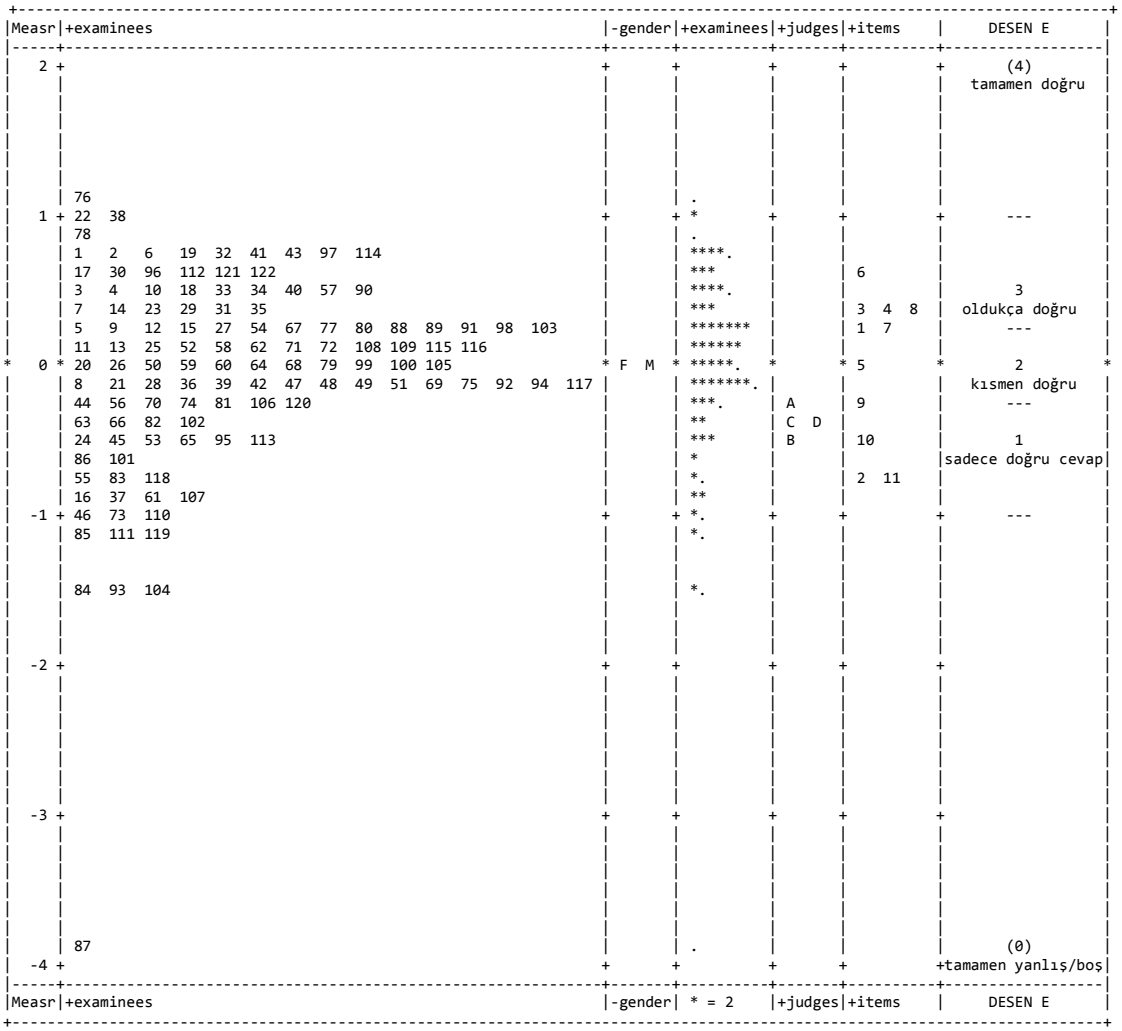
Şekil 8. Tamamlanmamış-bağlantılı olmayan puanlama deseni (Desen D) dereceli puanlama anahtarı ölçütlerine ait olasılık eğrileri

Olasılık eğrileri incelendiğinde düşük beceri düzeyindeki öğrencilerin 0, 1, 2 numaralı alt kategorilerde bulunma olasılığı yüksek; 3 ve 4 numaralı üst kategorilerde bulunma olasılığı düşüktür. Benzer şekilde yüksek beceri düzeyindeki öğrencilerin üst kategorilerde bulunma olasılığı yüksek; alt kategorilerde bulunma olasılığı ise düşüktür. Bunun sonucunda özel bütünsel dereceli puanlama anahtarı kategorilerinin açık uçlu maddeleri cevaplandırmadaki öğrenci başarılarını ayırt etmede iyi olduğu görülmüştür.

4.5. Tamamlanmamış-Bağlantılı Sarmal Puanlama Desenine (Desen E) İlişkin Bulgular

4.5.1. Tamamlanmamış - Bağlantılı Sarmal Puanlama Deseni (Desen E) Veri Kalibrasyon Haritasına İlişkin Bulgular

Puanlayıcı A ve C'nin her bir öğrencinin çift numaralı maddelerdeki cevabını puanladığı; Puanlayıcı B ve D 'nin her bir öğrencinin tek numaralı maddelerdeki cevabını puanladığı *[(2x5x122)+(2x6x122)=2684 gözlem]* Desen E'ye ilişkin kalibrasyon haritası Şekil 9'da sunulmuştur.



Şekil 9. Tamamlanmamış-bağlantılı sarmal puanlama deseni (Desen E) veri kalibrasyon haritası

Kalibrasyon haritası incelendiğinde matematik başarı testindeki açık uçlu maddeleri cevaplandırmada 76, 22, 38 nolu öğrencilerin yüksek beceriye, 87 nolu öğrencinin ise düşük

beceriye sahip oldukları görülmüştür. Logit ölçekte pozitif logit değerine sahip öğrenci sayısının negatif logit değere sahip öğrenci sayısından fazla olduğu ve birbirine yakın dağıldığı, öğrenci grubunun başarı açısından heterojen olduğu görülmüştür. Puanlayıcı yüzeyine bakıldığında tüm puanlayıcıların katı puanlama yapmasıyla birlikte puanlayıcı A'nın en cömert, puanlayıcı B'nin en katı puanlayıcı olduğu, puanlayıcı C ile D'nin birbirine yakın puanladığı anlaşılmıştır. Açık uçlu madde yüzeyine bakıldığında madde güçlüğü en yüksek Madde 6 ve madde güçlüğü en düşük Madde 2 ve Madde 11 olarak belirlenmiştir. Öğrenciler matematik başarı testindeki Madde 1, Madde 3, Madde 4, Madde 5, Madde 6, Madde 7 ve Madde 8'i kolay cevaplandırmış; Madde 2, Madde 9, Madde 10, Madde 11'i cevaplandırmada ise zorlanmışlardır. Öğrenci performanslarını puanlamada özel bütünsel dereceli puanlama anahtarı derecelendirmelerinin puanlayıcılar tarafından etkin ve geniş aralıkta kullanıldığı görülmüştür.

4.5.2. Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Puanlayıcı Yüzeyine İlişkin Bulgular

Desen E için puanlayıcı yüzeyine ilişkin uygunluk ölçütleri ve uyum yüzdeleri Tablo 31'de sunulmuştur.

Tablo 31

Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu

	Model		Uygunluk ölçütleri		Uyum yüzdeleri	
Puanlayıcı	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)	Gözlenen uyum	Beklenen uyum
A	-0,30	0,03	1,02	1,01	0,0	0,0
B	-0,51	0,04	1,07	1,14	65,6	39,4
C	-0,40	0,03	0,92	1,02	0,0	0,0
D	-0,39	0,04	0,91	0,93	65,6	39,4
Ortalama	-0,40	0,03	0,98	1,03		
SS (evren)	0,07	0,00	0,07	0,07		
SS (örneklem)	0,09	0,00	0,08	0,09		
Model (evren):	RMSE= 0,03	SS: 0,07	Ayırma oranı= 1,96		Ayırma indeksi= 2,94	
Güvenirlik (puanlayıcılar arası olmayan) = 0,79						
Model (örneklem):	RMSE= 0,03	SS: 0,08	Ayırma oranı= 2,33		Ayırma indeksi= 3,44	
Güvenirlik (puanlayıcılar arası olmayan) = 0,84						
Model, sabit etkili ki-kare: 19,1		s.d: 3	p= 0,00			
Model, rastgele ki-kare: 2,6		s.d: 2	p= 0,27			
Puanlayılar arası uyumun gözlenen değeri= % 65,6						
Puanlayıcılar arası uyumun beklenen değeri= % 39,4						
Puanlayıcılar arası güvenirlik (Rasch-Cohen'in kappa istatistiği)= 0,43						

Ayırma oranı=2,33, ayırma indeksi=3,44 ve ayırma indeksi güvenirliği=0,84 değerlerine bakıldığında puanlayıcıların katılık-cömertlik davranışlarının farklılaştığı ve puanlamaya puanlayıcıdan kaynaklanan hataların karıştığı düşünülmüştür. Ayırma oranı da puanlayıcı yüzeyinde ölçme hatası ile açıklanamayacak bir farklılığın oluştuğunu göstermiştir. Puanlayıcıların katılık ve cömertliğini gösteren RMSE değerinin 0,03 bulunmasıyla puanlayıcıların katılık davranışının daha kesin bir şekilde tahmin edildiği anlaşılmıştır. Sabit etkili ki-kare değeri $\chi^2(3)=19,1$, $p=0,00<0,05$ bulunmuş ve en az iki puanlayıcının katılık-cömertlik davranışları arasında istatistiksel olarak anlamlı fark olduğu görülmüştür.

Tamamlanmamış bağlantılı sarmal puanlama deseni için puanlayıcılar arası uyumun gözlenen değeri %65,6, puanlayıcılar arası uyumun beklenen değeri %39,4 bulunmuş ve Rasch-Cohen'in kappa istatistiği değeri 0,43 olarak hesaplanmıştır. Bu değerin +0,40'tan büyük olmasıyla öğrencilerin açık uçlu maddeleri cevaplandırma performanslarını değerlendirirken farklı puanlayıcıların birbirleriyle tutarlı puanlamalar yaptıkları, puanlayıcılar arası uyumun ve güvenirliğin modelden beklenen performanstan biraz daha fazla olduğu görülmüştür.

Logit değerlerine bakıldığında tüm puanlayıcıların negatif logit değerleri aldığı ve katı puanlama yaptığı görülmüştür. Puanlayıcı A en yüksek logit değerine (-0,30) sahip olduğundan en cömert, logit değeri sırasıyla -0,40, -0,39 olan puanlayıcı C ve puanlayıcı D'nin birbirlerine yakın puanlama yaptığı puanlayıcı B'nin (-0,51) en katı puanlamayı yaptığı belirlenmiştir. Puanlayıcıların iç uyum indeksleri 0,91 ile 1,07 arasında, dış uyum indeksleri 0,93 ile 1,14 arasında değişmektedir ve kabul edilebilir düzeydedir. Bu değerler puanlayıcıların model ile veri uyumunu sağladığını ve puanlayıcıların birbirlerinden farklı puanlama yapmalarına rağmen dereceli puanlama anahtarı ölçütlerinin tamamını göz önünde bulundurarak tutarlı puanlama yaptıklarını göstermiştir.

4.5.3. Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Öğrenci Yüzeyine İlişkin Bulgular

Öğrenci verilerinin tamamının gösterimi mümkün olmadığından logit değeri en yüksek ve en düşük 8 öğrenciye ilişkin değerler Tablo 32'de sunulmuştur.

Tablo 32

Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Öğrenci Yüzeyine İlişkin Ölçüm Raporu

	Ortalamalar		Model		Uygunluk ölçütleri	
Öğrenci sıra no	Gözlenen ortalama	Düzeltilmiş ortalama	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)
76	3,14	3,29	1,18	0,21	0,76	0,86
22	2,91	3,07	0,98	0,19	1,42	1,16
38	2,86	3,02	0,94	0,19	1,14	0,97
78	2,73	2,87	0,84	0,18	0,77	0,67
1	2,68	2,81	0,80	0,18	0,39	0,40
19	2,68	2,81	0,80	0,18	0,58	0,59
41	2,68	2,81	0,80	0,18	1,07	0,97
43	2,68	2,81	0,80	0,18	1,04	0,90
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
110	0,36	0,30	-0,98	0,30	0,83	0,70
85	0,32	0,26	-1,08	0,32	0,90	1,19
111	0,32	0,26	-1,08	0,32	1,33	1,17
119	0,32	0,26	-1,08	0,32	0,71	1,67
84	0,18	0,15	-1,50	0,44	0,64	1,12
93	0,18	0,15	-1,50	0,44	1,13	1,60
104	0,18	0,15	-1,50	0,44	0,66	0,89
87	0,00	0,01	-3,88	1,81	0,02	0,01
Ortalama	1,49	1,41	-0,03	0,21	0,98	1,02
SS (evren)	0,73	0,80	0,65	0,16	0,37	0,57
SS (örneklem)	0,74	0,81	0,66	0,16	0,37	0,57
Model(evren) Güvenirlik: 0,84	RMSE: 0,26	SS: 0,60	Ayırma oranı= 2,32		Ayırma indeksi= 3,42	
Model(örneklem) Güvenirlik: 0,84	RMSE: 0,26	SS: 0,60	Ayırma oranı= 2,33		Ayırma indeksi= 3,44	
Model, sabit etkili ki-kare: 706,3			s.d: 121		p= 0,00	
Model, rastgele ki-kare: 78,6			s.d: 120		p= 1,00	

RMSE: hata kareleri ortalamalarının karekökü SS: standart sapma s.d: serbestlik derecesi

Tablo 32 incelendiğinde ayırma oranı=2,33, ayırma indeksi=3,44 ve ayırma indeksi güvenirliği=0,84 olduğu görülmüştür. Bu değerler açık uçlu maddeleri cevaplandırma becerisi birbirlerinden farklı olan öğrencilerin güvenilir bir şekilde ayırt edilebildiklerini göstermiştir. Öğrenci performansları arasında ölçme hatası ile açıklanamayacak kadar farklılık oluştuğu görülmüştür. Öğrencilerin başarı ölçümlerinin RMSE değeri 0,26 olarak bulunmuş ve yüksek beceriye sahip öğrencilerin daha kesin bir şekilde tahmin edildiği anlaşılmıştır. Desen E için *([4.(3,44)+1)]/3=4,92)* öğrenci performanslarının beş seviyeye ayrıldığı söylenebilir. Sabit etkili ki-kare değerinin $\chi^2(121)=706,3$, $p=0,00<0,05$ olması en az iki öğrencinin yetenek düzeyleri arasında anlamlı fark bulunduğunu ve öğrenci grubunun

başarı açısından heterojen yapıda olduğunu göstermiştir. Logit değeri (1,18) olan 76 numaralı öğrenci en yüksek, logit değeri (-3,88) olan 87 numaralı öğrenci en düşük yetenek düzeyine sahiptir. Bireysel düzeydeki istatistiksel göstergeler olan iç uyum indeksleri 0,02 ile 1,94 arasında, dış uyum indeksleri 0,01 ile 4,72 arasında değişmiştir. Bazı öğrencilerin dış uyum indeksleri 2,00'den büyük olsa da iç uyum indekslerinin 2,00'den büyük olmaması model veri uyumunun bozulmadığını göstermiştir. Öğrencilerin % 90,98'inin iç ve dış uyum indeksleri açısından kabul edilebilir aralıkta (0,50-1,50) olduğu görülmüştür. Analiz sonucunda iç uyum aralığı dışında kalan on bir öğrenci için her bir açık uçlu maddeyi cevaplandırma performanslarının beklenen düzeyde olmadığı; dış uyum aralığı dışında da kalan on bir öğrencinin matematik başarı testi uygulamasında gösterdikleri genel performansın beklenen düzeyde olmadığı anlaşılmıştır. İç ya da dış uyum indeksi 1,00 her ve bir açık uçlu maddeyi cevaplandırma performansı ile matematik başarı testindeki genel performans uyumu mükemmel olan öğrenci bulunmamıştır.

4.5.4. Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Açık Uçlu Madde Yüzeyine İlişkin Bulgular

Desen E'ye ilişkin matematik başarı testindeki açık uçlu maddelerin madde güçlük ve uyum istatistikleri Tablo 33'te sunulmuştur.

Tablo 33

Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Açık Uçlu Madde Yüzeyine İlişkin Ölçüm Raporu

	Ortalamalar		Model		Uygunluk ölçütleri	
Madde no	Gözlenen ortalama	Düzeltilmiş ortalama	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)
1	1,80	1,55	0,20	0,05	1,06	1,09
2	0,49	0,38	-0,78	0,08	0,96	0,80
3	2,00	1,81	0,32	0,05	0,76	0,72
4	1,96	1,97	0,40	0,05	0,73	0,69
5	1,51	1,21	0,01	0,05	1,12	1,13
6	2,36	2,50	0,66	0,05	1,13	1,12
7	1,89	1,67	0,26	0,05	1,17	1,13
8	1,91	1,90	0,37	0,05	0,86	0,80
9	1,21	0,90	-0,19	0,05	0,73	0,75
10	0,67	0,52	-0,55	0,07	1,45	1,77
11	0,64	0,43	-0,69	0,07	0,92	1,28
Ortalama	1,49	1,35	0,00	0,06	0,99	1,02
SS (evren)	0,61	0,68	0,46	0,01	0,21	0,31
SS (örneklem)	0,64	0,71	0,49	0,01	0,22	0,32
Model (evren): Güvenirlik: 0,98	RMSE: 0,06	SS: 0,46	Ayırma oranı= 8,00		Ayırma indeksi= 11,00	
Model (örneklem): Güvenirlik: 0,99	RMSE: 0,06	SS: 0,48	Ayırma oranı= 8,40		Ayırma indeksi= 11,53	
Model, sabit etkili ki-kare: 599,7			s.d: 10		p= 0,00	
Model, rastgele ki-kare: 9,8			s.d: 9		p= 0,36	

RMSE: hata kareleri ortalamalarının karekökü *SS: standart sapma* *s.d: serbestlik derecesi*

Ayırma oranı=8,40, ayırma indeksi=11,53, ayırma indeksi güvenirliği=0,99 olan açık uçlu madde yüzeyine ilişkin bu değerler farklı güçlük düzeylerine sahip maddelerin yüksek güvenirlikle birbirlerinden ayırt edildiği göstermiştir. Ayırma oranına bakıldığında madde güçlük düzeylerinin ölçme hatası ile açıklanamayacak kadar birbirlerinden farklı olduğu görülmüştür. Sabit etkili ki-kare testi ile ($\chi^2(10)=599,7$, $p=0,00<0,05$) madde güçlük düzeyleri arasındaki farklılığın istatistiksel olarak anlamlı olduğu anlaşılmıştır.

Açık uçlu maddelerin güçlük düzeylerinin logit değerleri 0,66 ile -0,78 arasında değişmektedir. Tablo 33 incelendiğinde Madde 6 (logit=0,66)'nın en yüksek madde güçlüğüne ve Madde 2 (logit= -0,78)'nin en düşük madde güçlüğüne sahip olduğu görülmektedir. Logit değerlerine bakıldığında madde güçlüğü düşük maddeler, Madde 2, Madde 11, Madde 10, Madde 9; madde güçlüğü yüksek maddeler, Madde 6, Madde 4, Madde 8, Madde 3, Madde 7 ve Madde 1, Madde 5 şeklinde sıralanmıştır. Maddelerin güçlük/kolaylık durumunu gösteren RMSE değeri 0,06 bulunmuş ve madde güçlüğünün

daha kesin bir şekilde tahmin edildiği anlaşılmıştır. Standart hata değeri sırasıyla 0,08; 0,07; 0,07 olan Madde 2, Madde 10, Madde 11 dışındaki madde güçlükleri daha kesin tahmin edilmiştir. İç uyum indeksleri 0,73 ile 1,45 arasında, dış uyum indeksleri 0,69 ile 1,77 arasında değişmektedir. Tablo 33'te yer alan iç ve dış uyum değerleri genel olarak 0,50 ile 1,50 aralığında ve kabul edilebilir düzeydedir. Sadece Madde 10'nun iç uyum değeri (1,45) kabul edilebilir düzeyde olsa da dış uyum değeri (1,77) bulunmuştur. Dış uyum indeksi 2,00'den büyük olmadığından bu madde de kabul edilebilir düzeyde olarak belirlenmiştir. İç uyum indeks değerleri ilgili kazanımı ölçmede her bir maddenin iyi işlediğini ve dış uyum indeks değerleri başarı testinde yer alan maddelerin öğrenci matematik başarısını ölçmede iyi olduğunu göstermiştir.

4.5.5. Tamamlanmamış-Bağlantılı Sarmal Puanlama Deseni (Desen E) Özel Bütünsel Dereceli Puanlama Anahtarı Ölçütlerine İlişkin Bulgular

Desen E için dereceli puanlama anahtarı kategorilerine ilişkin değerler Tablo 34'te sunulmuştur.

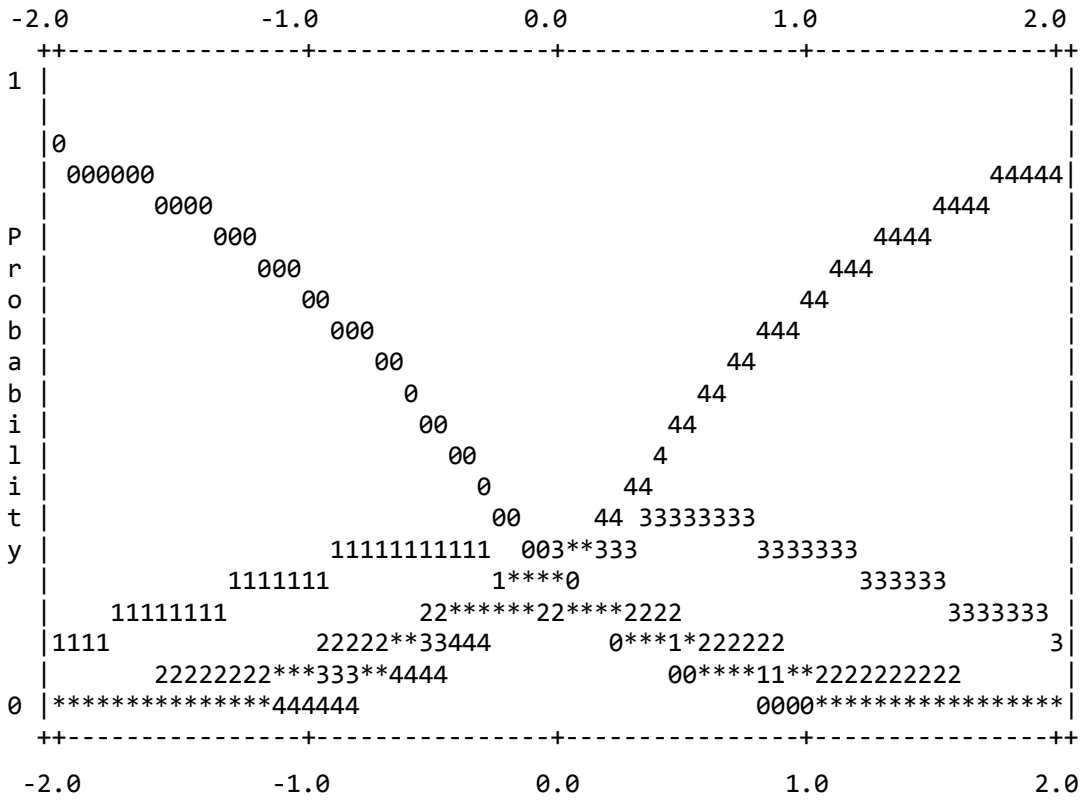
Tablo 34

Özel Bütünsel Dereceli Puanlama Anahtarı Puanlama Kategorilerine Ait Ölçüm Raporu

Puanlama Kategorileri	Frekans	Yüzde (%)	Birikimli yüzde (%)	Ortalama logit değeri	Beklenen logit değeri	Outfit (dış uyum)
0	1079	40	40	-0,92	-0,94	1,00
1	470	18	58	-0,57	-0,49	1,10
2	304	11	69	-0,19	-0,18	0,90
3	393	15	84	0,13	0,07	0,90
4	438	16	100	0,26	0,28	1,10

Frekans ve yüzde değerleri; puanlayıcıların puanlama yaparken en çok 0 numaralı kategoriyi tercih ettiklerini ve 1, 2, 3 ve 4 numaralı kategorileri birbirine yakın sayıda kullandıklarını göstermiştir. Öğrenci yüzeyi ölçüm raporunda öğrencilerin matematik başarı düzeylerinin yüksek güvenirlikle ayırt edildiğinden uç değer olan 0 kategorisinin daha fazla kullanılması merkezi eğilim etkisinden kaynaklanmamaktadır. Logit değerleri incelendiğinde 3 ve 4 numaralı kategorilerin tam ve oldukça doğru tanımlarına uygun bir şekilde pozitif değer aldığı; 0, 1, 2 numaralı kategorilerin boş, tamamen yanlış ve kısmen yanlış tanımlarına uygun olarak negatif değerler aldığı anlaşılmıştır. Ortalama ve beklenen logit değerlerinin birbirine yakın olduğu ve verilerin Rasch modeli beklentilerine uyduğu görülmüştür. Dış uyum indeksleri 0,90 ile 1,10 arasında değişmektedir ve kabul edilebilir düzeydedir.

Matematik başarı testindeki açık uçlu maddelerin değerlendirilmesinden elde edilen puanlarla özel bütünsel dereceli puanlama anahtarı kategorilerinin uyumlu ve dereceli puanlama anahtarının güvenilir bir ölçme aracı olduğu görülmüştür. 1,00 dış uyum indeks değerine sahip olan 0 numaralı kategori, matematik başarı testindeki öğrenci performanslarının değerlendirilme puanları ile mükemmel uyuma sahiptir.



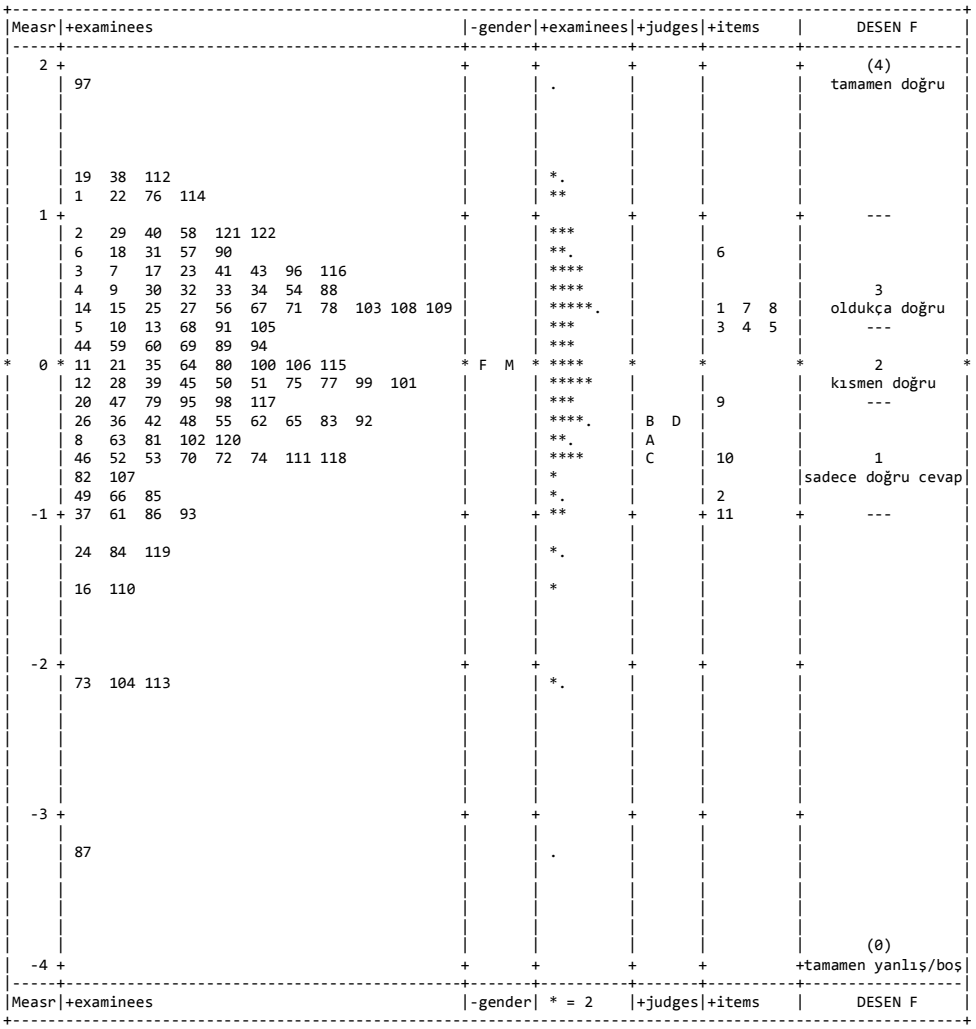
Şekil 10. Tamamlanmamış-bağlantılı sarmal puanlama desenine (Desen E) göre dereceli puanlama anahtarı ölçütlerine ait olasılık eğrileri

Beceri düzeyi düşük öğrencilerin 0, 1, 2 numaralı alt kategorilerde bulunma olasılığının yüksek; 3 ve 4 numaralı üst kategorilerde bulunma olasılığının düşük olduğu olasılık eğrisine bakıldığında görülmektedir. Benzer şekilde beceri düzeyi yüksek öğrencilerin üst kategorilerde bulunma olasılığı yüksek; alt kategorilerde bulunma olasılığı ise düşüktür. Bu durum özel bütünsel dereceli puanlama anahtarı kategorilerinin açık uçlu maddeleri cevaplandırmadaki öğrenci başarılarını ayırt etmede iyi olduğu göstermektedir.

4.6. Sarmal Puanlama Desenine (Desen F) İlişkin Bulgular

4.6.1. Sarmal Puanlama Deseni (Desen F) Veri Kalibrasyon Haritasına İlişkin Bulgular

Tharbai ve Young (2001) tarafından İngilizce branşındaki maddelerin kullanılan her bir öğrencinin her bir açık uçlu maddedeki cevabının farklı bir puanlayıcı tarafından puanlandığı ve bu puanlama tasarımının her dört öğrencide tekrarlanarak sarmal bir şekilde gerçekleştirildiği *[(4x11x122)/4=1342 gözlem)]* Desen F'ye ilişkin kalibrasyon haritası Şekil 11'de sunulmuştur.



Şekil 11. Sarmal puanlama deseni (Desen F) veri kalibrasyon haritası

Kalibrasyon haritası incelendiğinde matematik başarı testindeki açık uçlu maddeleri cevaplandırmada 97, 19, 38, 112 nolu öğrenci yüksek beceriye, 73, 104, 113, 87 nolu

öğrenciler ise düşük beceriye sahiptir. Pozitif logit değerine sahip öğrenci sayısı ile negatif logit değere sahip öğrenci sayısının logit ölçekte birbirine yakın dağıldığı, öğrenci grubunun başarı açısından heterojen olduğu görülmüştür. Puanlayıcı yüzeyine bakıldığında puanlayıcı B ve D'nin birbirine benzer puanlama yaptığı ve en cömert, puanlayıcı C'nin en katı puanlayıcı olduğu anlaşılmıştır. Açık uçlu madde yüzeyine bakıldığında madde güçlüğü en yüksek Madde 6 ve madde güçlüğü en düşük Madde 11 olarak belirlenmiştir. Öğrenciler matematik başarı testindeki Madde 6, Madde 1, Madde 7, Madde 8, Madde 3, Madde 4 ve Madde 5'i kolay; Madde 11, Madde 2, Madde 10, Madde 9'u zor cevaplandırmıştır. Öğrenci performanslarını puanlamada özel bütünsel dereceli puanlama anahtarı derecelendirmelerinin puanlayıcılar tarafından geniş aralıkta ve etkin şekilde kullanıldığı görülmüştür.

4.6.2. Sarmal Puanlama Deseni (Desen F) Puanlayıcı Yüzeyine İlişkin Bulgular

Desen F'ye ilişkin uygunluk ölçütleri ve uyum yüzdeleri Tablo 35'te sunulmuştur.

Tablo 35

Sarmal Puanlama Deseni (Desen F) Puanlayıcı Yüzeyine İlişkin Ölçüm Raporu

	Model		Uygunluk ölçütleri		Uyum yüzdeleri	
Puanlayıcı	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)	Gözlenen uyum	Beklenen uyum
A	-0,51	0,05	0,90	0,82	0,0	0,0
B	-0,41	0,05	1,06	1,12	66,4	39,2
C	-0,57	0,05	0,98	1,20	0,0	0,0
D	-0,33	0,05	0,91	1,01	66,4	39,2
Ortalama	-0,45	0,05	0,96	1,04		
SS (evren)	0,09	0,00	0,06	0,14		
SS (örneklem)	0,10	0,00	0,07	0,17		
Model (evren) RMSE= 0,05 SS: 0,08 Ayırma oranı=1,51 Ayırma indeksi= 2,35						
Güvenirlik (puanlayıcılar arası olmayan) = 0,70						
Model (örneklem) RMSE= 0,05 SS: 0,09 Ayırma oranı=1,84 Ayırma indeksi= 2,79						
Güvenirlik (puanlayıcılar arası olmayan) = 0,77						
Model, sabit etkili ki-kare: 13,2 s.d: 3 p= 0,00						
Model, rastgele ki-kare: 2,4 s.d: 2 p= 0,29						
Puanlayılar arası uyumun gözlenen değeri= %66,4						
Puanlayıcılar arası uyumun beklenen değeri= %39,2						
Puanlayıcılar arası güvenirlik (Rasch-Cohen'in kappa istatistiği)= 0,45						

RMSE: hata kareleri ortalamasının karekökü SS:standart sapma s.d:serbestlik derecesi

Ayırma oranı=1,84, ayırma indeksi=2,79 ve ayırma indeksi güvenirliği=0,77 olarak bulunmuş ve puanlayıcıların katılık-cömertlik davranışlarının birbirinden farklılaştığı ve puanlamaya puanlayıcıdan kaynaklanan hataların karıştığı düşünülmüştür. Ayırma oranı, puanlayıcı yüzeyinde ölçme hatasından kaynaklanabilecek bir farkın oluştuğunu göstermiştir. Puanlayıcıların katılık ve cömertliğini gösteren RMSE değerinin 0,05 bulunması puanlayıcıların katılık davranışının daha kesin bir şekilde tahmin edildiğini göstermektedir. Sabit etkili ki-kare değeri $\chi^2(3)=13,2$, $p=0,00<0,05$ bulunmuş ve en az iki puanlayıcının katılık-cömertlik davranışları arasında istatistiksel olarak anlamlı fark olduğu görülmüştür.

Sarmal puanlama deseni için puanlayıcılar arası uyumun gözlenen değeri %66,4, puanlayıcılar arası uyumun beklenen değeri %39,2 bulunmuştur. Rasch-Cohen'in kappa istatistiği değeri 0,45 olarak hesaplanmıştır. Bu değerin +0,40'tan büyük olması, farklı puanlayıcıların öğrencilerin açık uçlu maddeleri cevaplandırma performanslarını değerlendirirken birbirleriyle tutarlı puanlamalar yaptıklarını, puanlayıcılar arası uyumun ve güvenirliğin modelden beklenen daha fazla olduğunu göstermektedir.

Logit değerlerine incelendiğinde tüm puanlayıcıların negatif logit değerleri aldığı ve katı puanlama yaptığı görülmüştür. En yüksek sahip logit değerine (-0,33) sahip olan puanlayıcı D en cömert, en düşük logit değerine (-0,57) sahip olan puanlayıcı C en katı puanlayıcıdır. Puanlayıcıların iç uyum indeksleri 0,90 ile 1,06 arasında, dış uyum indeksleri 0,82 ile 1,20 arasında değişmektedir ve kabul edilebilir düzeydedir. Bu değerler puanlayıcıların desen ile veri uyumunu sağladığını ve puanlayıcıların birbirlerinden farklı puanlamalar yapsalar da öğrenci performanslarını dereceli puanlama anahtarı ölçütlerinin tamamını göz önünde bulundurarak tutarlı puanladıklarını göstermiştir.

4.6.3. Sarmal Puanlama Deseni (Desen F) Öğrenci Yüzeyine İlişkin Bulgular

Desen F için öğrenci yüzeyine ilişkin verilerin tamamının gösterimi mümkün olmadığından logit değeri en yüksek ve en düşük 8 öğrenci Tablo 36'da sunulmuştur.

Tablo 36

Sarmal Puanlama Deseni (Desen F) Öğrenci Yüzeyine İlişkin Ölçüm Raporu

	Ortalamalar		Model		Uygunluk ölçütleri	
Öğrenci sıra no	Gözlenen ortalama	Düzeltilmiş ortalama	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)
97	3,58	3,67	1,85	0,40	0,39	0,44
19	3,17	3,34	1,30	0,31	0,64	0,83
112	3,00	3,31	1,26	0,31	1,35	1,98
38	3,10	3,31	1,26	0,31	0,66	0,86
22	3,00	3,22	1,17	0,30	2,05	1,49
1	3,00	3,16	1,12	0,28	0,34	0,40
114	2,90	3,12	1,09	0,29	0,80	0,78
76	2,80	3,11	1,08	0,30	1,12	1,45
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
84	0,30	0,22	-1,20	0,49	0,65	1,62
119	0,25	0,21	-1,24	0,50	0,67	1,57
16	0,20	0,15	-1,50	0,62	0,90	3,44
110	0,20	0,15	-1,52	0,62	0,30	0,19
104	0,10	0,08	-2,07	0,93	0,78	0,68
73	0,08	0,07	-2,14	0,94	0,79	0,62
113	0,08	0,07	-2,14	0,94	0,93	2,67
87	0,00	0,02	-3,27	1,80	0,03	0,02
Ortalama	1,50	1,42	-0,03	0,32	0,94	1,04
SS (evren)	0,83	0,93	0,78	0,18	0,52	0,86
SS (örneklem)	0,84	0,94	0,78	0,18	0,52	0,87
Model (evren) Güvenirlik: 0,78	RMSE: 0,37	SS: 0,69	Ayırma oranı= 1,88		Ayırma indeksi= 2,85	
Model (örneklem) Güvenirlik: 0,78	RMSE: 0,37	SS: 0,69	Ayırma oranı= 1,89		Ayırma indeksi= 2,86	
Model, sabit etkili ki-kare: 475,1		s.d: 121		p= 0,00		
Model, rastgele ki-kare: 80,6		s.d: 120		p= 1,00		

RMSE: hata kareleri ortalamalarının karekökü SS: standart sapma s.d: serbestlik derecesi

Ayırma oranı=1,89, ayırma indeksi=2,86 ve ayırma indeksi güvenirliği=0,78 olarak bulunmuştur. Bu değerler açık uçlu maddeleri cevaplandırma becerisi birbirlerinden farklı olan öğrencilerin güvenilir bir şekilde ayırt edilebildiklerini göstermiştir. Ayırma oranına bakıldığında öğrenci performansları arasındaki farklılığın ölçme hatası ile açıklanabileceği düşünülebilir. Öğrencilerin başarı ölçümlerinin RMSE değeri 0,37 olarak bulunmuş ve yüksek beceriye sahip öğrencilerin daha kesin bir şekilde tahmin edildiği anlaşılmıştır. Desen F için *([4.(2,86)+1)]/3=4,15)* öğrenci performanslarının dört seviyeye ayrıldığı söylenebilir. Sabit etkili ki-kare değerinin $\chi^2(121)=475,1$, $p=0,00<0,05$ olması en az iki öğrencinin yetenek düzeyleri arasında anlamlı fark bulunduğunu ve başarı açısından öğrenci

grubunun heterojen yapıda olduğunu göstermiştir. Logit değeri (1,85) olan 97 numaralı öğrenci en yüksek, logit değeri (-3,27) olan 87 numaralı öğrenci en düşük yetenek düzeyine sahiptir. İç uyum indeksleri 0,03 ile 3,16 arasında, dış uyum indeksleri 0,02 ile 7,28 arasında değişmiştir. Bazı öğrencilerin dış uyum indeksleri 2,00'den büyük olsa da iç uyum indekslerinin 2,00'den büyük olmaması model veri uyumunun bozulmadığını göstermiştir. Öğrencilerin %82'sinin iç ve dış uyum indeksleri açısından kabul edilebilir aralıkta (0,50-1,50) olduğu görülmüştür. Uyum aralığı dışında kalan yirmi iki öğrenci için her bir açık uçlu maddeyi cevaplandırma performanslarının beklenen düzeyde olmadığı; dış uyum aralığı dışında da kalan yirmi iki öğrencinin matematik başarı testi uygulamasında gösterdikleri genel performansın beklenen düzeyde olmadığı anlaşılmıştır. Dış uyum indeksi 1,00 olan 120 numaralı öğrencinin matematik başarı testindeki genel performansı ile matematik başarısı arasındaki uyum mükemmeldir.

4.6.4. Sarmal Puanlama Deseni (Desen F) Açık Uçlu Madde Yüzeyine İlişkin Bulgular

Desen F için açık uçlu madde yüzeyine ilişkin uygunluk ölçütleri ve madde güçlük indeksleri Tablo 37'de sunulmuştur.

Tablo 37

Sarmal Puanlama Deseni (Desen F) Açık Uçlu Madde Yüzeyine İlişkin Ölçüm Raporu

	Ortalamalar		Model		Uygunluk ölçütleri	
Madde no	Gözlenen ortalama	Düzeltilmiş ortalama	Logit değeri	Standart hata	Infit (iç uyum)	Outfit (dış uyum)
1	1,93	1,81	0,36	0,08	1,03	0,91
2	0,54	0,32	-0,88	0,11	1,00	1,24
3	1,80	1,66	0,29	0,08	0,96	1,00
4	1,81	1,65	0,28	0,08	0,63	0,66
5	1,68	1,47	0,19	0,08	0,95	0,91
6	2,47	2,62	0,77	0,08	1,11	1,00
7	1,97	1,89	0,40	0,08	1,13	0,99
8	1,96	1,87	0,39	0,08	1,07	1,07
9	1,09	0,79	-0,25	0,08	0,68	0,76
10	0,80	0,49	-0,57	0,09	1,29	1,50
11	0,44	0,28	-0,99	0,12	0,61	1,37
Ortalama	1,50	1,35	0,00	0,09	0,95	1,04
SS (evren)	0,64	0,73	0,55	0,01	0,21	0,24
SS (örneklem)	0,67	0,76	0,58	0,02	0,22	0,25
Model (evren) Güvenirlik: 0,98	RMSE: 0,09	SS: 0,55	Ayırma oranı=6,31		Ayırma indeksi=8,75	
Model (örneklem) Güvenirlik: 0,98	RMSE: 0,09	SS: 0,58	Ayırma oranı=6,63		Ayırma indeksi=9,17	
Model, sabit etkili ki-kare: 359,2			s.d: 10		p= 0,00	
Model, rastgele ki-kare: 9,7			s.d: 9		p= 0,37	
RMSE: hata kareleri ortalamalarının karekökü			*SS: standart sapma*		*s.d: serbestlik derecesi*	

Ayırma oranı=6,63, ayırma indeksi=9,17, ayırma indeksi güvenirliği=0,98 olan açık uçlu madde yüzeyine ilişkin bu değerler farklı güçlük düzeylerine sahip maddelerin yüksek güvenirlikle birbirlerinden ayırt edildiği göstermiştir. Ayırma oranına bakıldığında madde güçlük düzeylerinin ölçme hatasından kaynaklanmayacak kadar birbirlerinden farklı olduğu görülmüştür. Sabit etkili ki-kare testi ile ($\chi^2(10)=359,2$, $p=0,00<0,05$) madde güçlük düzeyleri arasındaki farklılığın istatistiksel olarak anlamlı olduğu anlaşılmıştır.

Açık uçlu maddelerin güçlük düzeylerinin logit değerleri 0,77 ile -0,99 arasında değişmektedir. Tablo 37 incelendiğinde Madde 6 (logit=0,77)'nın en yüksek madde güçlüğüne ve Madde 11 (logit= -0,99)'nin en düşük madde güçlüğüne sahip olduğu görülmektedir. Logit değerlerine göre madde güçlüğü düşük maddeler, Madde 9, Madde 10, Madde 2, Madde 11; madde güçlüğü yüksek maddeler, Madde 6, Madde 7, Madde 8, Madde 1, Madde 3, Madde 4 ve Madde 5 şeklinde sıralanmıştır. Maddelerin güçlük/kolaylık durumunu gösteren RMSE değerinin 0,09 bulunması ile madde güçlüğünün daha kesin bir şekilde tahmin edildiği anlaşılmıştır. Standart hata değeri sırasıyla 0,12; 0,11; 0,09 olan Madde 11, Madde 2, Madde 10 dışındaki madde güçlükleri daha kesin tahmin edilmiştir. İç

uyum indeksleri 0,61 ile 1,29 arasında, dış uyum indeksleri 0,66 ile 1,50 arasında değişmektedir. İç ve dış uyum değerlerinin tamamı 0,50 ile 1,50 aralığında ve kabul edilebilir düzeydedir. İç uyum indeksleri ilgili kazanımı ölçmede her bir maddenin iyi işlediğini ve dış uyum indeksleri başarı testinde yer alan maddelerin öğrenci matematik başarısını ölçmede iyi olduğunu göstermiştir.

4.6.5. Sarmal Puanlama Deseni (Desen F) Özel Bütünsel Dereceli Puanlama Anahtarı Ölçütlerine İlişkin Bulgular

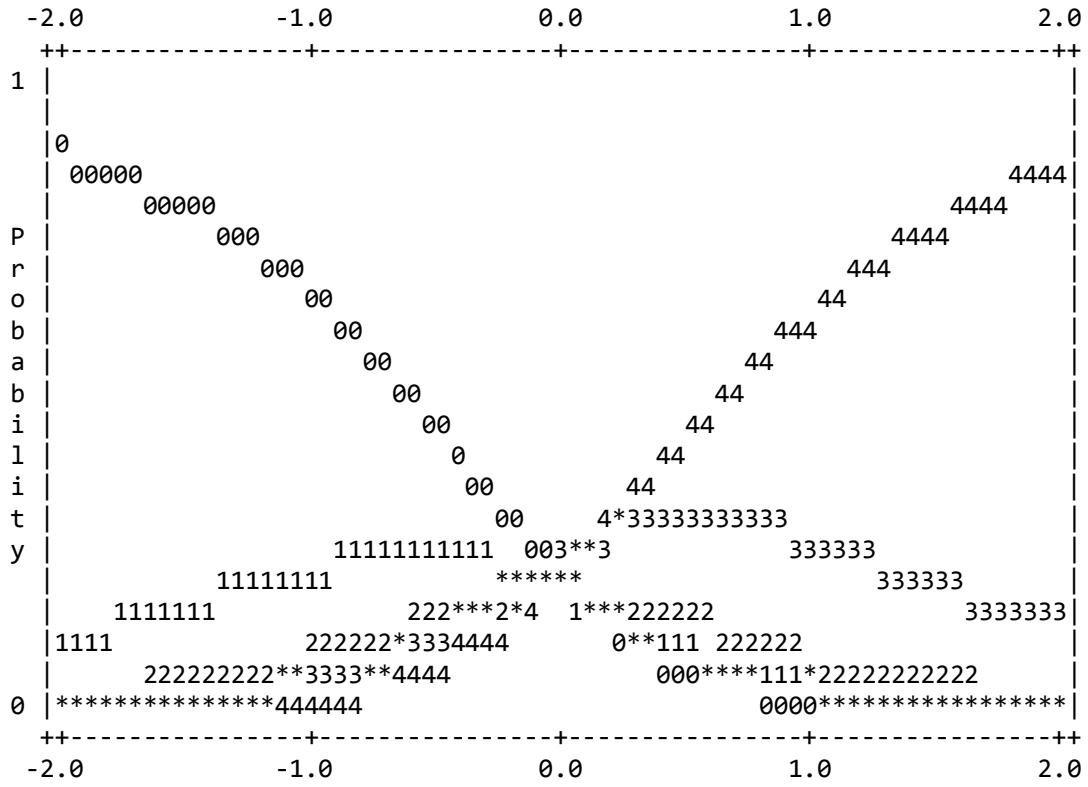
Desen F için dereceli puanlama anahtarı kategorilerine ilişkin değerler Tablo 38'de sunulmuştur.

Tablo 38

Özel Bütünsel Dereceli Puanlama Anahtarı Puanlama Kategorilerine Ait Ölçüm Raporu

Puanlama Kategorileri	Frekans	Yüzde (%)	Birikimli yüzde (%)	Ortalama logit değeri	Beklenen logit değeri	Outfit (dış uyum)
0	555	41	41	-1,11	-1,15	1,00
1	216	16	57	-0,70	-0,56	1,50
2	148	11	68	-0,20	-0,17	1,00
3	193	14	83	0,26	0,16	0,70
4	230	17	100	0,44	0,46	1,10

Frekans ve yüzde değerleri, puanlayıcıların puanlama yaparken en çok 0 numaralı kategoriyi tercih ettiklerini ve 1, 2, 3 ve 4 numaralı kategorileri birbirine yakın sayıda kullandıklarını göstermiştir. 0 kategorisinin daha fazla kullanılmasının sebebi öğrenci yüzeyi ölçüm raporunda öğrencilerin matematik başarı düzeylerinin yüksek güvenirlikle ayırt edildiğinden merkezi eğilim etkisininden kaynaklanmamaktadır. Logit değerleri incelendiğinde 3 ve 4 numaralı kategorilerin tam ve oldukça doğru tanımlarına uygun bir şekilde pozitif değer aldığı; 0, 1, 2 numaralı kategorilerin boş, tamamen yanlış ve kısmen yanlış tanımlarına uygun olarak negatif değerler aldığı görülmüştür. Ortalama ve beklenen logit değerlerinin birbirine yakın olduğu ve verilerin Rasch modeli beklentilerine uyduğu görülmüştür. Dış uyum indeksleri 0,70 ile 1,50 arasında değişmektedir ve kabul edilebilir düzeydedir. Matematik başarı testindeki açık uçlu maddelerin değerlendirilmesinden elde edilen puanlarla özel bütünsel dereceli puanlama anahtarı kategorilerinin uyumlu ve dereceli puanlama anahtarının güvenilir bir ölçme aracı olduğu görülmüştür. Dış uyum indeks değeri 1,00 olan 0 ve 2 numaralı kategoriler, matematik başarı testindeki öğrenci performanslarının değerlendirilme puanları ile mükemmel uyum içerisindedir.



Şekil 12. Sarmal puanlama desenine (Desen F) göre dereceli puanlama anahtarı ölçütlerine ait olasılık eğrileri

Olasılık eğrileri düşük beceri düzeyindeki öğrencilerin 0, 1, 2 numaralı alt kategorilerde bulunma olasılığının yüksek; 3 ve 4 numaralı üst kategorilerde bulunma olasılığının düşük olduğunu göstermiştir. Benzer şekilde yüksek beceri düzeyindeki öğrencilerin üst kategorilerde bulunma olasılığı yüksek; alt kategorilerde bulunma olasılığı ise düşüktür. Özel bütünsel dereceli puanlama anahtarı kategorilerinin açık uçlu maddeleri cevaplandırmadaki öğrenci başarılarını ayırt etmede iyi olduğu görülmüştür.

Tüm desenlerde puanlayıcıların katılık-cömertlik davranışları arasında anlamlı bir fark olduğu, katı puanlama yaptıkları, model ile veri uyumunu bozmadıkları, puanlamaya puanlayıcıdan kaynaklanan hataların karıştığı, puanlayıcılar arası güvenirliği gösteren Rasch-Cohen'in kappa istatistiğinin Desen A, B ve C ile Desen E ve F'nin birbirine yakın değerler aldığı anlaşılmıştır.

Tüm desenlerde öğrenci beceri düzeylerinin öğrenci beceri düzeyleri arasında anlamlı fark olduğu ve bu farklılığın yüksek güvenirlikle ayırt edildiği, model ve veri uyumunu bozan öğrencilerin bulunduğu, Desen C ile F'de ölçme hatasıyla açıklanabilecek farklılığın oluştuğu, Desen A ile B'de öğrencilerin logit değerlerinin benzer olduğu görülmüştür.

Tüm desenlerde açık uçlu maddelerin güçlük düzeyleri arasında anlamlı fark olduğu ve bu farklılığın yüksek güvenirlikle ayırt edildiği, model veri uyumunu bozan maddenin bulunmadığı sadece dış uyum indeks değerinin 1,50'den büyük olması nedeniyle Madde 10'un öğrencilerin matematik başarısını ölçmede verimsiz olduğu söylenebilir.

Tüm desenlerde dereceli puanlama anahtarının derecelendirmelerinden en çok 0 kategorisinin kullanıldığı ve bu durumun öğrenci beceri düzeylerinin düşük olmasından kaynaklandığı, becerisi düşük öğrencilerin 0, 1, 2 kategorilerinde, becerisi yüksek öğrencilerin 3, 4 kategorilerinde bulunma olasılığının yüksek olduğu anlaşılmıştır. Tablo 39'da puanlama desenlerine ilişkin bulgular karşılaştırılmıştır.

Tablo 39

Çok Yüzeyli Rasch Modeli Puanlama Desenlerine İlişkin Bulguların Karşılaştırılması

	Öğrenci beceri düzeyi		Puanlayıcı davranışı			Madde güçlüğü		Puanlama kategorileri
Desen adı	Yüksek	Düşük	Katı	Ilımlı	Cömert	Yüksek	Düşük	Çok kullanılan
Desen A	76, 22, 38	93, 87	C	A-B	D	6, 3-7, 1-4-8, 5	2, 10-11, 9	0
Desen B	76, 22, 38	85, 87	C	A-B	D	6, 3-7, 1-8, 4-5	2, 10-11, 9	0
Desen C	38, 76, 22	93,87	C	-	A-B-D	6, 3, 1-4-7, 5-8	2, 10, 11, 9	0
Desen D	22, 76	87	A-B	C	D	6, 1-3-7, 4-8, 5	2, 11, 10, 9	0
Desen E	76, 22, 38	87	B	C-D	A	6, 3-4-8, 1-7, 5	2-11, 10, 9	0
Desen F	97	87	C	A	B-D	6, 1-7-8, 3-4-5	11, 2, 10, 9	0

Açık uçlu maddeleri cevaplandırma beceri düzeyi yüksek öğrencilerin Desen A ve B'de aynı şekilde sıralandığı, beceri düzeyi en düşük öğrencinin tüm desenlerde aynı olduğu görülmüştür. Desen A ile B'deki katı, ılımlı ve cömert puanlama yapan puanlayıcılar aynı şekilde sıralanmıştır. Madde 6'nın öğrenciler tarafından kolay cevaplandığı başka bir deyişle madde güçlüğünün yüksek olduğu; Desen F dışında Madde 2'nin öğrenciler tarafından zor cevaplandığı, madde güçlüğünün düşük olduğu anlaşılmıştır. Sıralamaları farklı olsa da madde güçlüğü yüksek ve düşük maddelerin tüm desenlerde aynı olduğu; puanlayıcıların puanlama yaparken bütünsel dereceli puanlama anahtarı kategorilerinden en çok 0 derecelendirmesini kullandığı düşünülebilir. Bunun yanında öğrenci beceri düzeyi,

puanlayıcı davranışı, madde güçlüğü ve puanlama kategorileri bulgularına bakıldığında Desen A ile B'nin oldukça benzer olduğu söylenebilir.

BÖLÜM V

SONUÇLAR VE ÖNERİLER

Bu araştırma geniş ölçekli sınavlarda yer alan matematik alt testindeki ya da ilgili kazanımı ölçmek için hazırlanan matematik başarı testindeki açık uçlu maddelerin çok yüzeyli Rasch modeli puanlama desenlerinden hangisi ile değerlendirilmesinin daha geçerli ve güvenilir sonuçlar vereceği konusuna açıklık getirmek amacıyla gerçekleştirilmiştir. Eckes (2015) tarafından tanımlanan Desen A, Desen B, Desen C, Desen D ve Desen E ile Tharbai ve Young (2001) yılında kullandığı sarmal puanlama desenine ait (Desen F) logit değerleri, uyum yüzdeleri ve indeksleri, madde güçlük düzeyleri, Rasch-Cohen'in kappa istatistikleri karşılaştırılarak matematik alanındaki açık uçlu maddeler için en uygun puanlama deseni ortaya koyulmaya çalışılmıştır.

5.1. Sonuçlar

5.1.1. Veri Kalibrasyon Haritasına İlişkin Sonuçlar

Tüm puanlayıcıların tüm öğrencilerin tüm maddelerdeki cevaplarını puanladığı genel olarak kullanılan ve zaman alıcı puanlama deseni olan Desen A ile Desen B'de yüksek beceriye sahip öğrencilerin aynı şekilde sıralandığı, sıralamaları farklı olsa da düşük beceriye sahip öğrenci grubunun her iki desende aynı olduğu görülmüştür. Desen A, Desen B, Desen C, Desen D ve Desen E'de yüksek beceriye sahip öğrenci grubunun aynı öğrencilerden oluştuğu; düşük beceriye sahip öğrenci grubunun ise benzer olduğu anlaşılmıştır. Desen F'nin öğrenci başarı grupları diğer desenlerden farklıdır. Tüm maddeleri tüm puanlayıcılar tarafından 0 ile değerlendirilen 87 numaralı öğrenci tüm desenlerde logit ölçekte en düşük beceriye sahip öğrenci olarak gösterilmiştir.

Puanlayıcı yüzeyinde, Desen A ile Desen B'nin puanlayıcıların aynı şekilde sıralandığı, diğer desenlerdeki puanlayıcı sıralamasının Desen A'dan oldukça farklı olduğu görülmüştür. Açık uçlu madde yüzeyinde, madde güçlüğü yüksek maddeler ile madde güçlüğü düşük maddelerin tüm desenlerde ortak olduğu tespit edilmiştir. Açık uçlu maddelerin logit ölçekteki sıralamalarının Desen A ile Desen B'de bir madde dışında aynı, Desen A, B, C ve D'de en yüksek ve en düşük madde güçlüğüne sahip olan maddelerin aynı, Desen E ve F'de ise farklıdır. Tüm desenlerde puanlayıcıların öğrenci performanslarını puanlarken dereceli puanlama anahtarı puanlama kategorilerini geniş aralıkta kullandıkları görülmüştür.

5.1.2. Puanlayıcı Yüzeyine İlişkin Sonuçlar

Tüm desenlerde puanlayıcıların katı puanlama yaptığı, en az iki puanlayıcı arasında anlamlı fark olduğu sonucuna ulaşılmış; ayırma indeksi güvenirliklerinin 0,70'den büyük olması da puanlayıcılar arasındaki katılık cömertlik davranışlarının farklı olduğunu ve puanlamaya puanlayıcıdan kaynaklanan hatanın karıştığını göstermektedir. Desen A, B ve D'de puanlayıcıların katılık ve cömertlik davranışları arasındaki farkın benzer bir şekilde ayırt edildiği görülmüştür. Desen C ve F'de puanlayıcılar arasındaki katılık cömertlik davranışı farkının ölçme hatasından kaynaklanabileceği düşünülmüştür. Puanlayıcılar arasındaki uyumu gösteren Rasch-Cohen'in kappa istatistiği Desen A, B, C, D, E ve F için; 0,65, 0,66, 0,66, 0,93, 0,43, 0,45 şeklinde sıralanmıştır. Tüm desenlerde bu istatistiğin 0,00'dan büyük olması ile puanlayıcıların öğrenci performanslarını tutarlı bir şekilde değerlendirdiğini puanlayıcılar arası uyumun (güvenirliğin) sağlandığı görülmüştür. Desen A ile B'nin logit değerleri, iç ve dış uyum indeksleri, gözlenen ve beklenen uyum yüzdeleri, gözlenen ve beklenen uyum yüzdeleri arasındaki farka dayanan Rasch-Cohen'in kappa istatistik değerleri birbirine oldukça yakındır. Desen B ile C aynı Rasch-Cohen'in kappa değerine sahip olmasına rağmen logit ve uyum indeks değerlerinin, kalibrasyon haritalarının benzer olmaması bu iki desenin farklı olduğunu ve birbirlerine alternatif olamayacaklarını göstermiştir. Desen A ile Desen B'nin benzer Rasch-Cohen'in kappa istatistiğine sahip olmalarının yanında puanlayıcı logit değerleri, kalibrasyon haritası ve uyum indekslerinin birbirine oldukça yakın olması nedeniyle Desen B'nin Desen A'nın yerine kullanılabileceği sonucuna varılmıştır. Puanlayıcı A ve B'nin birinci alt kümeyi, Puanlayıcı C ve D'nin ikinci alt kümeyi puanlaması sonucu Desen D'de iki alt küme arasında bağlantılılık probleminin oluşması her alt küme içerisinde iki puanlayıcının verdikleri puanların daha tutarlı olmasına

neden olmuştur. Bu durum Desen D'nin Rasch-Cohen'in kappa istatistik değerinin yüksek başka bir deyişle puanlayıcılar arası uyumun oldukça fazla çıkmasının sebebi olarak düşünülebilir. Desen E ile Desen F'nin puanlayıcılar arası uyumunun (güvenirliğinin) benzer olduğu düşünülse de puanlayıcı yüzeyine ilişkin diğer değerler arasında benzerlik söz konusu değildir ve bu durum birbirlerine alternatif olamayacaklarını göstermektedir.

5.1.3. Öğrenci Yüzeyine İlişkin Sonuçlar

Desen F dışındaki desenlerde açık uçlu maddeleri cevaplandırmada yüksek beceriye sahip öğrenci grubu ile düşük beceriye sahip öğrenci grubunun benzer olduğu görülmüştür. Tüm desenlerde en az iki öğrenci performansı arasında anlamlı fark bulunmuş ve öğrenci performansları yüksek güvenirlikle ayırt edilmiştir. Öğrenci performansları Desen A ve B'de altı, Desen D ve E'de beş, Desen C ve F'de dört seviyeye ayrılmıştır. Desen C ve F'de öğrenci performansları arasında oluşan farkın ölçme hatasıyla açıklanabileceği anlaşılmıştır. Desen A, B ve E'de yüksek beceriye sahip öğrenciler logit ölçeğe aynı şekilde sıralanmıştır. Desen A ve B'de öğrencilerin performans seviye sayıları aynı, logit değerleri, sıralamaları, iç ve dış uyum değerleri oldukça benzerdir. Bu durum Desen B'nin Desen A'ya alternatif olabileceği görüşünü desteklemektedir. Diğer desenler arasında benzer bir ilişki bulunamamıştır.

5.1.4. Açık Uçlu Madde Yüzeyine İlişkin Sonuçlar

Tüm desenlerde madde güçlük düzeyleri arasındaki fark oldukça 0,98, 0,99 gibi yüksek güvenirlikle ayırt edilmiştir. En az iki maddenin güçlük düzeyleri arasındaki ölçme hatasıyla açıklanamayacak kadar büyük farkın oluştuğu ve bu farkın istatistiksel olarak anlamlı olduğu görülmüştür. Tüm desenlerde madde güçlüğü yüksek maddeler ile madde güçlüğü düşük maddeler aynıdır. Açık uçlu maddeler ilgili kazanımı ve matematik başarısını ölçmede çok iyi işlemektedir. Madde 6 tüm desenlerde öğrenciler tarafından en kolay cevaplanan ve madde güçlüğü en yüksek madde olarak belirlenmiştir. Desen F dışında madde güçlüğü en düşük ve öğrencilerin zor cevaplandırdığı maddenin Madde 2 olduğu görülmüştür. Desen A ve B'nin gözlenen ve düzeltilmiş ortalama, ölçmeye karışan standart hata ve logit değerlerinin, iç ve dış uyum değerlerinin birbirine çok yakın olduğu anlaşılmıştır.

5.1.5. Özel Bütünsel Dereceli Puanlama Anahtarı Kategorilerine İlişkin Sonuçlar

Puanlayıcılar öğrenci performanslarını puanlarken tüm desenlerde en fazla 0 kategorisini, birbirine yakın sayıda da 1, 2, 3, 4 kategorilerini kullanmışlardır. Kategori istatistiklerine bakıldığında merkezi eğilim etkisi olabileceği düşünülmüş ancak öğrenci yüzeyi ölçüm raporunda öğrencilerin başarı düzeylerinin tüm desenlerde güvenilir bir şekilde ayırt edildiği görülmüştür. Bu durum düşük seviyedeki öğrenci başarılarından kaynaklanmıştır. Tüm desenlerde derecelendirmeler performans tanımlarına uygun olarak pozitif ve negatif logit değerleri almıştır. Desen C ile E'nin matematik başarı testindeki açık uçlu maddelerin değerlendirilmesinden elde edilen puanlarla özel bütünsel dereceli puanlama anahtarı kategorilerinin uyumunu gösteren dış uyum değerleri aynıdır. Desen A ile B'nin ortalama ve beklenen logit değerleri, dış uyum değerleri birbirlerine oldukça yakındır. Desen C, D ve E'nin ortalama ve beklenen logit değerleri, dış uyum değerleri benzerdir. Desen F'ye ilişkin değerler ise diğer desenlerden oldukça farklıdır. Olasılık eğrilerinin tüm desenlerde benzer görünüme sahip olduğu görülmüştür. Düşük beceriye sahip öğrencilerin 0, 1, 2 numaralı alt kategorilerilerde yer alma olasılığı ile yüksek beceriye sahip öğrencilerin 3, 4 numaralı üst kategorilerde yer alma olasılığının tüm desenlerde yüksek olduğu görülmüştür.

5.1.6. Puanlama Desenlerine İlişkin Sonuçlar

Tüm puanlayıcıların tüm birey cevaplarını tüm açık uçlu maddelerde puanladığı tamamen çaprazlanmış desen (Desen A) ile her bir öğrencinin her bir açık uçlu maddedeki cevabının dört puanlayıcının üçü tarafından puanlandığı tamamlanmamış bağlantılı desenin (Desen B) puanlayıcı, öğrenci, açık uçlu madde yüzeyleri ile dereceli puanlama anahtarı kategorilerine ilişkin ölçüm raporlarının, puanlayıcılar arası güvenirliklerinin oldukça benzer olduğu ve birbirine alternatif olabilecekleri sonucuna varılmıştır. Geniş ölçekli sınavlarda kullanımı zor ve zaman alıcı olan Desen A yerine zaman ve iş yükü tasarrufu sağlayan Desen B matematik alt ya da başarı testlerindeki açık uçlu maddelerin puanlanmasında kullanılabilir.

Gözlem sayısı oldukça az olmasına rağmen bir puanlayıcının tüm bireyleri puanlaması gerektiği için geniş ölçekli sınavlarda kullanışlı olmayan Desen C ile Desen A'nın puanlayıcılar arası uyumları (güvenirlikleri) yakın olsa da yüzey ölçüm raporlarının benzer olmaması, Desen C'nin Desen A'nın alternatifi olamayacağını göstermiştir.

Puanlayıcılardan ikisinin öğrencilerin aynı maddedeki cevaplarını diğer ikisinin de kalan maddeleri puanladığı bağlantısı olmayan iki alt kümedeki ölçümler arasındaki farkın puanlayıcıların katılığı ya da öğrenci başarı düzeyinden kaynaklanıp kaynaklanmadığının karşılaştırılmasının mümkün olmadığı Desen D'nin puanlayıcılar arası güvenirliğinin, yüzey ölçüm raporlarının Desen A'dan oldukça farklı olduğu görülmüştür. Bu nedenle Desen D'nin Desen A yerine kullanılamayacağı sonucuna ulaşılmıştır.

Puanlayıcılardan ikisinin her bir öğrencinin çift numaralı, diğer ikisinin ise tek numaralı maddelerde puanlama yaptığı puanlayıcı iş yükü az ve maliyeti düşük olan Desen E ile Desen A'nın puanlayıcılar arası güvenirliğinin, yüzey ölçüm raporlarının birbirinden oldukça farklı olması Desen E'nin Desen A'nın alternatifi olamayacağını göstermiştir.

Açık uçlu maddelerin değerlendirilmesinde kullanılan her bir öğrencinin her bir açık uçlu maddedeki cevabının farklı bir puanlayıcı tarafından puanlandığı Desen F'nin puanlayıcılar arası güvenirliğinin, yüzey ölçüm raporlarının, puanlayıcı yüzeyine ilişkin istatistikleri tahmin etmede uygun olmaması, Desen A'dan oldukça farklı olması nedeniyle Desen F'nin de Desen A'nın alternatifi olamayacağı sonucuna varılmıştır.

5.2. Öneriler

5.2.1. Araştırmacılara Yönelik Öneriler

1. Bu çalışma ile matematik alt testinde yer alan açık uçlu maddelerin puanlanmasında kullanılabilecek güvenilir ve kullanışlı puanlama desenleri ortaya koyulmaya çalışılmıştır. Ancak sözel alanda hazırlanan bir testte bulunan açık uçlu maddeler daha çok yorum ve açıklama gerektirdiğinden farklı bir desen kullanmak söz konusu olabilir. Bu nedenle her bir branş için hazırlanan alt testler ya da başarı testlerindeki açık uçlu maddelerin değerlendirilmesinde güvenilir ve geçerli puanlama deseninin hangisi olduğu belirlenebilir.
2. Puanlayıcılara puanlayıcı eğitimleri verilerek eğitiminin etkisini görebilmek adına eğitim öncesi ve eğitim sonrası puanlayıcılar arası uyumun, güvenirliğin puanlama desenlerine göre nasıl değiştiği incelenebilir.
3. Geniş ölçekli sınavlarda düzey belirlemeye çalışıldığından bu çalışmada bütünsel dereceli puanlama anahtarları kullanılmıştır. Öğrencilere kazanımların ne kadarının

kazandırıldığının kontrolü için hazırlanan öğretmen yapımı testlerde yer alan açık uçlu maddeleri cevaplandırmadaki öğrenci performansları puanlayıcılar tarafından analitik dereceli puanlama anahtarı ile değerlendirilerek puanlama desenlerine göre puanlayıcılar arası güvenirlik durumlarına bakılabilir.

4. Örneklem büyüklükleri eşit olacak şekilde puanlama desenlerine göre öğrencilere öz ve akran değerlendirme yaptırılıp öğretmen değerlendirmeleri ile puanlayıcılar arası güvenirlikleri karşılaştırılabilir.
5. Puanlayıcı sayısı artırılarak ya da azaltılarak puanlayıcılar arası güvenirliğin puanlama desenlerine göre nasıl değiştiği kontrol edilebilir.
6. Matematik için hazırlanan testlerde yer alan açık uçlu madde sayısı artırılarak puanlayıcılar arası uyumun, güvenirliğin değişip değişmediği incelenebilir.
7. İlkokul ya da ortaöğretim kademesindeki öğrencilere seviyelerine uygun olarak açık uçlu maddelerden oluşan matematik ya da diğer branşlarda başarı testi uygulanıp puanlama desenlerinin puanlayıcılar arası güvenirlikleri karşılaştırılabilir.
8. Öğrenci sayısı artırılarak puanlama desenlerinin sonuçları ve puanlayıcı uyumları incelenebilir.
9. Alan yazınına katkı sağlamak amacıyla araştırmacı tarafından özgün puanlama desenleri oluşturularak bu desenin güvenirliği ve geçerliği kontrol edilebilir.

5.2.2. Uygulamaya Yönelik Öneriler

1. Tüm puanlama desenlerinde puanlamaya puanlayıcıdan kaynaklanan hataların karıştığı görülmüştür. Açık uçlu maddelerden oluşan öğretmen yapımı testlerde ya da geniş ölçekli sınavlarda daha güvenilir sonuçlar elde edebilmek için puanlayıcılara puanlayıcı eğitimi verilebilir ve puanlama konusunda öğretmen yeterliklerinin artırılması sağlanabilir.
2. Açık uçlu maddelerin değerlendirilmesinde puanlayıcılar arası uyumun sağlanması için desenin bir bölümünde puanlayıcı ve öğrenci yüzeylerinin örtüştüğü desenler tercih edilebilir. Puanlayıcı ve öğrenci yüzeyinin örtüşmediği durumlarda veriler arasında bağlantılılık sorunu oluşabilmekte ve tutarlı puanlamalar yapılsa da puanlayıcılar arası güvenirlikle ilgili problemler meydana gelebilmektedir.
3. Öğrenci yüzeyinden elde edilen bulgulara bakıldığında öğrencilerin açık uçlu maddeleri cevaplandırma becerilerinin yeterli olmadığı görülmüştür. Ayrıca üst bilişsel becerileri ölçme amacıyla hazırlanan çoktan seçmeli yeni nesil maddelerin

öğrenciler tarafından çözülmesinde sorunlar yaşandığı öğretmenlerce görülmektedir. Açık uçlu ve çoktan seçmeli maddeleri cevaplandırmada öğrenci beceri düzeyinin artırılması için derslerde üst bilişsel becerileri geliştiren etkinliklere yer verilebilir.

KAYNAKLAR

Anadolu Üniversitesi Açıköğretim, İktisat ve İşletme Fakülteleri Eğitim-Öğretim ve Sınav Yönetmeliğine Dair Değişiklik. (2017). *T.C. Resmi Gazete* (29952, 18 Ocak 2017).

Al Ruqeishi, M. (2015). An evaluation of alternative assessment tools used in grades 5–8 of Omani Basic Education schools as perceived by EFL teachers. In R. Al-Mahrooqi & C. J. Denman (Eds.), *Issues in english education in the arab world* (pp. 192–215). England: Cambridge.

Akgün, M. (2016). Yüksek öğretimde ideal öğretim elemanı nasıl olmalıdır. *Kaygı Uludağ Üniversitesi Fen-Edebiyat Fakültesi Felsefe Dergisi, 26*, 197-204.

Anastasi, A., & Urbina, U. (1997). *Psychological testing*. New Jersey: Prentice-Hall.

Andrich, D. (1988). *Rasch models for measurement*. California: Sage.

Arsan, N. (2012). *Buz pateninde hakem değerlendirmelerinin genellenebilirlik kuramı ve rasch modeli ile incelenmesi.* Doktora Tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

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

Atılgan, H. (2004). *Genellenebilirlik kuramı ve çok değişkenlik kaynaklı rasch modelinin karşılaştırılmasına ilişkin bir araştırma*. Doktora Tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Bahar, M. (2001). Çoktan seçmeli testlere eleştirel bir yaklaşım ve alternatif metotlar. *Kuram ve Uygulamada Eğitim Bilimleri, 1*(1), 24-38.

Bayram, N. (2009). *Sosyal bilimlerde SPSS ile veri analizi*. Bursa: Ezgi.

Bobcock, B. G. E. (2009). *Estimating a noncompensatory IRT model using a modified metropolis algorithm*. Unpublished Doctoral Dissertation, The University of Minnesota.

Brennan, R. L. (1992). *Elements of generalizability theory.* Iowa City, IA: ACT.

Bridgeman, B. (1992). A comparison of quantative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*(3), 253-271.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats. It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*, 385-395.

Brookhart, S. M. (1999). The art and science of classroom assessment: the missing part of pedagogy. *ASHE-ERIC Higher Education Report, 27*(1).

Brookhart, S. M., & Nitko, A. J.(2008). *From assessment and grading in classrooms.* Upper Saddle River. NJ: Pearson Education.

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

Büyüköztürk, Ş. (2008). *Sosyal bilimler için veri analizi el kitabı.* Ankara: PegemA.

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2012). *Bilimsel araştırma yöntemleri.* Ankara: Pegem.

Campbell, C. (2013). Research on teacher competency in classroom assessment. In J. H. McMillan (Eds.), *Sage handbook of research on classroom assessment* (pp. 71-84). USA: SAGE.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper & L. Odell (Eds), *Evaluating writing describing, measuring, judging (*pp. 3-31). Urbana: National Council of Teachers of English.

Craig, D. A. (2001). *Rater's perception of handwriting legibility and word-processing in L2 writing assessment.* Master's Thesis, Illinois Üniversitesi, Urbana.

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

Çepni, S. (2011). Performansların değerlendirilmesi. E. Karip (Ed.), *Ölçme ve değerlendirme* içinde (s. 233-292). Ankara: Pegem.

Çetin, Ş. (2019). Açık uçlu soruların çok yüzeyli rasch modeli ile analizi. *SETSCI Conference Proceedings, 4*(9), 108-110.

Deliceoğlu, G. (2009). *Futbol yetilerine ilişkin dereceleme ölçeğinin genellenebilirlik ve klasik test kuramına dayalı güvenirliklerinin karşılaştırılması*. Doktora Tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher, 27*, 14–24.

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

Demirel, Ö.(2004). *Kuramdan uygulamaya eğitimde program geliştirme*. Ankara: PegemA.

Doğru, Ş. C. (2019). *Karma testlerin psikometrik özelliklerini belirlemede klasik test kuramı ve rasch modelinin karşılaştırılması.* Yüksek Lisans Tezi, Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alfa to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399-412.

Dochy, F. (2001). A new assessment era: Different needs, new challenges. *Research Dialogue in Learning and Instruction, 10*(1), 11-20.

Eckes, T. (2009). Many-facet rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*, 1-52.

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

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment: Fundamental techniques* (pp.153-175). New York, NY: Routledge.

Elhan, A. H., & Atakurt, Y. (2005). Ölçeklerin değerlendirilmesinde niçin rasch analizi kullanılmalıdır. *Ankara Üniversitesi Tıp Fakültesi Mecmuası, 58*, 47-50.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Engelhard, G. J. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement, 1*(1), 19-33.

Engelhard, G. J. (2013). *Invariant measurement*: *Using rasch models in the social, behavioral, and health sciences*. Routledge: Taylor & Francis Group.

Farrokhi, F., Esfandiari, R., & Dalili, M. V. (2011). Applying the Many-Facet Rasch Model to detect centrality in self assessment, peer-assessment and teacher assessment. *World Applied Sciences (Innovation and Pedagogy for Lifelong Learning), 15*, 70-77.

Fernandez, M., & Siddiqui, A. M. (2017). Markers' criteria in assessing english essays: An exploratory study of higher secondary school certificate (HSCC) in Punjab province of Pakistan. *Language Testing in Asia, 7*(6).

Frederiksen, N. (1984). The real test bias: Influence of testing on teaching and learning. *American Psychologist, 39,* 193–202.

Gronlund, N. E. (1998). *Assessment of student achivement*. USA: By Allyn & Bacon Viacom Company.

Goodrich, H. (1997). Understanding rubrics. *Educational Leadership, 54*(4), 14-17.

Güler, N. (2008). *Klasik test kuramı genellenebilirlik kuramı ve rasch modeli üzerine bir araştırma.* Doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Güler, N. (2014). Analysis of open-ended statistics questions with many facet rasch model. *Eurasian Journal of Educational Research, 55,* 73-90.

Güler, N., & Gelbal, S. (2010). Açık uçlu matematik sorularının güvenirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri / Educational Sciences: Theory & Practice, 10*(2), 989-1019.

Gündoğan, N. (2017). *Anadolu üniversitesi açıköğretim fakültesi açık uçlu soru dönemi başlıyor.* http://egazete.anadolu.edu.tr/kampus/34983/anadolu-universitesi acikogretim -sisteminde-acik-uclu-soru-donemi-basliyor sitesinden erişilmiştir.

Hadenfeldt, J. C., Bernholt, S., Liu, X., Neumann K., & Parchmann I. (2013). Using ordered multiple-choice items to assess students' understanding of the structure and composition of matter - Leibniz Institute for science and mathematics education, *Journal of chemical education, 90*(12), 1602–1608.

Haiyang, S. (2010). An application of classical test theory and many facet rasch measurement in analyzing the reliability of an english test for non-english major graduates. *Chinese Journal of Applied Linguistics, 33*(2), 87-102.

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

Hambleton, R., Swaminathan, H., & Rogers, J. (1991). Concepts, models and features. In D. Foster (Eds.), *Fundamentals of item response theory* (pp. 7-31). California: Sage.

Hancock, C. L. (1995). Enhancing mathematics learning with open-ended questions. *The Mathematics Teacher, 88*(6), 496-499.

Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). *A simulation study of the effect of rater designs on ability estimation*. Princeton, NJ: Educational Testing Service.

İlhan, M. (2016). A Comparison of the results of many-facet rasch analyses based on crossed and judge pair designs. *Educational Sciences: Theory & Practice, 16*(2), 579-601.

İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *31*(2), 346-368.

Jönsson, A., & Panadero, E. (2016). The use and design of rubrics to support assessment for learning. In D. Carless, S. Bridges, C. Chan & R. Glofcheski (Eds.), *Scaling up assessment for learning in higher education* (pp. 99-111). Series Editor, Claire Wyatt-Smith.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-132.

Kan, A. (2013). Ölçme araçlarında bulunması gereken nitelikler. H. Atılgan (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (s. 23-80). Ankara: Anı.

Kaptan, F. (1994). Rasch modeli madde parametrelerini kullanarak en yüksek olabilirlik yöntemiyle yeteneğin kestirilmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *10*, 95-97.

Karakaya Özyer, K. (2018). Açıköğretim sistemindeki açık uçlu soruların çok yüzeyli rasch modeli analizi yöntemiyle puanlanması. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi, 4*(1), 61-77.

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

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis, 21*(4), 345–363.

Khitab, U., Zaman, A., & Ghaffar, A. (2014). Secondary school teachers' ability to understand scoring rubric and practices. *Journal of Applied Environmental and Biological Sciences, 4*(8), 153-159.

Kline, P. (1994). *An Easy Guide To Factor Analysis*. New York: Routledge.

Köse, İ. A., Usta, H. G., & Yandı, A. (2016). Sunum yapma becerilerinin çok yüzeyli rasch analizi ile değerlendirilmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 16*(4) , 1853-1864.

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

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking. *Educational Measurement: Issues and Practice*, *23*(3), 6–14.

Lee, H. S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge ıntegration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24*(2), 115–136.

Linacre J. M., & Wright B. D. (2002). Understanding rasch measurement: construction of measures from many-facet data. *Journal of Applied Measurement*, *3*(4), 486-512.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (1989). *Many-faceted rasch measurement.* Chicago: MESA.

Linacre, J. M. (1993). *Generalizability theory and many facet rasch measurement*. Paper presented at the Annual Meeting Of The American Educational Research Association, Atlanta.

Linacre, J. M. (1994). Constructing measurement with a many-facet rasch model. In M. Wilson (Eds.), *Objective measurement: Theory in practice* (pp. 129-144). Norwood, NJ: Abex.

Linacre, J. M. (2009). Local independence and residual covariance: A study of olympic figure skating ratings. *Journal of Applied Measurement, 10*(2),1-13.

Linacre, J. M. (2002). What do infit and outfit MNSQ standardized mean. *Rasch Measurement Transaction, 16,* 878.

Linacre, J. M. (2010). *A user's guide to FACETS Rasch-model computer programs.* http://www.winsteps.com/manuals.htm sayfasından erişilmiştir.

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

MacMillan, P. D. (2000). Classical, generalizability and multifaceted rasch detection of interrater variability in large sparse data sets. *The Journal of Experimental Education, 68*(2), 167-190.

Marcus, J. (1995). *Data on the impact of alternative assessment on students*. Unpublished manuscript. The Education Cooperative Wellesley, Mass.

McEwen, M. R. (2018). *The effects of ıncomplete rating designs on results from many-facets-rasch model analyses.* Doktora Tezi, Brigham Young University, Utah.

McDonald R. P. (1999). *Test theory: A unified treatment*. NJ: LEA.

McTighe, J. (2016). *Developing and using quality rubrics to evaluate and ımprove student performance.* https://www.vassonline.org/images/vass/resource-files/pba-resources/Quality%20Rubrics%20-%20Second%20AM%20Session.pdf sitesinden erişilmiştir.

Miller, R. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching.* New Jersey: Pearson Education.

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

Moskal, B. M. (2000). Scoring rubrics: What, when and how. *Practical assessment, Research & Evaluation, 7*(3).

Myford, C. M., & Wolfe, E.W. (2000). *Monitoring sources of variability within the test of spoken english assessment system.* NJ: Educational testing service.

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Myford, C. M., Marr, D. B., & Linacre, J. M. (1996). *Reader calibration and its potential role in equating for the test of written english*. NJ: Educational testing service.

Nunnally, J. C., & Bernstein, I. H. (1994): *Psychometric theory.* McGraw-Hill: New York.

Ölçme, Seçme ve Yerleştirme Merkezi (2015). *Yazılı sınav (açık uçlu sorularla sınav)*. http://www.osym.gov.tr/TR,721/yazili-sinav-acik-uclu-sorularla-sinav04022015.html sayfasından erişilmiştir.

Ölçme, Seçme ve Yerleştirme Merkezi (2017). *Açık uçlu sorular hakkında bilgilendirme ve açık uçlu soru örnekleri*. http: //www.osym.gov.tr/TR,12909/2017 -lisans-yerlestirme-sinavlari-2017-lys-acikuclu-sorular-hakkinda-bilgilendirme-ve-acik-uclu-soru-ornekleri-05012017.html sayfasından erişilmiştir.

Özçelik, D. A. (1998). *Ölçme ve değerlendirme*. Ankara: ÖSYM.

Özkan, Y. Ö. (2012). *Öğrenci başarılarının belirlenmesi sınavından (Öbbs) klasik test kuramı, tek boyutlu ve çok boyutlu madde tepki kuramı modelleri ile kestirilen başarı puanlarının karşılaştırılması.* Doktora Tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Rahman, A. A., Ahmad, J., Yasin, R. M., & Hanafi, N. M. (2017). Investigating central tendency in competency assessment of design electronic circuit: analysis using many facet rasch measurement (MFRM). *International Journal of Information and Education Technology, 7*(7).

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Denmark: Danish Institute for Educational Research; Chicago: MESA.

Reardon, V. (2017). Alternative assessment: Growth, development and future directions. In R. Al-Mahrooqi, C. Coombe, F. Al-Maamari & V. Thakur (Eds.), *Revisiting EFL assessment: Critical perspectives* (pp. 191–208). New York, NY: Springer.

Rodriguez, M. C. (1993). Choosing an item format. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 213–231). Hillsdale, NJ: Lawrence Erlbaum.

Rosenow, C. (2014). Collaborative design: Building task-specific rubrics in the honors classroom. *Journal of the National Collegiate Honors Council Online Archive, 441*.

Russell, M. K., & Airasian, P. W. (2012). *Classroom assessment: Concepts and applications.* New York, NY: McGraw-Hill.

Salmaner, R. (2015). *Yazma becerilerinin değerlendirmesinde öz, akran ve öğretmen puanlarının çok yüzeyli Rasch ölçme modeliyle incelenmesi*. Yüksek Lisans Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Sarıtaş Akyol, S., & Karakaya, İ. (2021). Çok Yönlü Rasch Ölçüm Modeli ile Problem Çözme Becerilerinin Değerlendirilmesinde Öğrenci ve Öğretmen Puanları Arasındaki Tutarlılığın İncelenmesi. *Avrasya Eğitim Araştırmaları Dergisi*, *91*, 281-300.

Schumacker, R. E. (1999). Many-facet Rasch analysis with crossed, nested and mixed designs. *Journal of Outcome Measurement*, *3*(4), 323-338.

Sharma, S.(1996). *Applied multivariate techniques*. New York: John Wiley and Sons.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14.

Shout, P. E., & Lane, S. P. (2012). Reliability. In H. Cooper (Eds.), *Handbook of researclı methods in psychology: foundations, planning, measures, aıud psychometrics* (pp. 643-660). USA: American Psychological Association.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9.*

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Eds.), *Handbook of Experimental Psychology* (pp. 22). New York: 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.

Tabachnick, G. B., & Fidell, L. S. (2013). *Using multivariate statistics*. London: Pearson.

Taghvafard, M. (t.y). *Inter-rater and intra-rater reliability*. https://www.winsteps.com/facetman/inter-rater-reliability.htm adresinden 04.04.2021 tarihinde erişilmiştir.

Taherbhai, H. M., & Young, M. J. (2001). The impact of rater effects on weighted composite scores under nested and spiraled scoring designs, using the multifaceted Rasch model. *Journal of Outcome Measurement, 5*(1), 819-838.

Tan, Ş. (2008). *Öğretimde ölçme ve değerlendirme*. Ankara: Pegem.

Tekin, H. (2014). *Eğitimde ölçme ve değerlendirme.* Ankara: Yargı.

Temizkan, M., & Sallabaş, M. E. (2011). Okuduğunu anlama becerisinin değerlendirilmesinde çoktan seçmeli testlerle açık uçlu yazılı yoklamaların karşılaştırılması. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, *30*, 207–220.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*.

Turgut, M. F., & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem.

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

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103–118.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. *Journal of Educational Measurement, 17*, 11-29.

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership, 49*(8), 26-34.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development, 45,* 197–210.

Wind, S. A., & Jones, E. (2019). The effects of incomplete rating designs in combination with rater effects. *Journal of Educational Measurement, 56*(1), 76–100.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA.

Yaman, S. (2016). Çoktan seçmeli madde tipleri ve fen eğitiminde kullanılan örnekleri. *Gazi Eğitim Bilimleri Dergisi, 2*(2), 151-170.

Yurdugül, H. (2005). Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması. *XIV. Ulusal Eğitim Bilimleri Kongresi bildirileri*, 771-774.

Yurdugül, H., & Bayrak, F. (2012). Ölçek geliştirme çalışmalarında kapsam geçerlik ölçüleri: kapsam geçerlik indeksi ve kappa istatistiğinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 2*, 264-271.

EKLER

EK 1. Açık Uçlu Maddelere İlişkin Belirtke Tablosu

KAZANIMLAR	BİLİŞSEL ALAN BASAMAKLARI						
	BİLGİ	**KAVRAMA**	**UYGULAMA**	**ANALİZ**	**SENTEZ**	**DEĞERLENDİRME**	**TOPLAM**
İki doğal sayının en büyük ortak bölenini (EBOB) ve en küçük ortak katını (EKOK) hesaplar, ilgili problemleri çözer.			X		X		**2**
Tam kare doğal sayılarla bu sayıların karekökleri arasındaki ilişkiyi belirler.				X			**1**
Kareköklü ifadelerde toplama ve çıkarma işlemlerini yapar.				X			**1**
Tam kare olmayan sayıların karekök değerlerinin hangi iki doğal sayı arasında olduğunu belirler.				X			**1**
Doğal sayılarla dört işlem yapmayı gerektiren problemleri çözer ve kurar.			X	X			**2**
Çok büyük ve çok küçük sayıları bilimsel gösterimle ifade eder ve karşılaştırır.				X			**1**
Tam sayıların, tam sayı kuvvetlerini hesaplar.						X	**1**
Ondalık gösterimlerle dört işlem yapmayı gerektiren problemleri çözer.				X X			**2**
						TOPLAM MADDE SAYISI	**11**

EK 2. Lawshe(1975)'nin Kapsam Geçerlik Oranı için Kritik Değerler Tablosu

	Tek Yönlü Test için Anlamlılık Düzeyi					
	.1	.05	.025	.01	.005	.001
	Çift Yönlü Test için Anlamlılık Düzeyi					
N	.2	.1	.05	.02	.01	.002
5	.573	.736	.877	.99	.99	.99
6	.523	.672	.800	.950	.99	.99
7	.485	.622	.741	.879	.974	.99
8	.453	.582	.693	.822	.911	.99
9	.427	.548	.653	.775	.859	.99
10	.405	.520	.620	.736	.815	.977
11	.387	.496	.591	.701	.777	.932
12	.370	.475	.566	.671	.744	.892
13	.356	.456	.544	.645	.714	.857
14	.343	.440	.524	.622	.688	.826
15	.331	.425	.506	.601	.665	.798
16	.321	.411	.490	.582	.644	.773
17	.311	.399	.475	.564	.625	.750
18	.302	.388	.462	.548	.607	.729
19	.294	.377	.450	.534	.591	.709
20	.287	.368	.438	.520	.576	.691
21	.280	.359	.428	.508	.562	.675
22	.273	.351	.418	.496	.549	.659
23	.267	.343	.409	.485	.537	.645
24	.262	.336	.400	.475	.526	.631
25	.256	.329	.392	.465	.515	.618
26	.251	.323	.384	.456	.505	.606
27	.247	.317	.377	.448	.496	.595
28	.242	.311	.370	.440	.487	.584
29	.238	.305	.364	.432	.478	.574
30	.234	.300	.358	.425	.470	.564
31	.230	.295	.352	.418	.463	.555
32	.227	.291	.346	.411	.455	.546
33	.223	.286	.341	.405	.448	.538
34	.220	.282	.336	.399	.442	.530
35	.217	.278	.331	.393	.435	.522
36	.214	.274	.327	.388	.429	.515
37	.211	.270	.322	.382	.423	.508
38	.208	.267	.318	.377	.418	.501
39	.205	.263	.314	.372	.412	.495
40	.203	.260	.310	.368	.407	.489

Not: KGO için kritik değerler 1,00 sınır değerinden büyük veya buna eşit değerler 0,99 olarak ayarlanmıştır. "Recalculation of the Critical Values for Lawshe's Content Validity Ratio" Wilson, F. R., Pan, W. & Schumsky, D. A., 2012, *Measurement and Evaluation in Counseling and Development, 45*(3), 197–210 kaynağından uyarlanmıştır.

EK 3. Matematik Başarı Testi Yönergesi

8. SINIFLAR İÇİN

MATEMATİK DENEME SINAVI

(BİRİNCİ BÖLÜM)

ADI: .....

SOYADI: .....

SINIF: ..... NO: .....

OKULU: .....

AÇIKLAMALAR:

1. Deneme sınavı iki bölümden oluşmaktadır. Birinci bölümde Matematik dersine ait 6 tane açık uçlu soru; ikinci bölümde ise 5 tane açık uçlu soru yer almaktadır.
2. Birinci bölümün bulunduğu bu kitapçık için sınav süresi 40 dakikadır. Sınavı zamanında bitirebilmek için doğru ve hızlı okuma kurallarına dikkat ediniz.
3. Soruları cevaplamaya istediğiniz sorudan başlayabilirsiniz.
4. Sorular kitapçığa rahatça çözüm yapabilmeniz için yeterli alanlar oluşturacak şekilde yerleştirilmiştir. Soruların çözümlerini belirlenen alanlara dışarı taşmayacak şekilde yazınız.
5. Soruları cevaplarken yaptığınız çözümün her bir adımını açıklamalar yaparak açık bir şekilde yazınız.
6. Açık uçlu soruların tek bir çözüm yolu olmadığından yazacağınız her cevap önemlidir. Bu nedenle soruları gerçek bir sınavı cevaplandırıyormuş gibi ciddiyetle cevaplayınız. Doğru sonuca ulaşamasanız dahi düşündüğünüz çözüm yolunu ilgili alana yazmaktan çekinmeyiniz.
7. Bu deneme sınavı not ile değerlendirilmeyecektir; yaptığınız çözümler sadece bilimsel bir araştırmaya katkı sağlamak için kullanılacaktır.

EK 4. Matematik Başarı Testi

Derin ile Serdar farklı büyüklüklerde ve dikdörtgen şeklindeki odalarının zeminini birbirine eş, kenar uzunluğu santimetre (cm) cinsinden tam sayı olan kare fayanslarla kaplamak istiyor. Dikdörtgen şeklindeki odalarının en ve boy uzunlukları aşağıdaki tablodadır.

Uzunluk / Kişiler	Odanın en uzunluğu (cm)	Odanın boy uzunluğu (cm)
Derin	**280**	**320**
Serdar	**420**	**450**

SORU 1: Yukarıdaki tablo değerlerine göre hangi bireyin oda zemininin kaplanmasında kullanılan fayansların boyutu daha küçüktür? Nedeniyle belirtiniz.

Cevap:

SORU 2: Derin ile Serdar'ın oda zeminleri eşit sayıda fayans ile kaplansaydı Serdar'ın odasının boy uzunluğunda ve kullanılan fayansın bir kenar uzunluğunda nasıl bir değişikliğin yapılması gerekirdi? Nedeniyle açıklayınız.

Cevap:

Bir dart oyununda her bir oyuncu bir turda 3 dart atar. Puanlar dartın isabet ettiği numaralardan elde edilir. Oyuncunun attığı dart puanları iç ya da dış bölgeye isabetlerine göre aşağıda belirtilen katsayılarla çarpılarak hesaplanır. En yüksek puana ulaşan kişi oyunu kazanır.



- **İç taraftaki halka içinde kalan kısımlar, atılan sayının 3 katı puan kazandırır.**
- **Dış taraftaki halka içinde kalan kısımlar, ise atılan sayının 2 katı puan kazandırır.**

(3. ve 4. soruları yukarıda verilen bilgiye göre cevaplandırınız.)

SORU 3: Nehir, Arda, Cihan, Selin ve Mert dart oyunu oynamak için bir araya gelmişlerdir. Oyuncuların 6 atış sonucunda isabet ettirdikleri dart puanları kareköklü sayı olarak aşağıdaki tabloda verilmiştir.

Bölgeler / Oyuncular	İç bölge	Dış bölge
Nehir	$\sqrt{25}$, $\sqrt{144}$, $\sqrt{256}$	$\sqrt{324}$, $\sqrt{16}$, $\sqrt{289}$
Selin	$\sqrt{196}$, $\sqrt{121}$, $\sqrt{381}$	$\sqrt{381}$, $\sqrt{324}$, $\sqrt{49}$
Arda	$\sqrt{381}$, $\sqrt{225}$, $\sqrt{144}$	$\sqrt{121}$, $\sqrt{4}$, $\sqrt{64}$
Cihan	$\sqrt{400}$, $\sqrt{100}$, $\sqrt{225}$	$\sqrt{144}$, $\sqrt{256}$, $\sqrt{169}$
Mert	$\sqrt{289}$, $\sqrt{144}$, $\sqrt{256}$	$\sqrt{400}$, $\sqrt{121}$, $\sqrt{144}$

Atışlar sonucunda elde edilen puanları toplayan hakem $\sqrt{196}$ ile $\sqrt{169}$ sayılarının değerlerini karıştırmıştır. Yapılan işlemler sonucunda birinciliğe en yakın isimler sırasıyla Mert, Selin, Cihan, Arda, Nehir olarak belirlenmiştir.

- Hakem $\sqrt{196}$ ile $\sqrt{169}$ değerlerini karıştırmasaydı sıralamada nasıl bir değişiklik olurdu? Nedeniyle belirtiniz.

Cevap:

Nehir ve Cihan puanlarını aşağıdaki gibi hesaplamıştır.



İç bölgeden aldığım puan

$\sqrt{25}+\sqrt{144}+\sqrt{256}=\sqrt{425}$

3. $\sqrt{425}=\sqrt{1275}$

Dış bölgeden aldığım puan

$\sqrt{324}+\sqrt{16}+\sqrt{289}=\sqrt{629}$

2. $\sqrt{629}=\sqrt{1258}$

Toplam puanım

$\sqrt{1275}+\sqrt{1258}=\sqrt{2533}=50$

olur.

İç bölgeden aldığım puan

$\sqrt{400}+\sqrt{100}+\sqrt{225}=20+10+15=45$

$3.45 = 135$

Dış bölgeden aldığım puan

$\sqrt{144}+\sqrt{256}+\sqrt{169}=12+16+13=41$

$2.41 = 82$

Toplam puanım

$135 + 82 = 217$ olur.

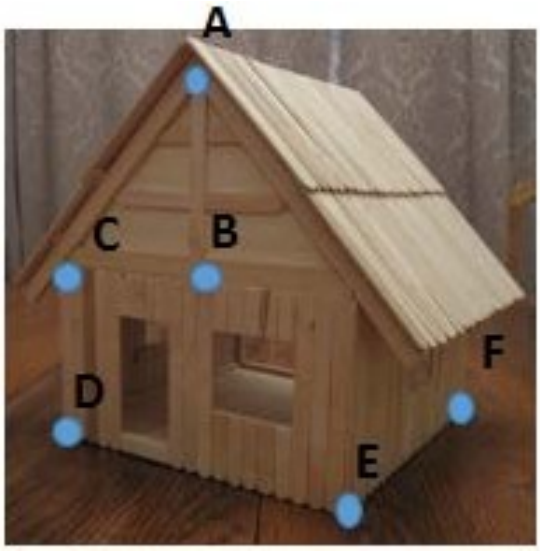


SORU 4: Atışlar sonucunda elde edilen puanları hesaplamada Nehir ve Cihan'ın kullandıkları yöntemlerin doğruluğunu tartışınız.

Cevap:

SORU 5

Önder farklı uzunluklardaki çıtaları birleştirerek şekildeki gibi bir ev maketi oluşturmuştur.



Çıtalar maketin ön yüzeyinde ikizkenar üçgen ve kare, diğer yüzeylerinde ise dikdörtgen oluşturacak şekilde birleştirilmiştir.

Maket üzerinde gösterilen noktalar arası uzaklıklar şöyledir:

IABI= 3 cm, ICDI=IDEI=8 cm ve IEFI= 13 cm dir.

Oluşturduğu ev maketini arkadaşlarına göstermek isteyen Önder maketini dikdörtgenler prizması şeklindeki kutuya <u>üstü tamamen kapanacak şekilde</u> yerleştirmek istemiştir.

- Önder ev maketini tabanının kısa kenar uzunluğu $\sqrt{48}$ cm, uzun kenar uzunluğu $\sqrt{148}$ cm ve yüksekliği $\sqrt{122}$ cm olan kutu içerisine yerleştirerek taşıyabilir mi? Nedeniyle belirtiniz.

Cevap:

SORU 6

USB bellek küçük, taşınabilir bir bilgisayar depolama aracıdır. İrfan'ın müzik ve fotoğraf yüklü bir USB belleği vardır. Bu belleğin kapasitesi 1 GB (1000 MB)'tır. Aşağıdaki grafik USB belleğin şu anki doluluk durumunu göstermektedir.



İrfan 350 MB'lık bir fotoğraf albümünü USB belleğine aktarmak istemektedir, fakat USB belleğinde yeterince boş alan bulunmamaktadır. İrfan, bellekteki fotoğrafları silmek istemezken en fazla iki adet müzik albümünü silmeyi tercih etmektedir. İrfan'ın USB belleğine yüklenmiş olan müzik albümlerinin büyüklüğü aşağıda gösterilmektedir:

Albüm	Büyüklük
Albüm 1	100 MB
Albüm 2	75 MB
Albüm 3	80 MB
Albüm 4	55 MB
Albüm 5	60 MB
Albüm 6	80 MB
Albüm 7	75 MB
Albüm 8	125 MB

İrfan'ın fotoğraf albümünü eklemek için gereken boş alanı en fazla iki müzik albümünü silerek oluşturması mümkün müdür? "Evet" ya da "Hayır" seçeneklerinden birini yuvarlak içine alınız ve yanıtınızı desteklemek için yaptığınız hesaplamaları gösteriniz.

Evet/ Hayır

Cevap:

SORU 7

TATİL EVİ

Ceren aşağıdaki satılık tatil evini internette bulmuştur. Tatilcilere kiralamak amacıyla bu tatil evini satın almayı düşünmektedir.

Oda sayısı:	1 x oturma ve yemek odası 1 x yatak odası 1 x banyo
Büyüklük:	60 metrekare (m²)
Otopark:	Evet
Şehir merkezine varış süresi:	10 dakika
Sahile uzaklık:	Düz bir yol üzerinden 350 metre (m)
Son 10 yıl içerisinde tatilciler tarafından kullanım ortalaması:	Yılda 315 gün

Fiyat: 200 000 zed



Tatil evinin fiyatını belirlemek için Ceren bir uzmana danışmıştır. Tatil evinin değerini belirlemek amacıyla uzman aşağıda verilen ölçütleri kullanmaktadır.

m² başına fiyat	**Taban fiyat:**	m² başına 2500 zed			
Ek değer ölçütleri	**Şehir merkezine varış süresi**	15 dakikadan fazla: +0 zed	5 ila 15 dakika arası: +10 000 zed	5 dakikadan az: +20 000 zed	
	Sahile uzaklık (düz bir yol üzerinden):	2 km'den fazla: +0 zed	1 ila 2 km arası: +5000 zed	0,5 ila 1 km arası: +10 000 zed	0,5 km'den az: +15 000 zed
	Otopark:	Hayır: +0 zed	Evet: +35 000 zed		

Uzmanın belirlediği değer ilanda verilen satış fiyatından fazla ise, evi satın almak isteyen Ceren için bu fiyat "çok iyi" olarak kabul edilmektedir.

- Uzmanın kullandığı ölçütlere bağlı olarak, Ceren için teklif edilen satış fiyatının "çok iyi" olduğunu gösteriniz.

Cevap:

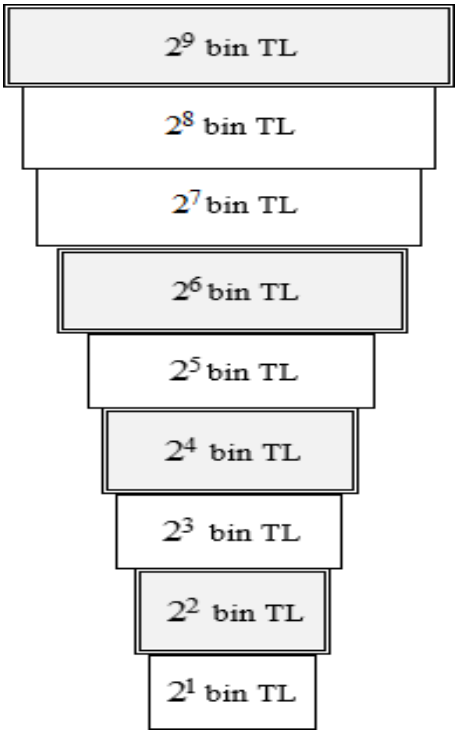
SORU 8

Bir uzman "Fen ve Teknoloji" ders kitabı "Atomun Yapısı" konu anlatımında bulunan "Protonun kütlesi 1673. 10^{-33} g'dır." **cümlesindeki sayı yazımının bilimsel olmadığını ve buyazımın düzeltilmesi gerektiğini belirtmiştir.**

- Sizce verilen cümledeki sayının doğru yazımı nasıl olmalıdır? Gerekçeleriyle açıklayınız.

Cevap:

Mehmet Bey 450 bin TL değerindeki bir apartman dairesi satın almak istemektedir. Bu nedenle her soruda bir önceki kazandığı paranın iki katı kadar para kazanabileceği 9 sorudan oluşan bilgi yarışmasına katılmak istemiştir.



- Koyu renk ile belirtilen baraj soruları değerlerine ulaşan yarışmacılar bir sonraki soruya yanlış cevap verseler dahi bu miktarı kazanarak yarışmadan ayrılmaktadır.
- Yarışmacı istediği soruda soru için belirtilen miktarı alarak yarışmadan ayrılabilmektedir.

SORU 9

Mehmet Bey, koyu renk ile belirtilen ilk üç baraj sorusuna doğru yanıt vermiş ancak son soruyu göremeden yarışmadan ayrılmıştır.

Bu bilgilere dayanarak Mehmet Bey'in son soruyu göremeden ayrılmasına neden olabilecek ihtimalleri belirtiniz ve bu ihtimallerden apartman dairesini satın alması için gerekli olan miktarı elde etmesine en yakın olanını hesaplamalarla açıklayınız.

Cevap:

Derya ve Seda resmi kurum belgelerinde kullanmak üzere vesikalık foto çektirmek istemişlerdir.



Derya

10 taneden fazla vesikalık fotoğrafa ihtiyacım var ve cebimde 25 TL bulunuyor.

Belirli sayıda fotoğrafa ihtiyacım yok ancak fotoğrafın tanesini en düşük fiyata çeken fotoğrafçıda vesikalık fotoğraf çektirmek istiyorum.



Seda

Bu nedenle fotoğrafçılar arasında fiyat araştırması yapmışlardır. Yaptıkları araştırma sonucunda aşağıdaki tabloda belirtilen fiyatları elde etmişlerdir.

	A FOTOĞRAFÇILIK	**B FOTOĞRAFÇILIK**	**C FOTOĞRAFÇILIK**
4 ADET	Tanesi 1,10 TL	5 TL	6 TL + 2 vesikalık bedava
8 ADET	14 TL + 4 vesikalık bedava	10 TL	Tanesi 1,10 TL
16 ADET	Tanesi 1,15 TL	18 TL + 2 vesikalık bedava	20 TL
20 ADET	25 TL	Tanesi 1,20 TL	26 TL + 4 vesikalık bedava

SORU 10

Tablodaki değerlere bakıldığında Derya ve Seda'nın aynı fotoğrafçıda vesikalık fotoğraf çektirmek isterlerse hangi fotoğrafçıya birlikte gidebileceklerini vesikalık fotoğraf adetini ve fotoğrafcıya ödeyecekleri miktarı hesaplayarak açıklayınız.

A Fotoğrafçılık / B Fotoğrafçılık / C Fotoğrafçılık (İşaretleyiniz.)

Cevap:

SORU 11

Derya ve Seda ayrı ayrı fotoğraf çektirselerdi hangi fotoğrafçıların Derya ve Seda için uygun fiyatlı olacağını vesikalık fotoğraf adedini belirterek sebebiyle açıklayınız.

DERYA: A Fotoğrafçılık / B Fotoğrafçılık / C Fotoğrafçılık (İşaretleyiniz.)

SEDA: A Fotoğrafçılık / B Fotoğrafçılık / C Fotoğrafçılık (İşaretleyiniz.)

Cevap:

EK 5. Uzman Değerlendirme Formu

AÇIK UÇLU MADDELER İÇİN UZMAN DEĞERLENDİRME FORMU

Değerli Uzman;

"Çok Yüzeyli Rasch Modeli Puanlama Desenlerine Göre Açık Uçlu Maddelerin Puanlayıcılar Arası Güvenirliklerinin Karşılaştırılması" adlı yüksek lisans tez çalışmam kapsamında 8. Sınıf öğrencilerine uygulanmak üzere matematik dersine yönelik hazırlanan açık uçlu maddeler ile bu maddelere ait özel bütünsel dereceli puanlama anahtarlarının uzman değerlendirmeleri çerçevesinde geliştirilmesi gerekmektedir. Hazırlanan açık uçlu maddeler ile puanlayıcıların maddeleri puanlamada kullanacağı özel bütünsel dereceli puanlama anahtarlarının değerlendirilmesi amacıyla uzman değerlendirme formu oluşturulmuştur. Değerli uzmanların katkılarıyla gerekli değişiklikler yapılarak ölçme araçlarının çalışmalarda kullanıma uygun hale gelmesi sağlanacaktır.

Açık uçlu maddeler ile özel bütünsel dereceli puanlama anahtarlarının geliştirilme sürecini formdaki ölçütler doğrultusunda değerlendirerek hazırlanan ölçme araçları ölçüte *uygun ise* (3), ölçme aracının ölçüte göre *düzeltilmesi gerekiyorsa* (2) ve *tamamen düzeltilmesi gerekiyorsa*
(1) işaretleyerek görüşlerinizi belirtiniz. Yapacağınız öneriler ve değerlendirmeler için şimdiden teşekkür ederiz.

Doç. Dr. İsmail KARAKAYA
Gazi Üniversitesi
Gazi Eğitim Fakültesi
Eğitim Bilimleri Bölümü
Eğitimde Ölçme ve Değerlendirme ABD

Zeynep GÖLCÜR
Gazi Üniversitesi
Gazi Eğitim Fakültesi
Eğitim Bilimleri Bölümü
Eğitimde Ölçme ve Değerlendirme ABD

1. Adı ve Soyadı:

2. Unvan:

3. Branş:

4. Eğitim düzeyi: Lisans( ) Yüksek lisans ( ) Doktora ( )

AÇIK UÇLU MADDE İNCELEME FORMU			
MADDE NO:			
ÖLÇÜTLER	**DERECELER**		
	UYGUN (3)	**DÜZELTİLMELİ (2)**	**TAMAMEN DÜZELTİLMELİ (1)**
Madde, belirtilen kazanımı ölçmek için uygundur. **Öneri:**			
Madde, öğrencilerin bilgilerini kullanmalarına olanak sağlamaktadır. **Öneri:**			
Madde öğrencilerin sınıf, yaş ve bilgi düzeyine uygun yazılmıştır. **Öneri:**			
Madde kökü ve yönergelerin yazımında açık ve anlaşılır bir dil kullanılmıştır. **Öneri:**			
Madde, günlük yaşama dayanan bir durumu yansıtmaktadır. **Öneri:**			
Madde, öğrencileri özgün cevaplar yazmaya ve farklı durumlar üzerinde düşünmeye teşvik etmektedir. **Öneri:**			

Üst bilişsel (analiz, sentez, değerlendirme) davranışları ölçmek için uygun bir maddedir. **Öneri:**			
Maddeyi cevaplandırmaları için öğrencilere verilen süre, öğrencilerin düzeylerine uygun ve yeterlidir. **Öneri:**			

DERECELİ PUANLAMA ANAHTARI İNCELEME FORMU			
MADDE NO:			
ÖLÇÜTLER	**DERECELER**		
	UYGUN (3)	**DÜZELTİLMELİ (2)**	**TAMAMEN DÜZELTİLMELİ (1)**
Dereceli puanlama anahtarının türü madde içeriğine uygundur. **Öneri:**			
Tam doğru cevaplar, öğrenci cevaplarından bağımsız hazırlanmıştır. **Öneri:**			
Kısmen doğru cevaplar kademeli bir şekilde sıralanmıştır. **Öneri:**			
Cevapların puanlama değerleri ve ölçüt tanımları tam doğru cevaptan uzaklıklarına göre sıralanmıştır. **Öneri:**			
Puanlama anahtarı, öğrenci cevaplarının tümünü kapsayacak şekilde hazırlanmıştır. **Öneri:**			
Puanlama anahtarı açık, anlaşılır ifadeler kullanılarak öğrencilerin ve puanlayıcıların anlayabilecekleri şekilde oluşturulmuştur. **Öneri:**			
Puanlama anahtarı; öğrencilere, puanlayıcılara geri bildirim sağlayacak biçimde doğru ve ayrıntılı hazırlanmıştır. **Öneri:**			

EK 6. Özel Bütünsel Dereceli Puanlama Anahtarları

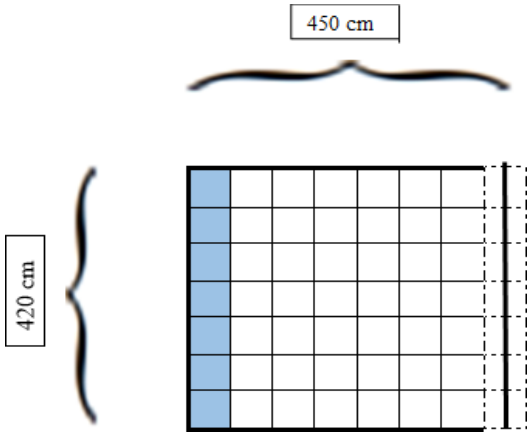
SORU 1

Başarı Puanı	**ÖLÇÜTLER**
4	Dikdörtgen şeklindeki oda zemini birbirine eş ve kenar uzunluğu cm cinsinden tam sayı olduğundan 280 ile 320'nin; 420 ile 450 sayılarının EBOB değeri bulunmalıdır. Fayansların bir kenar uzunluğu; asal çarpan algoritması, çarpan ağacı ya da sayıların bölenlerini bulup içerisinden ortak olanların en büyüğünü seçme yöntemlerinden biri kullanılarak bulunan $(280,320)_{ebob}=40$ cm ve $(420,\ 450)_{ebob}=30$ cm değerlerini karşılaştırarak Serdar'ın oda zemininde kullanılan fayansların boyutunun daha küçük olduğunu bulma ve "Serdar'ın oda zemininde kullanılan fayansların boyutu daha küçüktür."şeklinde açıklama yazma.
3	• İşlemlerin yanına yeterli açıklama yapılmadan sadece doğru işlemler yapılarak Serdar'ın oda zemininde kullanılan fayansların boyutunun daha küçük olduğunun bulunması. • İşlem yaparak Serdar ve Derin'in oda zemininde kullandıkları fayans boyutlarını karşılaştırmak için gerekli olan EBOB değerlerinin doğru bulunup açıklamanın yanlış yazılması.
2	• Verilen oda uzunluklarının EBOB değerinin hesaplanması gerektiğini belirtip temel işlemlerde hatalar yapma nedeniyle doğru sonuca ulaşamama ya da ulaşma. • Uygun çözümle başlayıp doğru işlemler yapıp küçük hatalardan dolayı doğru EBOB değerlerini bulamayarak doğru cevaba ulaşamama.
1	• Verilen oda uzunluklarının EBOB değerlerini hesaplamak için herhangi bir işlem yapılmadan sadece "Serdar'ın oda zemininde kullanılan fayans boyutunun daha küçük" olduğunun yazılması.

0	Boş ya da tamamen yanlış cevap verilmesi.

SORU 2

Başarı Puanı	ÖLÇÜTLER
4	**I.ÇÖZÜM YOLU:** Sadece Serdar'ın odasının boy uzunluğunda ve fayans boyutunda değişiklik yapılacağından Derin'in oda zemininin alanı, bir fayansın alanine $(280, 320)_{ebob} = 40$ cm fayansın bir kenar uzunluğu bölünerek kullanılan fayans sayısı bulunmalıdır. $\frac{280 \cdot 320}{40.40} = 56$ tane fayans kullanılmalıdır. Oda zeminlerini kaplarken eşit sayıda fayans kullanımı istendiğinden 56 tane fayans kullanılabilmesi için 420.480 işlemi 56 sayısına bölündüğünde $\frac{420 \cdot 480}{56} = 60 \cdot 60$ bulunarak (EBOB değeri) kullanılan fayansların bir kenar uzunluğunun 60 cm olarak yazılması. $(60 \cdot 7, 60 \cdot 8)_{ebob} = 60$ cm olduğunda odanın en ve boy uzunluklarının 420 cm ve 480 cm olması gerektiğinden Serdar'ın oda zemininin boy uzunluğunun (450 cm nin) **30 cm daha artırılarak** 480 cm olması gerektiğinin yazılması. **II.ÇÖZÜM YOLU:** Sadece Serdar'ın odasının boy uzunluğunda ve fayans boyutunda değişiklik yapılacağından Derin'in oda zemininin alanı, bir fayansın alanına $(280, 320)_{ebob} = 40$ cm fayansın bir kenar uzunluğu bölünerek kullanılan fayans sayısı bulunmalıdır. $\frac{280 \cdot 320}{40.40} = 56$ tane fayans kullanılmalıdır. Oda zeminlerini kaplarken eşit sayıda fayans kullanımı istendiğinden 56 tane fayans kullanılabilmesi için Serdar'ın odasının EBOB değeri olan 30 cm ve katları (30, 60, 90…)fayans boyutu olarak düşünülmelidir. Soruda fayans boyutunun değiştirilmesi istendiğinden **30 cm lik fayanslar kullanılamaz.** Bir sonraki katı olan 60 cm lik fayanslar kullanılmalıdır.

4	 Eşit sayıda fayans kullanılması istendiğinden 60 x 60 cm lik fayanslardan 56 tane kullanılmalıdır. Serdar'ın odasının en uzunluğu olan 420 cm de değişiklik yapılmayacaktır. Bu nedenle 60 x 60 cm lik fayanslardan odanın eni olan 420 cm lik uzunluğa 420 : 60 = 7 tane yerleştirilir. Dikdörtgen şeklindeki alana 56 tane fayans yerleştirileceğinden; odanın enine 7 tane fayans yerleştiriliyorsa (dikdörtgenin alanı = en uzunluğu x boy uzunluğu) 56 : 7 = 8 tane fayans da boy uzunluğuna yerleştirilmelidir. Her bir fayansın kenar uzunluğu 60 cm olduğundan ; Serdar'ın odasının boy uzunluğu 60 x 8 = 480 cm olmalıdır. Ancak Serdar'ın odasının boy uzunluğu 450 cm dir. 480 cm olması için odanın boy uzunluğunun 480 – 450 = **30 cm daha artırılması** gereklidir.
3	• <u>İşlemlerin yanına yeterli açıklama yapılmadan</u> sadece doğru işlemler yapılarak (EBOB ve fayans adedi doğru hesaplanarak) Serdar'ın oda zemininde kullanılacak fayansın bir kenar uzunluğunun 60 cm ve yeni durumda Serdar'ın oda zeminin boy uzunluğunun 30 cm artırılarak 480 cm olması gerektiğinin yazılması.

2	• Derin'in oda zeminin alanını, kare şeklindeki bir fayansın alanına bölerek fayans sayısını bulmayla başlayıp daha sonraki temel işlemlerde hatalar yapma nedeniyle doğru sonuca ulaşamama. • Uygun çözümle başlayıp doğru işlemler yapıp EBOB değerini ya da yeni durumda kullanılacak fayansın bir kenar uzunluğunu bulurken doğru işlemler yapıp küçük hatalardan dolayı işlemlerin sonucunu yanlış bularak doğru cevaba ulaşamama.
1	• EBOB değerlerini, kullanılacak fayans sayısını, yeni durumda Serdar'ın oda zemininde kullanılacak fayansın bir kenar uzunluğunu ve odanın boy uzunluğunda yapılması gereken değişikliği bulmak için herhangi bir işlem yapmadan sadece "kullanılan fayans sayısının 56, yeni durumda kullanılacak fayansın bir kenar uzunluğunun 60 cm ve odanın boy uzunluğunun 30 cm artırılarak 480 cm olmalı" şeklinde yazılması. • EBOB değerlerini, kullanılacak fayans sayısını, yeni durumda Serdar'ın oda zemininde kullanılacak fayansın bir kenar uzunluğunu ve odanın boy uzunluğunda yapılması gereken değişikliği bulmak için herhangi bir işlem yapmadan sadece "daha büyük fayansların kullanılması gerektiğini" belirtme.
0	Boş ya da tamamen yanlış cevap verilmesi.

SORU 3

Başarı Puanı	ÖLÇÜTLER
4	Tam kare doğal sayıların karekök değerleri hesaplanarak iç ve dış bölgeye ait katsayılarla çarpılır 196 ve 169 değerlerinin karıştırılmadan hesaplanmış birey puanları hesaplanır. Nehir: 3.(5+12+16) + 2. (18+4+17) = 177 Selin: 3.(**14**+11+19) + 2.(19+18+7) = 220 Arda: 3.(19+15+12) + 2. (11+2+8) = 180 Cihan: 3.(20+10+15) + 2.(12+16+**13**) = 217 Mert: 3.(17+12+16) + 2.(20+11+12) = 221 Bu durumda birinciliğe en yakın isimler : Mert, Selin, Cihan, Arda, Nehir şeklinde sıralanır.Yukarıdaki hesaplamaların yapılması ve son durumdaki sıralamanın "Mert, Selin, Cihan, Arda, Nehir " yazılması.
3	• İşlemlerin ve sıralamanın yanına yeterli açıklama yapılmadan sadece doğru işlemler yapılarak sıralamanın "Mert, Selin, Cihan, Arda, Nehir "yazılması. • Puan hesaplamalarını yaparken birkaç bireyin puanını yanlış bulduğu halde sıralamanın "Mert, Selin, Cihan, Arda, Nehir " şeklinde doğru yazılması.
2	• Çözüm yolu doğru olup tam kare doğal sayıların karekök değerlerinin yanlış yazılması gibi temel işlemlerde hatalar yapma nedeniyle doğru sonuca ve sıralamaya ulaşamama. • Tam kare doğal sayıların karekök değerlerini doğru yazıp doğru işlemler yapıp küçük hatalardan dolayı birey puanlarının yanlış hesaplanması ve sıralamanın doğru yapılamaması.
1	• Herhangi bir işlem yapılmadan sadece sıralamanın "Mert, Selin, Cihan,Arda, Nehir" şeklinde doğru yazılması. • Birey puanlarını hesaplamak için herhangi bir işlem yapmayıp sadece tam kare sayıların karekök değerlerini doğru yazıp sıralamanın"Mert,Selin, Cihan, Arda, Nehir"şeklinde doğru yazılması.

0	Boş ya da tamamen yanlış cevap verilmesi.

SORU 4

Başarı Puanı	ÖLÇÜTLER
4	Tam kare doğal sayıların karekök değerleri yazılarak iç bölgedeki atışlar 3 ile ve dış bölgedeki atışlar 2 ile çarpılıp toplanarak birey puanları hesaplanır. Nehir: 3.(5+12+16) + 2. (18+4+17) = 177 şeklinde puanını hesaplaması gerekirken kareköklü sayılarla toplama işlemini yanlış uygulayarak kök içerisindeki sayıları toplayıp kök içerisine yazmış ve bölgelerin katsayılarını kök içerisindeki sayıyla çarparak kareköklü sayılarla çarpma işlemini de yanlış uygulamıştır. Bulduğu kareköklü sayının değerini 50'ye yuvarlayarak toplam puanını 50 olarak bulmuştur. Cihan: 3.(20+10+15) + 2.(12+16+13) = 217 şeklinde kareköklü sayılarla toplama ve çarpma işlemlerini doğru bir şekilde yaparak toplam puanını doğru hesaplamıştır. Yukarıdaki açıklamalar yapılarak Nehir'in kullandığı yöntemin yanlış, Cihan'ın kullandığı yöntemin doğru olduğunun belirtilmesi.
3	• İşlemlerin yanına yeterli açıklama yapılmadan sadece Nehir ve Cihan'ın puan hesaplamalarının doğru yapılmış hali gösterilerek Nehir'in kullanmış olduğu yöntemin yanlış, Cihan'ın kullandığı yöntemin doğru olduğunun yazılması.
2	• Tam kare doğal sayıların karekök değerlerini doğru bilip kareköklü sayılarla toplama ve çarpma işlemlerini yanlış bilme gibi temel işlemlerde hatalar yapma nedeniyle bireylerden birinin kullandığı yöntemin doğruluğunu belirleyip diğeri hakkında yorum yapamama. • Tam kare doğal sayıların karekök değerlerini doğru bilip kareköklü sayılarla toplama ve çarpma işlemlerini doğru yapıp küçük hatalardan dolayı bireylerin toplam puanını yanlış bularak kullanılan yöntemlerin doğruluğunu yanlış yorumlama.
1	• Herhangi bir işlem yapılmadan sadece "Nehir'in kullandığı yöntemin yanlış, Cihan'ın kullandığı yöntemin doğru" olduğunun yazılması. • Herhangi bir işlem yapılmadan sadece "Nehir'in kullandığı yöntem yanlıştır."yazılması.

	• Herhangi bir işlem yapılmadan sadece "Cihan'ın kullandığı yöntem doğrudur."yazılması.
0	Boş ya da tamamen yanlış cevap verilmesi.

SORU 5

Başarı Puanı	**ÖLÇÜTLER**
4	**I. ÇÖZÜM YOLU: (**Tam kare olmayan sayıların karekök değerlerini tahmin ederek) Önder'in maketi kutu içerisinde taşıyabilmesi için kutu tabanının kısa kenar uzunluğu, uzun kenar uzunluğu ve yüksekliğinin; makete ait kısa kenar uzunluğu, uzun kenar uzunluğu ve yüksekliğinden **<u>büyük</u>** olması gerekir. Kutunun yüksekliği: $\sqrt{121} < \sqrt{122} < \sqrt{144} \rightarrow 11 < \sqrt{122} < 12$ cm dir. IABI+ICDI=11 cm olduğundan kutunun yüksekliğinden küçüktür ve maketin üstü kutu kapağı ile kapanabilir. Kutunun uzun kenar uzunluğu : $\sqrt{144} < \sqrt{148} < \sqrt{169} \rightarrow 12 < \sqrt{148} < 13$ cm'dir. IEFI=13 cm olduğundan maketin uzun kenarı kutunun içerisine yerleşemez. Kutunun kısa kenar uzunluğu : $\sqrt{36} < \sqrt{48} < \sqrt{49} \rightarrow 6 < \sqrt{48} < 7$ cm'dir. IDEI=8 cm olduğundan maketin kısa kenarı kutunun kısa kenarından büyük olduğundan kutunun içerisine yerleşemez. Yukarıdaki açıklamalar yapılarak Önder'in maketini kutu ile <u>taşıyamayacağının</u> belirtilmesi. **II.ÇÖZÜM YOLU**: (Kutunun ve maketin hacimlerini ayrı ayrı tahmin ederek) Maketin hacmi:Ev kısmının hacmi+Çatı kısmının hacmi= (8x8x13)+[(8x13x3)/2]=832+156=988 santimetreküp $\sqrt{121} < \sqrt{122} < \sqrt{144} \rightarrow 11 < \sqrt{122} < 12$ cm yaklaşık 11 cm $\sqrt{36} < \sqrt{48} < \sqrt{49} \rightarrow 6 < \sqrt{48} < 7$ cm yaklaşık 7 cm $\sqrt{144} < \sqrt{148} < \sqrt{169} \rightarrow 12 < \sqrt{148} < 13$ cm yaklaşık 12 cm Kutunun hacmi: Kısa kenar uz.x Uzun kenar uz.x Yükseklik=7x11x12=924 santimetreküptür. Kutunun hacmi, maketin hacminden küçük olduğundan

3	Önder'in maketini kutu ile taşıyamayacağının belirtilmesi. **I.ÇÖZÜM YOLU:** • İşlemlerin yanına yeterli açıklama yapılmadan sadece tam kare olmayan doğal sayıların karekökü olarak verilen kutuya ait yükseklik, uzun kenar uzunluğu ve kısa kenar uzunluğunun hangi doğal sayılar arasında bulunduğu gösterilerek Önder'in maketi kutu ile taşıyamayacağının yazılması. • İşlemlerin yanına yeterli açıklama yapılmadan sadece tam kare olmayan doğal sayıların karekökü olarak verilen kutuya ait yükseklik, uzun kenar uzunluğu ve kısa kenar uzunluğunun hangi doğal sayılar arasında bulunduğu gösterilerek maket ile kutu uzunluklarını karşılaştırıp sadece uzun kenarın uymadığını bularak (diğer uzunluklar hakkında yorum yapmama) Önder'in maketi kutu ile **taşıyamayacağının** yazılması. **II. ÇÖZÜM YOLU:** • İşlemlerin yanına yeterli açıklama yapılmadan sadece tam kare olmayan doğal sayıların karekökü olarak verilen kutuya ait yükseklik, uzun kenar uzunluğu ve kısa kenar uzunluğunun hangi doğal sayılar arasında bulunduğu gösterilerek kutu ve maketin hacimlerini ayrı ayrı doğru hesaplayarak Önder'in maketi kutu ile **taşıyamayacağının** yazılması.
2	**I.ÇÖZÜM YOLU**: • Tam kare olmayan doğal sayıların karekök değerlerinin hangi tam kare doğal sayıların karekökleri arasında olduğunu bulup hangi doğal sayılar arasında olduğunu bulamama gibi temel işlemlerde hatalar yapma nedeniyle Önder'in maketi kutu ile **taşıyabileceğini** yazma. • Tam kare olmayan doğal sayıların karekök değerlerinin hangi doğal sayılar arasında olduğunu doğru bilip maketin kısa, uzun kenar ve yükseklik uzunluklarının birbirine karıştırılması gibi küçük hatalardan dolayı Önder'in maketi kutu ile **taşıyabileceğini** yazma. • Tam kare olmayan doğal sayıların karekök değerlerinin hangi doğal sayılar arasında olduğunu **yanlış bulup** maket üzerindeki

2	uzunluklar ile bu değerleri karşılaştırarak işlemler yapıp Önder'in maketi kutu ile **taşıyabileceğini** yazma. • Tam kare olmayan doğal sayıların karekök değerlerinden bazılarının hangi doğal sayılar arasında olduğunu yazıp işlemlerden de bazılarını yapıp Önder'in maketi kutu ile **taşıyamayacağını** yazma. **II. ÇÖZÜM YOLU**: • Tam kare olmayan doğal sayıların karekök değerlerinin hangi tam kare doğal sayıların karekökleri arasında olduğunu bulup hangi doğal sayılar arasında olduğunu bulamama gibi temel işlemlerde hatalar yapma nedeniyle kutu ve maket hacimlerini yanlış bulması sonucu Önder'in maketi kutu ile **taşıyabileceğini** yazma. • Tam kare olmayan doğal sayıların karekök değerlerinden bazılarının hangi doğal sayılar arasında olduğunu yazıp hacim hesaplama işlemlerinden de bazılarını yapıp Önder'in maketi kutu ile **taşıyamayacağını** yazma.
1	• Herhangi bir işlem yapılmadan sadece "Önder maketi kutu ile taşıyamaz." yazılması. • Herhangi bir işlem yapılmadan sadece maket ya da kutu üzerindeki **yanlış uzunluğun** neden olarak gösterilerek "Önder maketi kutu ile taşıyamaz." yazılması. • Sadece tam kare olmayan doğal sayıların kareköklerinin hangi doğal sayılar arasında olduğu yazılarak "Önder maketi kutu ile taşıyamaz." yazılması.
0	Boş ya da tamamen yanlış cevap verilmesi.

SORU 6

Başarı Puanı	ÖLÇÜTLER
4	350 MB lık fotoğraf albümünü USB belleğe aktarabilmesi için var olan 152 MB lık boş alan dışında kalan 350-152=198 MB lık boş alan oluşturmak için müzik albümlerinden birini ya da ikisini silmesi gerekmektedir. Albümlerin içerisinde 198 MB lık bir albüm olmadığından toplamı 198 MB veya daha fazla olan iki albüm silinmelidir. Evet. Albüm 1 ile Albüm 8 silindiğinde 100+125=225 MB lık Albüm 7 ile Albüm 8 silindiğinde 75+125=200 MB lık Albüm 6 ile Albüm 8 silindiğinde 80+ 125=205 MB lık Albüm 3 ile Albüm 8 silindiğinde 80+125=205 MB lık Albüm 2 ile Albüm 8 silindiğinde 75+125=200 MB lık boş alan oluşur. "Evet" işaretlenerek yukarıdaki açıklamaların yapılıp silinmesi gereken albümlere en az iki tane örneğin verilmesi.
3	• "Evet" i işaretleyerek işlemlerin yanına yeterli açıklama yapılmadan neden 2 tane albümün silinmesi gerektiği söylenmeden sadece işlemlere yer verip silinebilecek albümlere iki tane örnek verilmesi. • "Evet" i işaretleyerek **açıklama yazarak** silinebilecek albümlere **bir tane** örneğin verilmesi.
2	• USB ye aktarılmak istenen fotoğraf albümünün alanından var olan boş alanın çıkarılması gibi (350-152=198 MB) temel işlemlerde hatalar yapma nedeniyle silinmesi gereken albüm alanının yanlış hesaplanması sonucu albümlerin silinmesiyle fotoğraf albümünün USB ye aktarılamayacağının yazılması.("Hayır "ın işaretlenmesi.) • USB den silinmesi gereken alanın doğru hesaplanıp (350-152=198 MB) silinmesi gereken albümlere yanlış örnek vererek fotoğraf albümünün USB ye aktarılamayacağının yazılması. ("Hayır "ın işaretlenmesi.) • USB den silinmesi gereken alanın doğru hesaplanıp (350-152=198 MB) "Evet" i işaretleyerek fotoğraf albümünün USB ye

	aktarılabileceğinin yazılarak örnek verilmemesi.
1	• Herhangi bir işlem yapılmadan sadece "Evet" i işaretleyerek fotoğraf albümünün USB ye aktarılabileceğinin yazılması. • "Evet" i işaretleyerek fotoğraf albümünün USB ye aktarılabileceğinin yazılarak **yanlış örnek verilmesi.**
0	Boş ya da tamamen yanlış cevap verilmesi.

SORU 7

Başarı Puanı	ÖLÇÜTLER
4	Tatil evinin metrekare fiyatı 2500 zed ve tatil evi 60 metrekare olduğundan 2500.60=150000 zed; otoparkı olduğu için 150000+**35000**=185000 zed olur. Sahile uzaklığı 350 metreden az olduğu için 185000+**15000**=200000 zed ; şehir merkezine varış süresi 10 dk olduğundan tatil evinin fiyatı uzman tarafından belirlenen ölçütlere göre 200000+**10000**=210000 zed olarak belirlenir. Uzmanın teklif ettiği hesaplanan fiyat (210000 zed), ilandaki satış fiyatı 200000 zedden fazla olduğundan Ceren için teklif edilen fiyatın **çok iyi** olduğunun yazılması.
3	• İşlemlerin yanına açıklama yapılmadan sadece toplama işlemleri yapılarak (2500.60+10000+35000+15000=210000 zed) Ceren için teklif edilen fiyatın **çok iyi** olduğunun belirlenmesi. • Uzman tarafından teklif edilen fiyatın 2500.60+10000+35000+15000=210000 zed olduğunun bulunup **teklif ile satış fiyatını (200000 zed) karşılaştırmadan** Ceren için teklif edilen fiyatın **çok iyi** olduğunun yazılması. • Uzman tarafından teklif edilen fiyatın 2500.60+10000+35000+15000=210000 zed olduğunun bulunup **teklif ile satış fiyatını (200000 zed) karşılaştırıp** Ceren için teklif edilen fiyatın **çok kötü** olduğunun yazılması.
2	• Verilen tablolardaki doğru değerler kullanılarak yapılan toplama ve çarpma işlemlerinin sonucunun yanlış bulunması gibi temel işlemlerde hatalar yapma nedeniyle Ceren için teklif edilen fiyatın **çok iyi olmadığının ya da teklif edilen fiyatın çok iyi olduğunun** yazılması. • Çarpma ve toplama işlemleri doğru yapıldığı halde verilen tablodaki kriter değerlerinden ilana uygun olmayan değerlerin seçilmesi sonucu Ceren için teklif edilen fiyatın **çok iyi** olmadığının yazılması.

1	• Herhangi bir işlem yapılmadan sadece tatil evi için Ceren'e teklif edilen fiyatın **çok iyi** olduğunun yazılması.
0	Boş ya da tamamen yanlış cevap verilmesi.

SORU 8

Başarı Puanı	ÖLÇÜTLER
4	"a ve n bir tam sayı olmak üzere a. 10^n şeklinde yazılan çok büyük ve çok küçük sayıların bilimsel gösteriminde virgülden önceki tam kısmın $1 \leq a < 10$ olması gerekir.Bilimsel gösterimlerde tam kısımda virgül sola doğru hareket ettirilirse üs sayı değeri olarak azalır (ya da üs büyür). Bilimsel gösterimde virgül sola doğru üç basamak hareket ettirildiğinden üs -33+3= -30 olmalıdır. Bu nedenle Miray Hanım kitapta yer alan $1673.\ 10^{-33}$ g yazımının düzeltilmesini istemiştir." Şeklinde açıklama yapılarak **protonun kütlesinin** $1,\ 673.10^{-30}$ **g** olarak yazılması.
3	İşlemlerin yanına açıklama yapılmadan sadece a. 10^n , $1 \leq a < 10$ ve protonun kütlesi $1,\ 673.10^{-30}$ g şeklinde yazılması.
2	• Virgülden önceki tam kısmın 1 ve 1 den büyük 10 dan küçük olması gerektiğini yazıp tam kısmı doğru yazıp 10'un kuvvetini yanlış yazma sonucu bilimsel gösterimi doğru yazamama. • Bilimsel gösterimde tam kısmın 1 ve 1 den büyük 10 dan küçük olması gerektiğini yazıp bilimsel gösterimde 10'un kuvvetini değiştirip tam kısmında virgül kullanmama ya da 1 den sonra virgül koyup 673 kısmını yazmama sonucu gösterimi doğru yazamama.
1	• Herhangi bir işlem ve açıklama yapılmadan sadece "**Protonun kütlesi** $1,\ 673.10^{-30}$ **g dır.**" şeklinde yazılması. • Herhangi bir işlem ve açıklama yapılmadan $1,673.10^{-x}$ g ifadesinde x değerini yanlış yazma.
0	Boş ya da tamamen yanlış cevap verilmesi.

SORU 9

Başarı Puanı	ÖLÇÜTLER
4	**TÜM İHTİMALLER:** • Mehmet Bey 7. soruyu görmüş ancak cevap vermeden 2^6 **bin TL =64 000TL alıp yarışmadan ayrılmış olabilir**. • **7.soruya yanlış cevap vermiş 64000 TL alarak ayrılmıştır.** • 7. soruya doğru cevap verdiyse 2^7 **bin TL=128 000TL almış ve 8. Soruyu görmeden ayrılmıştır.** • **8. soruyu görmüş cevaplayamadığı için yarışmadan çekilmiştir.** • 8. soruya yanlış cevap vermiş ve **64000 TL alarak yarışmadan ayrılmıştır.** • 8. soruya doğru cevap verdiyse 2^8 **bin TL=256 000 TL almış ve yarışmadan ayrılmıştır.** 450000 TL ye **en yakın** durum Mehmet Bey'in 8. soruya doğru cevap vererek **2^8 bin TL=256 000 TL** almasıdır.
3	• Yukarıdaki ihtimallerden 1-2 tanesini yazamayıp tüm üslü sayıların değerlerini doğru hesaplayarak 450 000 TL ye en yakın durumun Mehmet Bey'in 8.soruya doğru cevap vererek 256000 TL alması olduğunu yazma. • Yukarıdaki ihtimallerden 1-2 tanesini yazamayıp tüm üslü sayıların değerlerini doğru hesaplayarak sadece 450 000 TL ye en yakın durumu yanlış yazma.
2	• İhtimalleri doğru yazıp üslü sayıların değerlerini yanlış hesaplama sonucu 450000 TL ye **en yakın durumu yanlış bulma.** • Üslü sayıların değerlerini doğru hesapladığı halde ihtimalleri yanlış düşünme sonucu 450000 TL ye **en yakın durumu yanlış bulma.** • Üslü sayıların değerlerini doğru hesaplayarak ihtimallerden 1-2 tanesini doğru yazıp en yakın durumun Mehmet Bey'in 8.soruya doğru

	cevap vererek 256000 TL alması olduğunu yazma. • Üslü sayıların değerlerini doğru hesaplayarak ihtimallerden **iki ya da daha fazlasını doğru yazıp** en yakın durumu **yanlış** bulma ya da en yakın durumu **bulamama.**
1	• Herhangi bir işlem yapılmadan sadece "Mehmet Bey'in 450000 TL elde etmesine en yakın durum 8.soruya doğru cevap vererek yarışmadan 256 000 TL alarak ayrılması" şeklinde yazılması.
0	Boş ya da tamamen yanlış cevap verilmesi.

SORU 10

Başarı Puanı	ÖLÇÜTLER
4	Derya ve Seda'nın aynı fotoğrafçıda vesikalık çekilmeleri için ikisinin isteklerine cevap veren fotoğrafın tanesini en ucuza çeken, toplam ödenen ücretin 25 TL den fazla olmadığı ve fotoğraf sayısının 10 dan fazla olduğu fotoğrafçıda çekim yapılacaktır. Fotoğraf sayısının 10 dan fazla olması istendiğinden 16 ve 20 adet fiyatlarına bakıldığında; fotoğrafın tane fiyatını 1 TL ye çeken B fotoğrafçısı en düşük fiyatlı olanıdır. Derya ve Seda 16 + 2 vesikalık bedava olduğundan 18 adet fotoğraf için ayrı ayrı **18 TL** öderler.16 ve 20 adet için tüm fotoğrafçıların tane fiyatının hesaplanması ve B Fotoğrafçılık'ın işaretlenmesi gerekmektedir. Yukarıda Derya ve Seda için yapılan **tane fiyatı**, **ödenecek ücretin hesaplanması** ve her biri için seçilmesi gereken **fotoğrafçılığın işaretlenmesi.**
3	İşlemlerin yanına açıklama yapmadan sadece 16 ve 20 adet için tüm fotoğrafçıların tane fiyatını hesaplayarak B Fotoğrafçılık' ta çekilmelerinin uygun olduğunu ve ayrı olarak 18'er TL ödemeleri gerektiğini belirtme ve B Fotoğrafçılık'ı işaretleme.
2	• İstenen özelliklere uygun olacak şekilde 16 ve 20 adet için tane fiyatını yanlış hesaplama sonucu yanlış fotoğrafçıyı belirleme ve bunun sonucunda ödenecek ücreti farklı hesaplama. • Vesikalık fotoğraf tane fiyatlarını doğru hesaplayıp Derya ve Seda'nın isteklerine dikkate almadan **yanlış fotoğrafçıyı belirleme** ve bunun sonucunda **ödenecek ücreti farklı hesaplama.**
1	• Herhangi bir işlem yapılmadan sadece "En uygun fiyata sahip olan B Fotoğrafçılık'a birlikte giderek 18 adet fotoğrafa karşılık 18 TL öderler." şeklinde yazılması ve B Fotoğrafçılık'ın işaretlenmesi. • Toplam ödenecek ücret hesaplaması yapmayıp birkaç vesikalık tane fiyatı hesaplandıktan sonra sadece B Fotoğrafçılık'ın işaretlenmesi.
0	Boş ya da tamamen yanlış cevap verilmesi.

SORU 11

Başarı Puanı	ÖLÇÜTLER
4	• Derya'nın 10 fotoğraftan fazla ihtiyacı olduğu ve en fazla 25 TL ödeyebildiği için 16 ve 20 adet fiyatlarından uygun fiyat olanını seçmesi gerekmektedir. 16 adet vesikalık fotoğrafın tane fiyatları ayrı ayrı hesaplandığında sırasıyla; A fotoğrafçılık:1,15 TL B fotoğrafçılık:18/18=1 TL C fotoğrafçılık:20/16=1,25 TL 20 adet vesikalık fotoğrafın tane fiyatları ayrı ayrı hesaplandığında sırasıyla; A fotoğrafçılık:25/20=1,25 TL B fotoğrafçılık:1,20 TL C fotoğrafçılık:26/24=1,08 TL en uygun fiyattır ancak Derya'nın 25 TLsi olduğundan 26 TL yi ödeyemez. Bu nedenle uygun fiyatlı olması için ; **16 adet vesikalık fotoğrafı B fotoğrafçılık**ta çektirmesi ve **18 TL** ödemesi gerekmektedir. • Seda için fotoğraf sayısı önemli değil ancak tane fiyatının çok düşük olmasını istediğinden tüm fotoğrafçılardaki vesikalık fotoğrafın tane fiyatlarını hesaplandığında; A fotoğrafçılık: **1,10 TL**; 14/12=**1,17 TL**; **1,15 TL**; 25/20=**1,25 TL** B fotoğrafçılık: 5/4=**1,25 TL**; 10/8=**1,25 TL**; 18/18=**1 TL**; **1,20 TL** C fotoğrafçılık: 6/6=**1 TL**; **1,10 TL**; 20/16=**1,25 TL**; 26/24=**1,08 TL** Tane fiyatı 1TL olan **C Fotoğrafçılık'** ta **4 adet vesikalık fotoğraf** çekilmesi uygun fiyatlı olacaktır.**6 TL ödemesi gerekmektedir.** Derya için **B FOTOĞRAFÇILIK,** Seda için **C FOTOĞRAFÇILIK** işaretlenmelidir. Yukarıda Derya ve Seda için yapılan **tane fiyatı**, **ödenecek ücretin hesaplanması** ve her biri için seçilmesi gerekli **fotoğrafçılığın işaretlenmesi.**

3	Yukarıdaki hesaplamalardan birkaçını yazmayıp işlemlerin yanına açıklama yapmadan Derya ve Seda için tane fiyatı en uygun olan fotoğrafçıların hesaplamalarla belirlenerek Derya için B FOTOĞRAFÇILIK, Seda için C FOTOĞRAFÇILIK' ın işaretlenmesi ve çektirebilecekleri vesikalık fotoğraf sayısının, tane fiyatının açıkça yazılması.
2	• Tane fiyatını bulduran işlemler gibi temel işlemleri doğru yapıp ondalık gösterimleri sıralamada hata yapıldığı için vesikalık fotoğraf adedini ve gidecekleri fotoğrafçıyı yanlış belirleme. • Ondalık gösterimleri sıralamada hata yapmayıp vesikalık tane fiyatını doğru hesaplayıp hangi değerin hangi fotoğrafçıya ait olduğunu karıştırma gibi küçük hatalardan dolayı vesikalık fotoğraf adedini ve gidecekleri fotoğrafçıyı yanlış belirleme.
1	• Herhangi bir işlem yapılmadan sadece "Derya 16 adet vesikalık fotoğrafı B Fotoğrafçılık' ta Seda ise 4 adet vesikalık fotoğrafı C Fotoğrafçılık' ta çektirmelidir." şeklinde yazılması. • Herhangi bir işlem yapılmadan **fotoğraf adedi belirtilmeden** sadece Derya için B Fotoğrafçılık'ın, Seda için ise C Fotoğrafçılık'ın **işaretlenmesi.** • Herhangi bir işlem yapılmadan sadece **fotoğraf adedi belirtilerek** fotoğrafçılardan bitinin ya da ikisinin **yanlış işaretlenmesi.**
0	Boş ya da tamamen yanlış cevap verilmesi.

EK 7. Öğretmen Değerlendirme Formu

MATEMATİK BAŞARI TESTİNDEKİ AÇIK UÇLU MADDELERE İLİŞKİN ÖĞRETMEN DEĞERLENDİRME FORMU

Not: Açık uçlu maddelerin değerlendirilmesi için size verilen özel bütünsel dereceli puanlama anahtarındaki "4:Tam ve doğru cevap verilmesi", "3:Yapılan işlemlerin ve sonucun doğru olup sonuca ulaştıran işlemlerde yeterli açıklamanın yapılmaması","2: Çözüm yolu doğru olup temel işlemlerde hata yapılması nedeniyle doğru sonuca ulaşamama ya da uygun çözümle başlayıp doğru işlemler yapıp küçük hatalar dolayı doğru sonuca ulaşamama","1: Herhangi bir işlem yapmadan doğru cevabın yazılması.", "0: Boş ya da tamamen yanlış cevap verilmesi" derecelendirmelere uygun olarak maddelerin puanlanması gerekmektedir.

ÖĞRENCİ NO	MADDELER										
	1	2	3	4	5	6	7	8	9	10	11
1											
2											
3											
4											
5											
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											
30											
31											
32											
33											
34											
35											
36											
37											
38											
39											

EK 8. PISA 2012 Pilot Uygulama Soruları Kullanım İzni

Evrak Tarih ve Sayısı: 19.01.2021-E.10213

T.C.
MİLLÎ EĞİTİM BAKANLIĞI
Ölçme, Değerlendirme ve Sınav Hizmetleri
Genel Müdürlüğü

Sayı : E-57750415-605.01-19289351 18.01.2021
Konu : Bilimsel ve Eğitim Amaçlı İzin (Zeynep GÖLCÜR)

GAZİ ÜNİVERSİTESİ REKTÖRLÜĞÜNE
(Eğitim Bilimleri Enstitüsü Müdürlüğü)

İlgi : 12.01.2021 tarihli ve 80287700-302.08.01.08.01-E.6176 sayılı yazınız.

Enstitünüz Eğitim Bilimleri Ana Bilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalı Yüksek Lisans programı öğrencisi Zeynep GÖLCÜR'ün, Prof. Dr. İsmail KARAKAYA'nın danışmanlığında yürüttüğü "Çok Yüzeyli Rasch Modeli Puanlama Desenlerine Göre Açık Uçlu Maddelerin Puanlayıcılar Arası Güvenirliklerinin Karşılaştırılması' konulu tez çalışmasında PISA 2012 Pilot Uygulama Soruları Matematik Testi içerisinde yayımlanmış olan "USB Bellek" ile "Tatil Evi" isimli iki adet açık uçlu soruyu kullanma talebi yazımız ekinde gönderilen "Teşekkür ve Tez Teslim Taahhütnamesi"nin imzalanıp Genel Müdürlüğümüze iletilmesi şartıyla uygun görülmüştür.

Bilgilerini ve gereğini rica ederim.

Dr. Sadri ŞENSOY
Bakan a.
Genel Müdür

EK- Teşekkür ve Tez Teslim Taahhütnamesi (1 sayfa)

Bu belge güvenli elektronik imza ile imzalanmıştır.

Bu belge,güvenli elektronik imza ile imzalanmıştır.

EK 9. Etik Kurul Onayı

Evrak Tarih ve Sayısı: 28.12.2020-E.139862

T.C.
GAZİ ÜNİVERSİTESİ
Ölçme Değerlendirme Etik Alt Çalışma Grubu

Sayı : E-91610558-302.08.01-
Konu : Bilimsel ve Eğitim Amaçlı

EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

İlgi : 17.12.2020 tarih ve E.135491 sayılı yazı

İlgi yazınız ile göndermiş olduğunuz, Enstitünüz Eğitim Bilimleri Anabilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalı **Yüksek Lisans Öğrencisi Zeynep GÖLCÜR'ün, Prof.Dr.İsmail KARAKAYA'nın** danışmanlığında yürüttüğü ***"Çok Yüzeyli Rasch Modeli Puanlama Desenlerine Göre Açık Uçlu Maddelerin Puanlayıcılar Arası Güvenirliklerinin Karşılaştırılması"*** adlı tez çalışması ile ilgili konu Kurulumuzun **22.12.2020** tarih ve **13** sayılı toplantısında görüşülmüş olup,

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

Bilgilerinizi ve gereğini rica ederim.

Araştırma Kod No: 2020 - 703

Prof. Dr. İsmail KARAKAYA
Kurul Başkanı

Ek:1 Liste

Bu belge, güvenli elektronik imza ile imzalanmıştır.

Bu belge,güvenli elektronik imza ile imzalanmıştır.



GAZİLİ OLMAK AYRICALIKTIR…