



**T.C.
E G E Ü N İ V E R S İ T E S İ
Sosyal Bilimler Enstitüsü**

**AÇIK UÇLU SORULARIN PUANLAMA YÖNTEMLERİNİN
GENELLENEBİLİRLİK KURAMINA GÖRE İNCELENMESİ**

Yüksek Lisans Tezi

Pelin DAMLAR DEMİRCİ

**Eğitim Bilimleri Anabilim Dalı
Ölçme ve Değerlendirme
Yüksek Lisans Programı**

İZMİR

2019

T.C.
E G E Ü N İ V E R S İ T E S İ
Sosyal Bilimler Enstitüsü

AÇIK UÇLU SORULARIN PUANLAMA YÖNTEMLERİNİN
GENELLENEBİLİRLİK KURAMINA GÖRE İNCELENMESİ

Yüksek Lisans Tezi

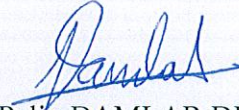
Pelin DAMLAR DEMİRCİ

Tez Danışmanı: Doç. Dr. Hakan ATILGAN

Eğitim Bilimleri Anabilim Dalı
Ölçme ve Değerlendirme Yüksek Lisans Programı

ETİK KURALLARA UYGUNLUK BEYANI

Ege Üniversitesi Sosyal Bilimler Enstitüsü Mdrlğne sunduğum Aık Ulu Soruların Puanlama Yntemlerinin Genellenebilirlik Kuramına Gre İncelenmesi adlı yksek lisans tezinin tarafımdan bilimsel, ahlak ve normlara uygun bir Őekilde hazırlandığını, tezimde yararlandığım kaynakları bibliyografyada ve dipnotlarda gsterdiğimi onurumla doğrularım.


Pelin DAMLAR DEMİRCİ



T.C.EGE ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ



YÜKSEK LİSANS

TEZ SAVUNMA TUTANAĞI

ÖĞRENCİNİN

Adı Soyadı : Pelin DAMLAR DEMİRCİ
Numarası : 92140002219
Anabilim Dalı : Eğitim Bilimleri (Eğitimde Ölçme ve Değerlendirme Bilim Dalı)
Tez Başlığı (Türkçe) : Açık Uçlu Soruların Puanlama Yöntemlerinin Genellenebilirlik Kuramına Göre İncelenmesi
Tez Başlığı (İngilizce) : The Investigation of Open-Ended Items Scoring Methods by Generalizability Theory

Tez Savunma Tarihi : 08.08.2019

Tez Başlığı Değişikliği Varsa Yeni Başlık:

JÜRİ ÜYELERİ

Jüri Başkanı

Unvan, Adı, Soyadı : Doç. Dr. Hakan ATILGAN (Danışman)
Karar : ☒ Başarılı ☐ Başarısız ☐ Düzeltme
İmza :

Jüri Üyesi

Unvan, Adı, Soyadı : Prof. Dr. Tuncay ÖĞRETMEN
Karar : ☒ Başarılı ☐ Başarısız ☐ Düzeltme
İmza :

Jüri Üyesi

Unvan, Adı, Soyadı : Doç. Dr. İrfan YURDABAKAN
Karar : ☒ Başarılı ☐ Başarısız ☐ Düzeltme
İmza :

TEZ HAKKINDA JÜRİNİN GENEL GÖRÜŞÜ

(Jüri Başkanı Tarafından Doldurulacaktır)

Tez savunması sonucunda öğrenci tarafından hazırlanan çalışma;

Oybirliğiyle ☒

Oy çokluğuyla ☐

Başarılıdır

Düzeltilmelidir

Başarısızdır

☒☐☐

- Bu tutanak üç (3) işgünü içerisinde jüri üyelerinin raporlarıyla beraber Anabilim Dalı Başkanlığı üst yazısıyla Enstitü Müdürlüğüne gönderilmelidir.
- Tezli yüksek lisans programlarında düzeltme alan öğrencinin 3 (üç) ay içerisinde yeniden savunmaya girmesi zorunludur.

ÖNSÖZ

Günümüzde değişen ve gelişen eğitim sisteminin en önemli parçalarından biri de ölçme ve değerlendirme süreçleridir. Bu süreçte kullanılan ölçme ve değerlendirme yöntemleri de eğitim sistemi ile doğru orantılı olarak değişmekte ve gelişmektedir.

Ölçme ve değerlendirme yaparken kullanılan yazılı sınavlar ve bu sınavlarda kullanılan açık uçlu maddeler şimdiye kadarki matematik öğretmenliği deneyimimde üst düzey bilişsel basamakları en iyi şekilde ölçebildiğim madde türleri olmuştur. Ayrıca açık uçlu maddelerin ülkemizin de katıldığı PISA ve TIMMS gibi uluslararası sınavlarda da kullanılması araştırmalarımın bu yönde ilerlemesini sağlamıştır.

Kuramsal çerçevede bu konuyu araştırdıkça açık uçlu maddelerin puanlama yöntemleriyle ilgili fazla sayıda çalışmanın olmaması dikkatimi çekmiştir. Bu nedenle araştırmamda kullandığım veriler hem serbest hem anahtarla puanlanarak incelenmiştir. Ayrıca ölçme sürecine farklı sayıda madde ve puanlayıcının katılmasının ölçme sonuçları üzerinde oluşturduğu etkiler de bu araştırmanın konusu olmuştur.

Bu araştırma öğretmenlik mesleğime ve araştırma süresince yaptığım alanyazın okumalarıyla bilgi birikimime katkı sağlaması nedeniyle benim için oldukça verimli bir süreçti.

Araştırmanın bilimselliğe, uygulayıcı ve araştırmacılara katkı sağlaması dileğiyle...

ÖZET

Bu araştırmanın amacı açık uçlu maddelerden oluşan matematik sınavının farklı puanlayıcılar tarafından serbest ve anahtarla puanlaması ile elde edilen verilerin karşılaştırılmasıdır.

Çalışma grubunu 2015-2016 eğitim öğretim yılında bir devlet ortaokulunda öğrenim gören 253 kişiden oluşan yedinci sınıf öğrencileri oluşturmaktadır.

Araştırmada toplanan veriler yansız seçilen beş farklı alan uzmanı tarafından farklı zamanlarda önce serbest daha sonra anahtarla puanlanmıştır. Her bir puanlayıcının (p) her bir maddeyi (m) tüm bireyler (b) için bağımsız olarak puanlaması ile $b \times m \times p$ deseni oluşturularak G ve K çalışmaları yapılmıştır.

Yapılan G ve K çalışmalarının sonuçları serbest ve anahtarla puanlamaya göre karşılaştırılmıştır. Elde edilen sonuçlara göre her iki puanlama türü için de güvenilir sonuçlar elde edildiği fakat anahtarla puanlamada beklendiği gibi serbest puanlamaya göre daha objektif puanlama yapıldığı görülmüştür. Ayrıca bu araştırmada yapılan K çalışması sonuçlarına göre her iki puanlama türü için de madde sayısının ve puanlayıcı sayısının artırılması G ve Phi katsayılarını arttırmıştır. Fakat işgücü, zaman ve ekonomi düşünüldüğünde madde sayısının artırılmasının puanlayıcı sayısının artırılmasından daha uygun olacağı görülmüştür.

Bunların yanında yapılan araştırmanın, eğitimde ölçme ve değerlendirmede başvuru, ortaöğretime geçiş ve yükseköğretime geçiş sınavlarında da kullanılması gündeme gelen ve puanlanma güvenirliği sıkça tartışılan açık uçlu maddelerden oluşan uzun yanıtlı sınavların güvenirliklerine ilişkin bilgi sağlayacağı düşünülmektedir.

Anahtar kelimeler: Genellenebilirlik Kuramı, Açık Uçlu Sorular, Puanlama Yöntemleri

ABSTRACT

The aim of this study is to compare the data obtained by the free and key scoring of the mathematics exam consisting of open-ended items by different raters.

The study group consisted of 253 seventh grade students studying at the state secondary school in 2015-2016.

The data collected the study were scored by five different field experts who were chosen impartially at different times before they were released and then with the key. Each rater (r) scored each item (i) independently for all people (p) and G and K studies were conducted with $r \times i \times p$ pattern. The results of G and K studies were compared according to free and key scoring. According to the results, reliable results were obtained for both scoring types but as expected in key scoring, more objective scoring was done compared to free scoring. Furthermore, according to the results of the K study, increasing the number of items and the number of items and the number of raters for both scoring types increased the G and Phi coefficients. However, considering the labour force, time and economy, increasing the number of items would be more appropriate than increasing the number of raters.

In addition to this, it is thought that long-term exams consisting of open-ended items which are frequently discussed and used in the measurement and evaluation in education and used in the measurement and evaluation in education and transition to secondary education and transition to higher education examinations in the recent years will provide information about the reliability of long-term exams.

Keywords: Generalizability Theory, Open- Ended Items, Scoring Methods

İÇİNDEKİLER

ÖNSÖZ.....	i
ÖZET.....	ii
ABSTRACT	iii
İÇİNDEKİLER	iv
TABLolar LİSTESİ.....	vi
ŞEKİLLER LİSTESİ.....	vii
KISALTMALAR	viii
BÖLÜM I	1
GİRİŞ	1
1.1.Problem Durumu	21
1.2. Problem Cümlesi	22
1.2.1. Alt Problemler.....	22
1.3. Araştırmanın Amacı ve Önemi.....	23
1.4. Sınırlılıklar.....	23
1.5. Konu İle İlgili Yapılan Çalışma ve Araştırmalar	23
1.5.1 Yurt İçinde Yapılan Çalışmalar	23
1.5.2. Yurt Dışında Yapılan Çalışmalar.....	29
BÖLÜM II	31
YÖNTEM.....	31
2.1. Araştırmanın Türü	31
2.2. Çalışma Grubu, Evren ve Örneklem	31
2.3. Araştırma Verileri.....	31
2.4. Verilerin Analizi	32
2.5. Yapılan G Çalışması Sonucu Elde Edilen İstatistiklerin Yorumlanması	33
BÖLÜM III.....	37
BULGULAR.....	37
3.1. Alt Problem 1.	37
3.1.1. Tek değişkenli G çalışması sonuçları nasıldır?.....	37
3.1.2. Tek değişkenli K çalışması sonuçları nasıldır?.....	38
3.1.3. Çok değişkenli G çalışması sonuçları nasıldır?	42
3.1.4. Çok değişkenli K çalışması sonuçları nasıldır?	46
3.2.Alt Problem 2.	55

3.2.1. Tek deęiřkenli G alıřması sonuları nasıldır?.....	55
3.2.2. Tek deęiřkenli K alıřması sonuları nasıldır?.....	56
3.2.3. ok deęiřkenli G alıřması sonuları nasıldır?	61
3.2.4. ok deęiřkenli K alıřması sonuları nasıldır?	64
3.3.Alt Problem 3.	73
3.3.1. Tek deęiřkenli G alıřması sonuları nasıldır?.....	73
3.3.2. Tek deęiřkenli K alıřması sonuları nasıldır?.....	75
3.3.3. ok deęiřkenli G alıřması sonuları nasıldır?	81
3.3.4. ok deęiřkenli K alıřması sonuları nasıldır?	87
BÖLÜM IV	96
SONU VE ÖNERİLER.....	96
4.1. Sonular.....	96
4.2. Öneriler.....	103
KAYNAKA	104
EKLER.....	111
TEŐEKKÜR	117
ÖZGEMİŐ.....	118

TABLÖLAR LİSTESİ

Tablo 1. Açık Uçlu Matematik Sorusunun Puanlanmasına Uygun Rubrik Örneği 1	8
Tablo 2. Açık Uçlu Matematik Sorusunun Puanlanmasına Uygun Rubrik Örneği 2	10
Tablo 3. Araştırmada Kullanılan Ölçme Aracına ait Belirtke Tablosu.....	32
Tablo 4. Çaprazlanmış İki Yüzeyli $b \times m \times p$ Deseni için Matris.....	35
Tablo 5. Açık Uçlu Matematik Sınavının Serbest Puanlanmasıyla Elde Edilen Tek Değişkenli G Çalışması Sonuçları	37
Tablo 6. Açık Uçlu Matematik Sınavının Serbest Puanlanmasında Bileşik Testte Bulunan Farklı Madde ve Puanlayıcı Sayısı Senaryoları ile Yapılan K Çalışmasının Sonuçları.	39
Tablo 7. Açık Uçlu Matematik Sınavının Serbest Puanlanmasıyla Elde Edilen Çok Değişkenli G Çalışması Sonuçları	43
Tablo 8. Açık Uçlu Matematik Sınavının Serbest Puanlanmasında Alternatif Madde ve Puanlayıcı Sayıları Senaryoları ile Yapılan Çok Değişkenli K Çalışması Sonuçları	47
Tablo 9. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasıyla Elde Edilen Tek Değişkenli G Çalışması Sonuçları	55
Tablo 10. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasında Bileşik Testte Bulunan Farklı Madde ve Puanlayıcı Sayısı Senaryoları ile Yapılan K Çalışmasının Sonuçları.	57
Tablo 11. Açık Uçlu Matematik Sınavının Serbest Puanlanmasıyla Elde Edilen Çok Değişkenli G Çalışması Sonuçları.....	61
Tablo 12. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasında Alternatif Madde ve Puanlayıcı Sayıları Senaryoları ile Yapılan Çok Değişkenli K Çalışması Sonuçları.	65
Tablo 13. Açık Uçlu Matematik Sınavının Anahtarla ve Serbest Puanlanmasıyla Elde Edilen Tek Değişkenli G Çalışması Sonuçları.....	73
Tablo 14. Açık Uçlu Matematik Sınavının Anahtarla ve Serbest Puanlanmasında Bileşik Testte Bulunan Farklı Madde ve Puanlayıcı Sayısı Senaryoları ile Yapılan K Çalışmasının Sonuçları	76
Tablo 15. Açık Uçlu Matematik Sınavının Serbest ve Anahtarla Puanlanmasıyla Elde Edilen Çok Değişkenli G Çalışması Sonuçları	82
Tablo 16. Açık Uçlu Matematik Sınavının Serbest ve Anahtarla Puanlanmasında Alternatif Madde ve Puanlayıcı Sayıları ile Yapılan Çok Değişkenli K Çalışması Sonuçları	89

ŞEKİLLER LİSTESİ

Şekil 1. Tipik Bir Madde Karakteristik Eğrisi	16
Şekil 2. Açık Uçlu Matematik Sınavının Serbest Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan K Çalışması Sonucunda G Katsayısındaki Değişim Grafiği.	41
Şekil 3. Açık Uçlu Matematik Sınavının Serbest Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan K Çalışması Sonucunda Phi Katsayısındaki Değişim Grafiği.....	41
Şekil 4. Açık Uçlu Matematik Sınavının Serbest Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan Çok Değişkenli K Çalışması Sonucunda Alt Testlere Göre G Katsayısındaki Değişim Grafiği	48
Şekil 5. Açık Uçlu Matematik Sınavının Serbest Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan Çok Değişkenli K Çalışması Sonucunda Alt Testlere Göre Phi Katsayısındaki Değişim Grafiği.....	49
Şekil 6. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan K Çalışması Sonucunda G Katsayısındaki Değişim Grafiği	58
Şekil 7. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan K Çalışması Sonucunda Phi Katsayısındaki Değişim Grafiği.....	58
Şekil 8. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan Çok Değişkenli K Çalışması Sonucunda Alt Testlere Göre G Katsayısındaki Değişim Grafiği.	66
Şekil 9. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan Çok Değişkenli K Çalışması Sonucunda Alt Testlere Göre Phi Katsayısındaki Değişim Grafiği.....	67
Şekil 10. Açık Uçlu Matematik Sınavının Anahtarla ve Serbest Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan K Çalışması Sonucunda Hesaplanan G Katsayısındaki Değişim Grafiği	77
Şekil 11. Açık Uçlu Matematik Sınavının Anahtarla ve Serbest Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan K Çalışması Sonucunda Hesaplanan Phi Katsayısındaki Değişim Grafiği	78

KISALTMALAR

KTK: Klasik Test Kuramı

MTK: Madde Tepki Kuramı

G Kuramı: Genellenebilirlik Kuramı

G Çalışması: Genellenebilirlik Çalışması

K Çalışması: Karar Çalışması

G Katsayısı: Genellenebilirlik Katsayısı

Phi Katsayısı: Güvenirlilik Katsayısı

b: birey

m: madde

p: puanlayıcı

b x m: birey x madde ortak etkisi

b x p: birey x puanlayıcı ortak etkisi

m x p: madde x puanlayıcı ortak etkisi

b x m x p: birey x madde x puanlayıcı ortak etkisi

BÖLÜMÜ

GİRİŞ

Eğitim girdiler, süreç, çıktılar ve değerlendirme olmak üzere dört temel öğeden oluşmaktadır (Baykul, 2000). Bu öğelerden girdileri öğrenci, öğretmen özellikleri eğitim programları vb. oluştururken, eğitim programındaki hedeflere ulaşmak için geçirilen yaşantılar ise süreci oluşturur. Süreç sonunda öğrenmenin gerçekleştiği kişilerde gözlenen öğrenme ürünlerinin tümü çıktıları oluşturmaktadır. Değerlendirme ise yapılan ölçmelerin sonuçlarını bir ölçütle karşılaştırarak karara varma olarak tanımlanabilir (Demirel ve Ün, 1987:44). Bu nedenle değerlendirme, eğitimin aksayan yönlerinin giderilmesi rolünü de üstlenmektedir. Ölçme ve değerlendirme bu özelliğiyle birlikte düşünüldüğünde sistemin önemli öğelerinden biri olduğunu söylenebilir. Ölçme günlük hayatta da önemli bir yer tutar. Birçok şeyin sayısal büyüklüğü hakkında bilgi sahibi olmak için ölçme araçlarından yararlanılır. Ölçme, sahip olduğu öğeler arasındaki dinamik ilişkiler bütünü olarak tanımlanabilir. Bu nedendir ki ölçme bir model olarak ele alınıp değerlendirilebilir (Atılgan, Kan ve Doğan, 2007). Ayrıca ölçme ve değerlendirme günlük yaşamın ayrılmaz bir parçası olan bilimsel çalışmaların da temelini oluşturmaktadır. Birçok bilim dalındaki gelişmelerin hız kazanması o bilim dalına özgü ölçme ve değerlendirme araçlarının geliştirilmesiyle mümkün olur.

Ölçmenin çoğu durumda üç basamakta gerçekleştiğini söyleyebiliriz. Birinci basamakta ölçme objesinin hangi niteliklerinin ölçüleceği belirlenir. İkinci basamakta, ölçme objesinin yapısına ve yapacağımız ölçmenin amacına uygun bir sayı veya sembol kümesi seçilir. Üçüncü basamakta ise ölçme kuralı veya puanlama kuralı kararlaştırılır. Seçilen bir ölçme aracı ile öğrenmenin ne oranda gerçekleştiğine dair bilgi toplanır. Öğrencilerden toplanan bilgiler, program, öğretimin kalitesi gibi başka eğitim kararlarında da kullanılır. Günümüzde eğitimde değerlendirmeye, öğrencilerin gelecekteki eğilimlerini belirleme ve öğretimi iyileştirme amacıyla bilgi toplama ve kullanma olarak bakılmaktadır. Bu bakış açısında ölçmeyi kullanma bir yargı ile sonuçlanır (Turgut ve Baykul, 2013).

Ölçme yapılırken farklı ölçme araçlarından yararlanılmaktadır. Bunun nedeni

yeni gelişen öğrenme teorilerinin farklı öğrenme modellerini oluşturmasıdır. Bu durum ölçme ve değerlendirmede de bazı anlayışların değişmesine neden olmuştur. Artık ölçme ve değerlendirmede sonuç yerine sürecin ölçülmesi, bilginin hatırlanması yerine bilginin kullanılması, günlük hayat problemlerinin çözülmesi, tek bir ölçme yöntemine bağlı kalmayıp birden fazla ölçme tekniği kullanılması önem kazanmıştır (McMillian, 1997). Bu nedenle farklı ölçme durumları için en uygun ölçme aracının kullanılması yapılan ölçme sonucunda doğru kararlara varmak için önemlidir.

Okulda öğrenilen derslere ilişkin kazanımların ölçülmesinde farklı ölçme araçları kullanılır. Bunlar, uzun yanıtı, kısa yanıtı, sözlü, doğru-yanlış eşleştirmeli ve çoktan seçmeli sınavlar olabilir. Bunların yanı sıra portfolyo ve performans değerlendirme gibi değerlendirmeler için ise dereceli puanlama anahtarı gibi araçlar kullanılmaktadır. Ölçme araçlarından hangisinin kullanılacağını belirlemede etkili faktörler ölçtüğümüz davranışların nitelikleri, uyguladığımız grubun yaşı, uygulanan gruptaki öğrenci sayısı olabilir. Bahsedilen ölçme değerlendirme araçlarından her biri farklı bir amaca hizmet eder. Örneğin öğrencinin kişisel çabalarını, ilerlemelerini ve başarılarını bir veya daha çok öğrenme aralıklarıyla ortaya koyması isteniyorsa portfolyolar bu amaç için uygundur.

Çoktan seçmeli sınavlar ise özellikle büyük bir kitleyi değerlendirip yerleştirmeyi esas alan sınavlarda kullanılır. Çoktan seçmeli maddeler, madde kökü ve seçenek olmak üzere iki kısımdan oluşur. Cevabın şansla bulunması olasılığı vardır. İstatistiksel yollarla şans başarısını azaltmak mümkünse de tamamen ortadan kaldırılamaz. Bu tip sınavlarda soru hazırlama diğerlerine göre zaman alıcıdır. Çünkü soru hazırlayan kişinin çoktan seçmeli soru hazırlamada beceri sahibi olması gerekir. Çok değişik madde tipleri olmasından ve farklı öğrenim düzeylerinde başarıyla kullanılması nedeniyle sıkça tercih edilir (Turgut ve Baykul, 2013).

Eğitimde çoktan seçmeli testlerin yanında yazılı sınavlar da kullanılır. Bu sınavlarda kullanılan soruların bir diğer adı da serbest yanıtı ya da açık uçlu sorulardır. Açık uçlu sorular kısmi bilgiyi puanlamaya imkân verir ancak puanlaması güçtür ve puanlayıcı kanısından etkilenir (Bağcan, Büyükturan ve Çıkrıkçı Demirtaşlı, 2013). Açık uçlu sorular, yaratıcı düşünme gücü bilgiyi örgütleme yeteneğinin düşüncelerin iç ve dış tutarlılığını değerlendirme gücü gibi özelliklerin kazanılıp kazanılmadığının

ölçülmesinde en uygun soru türüdür (Atılgan, Kan ve Doğan, 2007).

Açık uçlu maddeler seçenekleri olmayan öğrencin cevabı kendisinin yapılandığı ve bazı durumlarla gerekçeleriyle açıkladığı düşüncelerini daha özgür bir şekilde ifade etme imkânı bulduğu soru çeşididir (Gronlund, 1998). Açık uçlu maddeler seçeneksiz çoktan seçmeli maddeler değildir. Bu soru tipinde tek bir doğru cevap yoktur fakat doğru cevaplar önceden belli ve sınırlıdır. Açık uçlu maddelerden oluşan testler ile öğrencinin bilgiyi kullanma, gerekçelendirme ve algılama becerilerinin ne düzeyde olduğu çoktan seçmeli testlere göre daha iyi ölçülebilir ve öğretime rehberlik edilebilir (Badger ve Thomas, 1992).

Açık uçlu sorular, sınavın uygulandığı gruba cevabın niteliği, içeriği ve uzunluğunu belirlerken özgürlük tanır. Bunun yanında kısa cevaplı testlerle veya çoktan seçmeli sınavlarla öğrenci yeteneği belli bir bilişsel basamağa kadar ölçülürken, açık uçlu sınavlar daha üst bilişsel basamakları da ölçmeye imkân tanır. Çoktan seçmeli olarak sorulan bazı maddelerde öğrencinin cevabı olası doğru yanıt seçenekleri deneyerek bulma imkânı varken aynı madde açık uçlu olarak sorulduğunda böyle bir şansı olmamaktadır. Örneğin ' $2x+5=11$ ifadesinde x değeri kaçtır?' şeklinde sorulan bir matematik problemi açık uçlu olarak sorulduğunda işlem basamakları açıklanarak problem çözülmelidir. Oysaki aynı soru çoktan seçmeli olarak sorulduğunda seçenekler deneyerek de doğru cevap bulunabilmektedir. (Braun, Bennett, Frye ve Soloway, 1990).

Açık uçlu sorular, öğrencilerin öğrendiklerini düzenlemesi, önceki öğrendikleriyle birleştirme yeteneğinin ölçülmesi, bilgilerini yeni karşılaştığı problem durumları için kullanması ve yaratıcı fikirler üretmesi için elverişli ölçme aracıdır. Değerlendirme basamağındaki davranış ya da öğrenmeleri test etmek için de kullanılır. Çoktan seçmeli testlerde öğrenci bulduğu cevap seçeneklerde olmayınca işleminin yanlış olduğunu anlayıp tekrar yapma imkânı bulur fakat açık uçlu maddelerde bu durum söz konusu değildir (Bridgeman, 1991). Bunun yanında matematik için bazı çoktan seçmeli test maddelerinde özellikle denklemin değerini yerine koyup bulmayı gerektiren sorularda seçenekler deneyerek doğru yanıt bulunabilir. Fakat açık uçlu maddelerde bu durum ortadan kalkar. Sınavı alan kişinin ölçülmek istenen özelliğe sahip olup olmadığı anlaşılır. Ayrıca bir sorunun çoktan seçmeli ve açık uçlu formları çok farklı beceriler gerektirebilir.

Örneğin ‘Saatteki hızı ortalama 80 km/sa olan bir otomobil K kilometreyi kaç saniyede alır?’

A) K/80 B)K/45 C)45K D)288.000K

Sorusuna sayısal olarak nasıl çözeceği konusunda hiçbir fikri olmayan öğrenci saatte 80 km/sa hızın karayolunda bir araba için ortalama bir hız olduğunu bilip 1 yerine K koyarak doğru cevabı elde edebilir. Araba 1 km’yi 1/80 saniyede mi 1/45 saniyede mi yoksa 288.000 saniye de mi gider? Doğru cevap olabilecek tek seçenek arabanın 1 kilometreyi 45 saniyede almasıdır. Bu strateji hiçbir cevap seçeneği verilmeyen ve kesin bir sayısal cevap gerekli olan açık uçlu soru formatında doğru cevabı bulabilmek için işe yaramazdı (Bridgeman,1991).

Açık uçlu maddeler;

- Durumları karşılaştırma,
- Neden-sonuç ilişkisini açıklama,
- Verileri düzenleme
- Güçlü ve zayıf yönleri belirleme
- Problem için çözüm üretme gibi durumlarda kullanılabilir.

Açık uçlu maddelerin birtakım sınırlılıkları da vardır. Açık uçlu maddelerin puanlanması aşamasında, maddelere yönelik cevapların yazılı olarak verilmesi zorunluluğu, aslında ölçülmek istenmeyen özelliklerin (yazı güzelliği, düzenliliği gibi) de puanlamaya yansımaya sebep olabilir (Temizkan ve Sallabaş, 2011). Maddelere verilen cevapların objektif olarak puanlanmasında yaşanan bu zorluk ölçmenin güvenilirliğinin düşmesine neden olur. Açık uçlu soruların diğer bir sınırlılığı da kapsam geçerliğinin yeterli soru sayısı sağlanamadığı durumlarda düşük olmasıdır. Sınav süresinin kısıtlı olması nedeniyle açık uçlu maddelerden oluşan sınavlarda sorulabilecek soru sayısı çoktan seçmeli testlere göre daha azdır. Bu durum açık uçlu maddelerden

oluşan sınavlarda kapsam geçerliliğinin sağlanmasını zorlaştırmaktadır; açık uçlu maddelerden oluşan sınavlarda kapsam geçerliliğini düşürmektedir (Özçelik, 2011).

Açık uçlu maddelerin puanlamaları anahtarla, rubrikle ve serbest puanlama gibi farklı puanlama yöntemleriyle yapılabilir. TIMMS, PISA, PIRLS, ICILS gibi birçok uluslararası sınavda açık uçlu sorular kullanılmaktadır. Bu tür uluslararası sınavlarda uzun yanıt gerektiren maddelere verilen yanıtlar alanlarında uzman puanlayıcılar tarafından kodlanır. Açık uçlu uzun yanıtlı maddeler daha önceden bu amaç için hazırlanmış kodlama anahtarındaki yönergelerden yararlanılarak kodlanır. Tüm açık uçlu maddeler için verilen kodlar veri giriş yazılımına yüklenerek her bir öğrenci için puana dönüştürülür.

Bunun yanında ülkemizde yükseköğretime geçiş sınavlarında da kullanılması gündeme gelen açık uçlu sorular hazırlanırken, bu soruların nasıl puanlanacağını gösteren puanlama anahtarları her biri kendi alanında uzman akademisyenler tarafından hazırlanır. Hazırlanan puanlama anahtarlarında soruya verilebilecek tüm olası cevaplarla bu cevaplara verilebilecek puanlar yer alır. Sorunun cevabı olarak kabul edilemeyecek yanıt örüntüleri de puanlama anahtarında belirtilir. Hazırlanan sorular verilen cevaplara tam puan verilebileceği gibi bazı yanıtlara kısmi puan verilmesi de söz konusudur. Hazırlanan puanlama anahtarlarında sorunun çözüm aşamalarından hangi aşamaya ne kadar puan verileceği önceden bellidir. Değerlendirme sırasında, her bir soru için; soru metni adayın verdiği cevap ve soruya ait puanlama anahtarı puanlayıcıya bilgisayar ekranından gösterilmektedir. Verilen puanlar açık ve net olduğu için yanlış bir puanlama yapılması söz konusu değildir (ÖSYM, 2015).

Açık uçlu sorular olası cevapların öne sürülmediği, yanıtlayanın kendi cümlelerini kullanarak yanıtladığı sorulardır. Bu sorular cevabı açıklamaya teşvik edecek şekilde çerçevelendirilir. Sorunun cevabı cevaplayıcıya bağlı olarak bir cümle, bir paragrafla veya bir sayfa sürebilir. Bu nedenle açık uçlu soruların açık uçlu olmayan sorulara göre daha nesnel olma eğiliminde olduğu söylenebilir. Açık uçlu sorular genellikle nasıl, ne, ne zaman, nerede, neden gibi soru zarflarıyla başlayarak niceliksel bilgiden çok niteliksel bilgi sağlar. Açık uçlu soruların cevap örüntüleri çok çeşitlidir. Kuşkusuz tüm cevaplayıcıların aynı tarzda ifade kabiliyetine sahip olmamaları bu duruma nedendir. Yanıtlayıcının soruyu birden fazla olası cevapla yanıtlayabilmesi

durumunda açık uçlu sorular kapalı uçlu sorulara iyi bir alternatif olabilir. Buna karşın açık uçlu sorulara verilen yanıtların puanlarının güvenilirliği kapalı uçlu sorulara göre daha düşük olacaktır (Popping, 2008). Buna ek olarak açık uçlu maddelerin ölçme değerlendirmede kullanılan diğer maddelere göre uygulanması ve puanlanmasının zaman alıcı olduğu söylenebilir.

Açık uçlu maddelerden oluşan sınavlar sıralama, sınıflama, analitik ve genel izlenimle puanlama yöntemlerinden biri ile puanlanabilir. Kullanılan bu yöntemler arasında genel izlenimle puanlamada, puanlayıcı, bir cevap kâğıdının tamamını okuduktan sonra onun kendisinde bıraktığı izlenime göre, onu daha önceden belirlediği birkaç puandan grubundan birine koyar. Bu puanlama şekli, puanlama standartları açıkça belli olmadığından dolayı öğrencinin puanına neden eksik puan verildiği konusunda bir bilgi sağlamaz. Sorular puanlanırken belirlenen puanlama standartları öğretmenlerden öğretmene ve hatta bir öğretmende değişik zamanlarda farklılıklar gösterir. Bu nedenle genel izlenimle puanlama yöntemiyle elde edilen puanların güvenilirliğinin düşük olması beklenir (Tekindal, 1997).

Anahtarla puanlamada ise sorulara verilebilecek cevaplar önceden hazırlanır ve her cevap en ince ayrıntılarına kadar analiz edilir. Her bir öğeye verilebilecek en yüksek ve en düşük puanlar belirlenir, cevaplayıcının yazdığı cevaplar puanlama anahtarıyla karşılaştırılarak cevaplara ne derece uyduğuna göre onlara bir puan verilir. Anahtarla puanlama yönteminin diğer yöntemlere göre daha güvenilir puanlar vermesi beklenir.

Görüldüğü gibi ölçme aracının yapısı ve puanlamanın biçimi, puanlamanın hatasız yapılıp yapılmamasına etki eder. Kısa cevap gerektiren veya uzun cevap gerektiren sınavlarda olduğu gibi, bir cevabın doğru olup olmadığını tayin etmek puanlayıcının takdirine kalırsa, puanlama yöntemi objektif olmaz. Bu gibi durumlarda puanlamalara takdir farklılıklarından kaynaklanan puanlama hataları karıştığı için puanlama güvenilirliği sağlanamayabilir. Bu anahtarla puanlanabilen seçmeli testlerde olduğu gibi, puanlayıcının görevi, seçilen cevabın anahtara uygun olup olmadığını tayin etmekten ibaret ise, puanlama yöntemi objektiftir.

Açık uçlu sorulara verilen cevaplarda puanlayıcılar içi ve puanlayıcılar arası güvenilirliği arttırmak için cevapların puanlandırılmasında dereceli puanlama anahtarı (rubrik) kullanmak tercih edilen bir yöntemdir. Valenza'ye (2000) göre dereceli

puanlama anahtarı bir performans açısından önemli olan davranış ve görevleri örneklendiren puanlama araçları iken, Jonsson ve Svingby'a (2007) göre karmaşık yapıda öğrenci görevlerinin puanlanmasında kullanılan puanlama araçlarıdır. Dereceli puanlama anahtarında kullanılacak ölçek bütünsel ya da analitik olabilir. Puanlamanın objektif olmadığı sınavlarda, analitik veya bütünsel olarak hazırlanan bir puanlama anahtarı puanlamanın objektifleşmesine katkıda bulunur. Ürünlerin bir bütün olarak değerlendirilmesi durumunda kullanılması uygun olan bütünsel puanlama anahtarı, öğrenci performanslarına tek bir puan verir (Turgut,1997). Analitik puanlama anahtarı ise öğrencinin hangi performans basamağında eksiği olduğunu açıklayabilir. Çünkü performansın alt bölümlerinin tek tek puanlanmasından sonra toplam bir puan hesaplanması esasına dayanır (Popham, 1997). Bütünsel ve analitik puanlama anahtarlarının farklı kullanım alanları vardır. Verilen görevin öğrencilerin yaratıcılığını yansıtmaması gerektiğinde ya da kesin ve açık bir doğru olmadığı durumlarda bütünsel bir değerlendirme daha uygundur. Görevin bir kısmındaki hatalar genel sonucu etkilemiyorsa da bütüncül puanlama anahtarı kullanılabilir (Kan, 2007). Puanlayıcı sayısı göz önüne alındığında, öğrenci çalışmalarını (ürün) birden fazla öğretmen puanlayacaksa, her bir öğretmen ortaya çıkan ürünün kabul edilebilir koşulları hakkında farklı fikirlere, yorumlara ve anlayışa sahiptir. Bu nedenle görevde bulunması gereken bileşenlerin hangi açıdan değerlendirileceği konusunda standart olması önem kazanır. Bu gibi durumlarda ise analitik puanlama anahtarı bütünsel puanlama anahtarına tercih edilir (Atılğan, Kan ve Doğan, 2017).

Aşağıda verilen Tablo 1 ve Tablo 2’de sırasıyla ortaokul ve ilkokul düzeyinde iki farklı matematik sorusu ve bu soruları puanlamak için kullanılan puanlama anahtarı (rubrik) örneği verilmiştir.

SORU: Pazarlamacı Aslı Hanım, maliyeti a TL olan bir malı b TL ‘ye satmaktadır. a ile b arasında $b=6a-60$ bağıntısı vardır. Aslı Hanım’ın bu pazarlamadan kar ettiği biliniyor. Buna göre malın maliyet fiyatı olan a sayısının alabileceği değerler için ne söyleyebilirsiniz, işlem yaparak ve açıklayarak yazınız.

Tablo 1. Açık Uçlu Matematik Sorusunun Puanlanmasına Uygun Rubrik Örneği 1

3 Puan	Problem tam olarak anlaşılmış ve uygun çözüm yolu kullanılarak doğru cevaba gidilmiştir. -Öğrenci Aslı Hanım'ın kar edebilmesi için satış fiyatı olan b 'nin alış fiyatı olan a ' dan büyük olması gerektiğinin farkındadır. $b=6a-60$ bağıntısını kullanarak $6a-60>b$ eşitsizliğini yazmıştır. Bu eşitsizliği hatasız bir biçimde çözerek satış fiyatının alış fiyatından fazla olması için olması için $a>12$ olmalı diye yazarak doğru cevaba ulaşmıştır.- Aslı Hanım'ın kâr zarar durumuna ilişkin yapılan işlemler açık, ayrıntılı ve anlaşılırdır.
2 Puan	Problem çoğunlukla anlaşılmıştır. -Çözüm genel olarak doğru olup yalnızca küçük hatalar bulunmaktadır. Aslı Hanım'ın kar edebilmesi için $b>a$ olmalıdır şeklinde uygun strateji kullanılmıştır. $b=6a-60$ bağıntısı kullanarak çözüm $6a-60>a$ şeklinde devam ettirilmiştir. Ancak eşitsizliği çözerken küçük işlem hatalarından veya anlaşılamayan nedenlerden dolayı sonuca ulaşılammış ya da yanlış sonuca ulaşılmıştır.
1 Puan	Problem bazı yönlerden anlaşılmıştır. -Öğrenci yalnızca uygun çözüm yolu ile başlamış ama devamını getirememiştir. Örneğin, "Aslı Hanım'ın kar edebilmesi için $b>a$ olmalıdır ya da satış fiyatı alış fiyatından fazla olmalı" ifadelerini kullanarak çözüme uygun şekilde başlamış olabilir. Ancak çözüm yolu devam ettirilerek $6a-60>a$ eşitsizliği kurulup çözülememiş olabilir. - Uygun çözüm yolu ile $b>a$ eşitsizliği kurulmuştur. Ancak, $b>a$ eşitliği kurulduktan sonra problemin çözümüne yönelik doğru işlemler yapılmamıştır ya da yapılan işlemlerde birinci dereceden bir bilinmeyenli eşitsizliklerin çözümü ile ilgili önemli işlem hataları bulunmaktadır.

0 Puan	Öğrenci problemi anlamamıştır -“Problemde iki bilinmeyen olduğu için işlem yapamayacağına dair ifadeler kullanmıştır.- Öğrenci problemin çözümünü bulmaya yönelik herhangi bir işlem yapmamıştır.- Öğrenci kâr zarar durumunu açıklamaya yönelik herhangi bir işlem yapmamıştır.- Problemin tekrarı niteliğinde ifadeler kullanmıştır.- Sonuca ulaşmamız mümkün olmayan yanlış eşitlikler kurup onları çözmeye çalışmıştır.
--------	---

SORU: Deniz kolye yapmak için çeşitli boncukları aşağıda gösterildiği gibi ipe dizmektedir.



- Deniz kolyeye devam ederse dizeceği ilk üç boncuk hangi geometrik şekilde olur.
- Deniz yukarıdaki şekle 14 daire şeklinde boncuk daha dizmek için kaç tane daha kare şeklinde boncuk kullanılabilir? Çözüm yolunuzu göstererek cevabı açıklayınız.

Tablo 2. Açık Uçlu Matematik Sorusunun Puanlanmasına Uygun Rubrik Örneği 2

3 Puan	Cevaplayıcı Deniz'in dizeceği ilk üç boncuğu daire, daire, kare şeklinde yazar veya şekillerini çizerse: 7, çünkü her 2 daire için 1 tane kare kullanılır. ($14: 2 = 7$) ya da 6, çünkü DDK DDKDDKDDKDDKDDK DD şeklinde de dizilebilir şeklinde cevaplarsa 3 (tam) puan verilir.
2 Puan	Cevaplayıcı Deniz'in dizeceği ilk üç boncuğu doğru şekilde dizer veya söylerse; Deniz'in kullanacağı kare şeklindeki boncuk sayısını 6 veya 7 olarak söyler ama yanlış veya eksik bir çözüm yolu izlerse; Cevaplayıcı Deniz'in dizeceği ilk üç boncuğu doğru şekilde dizer veya söylerse; Deniz'in kullanacağı kare şeklindeki boncuk sayısını yanlış hesaplar fakat doğru bir çözüm stratejisi uygularsa; Cevaplayıcı Deniz'in dizeceği ilk üç boncuğu doğru şekilde dizer veya söylerse; Deniz'in kullandığı toplam boncuk sayısını 14 boncuk olarak düşünerek bunların içlerinde 4 ya da 5 tane kare şeklinde boncuk vardır cevabını verirse 2 puan verilir.
1 Puan	Cevaplayıcı Deniz'in dizeceği ilk üç boncuğu doğru şekilde dizer veya söylerse; Deniz'in 6 veya 7 daire şeklindeki boncuk kullanması gerektiğini söyler fakat bunu yanlış bir yol izleyerek gösterirse 1 puan verilir.
0 Puan	Verilen cevapta problemin matematiksel olarak anlaşılamadığını gösteren Temel hatalar var ise 0 puan verilir.

Açık uçlu sınavlarda puanlamada belli bir anahtar kullanılsa da farklı puanlayıcılar tarafından yapıldığı için ölçme sonuçlarına hata karışabilir. Kaynağı tam olarak belirlenemeyen bu hatalara tesadüfi hatalar denir. Bir ölçme aracının tesadüfi hatalardan arınlık düzeyi ise güvenilirlik olarak adlandırılır. Açık uçlu sorularda soruların yanı sıra puanlayıcılardan da ölçümlere hatalar karışır ki hata kaynağı olarak puanlayıcılar ve puanlayıcılarla maddelerin ortak etkileri güvenilirlik için önemli bir sorun oluşturmaktadır (Brennan, 2001).

Ölçme sonuçları tesadüfi hatalardan ne kadar arınıksa ölçme aracı o kadar güvenilirlerdir (Baykul, 2000). Güvenilirlik, aynı testin tekrarlanan uygulamaları veya testin paralel formları üzerindeki değişen performansların tutarlılığını ifade eder (Crocker, Algina, 1986). Güvenirlik katsayısı 0 ile 1 arasında değerler alır. 0'a yaklaştıkça tesadüfi hata oranı artar; 1'e yaklaştıkça azalır. Ölçmede güvenilirlik; tutarlılık, duyarlılık ve kararlılık, objektiflik anlamlarını taşıyabilir. Bir testin iki formu aynı özelliği tutarlı bir şekilde ölçüyorsa, uygulandığı grupta yer alan bireylerin iki formdan elde ettiği puanlar arasında mükemmel bir korelasyon olacaktır. Bu durum tutarlılık anlamında güvenirliliği yansıtır. Duyarlılık, daha çok ölçme aracının birimleri ile ilgilidir. Birimleri küçük olan ölçme aracı veya sonucu, birimi büyük olana göre daha duyarlı olması nedeniyle daha güvenilirlerdir. Güvenirliğin kararlılık anlamı ise bir özelliğin, aynı ölçme aracının benzer durumlarda birden fazla ölçme durumunda kullanılması sonucu elde edilen ölçüm sonuçlarının birbiriyle tutarlı olmasıdır. Yani tekrarlı ölçümlerde, aynı özelliğin ölçülmesinde elde edilen sonuçlar arasındaki fark büyüdükçe güvenilirlik azalır (Atılğan, Kan ve Doğan,2007).

Güvenirlik uygulanan test sonuçlarından elde edilen standart sapmaların, z puanlarının aynı testin tekrar uygulanması veya alternatif bir testin uygulanması sonucunda elde edilen puanlarla tutarlı olması olarak tanımlanabilir. Test puanlarının güvenirliliğini düşüren nedenlerden biri de test maddelerinin ölçülmek istenen davranışları iyi örnekleyememesidir. Bu nedenle test puanları bu koşullar altında yanıltıcı olabilir ve ölçme hatası oluşabilir. Ölçme hataları sistematik hata, sabit hata ve tesadüfi hata olarak sınıflandırılabilir (Crocher ve Algina, 2006). Kaynakları iyi bilinmeyen ve ölçmeye hangi yönde karıştıkları belli olmayan hatalar tesadüfi hata olarak adlandırılır. Ölçmede sınav kâğıtlarındaki cevapları dikkatsizce okuyan bir

puanlayıcının puanlarında tesadüfi hatalar bulunur. Belli bir gruba veya belli bir özelliğe daha fazla puan verilmesi de sistematik hata olarak nitelendirilebilir. İstatistik dilinde bu hataya yanlılık da denir. Sabit hata ise ölçmeden ölçmeye değişmez, kaynağı bellidir ve ayıklanabilir. Örneğin bir sınavda sınavın uygulandığı tüm bireylere beş puan eklenmesi sabit hataya neden olur. Bir ölçme sonucunda bu hatalardan herhangi biri bulunabileceği gibi üçü de bulunabilir. Sistematik ve sabit hataların kaynağı ve yönü bellidir fakat tesadüfi hatalar kaynakları kesin olarak bilinemeyen ve miktarları hesaplanamayan, ancak kestirilebilen hata türüdür (Baykul ve Turgut, 2013). Güvenilirlikte, hata varyansının kaynağı maddeler arası tutarsızlık, paralel maddeler ve testler arasındaki performans farklılıkları, benzer ya da paralel testlerden elde edilen performansın zamana göre farklılık göstermesi olabileceği gibi iki veya daha fazla puanlayıcının her bir maddeyi puanlamaları arasındaki farklılıklar da olabilir. Bu nedenle eğitimde açık uçlu soruların puanlanmasına öznellik karışması durumunda puanlayıcıların verdiği öznel puanları ve bu puanlara dayanarak verilen kararların doğruluğunu değerlendirmek için farklı puanlayıcıların maddelere verdikleri puanların birbiriyle ne derece tutarlı olduğunun belirlenmesi gereklidir (Atılğan, Kan ve Doğan, 2007). Klasik Test Kuramında, uyuşma yüzdesi, Kappa istatistiği, Pearson korelasyon katsayısı ve ortalamaların karşılaştırılması gibi farklı teknikler puanlayıcı güvenilirliğini tespit ederken kullanılabilir (Goodwin, 2001; İlhan, 2016). Pearson korelasyon katsayısı muhtemelen en yaygın olarak kullanılan korelasyonel istatistik tekniğidir. Bu teknik her iki değişkende eşit oran veya eşit aralık ölçeğindeyse uygulanması uygundur (Goodwin, 2001).

Klasik test kuramına göre güvenilirlik tanımını aşağıdaki eşitliklerden yararlanarak yapılabilir. Bir ölçme aracındaki herhangi bir maddeye verilen yanıtlar (X_i); öğrencinin bu maddenin ölçmeye çalıştığı özelliğe sahip olma düzeyi (T_i) ve hata terimleri (E_i) ile açıklanır. Bu bağıntı klasik test kuramının modeliyle ifade edilir.

$$X_i = T_i + E_i \quad (1)$$

Eğer i. madde için rastgele hatalardan arınık bir ölçme yapıldıysa öğrencinin ölçülmek istenen özelliği ölçme aracından elde edilen puanlarla aynı ($X_i = T_i$) olacaktır.

Bu durumda elde edilen ölçme sonuçları tam güvenilir (perfectreliability) olarak tanımlanır (Yurdugül, 2006). Diğer bir deyişle gözlenen puanların varyansı, gerçek puanların varyansı ile hata puanlarının varyansına eşittir.

Eşitlik 2 gerçek puan varyansının hata varyansı ile gözlenen puan varyansına bağlı olduğunu gösterir. Hem gerçek hem de hata puanlarındaki ortalama puanın ve standart sapmanın aşağıda gözlenen miktarlarla ifade edilebileceği gösterilmiştir.

$$M_T = M_X$$

$$S_T = S_X \sqrt{r_{X_g \cdot X_h}}$$

$$s_e = S_X \cdot \sqrt{1 - r_{X_g \cdot X_h}}$$

(3)

Eşitlik 3 hem gerçek puanın hem de hata puanının ortalaması ve standart sapmasının yukarıdaki gibi gözlemlenebilir miktarlarla ifade edilebileceğini göstermektedir.

$$r_{xt} = \sqrt{r_{X_g} \cdot r_{X_h}}$$

(Güvenirlilik indeksi)

(4)

$$r_{xe} = \sqrt{1 - r_{X_g} \cdot r_{X_h}}$$

Eşitlik 4 gözlemlenen puanlar ile gerçek ve hata puanları arasındaki korelasyonun yukarıdaki gibi gözlemlenebilir miktarlarla ifade edilebileceğini göstermektedir

$$T_{ig} = T_{ih}$$

$$S_{t_g} = S_{t_h}$$

$$S_{e_g} = S_{e_h}$$

(5)

Eşitlik 5’ de gerçek puan tanımından, bir kişinin aldığı belirli testlerden bağımsız olarak puanın aynı olduğunu ve sonuç olarak bir dizi paralel test için gerçek varyansın aynı olduğu görülür. Gözlenen puanın varyans değeri ve gerçek puanların varyansları bir dizi paralel testin her biri için aynı olduğundan, her bir testin hata varyansının aynı olması gerektiği anlaşılmaktadır (Gulliksen, 1950).

Yukarıda açıklandığı gibi klasik test kuramı güvenilirlik tanımını bireylerin gerçek puanları üstünden yapmıştır. Bireylerin gerçek puanını bilmediğimiz için bu tanım kullanışlı değildir (Webb, Shavelson ve Haertel, 2006). Klasik güvenilirlik kuramı, tek bir gerçek puana sahip her bir gözlem ya da test puanının paralel test puanlarının bir grubuna ait tek bir güvenilirlik katsayısı gelmesi fikri etrafında toplanır. Bu model, paralel formlar eşitlendiğinde makulken, örneğin puanlayıcıların merkezi eğilim ve varyansları farklı olduğunda, gözlemlerden oluşan kapsama bağlı ve yapılar açık olarak heterojen olduğunda gerçekçi olmaz ve oldukça kısıtlayıcıdır (Baykul, 2000; Lord ve Novic, 1968).

Öğrencinin cevabı kendisinin yapılandığı, üst düzey zihinsel süreçlerin ölçüldüğü açık uçlu maddelerden oluşan testlerin veya öğrencilerin kendi başlarına özgün bir ürünü ortaya koyabildikleri performans görevlerinde güvenilirlik belirlemede klasik test kuramı, madde tepki kuramı ve genellenebilirlik kuramı gibi yöntemlere başvurulabilir. Bunlardan klasik test kuramı son yıllarda daha güvenilir sonuçlar vermesinden dolayı yerini genellenebilirlik kuramı ve madde tepki kuramına bıraksa da temel yaklaşımlardan biridir.

Klasik test kuramında güvenilirlik katsayısı, gerçek puan varyansının gözlenen puan varyansına oranı olarak tanımlanır. Gözlenen puan varyansı, gerçek puan varyansı ve hata puanı varyansı olmak üzere iki farklı varyans kaynağından oluşmaktadır. Gerçek varyans dışındaki tüm varyansın farklı hata kaynaklarından gelebileceği düşünülür (Güler, 2008). KTK güvenilirlik kestiriminde gerçek puanlardan hata puanlarının ayırt edilmesi için çeşitli yollar sağlar. Bir kişinin gerçek puanı bu testi sonsuz sayıda alması ile elde edilen puanların ortalaması alınarak elde edilebilir. Sınırsız sayıda test puanı elde etmek imkânsız olduğu için gerçek puan kavramı hipotetiktir (Kline, 2005). KTK adı altında güvenilirlik, iki puan seti arasındaki korelasyonun hesaplanmasıyla veya

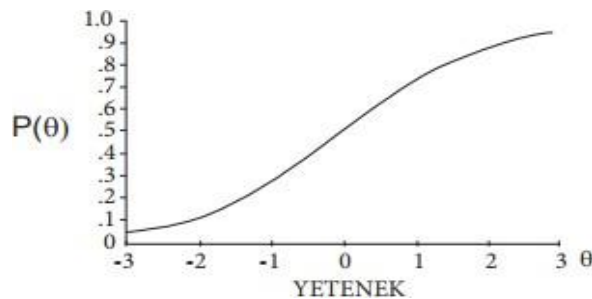
farklı puan setlerinin varyansına dayanan Cronbach alfa katsayısı hesaplanarak tahmin edilebilir (Bachman, 1990). Bu nedenle Cronbach alfa değeri ne kadar yüksek olursa, testin tutarlılık düzeyi o kadar iyi olur. Eğer ölçme aracı, doğru yanıtla 1 ve yanlış yanıtla 0 puan vererek, iki kategorili puanlanabilen maddelerden oluşuyorsa KR-20 güvenilirlik katsayısı hesaplanabilir. Aynı mantık üzerine kurulu ve ona benzeyen Cronbach alfa güvenilirlik katsayısı ise çoklu puanlanabilen maddelerden oluşan testlere de uyarlanabilir (Crocker, Algina, 1986; Shavelson and Webb 1991; Atılğan, Kan ve Doğan, 2014)

KTK'da gerçek puan sonsuz çoklukta ölçmelerin ortalaması alınarak oluştuğu için bu ölçme durumlarında ölçülen bireylerin aynı kaldığı, değişmediği varsayılır. KTK hatayı en aza indirebilmek için aynı özelliği birden fazla ölçer. Bunun nedeni hatanın tesadüfi bir değişken olması ve ancak sayısız ölçme durumunda ortalamasının sıfır olmasıdır (Hambelton ve Swaminathan, 1985). Ayrıca KTK puanlayıcılar arası güvenilirlik hesaplamak için Kappa istatistiği, Uyuşma Yüzdesi, Pearson Momentler Çarpımı Korelasyonu ile ortalamaların karşılaştırılması gibi farklı yöntemler kullanılmaktadır (Goodwin, 2001). Performansın ölçülmesinde önemli bir hata kaynağı olan puanlayıcıya bağlı olarak güvenilirlik hesaplamasında puanlayıcı-içi güvenilirlik ve puanlayıcılar-arası güvenilirlik hesaplama yöntemleri olarak ikiye ayırabiliriz (Alharby, 2006; Güler, 2008). Puanlayıcı-içi güvenilirlik, bir puanlayıcının farklı zamanlarda aynı bireylerin kâğıtlarının birden fazla puanlanmasıyla elde edilen puanlar arası korelasyonun hesaplanması ile elde edilir (Güler,2008). Bir puanlayıcının iki kez puanladığı performanslarının uyumunun basit yüzdesi tutarlılık anlamında güvenilirlik için bir ölçüttür. Farklı puanlayıcıların aynı bireyleri puanlaması sonucu ortaya çıkan puanların korelasyonu ise puanlayıcılar-arası güvenirliliği verir. Puanlayıcılar arası güvenilirlik hesaplamak için Kappa istatistiği, Uyuşma Yüzdesi, Pearson Momentler Çarpımı Korelasyonu ile ortalamaların karşılaştırılması gibi farklı yöntemler kullanılmaktadır (Goodwin,2001).

Klasik Test Kuramının güvenilirlik tahminleri, söz konusu test puanlarının güvenirliliğinin ve kalitesinin saptanmasında yararlıdır. Bununla birlikte bu kestirimlerin bazı sınırlılıkları vardır. Öncelikle, her Klasik Test Kuramı tahmini, aynı anda sadece bir ölçüm hatası kaynağını belirleyebilir ve bu nedenle, birden fazla hata kaynağının

etkileri ve bunların nasıl farklılık gösterdiği hakkında bilgi sağlayamaz. İkinci olarak Klasik test Kuramı tüm hataları rastgele veya "tek boyutlu" olarak değerlendirir ve bu nedenle Klasik Test Kuramı güvenilirlik tahminleri sistematik ölçme hatalarını rastgele ölçme hatalarından ayırmaz. Son olarak, Klasik Test Kuramı sınavı alan tüm bireyler için standart ölçme hatası için tek bir kestirime sahiptir. Klasik Test Kuramının bu sınırlamaları, madde tepki kuramı (MTK) ile ele alınmaktadır (Haiyang, 2010).

Madde Tepki Kuramı, testi alan bireyin o teste ilişkin performansının, yeteneklerinin tanımlanarak; bunlara bağlı olarak bireylerin yetenek puanlarının tahmin edilerek madde ve test performanslarının kestirilebileceğini söyler (Hambelton ve Swaminathan, 1985). MTK' ya göre kişilerin test maddelerine verdikleri yanıtlar ile doğrudan gözlenemeyen yetenekleri veya özellikleri arasında matematiksel bir ilişki vardır. Madde Tepki Kuramına göre kişilere uygulanan testler sonucu elde edilen yetenek kestirimleri, bireylere uygulanan testlerden bağımsız olarak elde edilebilmektedir. Başka bir ifadeyle seçilen modelle veri seti arasında uyum sağlandığında Madde Tepki Kuramı modelleri değişmez yetenek kestirimleri ve madde parametreleri elde etmemizi sağlar (Hambelton ve Swaminathan, 1985; Çelen, 2008). Madde Tepki Kuramı bireyin örtük yetenekleri ile ölçme aracında kullanılan maddeler arasındaki ilişkiyi matematiksel bağıntılarla göstermeye çalışan bir kuram olarak geliştirilmiştir. Madde tepki kuramının temel kavramlarından bir tanesi de madde karakteristik eğrisidir. Madde karakteristik eğrisinin grafiği testi alan kişinin bir soruya doğru cevap verme olasılığının farklı yetenek düzeylerine göre dağılımını gösteren fonksiyonun grafiğidir (Crocker ve Algina, 1986).



Şekil 1. Tipik bir madde karakteristik eğrisi

Madde karakteristik eğrisinde maddeyi tanımlamak için kullanılan iki teknik özellikten birincisi madde güçlüğü, ikincisi ise madde ayıricılık indeksidir. Madde tepki kuramında madde güçlüğü, maddenin yetenek ölçeğinin hangi noktasında işlevsel olduğunu ifade eder. Örneğin düşük yetenek düzeyindeki katılımcılar için kolay bir madde daha fonksiyonelken, yüksek yetenek düzeyindeki katılımcılar için zor bir madde daha fonksiyoneldir. Madde ayırt ediciliği ise maddenin bulunduğu yerin altındaki yetenek düzeylerine sahip bireyler ile üzerindeki yetenek düzeylerine sahip bireyleri birbirinden ne kadar iyi ayırt edilebildiğini yansıtır. Şekil 1 incelendiğinde madde ayırt ediciliği, madde karakteristik eğrisinin orta noktasındaki dikliği ifade eder. Daha dik olan eğriler, madde ayırt ediciliğinin daha yüksek olduğunu gösterir. Düzeye daha yakın olan eğriler ise maddenin ayırt ediciliğinin düşük olduğuna işaret eder (Baker, 2016).

Güvenirlilik belirlenirken hata kaynaklarını ve bu hata kaynakları arasındaki etkileşimleri bir arada değerlendirebilen yöntemlerden biri de Genellenebilirlik (G) Kuramıdır. Genellenebilirlik kuramı davranışsal ölçmenin güvenirliğiyle ilgili istatistiksel bir teoridir (Shavelson ve Webb, 1991).

G kuramı, ölçüm koşullarının testler ve ölçme araçları üzerindeki etkilerini kestirebilmenin çok yönlü ve etkili bir yoludur. Tek bir desenle, araştırmacı eşzamanlı olarak maddeler, puanlayıcılar, ölçme koşulları gibi yönlerden kaynaklanan hata varyanslarını tahmin edebilmektedir. Özellikle, ortak etkileri de değerlendirilebilir. Buna karşın, klasik test teorisi aynı anda sadece bir kaynaktan gelen ölçüm hatalarını değerlendirmeye olanak tanır. G kuramının diğer birincil katkısı araştırmacıyı genellikle göz ardı edilen ölçüm esaslarına yöneltmesidir. Örneğin, araştırmacı, genellemek istediği kabul edilebilir gözlemler evrenini açıkça anlamalıdır. Bunu yaparken, araştırmacı, herhangi bir kişinin puanını etkileyebilecek ölçüm koşullarını açıklamalıdır. Bu koşullar kabul edilebilir gözlemler evrelerini, ölçme aracının genellenebilirliğinin sınırlarını tanımlar ve araştırmacıyı ölçmeden kaynaklı hata kaynaklarına karşı uyarır (Kraiger, 1989).

Klasik test kuramının farklı yöntemleriyle farklı anlamlarda elde edilen güvenirlik katsayılarının birbirinden farklı, çelişir ve çoğu zaman açıklanmasının güçlüğünden hareketle, Cronbach ve arkadaşları, tesadüfi etki varyans bileşenleri

analizini kullanarak G kuramını geliřtirmişlerdir. Alternatif olarak puanlayıcı, zaman, test formu, madde, görev gibi bir ölçmedeki bütün potansiyel hata kaynaklarından gelen potansiyel hataları birlikte ve eş zamanlı değerlendirebilen bir yaklaşımdır.

G kuramında kullanılan temel kavramlardan biri yüzey diğeri ise koşuldur. G kuramında yüzey (facet), zaman, madde ya da görev puanlayıcı gibi ölçmenin benzer durumlarının setine verilen isimdir (Brennan, 2001). Bu çalışmada ise puanlayıcılar ve maddeler birey yüzeyken her bir madde ve her bir puanlayıcı da koşulları oluşturmaktadır. Ayrıca G kuramında bağıl ve mutlak olmak üzere iki farklı değerlendirmeden söz edilebilir. Bağıl değerlendirme, bireylerin ölçme sonuçlarına göre, ölçmenin uygulandığı grup içinde sıralanması olarak tanımlanırken; mutlak değerlendirme bireyin yapılan bir ölçme sonucuna göre aldığı puanı, daha önceden belirlenmiş bir ölçüte göre açıklayan değerlendirme türüdür. Bağıl değerlendirmelerde kullanılan G katsayısı evren puanı varyansının gözlenen puan varyansına oranıdır (Shavelson ve Webb, 1981). Bu amaçla öncelikle bağıl hata varyansının belirlenmesi gereklidir. Bağıl değerlendirme üç tür hata kaynağından etkilenir, (a) birey-madde ortak etkisi, (b) birey-puanlayıcı ortak etkisi ve (c) kalan varyans. Bağıl hata terimi “ δ ” sembolü ile gösterilmek üzere bağıl hata varyansı;

$$\sigma_{\delta}^2 = \frac{\sigma_{bm}^2}{n_m} + \frac{\sigma_{bp}^2}{n_p} + \frac{\sigma_{bmp,e}^2}{n_m n_p}$$

eşitliği ile hesaplanır. G katsayısı ise;

$$G = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_{\delta}^2}$$

eşitliği ile hesaplanır (Atılğan, 2005).

Gözlenen puan varyansı da evren puanı varyansı ve bağıl hata varyansının toplamına eşittir ise, evren puanı varyansının gözlenen puan varyansına oranıdır (Shavelson ve Webb, 1991; Brennan, 2001). Bu amaçla öncelikle mutlak hata teriminin belirlenmesi gereklidir. Mutlak değerlendirmenin etkilendiği varyans bileşenleri; (a) madde (b) puanlayıcı ana etkileri, (c) madde-puanlayıcı, (d) birey-madde, (e) birey-puanlayıcı ortak etkileri ve (f) kalan varyanstır. Mutlak hata terimi “ Δ ” sembolü ile

gösterilmek üzere mutlak hata varyansı;

$$\sigma_{\Delta}^2 = \frac{\sigma_m^2}{n_m} + \frac{\sigma_p^2}{n_p} + \frac{\sigma_{bm}^2}{n_m} + \frac{\sigma_{bp}^2}{n_p} + \frac{\sigma_{mp}^2}{n_m n_p} + \frac{\sigma_{bmp,e}^2}{n_m n_p}$$

eşitliği ile gösterilir. Phi katsayısı ise;

$$\phi = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_{\Delta}^2}$$

eşitliği ile hesaplanabilir (Atılğan, 2005).

G kuramında verilerin çaprazlanmış (crossed) ya da yuvanmış (nested) olmaları da desenin belirlenmesi ve analizlerin yapılmasında farklılıklar doğurur. Bir yüzeyin bütün koşulları diğer bir yüzeyin bütün koşullarını gözlüyorsa çaprazlanmış olarak adlandırılır ve yüzeyler arasına “x” işareti konularak gösterilir (Shavelson ve Webb, 1991; Brennan, 2001). On maddeden oluşan ve 25 kişiye uygulanan ölçme aracının iki öğretmen tarafından puanlanmasında on maddenin her birini 25 öğrencinin tamamının yanıtlaması söz konusu ise ve 25 öğrencinin yanıtladığı on maddenin her birini her bir öğrenci için iki öğretmen de puanlıyorsa bu ölçme durumunda çaprazlanmış desen bulunur. Çaprazlanmış desenlerden farklı olarak bir değişkenlik kaynağının sadece bazı koşulları diğer bir yüzeyin sadece bazı koşullarınca gözleniyorsa yuvalanmış (nested) olarak adlandırılır ve iki yüzey arasına “:” işareti konularak gösterilir (Shavelson ve Webb, 1991; Brennan, 2001).

G kuramında değişkenlik kaynaklarının özelliklerine göre sabit ve tesadüfî etkiden söz edilebilir. Araştırmacının örneklemdaki koşulları genellenen evrendeki diğer koşullarla değiştirebilir ise, bu yüzey tesadüfî yüzey olarak adlandırılabilir (Kieffer, 1998). Eğer araştırmacının amacı örneklemin ötesinde bir genelleme yapmak değilse ele alınan değişkenlik kaynağı sabit olarak tanımlanır (Güler, Uyanık ve Teker; 2012). Yüzeyin sabit ya da tesadüfî seçilmesi genellenebilirliği etkiler. Bazı yüzeylerin sabit olarak ele alınması hata varyansını azaltıp G ve Phi katsayılarını artırır fakat yapılan ölçmenin sonundaki yorumlarımız daha sınırlı kalır (Brennan,2011).

Güvenirlilik hesaplanırken G kuramı ile yüzey olarak tanımlanan bütün değişkenler (zaman, maddeler, puanlayıcılar) dikkate alınarak ve birbirleri ile olan etkileşimleri dikkate alınabilmektedir. G Kuramıyla güvenirlik hesaplanırken bir hata kaynağının olduğu durumlar tek yüzeyli evren; iki hata kaynağının olduğu durumlar iki

yüzeyle evren; üç veya daha fazla hata kaynağının olduđu durumlar ise üç veya daha fazla yüzeyle evrenler olarak tanımlanır.

G Kuramında Genellenebilirlik (G) çalışması ve Karar (K) çalışması olarak iki temel çalışma güvenilirliğin araştırılmasında kullanılmaktadır. G Kuramında G çalışması ile güvenilirlik katsayılarının (G ve Phi) hesaplanması için gerekli değerler sağlanmakta ve aynı zamanda varyansların bağıl büyüklüklerine dayalı olarak ana ve ortak etkilerin yorumlanması sağlanmaktadır. Diğer yandan K çalışmasıyla ise madde ve/veya puanlayıcı sayılarının arttırılıp azaltılmasıyla güvenilirliği yüksek ve en verimli ölçme desenleri belirlenmektedir (Atılğan, 2004). G çalışmasından elde edilen bilgileri kullanan K çalışmasının temel amacı, ölçme işlemindeki hataları en aza indirmenin yollarını araştırarak, belli bir amaçla yapılan ölçmenin en uygun desenine ulaşılmasını sağlamaktır. Bir G çalışması, kabul edilebilir gözlemlerin evreni, araştırmacılara çok çeşitli varyans bileşen tahminleri sunmak için olabildiğince geniş olarak tanımlanır. Bir K çalışması ise özellikle belirli bir amaç için yalnızca bazı yönleri seçer, böylece puan yorumlamayı bir genelleme evrenine daraltılır (Shavelson, Webb; 1991).

Bir K çalışmasını planlamada ilk olarak genellemenin evrenini tanımlamak için, genellenmek istenilen yüzeylerin genişlik ve sayısı belirlenir. Sonrasında ölçme hatalarının ve hata kaynaklarının tanımlanması için gerekli olan ölçme sonuçlarının bağıl mı yoksa mutlak mı yorumlanacağı belirlenir. Daha sonra farklı ölçme hatalarının kaynakları hakkında G-çalışması ile elde edilen bilgileri kullanarak, hataları en aza indirmek ve güvenilirliği en üst düzeye çıkarmak için alternatif desenlerin etkisi değerlendirilir (Shavelson ve Webb, 1991). Son aşamada belirtilen, hataları en aza indirmek ve güvenilirliği en üst düzeye çıkarmak için alternatif desenlerin etkisini değerlendirmek, G-kuramının önemli bir avantajıdır. Bu özellik, Klasik Test Kuramındaki Sperman-Brown formülüne benzer. Bu sayede yüksek güvenilirlik elde etmek için yüzeylerin koşullarının en uygun sayısı kullanılarak, ölçme aracının geliştirilmesi ya da bir dahaki uygulamalarda ölçme durumunun ne olması gerektiğine karar verilmesini sağlar.

G çalışmalarının amacı varyans bileşenlerinin toplam varyans üzerindeki etkisini kestirmek iken, K çalışmalarının amacı maksimum güvenilirliği sağlamak adına, G çalışmalarında toplanan verileri yorumlamaktır (Güler, Uyanık ve Teker, 2012).

1.1. Problem Durumu

Eğitimin birçok alanında puanlayıcıların ölçme sürecine katıldığı bilinmektedir. Puanlayıcıların ölçme sürecine katıldığı durumlarda farklı değişkenlik kaynaklarının bir arada değerlendirilmesi gerekir. Klasik ölçme yaklaşımlarının sınırlı kalması nedeniyle farklı değişkenlik kaynaklarının bir arada değerlendirildiği G Kuramının kullanıldığı çalışmalara gereksinim duyulmaktadır.

Son yıllarda eğitimde seçme ve yerleştirme sınavlarında açık uçlu maddelerin kullanması sıkça gündeme gelmiştir. Açık uçlu maddelerin puanlamasının birden fazla puanlayıcı tarafından yapılması nedeniyle puanlayıcı güvenilirliği bu tür sınavlarda oldukça önemlidir. Puanlayıcı sayısı gibi açık uçlu maddelerin puanlanmasında kullanılan serbest ve anahtarla puanlama gibi puanlama yöntemleri de güvenilirliğe etki etmektedir. Serbest puanlama; alan uzmanının, önceden hazırlanmış hiçbir ölçüte bağlı kalmadan kendi puanlama anahtarını oluşturarak puanlama yapmasıdır. Anahtarla puanlamada ise puanlayıcı açık uçlu maddelere verilmesi beklenen cevap örüntülerini listeler. Açık uçlu maddelere verilmesi muhtemel cevaplardan oluşan bu listeye cevap anahtarı ya da ölçütler listesi adı verilebilir. Sonra bu listedeki cevapların her birine kaç puan verileceğini, hangi cevaplara puan verilmeyeceğini ya da kısmi puan verileceğini belirleyerek puanlama anahtarını oluşturur. Öğrencilerin verdiği yanıtlar puanlama anahtarı ile karşılaştırılarak uygun olan puanlar verilir (Atılğan, Kan ve Doğan, 2007).

Bu araştırmada serbest puanlama ve anahtarla puanlama yöntemlerinin ve puanlayıcı sayılarının açık uçlu sorulardan oluşan sınavların güvenilirliğinde yarattığı etki incelenecektir. Serbest ve anahtarla puanlamalarda en az kaç puanlayıcı ile güvenilir bir puanlama yapılacağı; madde sayısının arttırılmasının hangi puanlama türünde G katsayısını ne kadar arttıracak gibi araştırma sorularına yanıt aranacaktır. Puanlayıcı sayısının arttırılması durumunda hangi puanlama türünde güvenilirlik katsayısında daha fazla artış olduğu incelenecektir. Farklı puanlama türlerinde maksimum güvenilirlik elde etmek için en uygun madde ve puanlayıcı sayısının ne olduğu, hangi puanlama türünün zaman ve işgücü açısından daha ekonomik olduğunun araştırılması için aşağıdaki problem cümlesi ve alt problemlere yanıt aranmaktadır.

1.2. Problem Cümlesi

Açık uçlu matematik sorularının birden çok puanlayıcı tarafından serbest ve anahtarla puanlanması sonucu elde edilen G ve K çalışmalarının sonuçları nedir ve nasıl değişmektedir?

1.2.1. Alt Problemler

1. Tüm bireyleri (b), tüm maddelerde (m) aynı puanlayıcıların (p) serbest puanladığı açık uçlu matematik sorularının kullanıldığı ölçme durumunda;
 - 1.1. Tek değişkenli G çalışması sonuçları nedir?
 - 1.2. Tek değişkenli K çalışması sonuçları nedir?
 - 1.3. Çok değişkenli G çalışması sonuçları nedir?
 - 1.4. Çok değişkenli K çalışması sonuçları nedir?
2. Tüm bireyleri (b), tüm maddelerde (m) aynı puanlayıcıların (p) anahtarla puanladığı açık uçlu matematik sorularının kullanıldığı ölçme durumunda;
 - 2.1. Tek değişkenli G çalışması sonuçları nedir?
 - 2.2. Tek değişkenli K çalışması sonuçları nedir?
 - 2.3. Çok değişkenli G çalışması sonuçları nedir?
 - 2.4. Çok değişkenli K çalışması sonuçları nedir?
3. Tüm bireyleri (b), tüm maddelerde (m) aynı puanlayıcıların (p) serbest ve anahtarla puanladığı açık uçlu matematik sorularının kullanıldığı ölçme durumunda;
 - 3.1. Tek değişkenli G çalışması sonuçları nedir?
 - 3.2. Tek değişkenli K çalışması sonuçları nedir?
 - 3.3. Çok değişkenli G çalışması sonuçları nedir?
 - 3.4. Çok değişkenli K çalışması sonuçları nedir?

1.3. Araştırmanın Amacı ve Önemi

Bu araştırmanın amacı açık uçlu maddelerle hazırlanan matematik sınavının iki farklı puanlama yöntemi (anahtarla ve serbest puanlama) kullanılarak elde edilen verilerin G kuramına göre analiz edilerek karşılaştırılmasıdır. Ayrıca araştırmada puanlayıcı ve madde sayılarının değiştirilmesi ile güvenilirliğin nasıl etkilendiği her iki puanlama türü için de incelenmiştir. Bu şekilde en güvenilir sonuçları elde etmek için anahtarla ve serbest puanlama için uygun madde ve puanlayıcı sayısına ulaşmaya çalışılmıştır. Araştırma açık uçlu soruların puanlanmasında uygun puanlama yönteminin ve her yöntem için en uygun puanlayıcı sayısının belirlenmesi açısından önem taşımaktadır.

Bunların yanında araştırmanın eğitimde ölçme ve değerlendirmede başvuru, son yıllarda ortaöğretime geçiş ve yükseköğretime geçiş sınavlarında kullanılması gündeme gelen ve puanlanma güvenirliliği sıkça tartışılan açık uçlu maddelerden oluşan uzun yanıtli sınavların güvenirliliklerine ilişkin bilgi sağlayacağı düşünülmektedir. Yine araştırmanın açık uçlu sorulardan oluşan sınavların hangi yöntemle ve kaç puanlayıcı ile puanlanmasının daha güvenilir sonuçlar vereceğine yönelik bulgular sunacağı düşünülmektedir. Bunların dışında G çalışmalarında kullanılan K çalışmasının sonuçlarının G çalışmasını ne derece yansıttığı hakkında bilgi vereceği ve tüm bunların sonucunda açık uçlu sorulardan oluşan sınav kullanmak isteyen eğitimcilere puanlama yöntemi ve puanlayıcı sayısı ile ilgili tavsiyede bulunulabileceği düşünülmektedir.

1.4. Sınırlılıklar

Bu araştırma, 2015-2016 eğitim-öğretim yılı Ödemiş 50. Yıl Ortaokulu yedinci sınıf öğrencilerine uygulanan açık uçlu soruları yanıtlama becerilerinin ölçüldüğü matematik sınavına katılan öğrencilerden cevap kâğıdını uygun bir şekilde yanıtlayan öğrencilerle sınırlıdır.

1.5. Konu ile İlgili Yapılan Çalışma ve Araştırmalar

1.5.1 Yurt İçinde Yapılan Çalışmalar

Bu bölümde yurt içinde yapılan genellenebilirlik kuramına ilişkin ve açık uçlu maddeler ve bu maddelerin puanlanmasına ilişkin literatürdeki çalışmalara yer verilmiştir.

Atılgan (2004), İnönü Üniversitesi Eğitim Fakültesi Güzel Sanatlar Eğitimi Bölümü Müzik Öğretmenliği Özel Yetenek Seçme Sınavları verilerinde birden çok değişkenlik kaynağını birlikte değerlendiren G kuramı ve Çok Değişken Kaynaklı Rasch Modeli ile kestirilen parametreler incelenmiş ve karşılaştırılmıştır. G ve K çalışmasıyla aynı amaçla yapılan iki sınavdan kestirilen parametreler karşılaştırılarak kuramın tutarlılığı da incelenmiştir. Ayrıca G tek değişkenli ve çok değişkenli modelleri ile aynı veriler için kestirilen istatistikler karşılaştırılmıştır. Yapılan karşılaştırmalar sonucunda G Kuramı ve Çok Değişkenlik Kaynaklı Rasch Model yaklaşımları sonuçlarının her noktada aynı olmadığı fakat puanlayıcı ve görev değişkenlik kaynağı için bu iki yaklaşımın tutarlı olduğu sonucuna varılmıştır.

Yelboğa (2007), "Klasik Test Kuramı ve Genellenebilirlik Kuramına Göre Güvenirliğin Bir İş Performansı Ölçeği Üzerinde İncelenmesi" adlı çalışmasında 2005 ve 2006 yıllarında gerçek veriler kullanarak Klasik Test Kuramı ve G Kuramı analizlerini gerçekleştirmiştir. Her iki kuramdan elde ettiği güvenilirlik katsayılarını karşılaştırmıştır. Klasik test kuramından puanlayıcılar arası güvenilirlik için Kendall'ın konkordans katsayısı, test tekrar test ve Cronbach alfa güvenilirlik katsayıları; Genellenebilirlik kuramından çok değişkenle modelle G ve Phi katsayıları hesaplanmıştır. Elde edilen katsayılar karşılaştırıldığında İş Performansı Ölçeğinde aynı ölçme durumu için her iki kuramla hesaplanan katsayıların birbiriyle uyumlu olduğunu gözlemiştir. Araştırma sonucuna göre araştırmada kullanılan ölçme aracının geçerliliği ve güvenilirliği yüksek bulunmuştur. Ayrıca yine araştırmadan elde edilen verilere göre güvenirliliğin iş performansı ölçeğinde değerlendirilmesinde Genellenebilirlik Kuramı kullanımının daha uygun olduğunu belirtmiştir.

Güler (2008), matematik başarısının ölçülmesine yönelik çalışmasında Klasik Test Kuramı, Genellenebilirlik Kuramı ve Rasch Modeli karşılaştırmıştır. Matematik başarısının ölçülmesinden elde edilen puanların güvenilirliklerini bu üç kuram için hesaplayıp karşılaştırmıştır. Araştırma sonuçlarına göre, matematik başarısının ölçüldüğü ölçme aracının güvenilir sonuçlar verdiği, puanlayıcıların puanlarının birbiriyle uyumlu olduğu belirlenmiştir. Araştırma sonucunda puanların kullanılma amacına bağlı olarak üç kuramdan hangisinin güvenilirlik belirlemede uygun olacağını açıklamış ve en az iki kuramdan yararlanılması gerektiğini belirtmiştir.

Deliceoğlu (2009), Futbol Yetilerine İlişkin Dereceleme Ölçeği kullanarak 4 puanlayıcı ile yaptığı puanlamalar sonucunda Genellenebilirlik ve Klasik Test Kuramına göre belirlenen güvenirlik katsayılarının düzeylerini saptamış ve birbirleriyle karşılaştırmıştır. Alt boyutlar incelendiğinde en yüksek varyans bileşeninin birey temel etkisine ait olduğunu söylemiştir. Bunu $b \times m \times p$ ortak etkisi varyans bileşeni takip etmiştir. Bunların dışındaki temel ve ortak etkilerin sıfıra yakın olduğunu belirtmiştir. Futbol Yetilerine İlişkin Dereceleme Ölçeğinin puanlanmasından elde edilen güvenirlik katsayısı iç ölçütlerle karşılaştırıldığında G ve Cronbach alfa katsayılarının beklenen değerden yüksek olduğu, Phi katsayısı ve Kendall W güvenirlik katsayılarının beklenen değerden daha düşük olduğunu söylemiştir. Bu sonuçlara göre potansiyel hata kaynaklarının fazla olduğu benzer çalışmalarda Genellenebilirlik Kuramının Klasik Test Kuramına alternatif olabileceği belirtilmiştir.

Nalbantoğlu (2009), Performans Ölçümlerinde Genellenebilirlik Kuramıyla Farklı Desenlerin Karşılaştırılması” adlı çalışmasında 2006-2007 eğitim öğretim yılı Hacettepe Üniversitesi üçüncü sınıf öğrencilerinden tesadüfi olarak seçilen 48 öğrenci ve üç puanlayıcı ile çalışmıştır. Yapılan araştırmada öğrencilerin puanlayıcılar tarafından aynı iletişim becerileri değerlendirme formuyla 15 görev doğrultusunda birlikte ve dönüşümlü puanlamasıyla oluşan iki farklı desen doğrultusunda ($\bar{o} \times g \times p$ ve $(\bar{o}:p) \times g$) G ve K çalışması yapmıştır. Her iki desenle yapılan analizler sonucunda her iki desende değişkenler için kestirilen varyans değerlerinin birbiriyle uyumlu olduğunu belirtmiştir. Her iki desenle yapılan K çalışması sonucunda G ve Phi katsayıları arasında çok fark olmadığı fakat $(\bar{o}:p) \times g$ deseninde katsayıların daha büyük çıkma eğiliminde olduğu sonucuna varmıştır.

Taşdelen (2009), Nedelsky ve Angoff standart belirleme yöntemleri ile elde edilen kesme puanlarını Genellenebilirlik Kuramı kullanılarak karşılaştırmıştır. Çalışmada kullanılan standart belirleme yöntemlerinin uygulanması için gereken en uygun puanlayıcı sayısı belirlenmeye çalışılmıştır. Yapılan çalışmada Ankara ilinde görev yapan 40 öğretmen 2008 altıncı sınıf Seviye Belirleme Sınavı’nda yer alan Fen ve Teknoloji testindeki 16 maddeyi Nedelsky ve Angoff standart belirleme yöntemlerine uygun olarak puanlamıştır. Araştırma verilerinden elde edilen bulgulara göre G çalışmasında her iki yöntemde de puanlayıcılar arasında tutarlılık olduğu ve maddeler

arasındaki farklılıkların ortaya çıkarılabildiği belirtilmiştir.

Güler (2011), ‘Rastgele Veriler Üzerinde Genellenebilirlik Kuramı ve Klasik Test Kuramına Göre Güvenirliğin Karşılaştırılması’ adlı çalışmasında 125 öğrenciye, 18 satır ve 4 sütun olmak üzere, 18×4 hücreden oluşan tablolar dağıtmıştır. Öğrencilerden dağıtılan tablolardaki tüm hücreleri 0 ile 5 arasındaki sayılarla tamamen rastgele olarak doldurmaları istenmiştir. Böylece araştırmada tümüyle çaprazlanmış desen ($b \times m \times p$) kullanılması amaçlanmıştır. Araştırmanın sonucunda tek değişkenlik kaynaklı desenler için Klasik Test Kuramı ve Genellenebilirlik Kuramına göre hesaplanan güvenilirlik değerleri bu çalışmada yer alan tamamen rastgele oluşturulmuş veriler için de aynı değerlerde çıkmıştır. Puanlayıcı, durum vb. gibi birden fazla değişkenlik kaynağının bulunduğu durumlara ilişkin yapılacak güvenilirlik çalışmalarında tek bir seferde her bir değişkenlik kaynağı ve bunların etkileşimlerine dayalı güvenilirliği belirleyebilmemizi sağlayarak Genellenebilirlik Kuramının Klasik Test Kuramına alternatif olduğunu vurgulamıştır.

Nalbantoğlu Yılmaz (2012), ‘Genellenebilirlik Kuramında Dengelenmiş ve Dengelenmemiş Desenlerin Karşılaştırılması’ adlı çalışmasında Tıp Fakültesi Yapılandırılmış Objektif Klinik Sınav’daki İntramusküler Enjeksiyon yapma istasyonu verileri kullanılmıştır. Puanlanan öğrenci sayısı değişmemek şartıyla puanlayıcıların her birinin birbiriyle eşit ve birbirinden farklı sayıda öğrenciyi aynı görevler doğrultusunda puanlamasıyla oluşturulmuş dengelenmiş ve dengelenmemiş desenlerle yapılan Genellenebilirlik Kuramı sonuçlarını karşılaştırmıştır. Araştırmasında puanlayıcıların her birinin araştırmaya katılan öğrencilerden sadece bir kısmını puanlamasıyla, öğrenci (ö) ve puanlayıcı (p) değişkenlerinin yuvalanmış, tüm öğrenciler için aynı olan görevlerin (g) ise bu değişkenlerle çaprazlandığı (ö:p) x g deseni dengelenmiş ve dengelenmemiş durumlara göre incelemiştir. Dengelenmiş ya da dengelenmemiş desenlerle yapılan G çalışması sonucunda puanlayıcıların her biri için hesaplanan varyans değerlerinin paralellik gösterdiği, senaryolara göre toplam veri sayısı aynı olmak şartıyla dengelenmiş ve dengelenmemiş (ö:p) x g desenleri ile kestirilen güvenilirlik katsayılarının her iki desende de yakın sonuçlar verdiği sonucuna varmıştır.

Özberk (2012), ‘Genellenebilirlik Kuramı Karar Çalışmalarında Kullanılan Farklı Katsayıların Karşılaştırılması’ adlı çalışmasında Wiley (2001) tarafından önerilen

iki farklı varyans bileşenlerini belirleme yöntemine göre kestirilen standart hata değerleri ile güvenirlik ve geçerlik hakkında karşılaştırmalar yapmıştır. Simülasyon yardımıyla tek değişkenlik kaynaklı desene göre üretilen verilerin tüm değişkenlik kaynaklarına göre ANOVA ve bootstrap yöntemleri kullanılarak standart hatalarını, varyans bileşenlerini, mutlak ve bağıl hatalarını hesaplamıştır. Araştırma sonucuna göre normal dağılım göstermeyen ve 1-0 şekilde puanlanan veriler üzerinde hesaplanan evren puanı-hata değerinin en iyi *boot-b* prosedüründe kestirildiği, hata tolerans değerinin ise en iyi *boot-m* prosedüründe kestirildiği belirtilmiştir. Bu bakımdan *boot-m* prosedürünün geçerlik hakkında daha fazla bilgi verdiği, *boot-b* prosedürünün de G Kuramı çalışmalarında evren puanlarını belirlemede daha kesin kestirimler yaptığını söylemiştir.

Gültekin, Demirtaşlı (2012), Yaptıkları çalışmada TIMSS 2007 matematik testine verilen yanıtlardan oluşan verileri kullanmışlardır. Matematik başarısına ilişkin benzer bilişsel becerileri ölçen çoktan seçmeli ve sınırlı yanıtlu açık uçlu maddeleri tek başına ve birlikte kullanarak ve bu iki madde tipinin test içinde yer alma yüzdelerinin farklı olduğu testlerden Madde Tepki Kuramına göre kestirilen madde ve test bilgi fonksiyonlarının göreceli etkililik düzeylerini araştırmışlardır. Araştırma verilerine göre, sınırlı yanıtlu açık uçlu maddelerden oluşan testin öğrencilerin matematik başarısı hakkında çoktan seçmeli maddelerden oluşan teste göre daha fazla bilgi verdiği, göreceli etkililik değerleri, açık uçlu maddelerden oluşan testin diğer testlerden etkili olduğu belirtilmiştir.

Başat (2015), ‘6. Sınıf Fen Bilimleri Dersi Elektrik İletim Ünitesinin Değerlendirilmesinde Kullanılan Ölçme Araçlarının Güvenirliğinin Genellenebilirlik Kuramına Göre Analizi’ adlı çalışmada araştırmacı tarafından geliştirilen çoktan seçmeli test, yapılandırılmış grid ve performans görevini altıncı sınıf (40 kişi) öğrencilerine uygulamıştır. Uygulanan ölçme araçları üç puanlayıcıya puanlatılarak ölçme sonuçları üstünde G ve K çalışması yapılmıştır. Yapılan G ve K çalışmalarının sonuçlarına göre uygulanan çoktan seçmeli testin G katsayısının .68, Phi katsayısının .64; yapılandırılmış gridin G katsayısının .91 ve Phi katsayısının .83; performans görevinin G katsayısının .89 ve Phi katsayısının .86 olarak elde edildiğini belirtmiştir. Ölçme araçları ile yapılan karar çalışmaları sonucunda her üç ölçme aracı için bulunan

G ve Phi katsayılarının paralellik gösterdiğini söylemiş, öğrenci puanlarının puanlayıcıdan puanlayıcıya farklılık göstermediği sonucuna varmıştır.

İlhan (2016), ‘Açık Uçlu Sorularla Yapılan Ölçmelerde Klasik Test Kuramı ve Çok Yüzeyle Rasch Modeline Göre Hesaplanan Yetenek Kestirimlerinin Karşılaştırılması’ adlı çalışmasında açık uçlu altı matematik sorusundan oluşan bir başarı testi kullanmıştır. Sekizinci sınıfa devam eden 100 öğrenciye uyguladığı bu test verilerini bütüncül bir rubrik kullanarak dört puanlayıcıya puanlatmış; elde edilen verileri hem KTK (Klasik Test Kuramı) ve ÇDKRM’ye (Çok Değişken Kaynaklı Rasch Model) göre analiz etmiştir. Dört puanlayıcının verdiği puanların ortalaması alınarak, KTK’ye göre yetenek kestirimlerini hesaplamış ve üç yüzeyle bir desen (puanlayıcı, birey, madde) ile çok yüzeyle Rasch analizi uygulamıştır. Araştırma sonucunda; KTK ve ÇYRM’ ye göre hesaplanan yetenek kestirimleri arasındaki görel uyumun son derece yüksek olduğunu belirterek iki kurama göre hesaplanan yetenek kestirimlerine ait ortalamalar arasında anlamlı fark bulunduğunu ve dolayısıyla mutlak bir uyumun söz konusu olmadığını saptamıştır. Ayrıca ÇYRM’ de rapor edilen yetenek kestirimlerinin ölçüt geçerliğinin KTK ile elde edilen yetenek kestirimlerine göre daha yüksek olduğunu da belirtmiştir.

Anadol (2017), ‘Dereceli Puanlama Anahtarlarının Güvenirliğinin Farklı Deneyim Yıllarına Sahip Puanlayıcıların Kullanıldığı Durumlarda İncelenmesi’ adlı çalışmasında İngilizce yazılı anlatım becerisi puanlamada dereceli puanlamam anahtarı kullanmaya ilişkin üç grup puanlayıcıdan topladığı verilerin G çalışması sonuçlarını karşılaştırmıştır. Araştırmanın çalışma grubunu amaçlı örneklem yöntemi ile Ankara’da bulunan bir vakıf üniversitesinin hazırlık sınıfına devam eden öğrenciler ve bu öğrencilere eğitim veren okutmanlar oluşturmuştur. Araştırmada (b:p) x m deseni ile dereceli puanlama anahtarı kullanmaya ilişkin farklı deneyim yıllarına sahip puanlayıcılar araştırmaya katılan öğrencileri 6 görev doğrultusunda puanlamıştır. Elde edilen veriler ile yapılan G çalışmaları sonuçları karşılaştırılarak farklı durumlar için hesaplanan varyans değerleri, G ve Phi katsayıları, mutlak ve bağıl hataların paralellik gösterdiğini söylemiştir. Elde edilen veriler sonucunda puanlayıcıların dereceli puanlama anahtarı kullanmaya ilişkin farklı deneyim yıllarına sahip olmasının dereceli puanlama anahtarının güvenirliğini etkilemediğini belirtmiştir.

1.5.2. Yurt Dışında Yapılan Çalışmalar

Kieffer (1998) yaptığı çalışmada klasik test kuramı yerine genellenebilirlik kuramının kullanılmasının sağlayacağı avantajları tartışmıştır. Çalışmasında genellenebilirlik kuramıyla klasik test kuramından farklı olarak farklı değişkenlik kaynaklarının ve bu değişkenlik kaynaklarının etkileşiminin büyüklüğünü tek bir analizle bir arada değerlendirmenin mümkün olduğunu açıklamıştır. Hipotetik bir örnekte elde edilen sonuçlarla klasik test kuramının sınırlılıklarını ve genellenebilirlik kuramının güçlü yanlarını açıklamıştır.

Wang (2005) yaptığı çalışmada geleneksel yöntem ve genellenebilirlik kuramıyla hesaplanan katsayıları karşılaştırmak için öğrencilerin İngilizce yazılı anlatım becerilerini ölçen sınav verileri kullanmıştır. Bu çalışmada sınavı alan bireyler (p), maddeler (i) ve beceri (t) olmak üzere $p \times (i:t)$ yuvalanmış deseni kullanılmıştır. Yapılan G çalışması sonuçlarına göre hesaplanan katsayıların birbirine yakın olduğu sonucuna varmıştır.

Kantor ve Lee (2007) yaptıkları çalışmada TOEFL sınavında yazma becerilerinin ölçüldüğü bölümde Genellenebilirlik Kuramı kullanarak alternatif değerlendirme yöntemleri belirlemeyi amaçlamıştır. Yaptıkları çalışmada farklı senaryolara bağlı karar çalışması ile puan güvenilirliğini yükseltmek için gerekli görev ve puanlayıcı sayısını belirlemeye çalışmışlardır. Yaptıkları çalışmadan elde edilen sonuçlara göre maksimum puan güvenilirliği için görev sayısının arttırılmasının görev başına düşen puanlayıcı sayısının arttırılmasından daha etkili olduğu sonucuna varmışlardır.

Huang (2012) çalışmasında Genellenebilirlik (G) teorisini kullanarak, Kanada'daki İngilizce sınavlarında orta öğretim ESL (İngiliz ulusundan olmayan) öğrencilerine verilen yazım (writing) puanlarının güvenilirlik ve geçerliliğini incelemiştir. Bu araştırmanın temelini oluşturan araştırma problemi şudur: ESL (İngiliz ulusundan olmayan) ve NE (İngiliz ulusundan olan) öğrencilerine üç yıl boyunca İngilizce yazma sınavı için verilen analitik puanlardaki doğruluk ve yapı geçerliliği arasında herhangi bir fark var mı? Bu araştırma problemi doğrultusunda bir dizi G ve K çalışması yapılarak ESL öğrencileri için gözlemlenen G katsayılarının üç yılda da NE öğrencilerinden daha düşük olduğunu bu nedenle ESL öğrencilerine verilen yazma

puanlarının güvenilirliđinin daha düşük, hatasının daha yüksek olduđunu belirtmiřtir. Ayrıca ELS öğrencilerine uygulanan sınavın diđer öğrencilere uygulanan sınava göre yapı geçerliliđinin üç yılda da daha düşük olduđu sonucuna varmıřtır.

Graham, Hebert, Sandbank, Harris (2016), yaptıkları çalışmada öğretmenleri tarafından özel yardıma ihtiyacı olan yazarlar olarak belirlenen ve normlara dayalı hikâye yazma testinde yüzde 25 veya daha düşük puan alan ikinci ve üçüncü sınıf öğrencileri (81 erkek, 56 kız); bir hikâye, kişisel anlatı, fikir kompozisyonu ve bilgilendirici metin olmak üzere dört kompozisyon yazmıřtır. Dört kompozisyon için genellenebilirlik teorisini (G- teorisi) uygulayarak öğrencilerin üç yazım ölçüsüne (toplam kelime sayısı, kelime çeřitliliđi ve yazı kalitesi) göre skorları öğrenciler, yazım görevleri ve öğrenciler ile yazma görevlerinin etkileřimi olmak üzere üç varyans bileřenine ayrılmıřtır. Yapılan çalışmanın sonucunda bu öğrencilerin yazma yeteneklerini toplam kelime sayısı, kelime çeřitliliđi ve yazı kalitesi bakımından güvenilir bir řekilde elde etmek için sırasıyla 14, 8, 11 kompozisyona ihtiyaç duyulduđunu belirtmiřtir.

BÖLÜM II

YÖNTEM

2.1. Araştırmanın Türü

Araştırma, gerçek verilerin G kuramı ile analiz ile sonuçlar elde edilmesine ve G kuramının bu araştırmadaki avantajlarının ve sınırlılıklarının ortaya konulmasına dayanır.

2015-2016 eğitim-öğretim yılında uygulanan açık uçlu matematik sorularından ve alt testlerden oluşan ölçme aracının G Kuramı ile niteliklerinin belirlenmesi ve durum saptamaya yönelik olması nedeniyle bu çalışma betimsel bir araştırmadır.

2.2. Çalışma Grubu, Evren ve Örneklem

Yapılan araştırmanın çalışma grubunu; 2015-2016 eğitim-öğretim yılında İzmir ilinin Ödemiş ilçesinde öğrenim gören yedinci sınıf öğrencileri oluşturmaktadır. Bu sınavı puanlayabilecek niteliklere sahip olan tüm ilköğretim matematik öğretmenleri puanlayıcı evrenini; uygulanan bu sınavda kullanılabilecek tüm maddeler ise madde evrenini oluşturmaktadır.

Araştırmanın örneklemini, 2015-2016 eğitim-öğretim döneminde İzmir ilinin Ödemiş ilçesinde bulunan bir devlet okulunda öğrenim gören 253 kişilik yedinci Sınıf öğrencisine uygulanan açık uçlu matematik sınavında kullanılan maddelerden ve bu sınavı puanlayan beş puanlayıcıdan oluşmaktadır. Araştırmada yer alan puanlayıcılar seçkisiz olarak seçilmiştir Puanlayıcıların dördü İlköğretim Matematik Öğretmenliği Lisans mezunu, biri Matematik Eğitimi Yüksek Lisans mezunudur. Araştırmada çalışan puanlayıcılarla ilgili ayrıntılı bilgi Ek. 3' de verilmiştir.

2.3. Araştırma Verileri

Araştırmacı ve puanlayıcılardan farklı iki alan uzmanı tarafından geliştirilen ölçme aracı; üç farklı alt öğrenme alanı için üç alt test ve her bir alt testte beş madde olmak üzere on beş maddeden oluşmaktadır. Ölçme aracı hazırlandıktan sonra bir ölçme değerlendirme uzmanı tarafından kontrol edilmiştir. Kullanılan ölçme aracı Ek. 1 de verilmiştir. Aşağıdaki tabloda ölçme aracının hangi bilişsel düzeyde hangi davranışları ölçtüğü belirtke tablosu ile verilmiştir.

Tablo 3. Araştırmada kullanılan Ölçme Aracına ait Belirtke Tablosu

Hedef- dav. Konular	Bilgi	Kavrama	Uygulama	Analiz	Sentez	Değerlendirme	Toplam
Eşitlik ve Denklem		1	2	1	1		5
Rasyonel Sayılarla İşlemler		1	3	1			5
Doğrusal Denklemler		1	2	1	1		5

Hazırlanan ölçme aracı 253 kişilik yedinci sınıf öğrencisine uygulanmıştır. Uygulanan sınav yansız seçilen beş farklı alan uzmanı (en az ilköğretim matematik öğretmenliği lisans mezunu) tarafından farklı zamanlarda olmak üzere önce serbest, daha sonra anahtarla puanlanmıştır.

Serbest puanlamada puanlayıcılar genel izlenimle puanlama yaparken, anahtarla puanlamada ise önceden araştırmacı tarafından puanlayıcılardan farklı alan uzmanları ile birlikte hazırlanmış belli bir puanlama anahtarından yararlanılarak puanlama yapılmıştır. Puanlayıcıların kullandıkları bu puanlama anahtarı Ek. 2 de verilmiştir. Tüm puanlayıcılara aynı puanlama anahtarı serbest puanlama yaptıktan bir süre sonra dağıtılarak puanlamalarını tekrar bu puanlama anahtarına göre yapmaları istenmiştir. Araştırmada her bir puanlayıcı (p) her bir maddeyi (m) tüm bireyler (b) için birbirlerini etkilemeden bağımsız olarak puanlamıştır.

2.4. Verilerin Analizi

Araştırma verileri G Kuramının desenine uygun olarak analiz edilmiştir. G Kuramı için ana etkilerin ve ortak etkilerin toplam varyansa etkilerinin kestirilmesi, puanların güvenilirliğinin yorumlanabilmesi için G ve Phi katsayıları hesaplanırken

GENOVA (Crick ve Brennan, 1983) ve mGENOVA versiyon 2.1 kullanılmıştır (Brennan, 2001b). Ayrıca araştırmacının uyguladığı ölçme aracının üç farklı alt testten oluşması nedeniyle hem Çok Değişkenli hem de Tek Değişkenli G Kuramı modelleri ile analizler gerçekleştirilmiştir.

Araştırmanın deseninde G Kuramına göre iki değişkenlik kaynaklı (birey x madde x puanlayıcı) tamamen çaprazlanmış desen kullanılarak G katsayısı ve güvenilirlik katsayısı hesaplanmıştır. Bu nedenle araştırma desenin yüzeyleri birey (b), madde (m) ve puanlayıcıdır (p). Tamamen çaprazlanmış desen her puanlayıcının tüm bireylerin yanıtladıkları maddelerin hepsini puanlamaları anlamına gelir. Araştırmada 253 kişiye uygulanan ölçme aracı beş puanlayıcı tarafından önce serbest sonra da belli bir uzman tarafından hazırlanmış puanlama anahtarına göre değerlendirilmiştir. Yapılan G Kuramı analizleri sonucu kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri incelenmiştir. Varyans bileşenlerine dayanarak mutlak hata varyansı, göreceli hata varyansı hesaplanmıştır. Mutlak hata varyansı kullanılarak Phi katsayısı, göreceli hata varyansı kullanılarak da G katsayısı hesaplanmıştır.

Araştırmada her bir alt testteki madde sayılarının 5, 6, 7, 8, 9, 10 olduğu, puanlayıcı sayılarının 2, 3, 4, 5 olduğu senaryolar için yapılan Karar (K) çalışmasının sonucunda elde edilen G ve Phi katsayıları hem anahtarla puanlama hem serbest puanlama için ayrı ayrı hesaplanmıştır. Bu katsayılar hazırlanması ve puanlanması en ekonomik aynı zamanda en güvenilir ölçme aracı için gereken uygun madde ve puanlayıcı sayılarını belirlemede kullanılmıştır.

2.5. Yapılan G Çalışması Sonucu Elde Edilen İstatistiklerin Yorumlanması

Yapılan G çalışmasında önce yüzey sayısına ve ölçme desenine karar verilir. G kuramı varyans analizi (ANOVA) kullanarak puanları varyans bileşenlerine ayırır. G çalışması ile kestirilen varyans bileşenleri; test puanından genellemedeki hatasını değil, bireyin madde örnekleminin sadece herhangi bir maddesinden aldığı puanından, evrendeki bütün maddeler üzerinden bu bireyin ortalamasına genellemede hatanın büyüklüğünü gösterir (Atılğan, 2005). G çalışması, ölçülen yapıdaki her bir yüzeyin toplam ölçme hatasındaki yüzdesini elde etmemizi sağlar. Ayrıca G çalışması sonucu bağıl hata varyansı ve mutlak hata varyansı elde edilir. Eğer araştırmacı her bir öğrencinin ölçmenin uygulandığı grup içindeki başarısını öğrenmek istiyorsa, yani

öğrencinin başarısı gruptaki diğer öğrencilerin gösterdikleri performansa göre değişiyorsa bağıl hata varyansı kullanılır. Bağıl hata varyansı ölçme objesinin ham puan büyüklüklerindeki değişime bakmadan yüzeylerin sıralamada yerlerini koruma derecelerini gösterir. Bağıl hata varyansı klasik test kuramında yer alan güvenilirlik katsayısı ile aynı anlamı taşımaktadır. Elde edilen puanların gerçek değerleri araştırmacı için daha önemli veya faydalı ise bağıl hata varyansı kullanılabilir (Mushquash&O'Connor, 2006).

Araştırmacının amacı, öğrencilerin başarısını mutlak bir ölçüte bağlı olarak değerlendirmek ise mutlak hata varyansı kullanılabilir. Bu durumda öğrencilerin başarısı içinde bulundukları grubun başarısından bağımsızdır. Öğrencilerin başarısı önceden belirlenmiş ölçüte göre değerlendirilir. Bu nedenle mutlak hata varyansı bağıl hata varyansına göre daha katıdır (Gelbal, Güler, 2007).

Sosyal bilimlerin farklı dallarında yapılan ölçmeler, bir yüzeyli desenlerden daha karmaşık yapıya sahip olabilirler. Örneğin birçok maddeden oluşan bir sınavda puanlamanın farklı puanlayıcılar tarafından yapılması sonucu maddelerden gelebilecek hataların yanı sıra puanlayıcılar ya da sınavın farklı iki zamanda aynı öğrenci grubuna uygulanması sonucu zaman hata kaynağı oluşturabilir.

Bir madde örnekleminden seçkisiz (random) olarak seçilen maddelerden oluşan bir sınavın bütün bireylere bütün maddeleri ile uygulandığı ve puanlayıcı evreninden örneklenmiş bütün puanlayıcıların bütün bireyler için bütün maddeleri puanladıkları ölçme durumları, seçkisiz çaprazlanmış desen olarak isimlendirilir. Bireyler b , maddeler m ve puanlayıcılar p olmak üzere iki yüzeyli tamamı çaprazlanmış desen $b \times m \times p$ biçiminde gösterilir.

Yapılan bu çalışma seçkisiz tümüyle çaprazlanmış iki yüzeyli bir G çalışmasıdır. Yapılan analizler sonucunda üç ana etki (birey, madde, puanlayıcı) üç ortak etki (birey x madde, madde x puanlayıcı, birey x puanlayıcı) ve artık/hata varyansı elde edilmektedir. G çalışması sonucunda her birinin toplam varyansa katkıları elde edilir.

Araştırmada kullanılan $b \times m \times p$ deseni için Tablo 4'teki gibi bir matris düzenlenebilir. Aşağıdaki matrise göre herhangi bir bireyin, herhangi bir madde üzerinden, herhangi bir puanlayıcının puanlaması sonucu oluşan puanı gözlenen puan (X_{bgp}) olarak adlandırılabilir.

Tablo 4. Çaprazlanmış İki Yüzeyle $b \times m \times p$ Deseni için Matris

p1			p2			Np		
1	2	g	1	2	g	1	2	g
111	121	1g1	112	122	1g2	11p	12p	1gp
211	221	2g1	212	222	2g2	21p	22p	2gp
311	321	3g1	312	322	3g2	31p	32p	3gp
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
b11	b21	bg1	b12	b22	bg2	b1p	b2p	bgp

Bu varyans bileşenler içinde toplam varyansa en çok katkı yapması beklenen varyans bileşeni birey varyansıdır. Birey için kestirilen varyans bileşeni evren puan varyansıdır ve klasik test kuramındaki bireyin gerçek puanına karşılık gelir. Birey varyansının toplam varyansın içinde en büyük değeri alması ölçmenin uygulandığı grubun bireysel farklılıklara sahip (heterojen) olduğunu veya uygulanan ölçmenin bireyler arasındaki farklılıkları ortaya koyabildiğinin göstergesidir.

Yapılan analizde en düşük çıkması beklenen varyans bileşeni ise artık hata varyansıdır. Bir yüzey, seçkisiz yerine sabit (fixed) olarak alınırsa, ölçüm hatasının derecesi her zaman daha küçük olur; çünkü bireyler ve sabit bir yüzey arasındaki etkileşim için gözlemlenen puan değişimi, hata varyansı yerine evren varyansına katkıda bulunur (Brennan, 2001). Yapılan ölçme durumunda uygulanan testteki maddelerin farklı güçlük düzeylerinde olması nedeniyle madde değişkenlik kaynağı oluşur. Madde güçlüklerinin değişmesi ise madde örnekleminin evrene genellenmesinin doğruluğunu azaltacaktır (Güler, Uyanık ve Teker, 2012). Uygulanan ölçmede puanlamayı farklı farklı puanlayıcıların yapması durumunda puanlayıcıların yaptığı puanlamalar arasındaki tutarsızlıktan kaynaklanan farklılıklar olabilir. Bazı puanlayıcı bazı bireylerin puanlamasında katı davranırken bazı bireylerin puanlamasında ise

cömert davranabilir. Bu durum puanlayıcı örnekleminin evrene genellenmesini etkileyebilir.

Ortak etki varyanslarından birey x madde varyans bileşeni uygulanan testteki bazı maddelerin kimi bireylere kolay kimi bireylere zor gelmesinden kaynaklanır. Bu nedenle farklı başarı sıralamalarının bireyler ve maddeler arasındaki bu etkileşimden kaynaklandığı söylenebilir. Bir maddenin belirli bir ifadesi bazı bireylerin diğer bireylere göre daha olumsuz değerlendirilmesine yol açabilir. Bu durum birey x madde varyans bileşeninden kaynaklanmaktadır. Madde x puanlayıcı varyans bileşeni ise puanlayıcıların bazı maddeleri diğerlerine göre daha katı ya da cömert puanlamaları durumundan kaynaklanır ve hata varyansını etkiler. Tüm bunlar göz önünde bulundurularak; G Kuramı analizleri sonucunda oluşturulan tablo ve şekiller dikkate alınarak alt problemler yorumlanmıştır.

BÖLÜM III

BULGULAR

3.1. Alt Problem 1. Tüm bireyleri (b), tüm maddelerde (m) aynı puanlayıcıların (p) serbest puanladığı açık uçlu matematik sorularının kullanıldığı ölçme durumunda;

3.1.1. Tek değişkenli G çalışması sonuçları nasıldır?

Açık uçlu matematik sınavının puanlayıcılar tarafından serbest puanlanmasıyla elde edilen veriler üzerinde tek değişkenli G Kuramının $b \times m \times p$ deseni için G çalışması yapılmıştır. Bu çalışma ile varyans kaynakları için elde edilen varyans büyükleri ile toplam varyansı açıklama oranları Tablo 5'te verilmiştir.

Tablo 5. Açık Uçlu Matematik Sınavının Serbest Puanlanmasıyla Elde Edilen Tek Değişkenli G Çalışması Sonuçları

Varyans Kaynağı	Varyans	Toplam Varyans İçindeki Payı (%)
Birey (b)	561,008	18,720
Madde (m)	1574,682	52,543
Puanlayıcı (p)	783,837	26,155
Birey-Madde ($b \times m$)	42,595	1,421
Birey-Puanlayıcı ($b \times p$)	10,313	0,344
Madde-Puanlayıcı ($m \times p$)	21,356	0,713
Birey-Madde-Puanlayıcı ($b \times m \times p$)	3,120	0,104
Toplam:	2996,911	100

Tablo 5 incelendiğinde açık uçlu matematik sınavının serbest puanlanmasıyla elde edilen verilerle yapılan tek değişkenli G çalışması sonucunda kestirilen en büyük varyans 1574,682 ile toplam varyansın %52,543'ünü açıklayan madde (m) ana etkisi varyansıdır. Tüm varyans bileşenleri içinde en büyük olan madde varyansı testin maddelerinin birbirinden oldukça farklılaştığını, madde güçlüklerinin çok farklı olduğunu göstermektedir. Büyüklük bakımından ikinci sırada olan puanlayıcı(p) ana etkisi için kestirilen varyans bileşeni 783,837 olarak toplam varyansın %26,155'ini açıklamaktadır. Puanlayıcı(p) ana etkisi için kestirilen varyans bileşeninin toplam varyansı açıklama yüzdesinin büyük olması puanlayıcıların puanlamalarında

farklılıkların çok olduğunun, puanlayıcıların bütün bireyler boyunca yaptıkları puanlamalarda oldukça farklılaştıklarının göstergesidir. Tek değişkenli G çalışmasında 561,008 olarak kestirilen birey (*b*) ana etkisi varyansı toplam varyansın %18,720'sini açıklamaktadır. Birey ana etkisi klasik test kuramındaki gerçek puana karşılık gelmektedir ve G kuramında ölçmenin amacı olarak tanımlanmaktadır. Bu nedenle bu varyansın en büyük varyans olması beklenir. Ortak etki varyansı olan birey x madde (*bm*) varyansı(42,595) toplam varyansın%1,421'ini açıklamaktadır. Bu ortak etki belli bir bireyin bir maddeden diğer maddeye bağlı durumunun farklı olup olmadığını gösterir (Atılğan 2004; Brennan 2001; Shavelson ve Webb 1991). Birey x madde ortak etkisinin toplam varyansı açıklama yüzdesine bakıldığında küçük olan bu varyans bireylerin bağlı durumlarının bir maddeden diğer maddeye çok farklılaşmadığını gösterir. Birey x puanlayıcı (*bp*) ortak etkisivaryansı10,313 olarak kestirilmiştir ve toplam varyansın %0,344'ünü açıklamaktadır. Bu ortak etki varyansının toplam varyans içinde payının düşük olması; belli puanlayıcıların belli bireyler için puanlamalarının diğer puanlayıcılara göre fazlaca değişmediğini gösterir. Tek değişkenli olarak yapılan G çalışması ile 21,357 olarak kestirilen madde x puanlayıcı (*mp*) ortak etkisi varyansı toplam varyansın %0,713'ünü açıklamaktadır. Bu varyans bileşeninin payının toplam varyans içinde küçük olması; puanlayıcıların bir maddeden diğer maddeye puanlamalarının farklılaşmadığını gösterir. Son olarak Birey x madde x puanlayıcı ortak etkisi varyansı 3,120 olarak hesaplanmıştır. Bu varyans bileşeni tüm varyanslar içinde %0,1041 ile en küçük varyanstır. Bu varyansın yüzdesinin sıfıra yakın olması birey x madde x puanlayıcı ortak etkisinden kaynaklanan hataların az olduğunu göstermektedir.

3.1.2. Tek değişkenli K çalışması sonuçları nasıldır?

Açık uçlu matematik sınavının serbest puanlanmasıyla elde edilen puanlarda tüm değişkenlerin çaprazlandığı $b \times m \times p$ deseninde birey (*b*) ölçmenin nesnesi (object of measurement), *m*, *p* ise yüzey (facet) olarak ele alınır. Elde edilen varyans bileşenlerinden hesaplanan mutlak ve bağlı hata varyanslarına ve görev puanlayıcı sayılarının artırılıp azaltılmasına dayalı senaryolar için kestirilen genellenebilirlik katsayısı (*G*), phi (ϕ) katsayısına ait değerler Tablo 6'da verilmiştir.

Tablo 6. Açık Uçlu Matematik Sınavının Serbest Puanlanmasında Bileşik Testte Bulunan Farklı Madde ve Puanlayıcı Sayısı Senaryoları ile Yapılan K Çalışmasının Sonuçları.

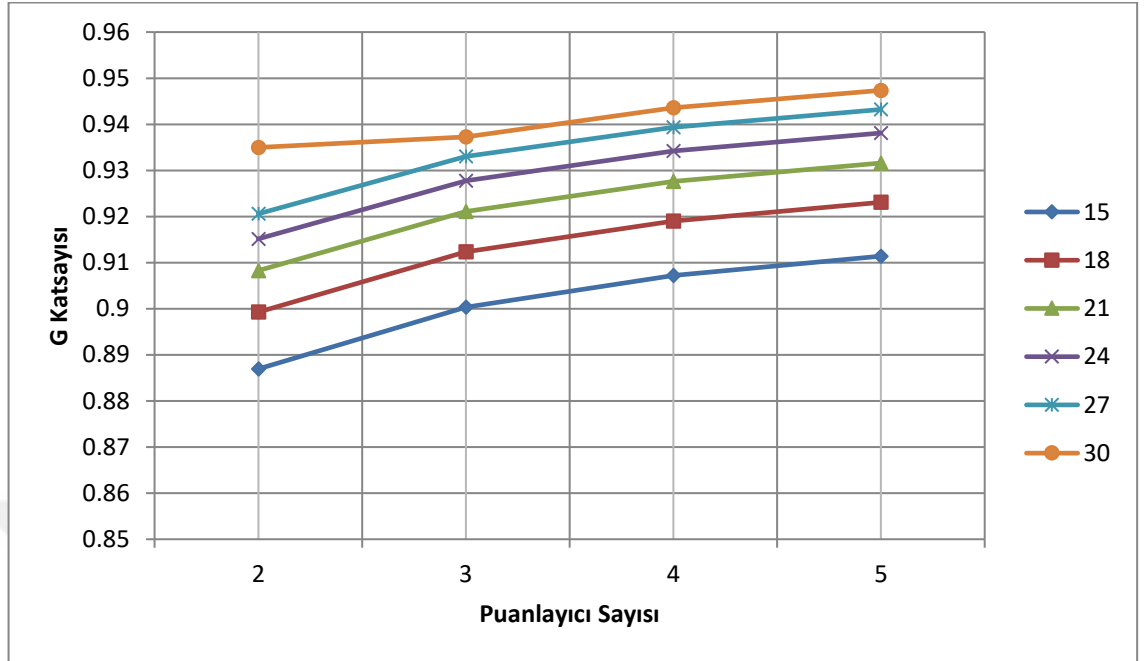
n_p	N_m	G Katsayısı	Phi Katsayısı
2	15	0,88694	0,86646
2	18	0,89926	0,87976
2	21	0,90827	0,88952
2	24	0,91515	0,89699
2	27	0,92058	0,90288
2	30	0,93498	0,91856
3	15	0,90036	0,88313
3	18	0,91234	0,89625
3	21	0,92109	0,90585
3	24	0,92776	0,91320
3	27	0,93302	0,91899
3	30	0,93727	0,92368
4	15	0,90723	0,89171
4	18	0,91902	0,90472
4	21	0,92763	0,91425
4	24	0,93420	0,92153
4	27	0,93937	0,92727
4	30	0,94355	0,93191
5	15	0,91140	0,89694
5	18	0,92308	0,90988
5	21	0,93160	0,91936
5	24	0,93811	0,92660
5	27	0,94322	0,93230
5	30	0,94736	0,93692

Not: Koyu yazılmış olanlar sınavda alt testlere göre kullanılan orijinal görev ve puanlayıcı sayısıdır.

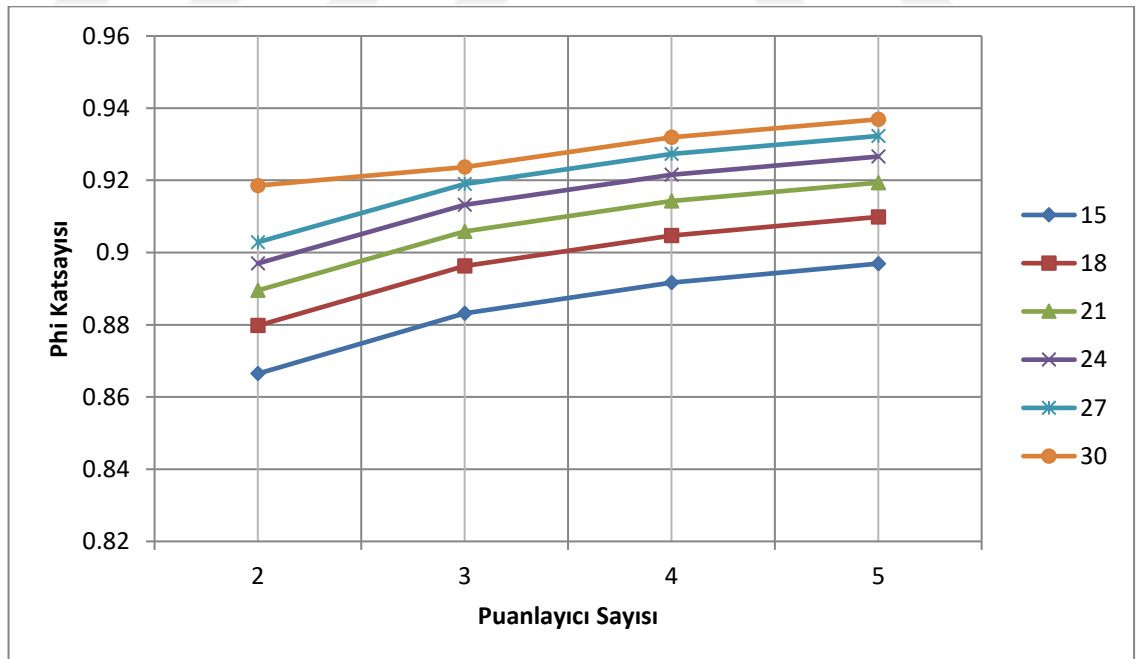
Tablo 6 incelendiğinde 15 madde ve 5 puanlayıcıdan oluşan orijinal bileşik testin tek değişkenli modelle kestirilen G katsayısının 0,91140 olduğu görülmektedir. Puanlayıcı sayısı sabit tutulup bileşik testte bulunan madde sayısı üç arttırıldığında ($n_m=18$) kestirilen G katsayısı 0,92308; bileşik testte bulunan madde sayısı altı

arttırıldığında ($n_m=21$) kestirilen G katsayısı 0,93160 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken ($n_p=5$) her bir alt testte bulunan madde sayısı üç; bileşik testte bulunan madde sayısı dokuz ($n_m=24$) arttırıldığında kestirilen G katsayısı 0,93811; bileşik testte bulunan madde sayısı on beş arttırıldığında ($n_m=30$) kestirilen G katsayısı 0,94736 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken madde sayısının arttırılması durumunda G katsayısı artmıştır. Örneğin puanlayıcı sayısı sabit kalmak koşuluyla toplam testte bulunan madde sayısı üç arttırıldığında G katsayısı 0,01168 artmış; madde sayısı altı arttırıldığında G katsayısı 0,0202 artmış ve madde sayısı on beş arttırıldığında G katsayısında 0,03596 artış olmuştur.

Çalışmada madde sayısının sabit tutulup ($n_m=15$) puanlayıcı sayısının bir azaltılması durumunda kestirilen G katsayısı 0,90723; puanlayıcı sayısının iki azaltılması durumunda kestirilen G katsayısı 0,90036; puanlayıcı sayısının üç azaltılması durumunda kestirilen G katsayısı ise 0,88694 olarak hesaplanmıştır. Puanlayıcı sayısının azaltılması durumunda G katsayısında beklendiği gibi düşüş olmuştur. Örneğin puanlayıcı sayısının bir azaltılması durumunda G katsayısı 0,00417; puanlayıcı sayısının iki azaltılması durumunda G katsayısı 0,01104; puanlayıcı sayısının üç azaltılması durumunda G katsayısı 0,02446 azalmıştır. Puanlayıcı sayısı sabitken ($n_p=5$) madde sayısının on beş arttırılması durumunda kestirilen G katsayısı 0,94736 olmuştur. Testin diğer alt boyutlarında da olduğu gibi bileşik test açısından da kestirilen G katsayısının arttırmak için madde sayısını arttırmak, puanlayıcı sayısını arttırmaktan daha uygun olacaktır.



Şekil 2. Açık Uçlu matematik sınavının serbest puanlanmasında farklı madde ve puanlayıcı sayılarına göre yapılan K çalışması sonucunda G katsayısındaki değişim grafiği.



Şekil 3. Açık Uçlu matematik sınavının serbest puanlanmasında farklı madde ve puanlayıcı sayılarına göre yapılan K çalışması sonucunda Phi katsayısındaki değişim grafiği.

Açık uçlu matematik sınavının serbest puanlanması ile elde edilen verilerin Tek Değişkenli G çalışmasıyla analizinden elde edilen sonuçlarının gösterildiği Tablo 6 ve Şekil 3 incelendiğinde; orijinal madde ve puanlayıcı sayısı ile yapılan K çalışması sonucu kestirilen Phi katsayısı 0,89694 olarak hesaplanmıştır. Madde sayısı sabit kalmak şartı ile ($n_m=15$) puanlayıcı sayısının bir azaltılması ($n_p=4$) ile kestirilen Phi katsayısında 0,00523 azalma olmuştur. Yine madde sayısı sabitken ($n_m=15$) puanlayıcı sayısının üç azaltılması ile hesaplana Phi katsayısı orijinal madde ve puanlayıcı sayısına göre yapılan K çalışmasında hesaplanan Phi katsayısına göre 0,03048 azalmıştır.

Puanlayıcı sayısı sabit kalmak şartıyla ($n_p=5$) toplam madde sayısının üç artırılması ($n_m=18$) ile yapılan K çalışmasında kestirilen Phi katsayısı 0,90988; toplam madde sayısının iki artırılması ($n_m=21$) ile yapılan K çalışmasında kestirilen Phi katsayısı 0,91936 olarak hesaplanmıştır. Puanlayıcı sayısı değiştirilmeden her bir alt testte bulunan madde sayısının beş, toplam madde sayısının on beş artırılması ile yapılan K çalışması sonucu kestirilen Phi katsayısı 0,93692 olarak hesaplanmıştır. Puanlayıcı sayısını değiştirmeden açık uçlu matematik sınavında bulunan toplam madde sayısı on beş arttırıldığında Phi katsayısında orijinal madde ve puanlayıcı sayısı ile yapılan K çalışmasında kestirilen Phi katsayısına göre 0,02704 artış hesaplanmıştır.

Puanlayıcı sayısının bir azaltılıp ($n_p=4$) her bir alt testte bulunan madde sayısı bir arttırılmak üzere toplam madde sayısı üç arttırıldığında hesaplanan Phi katsayısı orijinal madde ve puanlayıcı sayısı ile yapılan K çalışmasına göre hesaplanan Phi katsayısından 0,00516 fazladır. Puanlayıcı sayısı iki azaltılıp ($n_p=3$) toplam madde sayısının altı arttırılması ($n_m=21$) ile yapılan K çalışması sonucu kestirilen Phi katsayısı orijinal madde ve puanlayıcı sayısına göre hesaplanan Phi katsayısından 0,00403 daha yüksek hesaplanmıştır. Açık uçlu matematik sınavının serbest puanlanmasında bileşik testte bulunan madde sayısının arttırılması puanlayıcı sayısının arttırılmasına göre işgücü ve zaman açısından daha uygun olacaktır.

3.1.3. Çok değişkenli G çalışması sonuçları nasıldır?

Açık uçlu matematik sınavının puanlayıcılar tarafından serbest puanlanmasıyla elde edilen veriler üzerinde çok değişkenli G Kuramının $b \cdot x_m \cdot x_p$ deseni için G çalışması yapılmıştır. Bu çalışma ile varyans kaynakları için elde edilen varyans

büyükleri ile toplam varyansı açıklama oranları Tablo 7’de verilmiştir.

Tablo 7. Açık Uçlu Matematik Sınavının Serbest Puanlanmasıyla Elde Edilen Çok Değişkenli G Çalışması Sonuçları

Varyans Kaynağı*	Çok Değişkenli G Çalışması					
	Alt Test 1		Alt Test 2		Alt Test 3	
	Varyans	%	Varyans	%	Varyans	%
B	226,933	7,874	205,795	11,374	274,624	25,344
M	2446,319	84,878	999,627	55,246	545,292	50,322
P	154,782	5,370	527,177	29,135	206,718	19,077
Bxm	39,841	1,382	39,729	2,196	32,927	3,039
Bxp	4,398	0,153	7,712	0,426	5,321	0,491
Mxp	6,959	0,241	25,770	1,424	15,808	1,459
bxm xp,e	2,919	0,101	3,593	0,198	2,9079	0,268
Toplam	2882,151	100	1809,403	100	1083,598	100

* b: Birey, m: Madde, p:Puanlayıcı

Alt Test 1: Bu boyutta çok değişkenli analizle kestirilen en büyük varyans bileşeni 2446,31921 ile madde (m) ana etkisi için kestirilen varyans bileşenidir ve toplam varyansın%84,878’ini açıklamaktadır. Madde değişkenlik kaynağı madde güçlükleri arasındaki farklılıktan ortaya çıkar. Değişen madde güçlükleri, madde örneklerinin evrene genellenmesinin doğruluğunu etkileyecek ve hatta azaltacaktır (Güler, Uyanık ve Teker, 2012). Bu nedenle büyük olan madde varyansına dayalı olarak bazı maddelerin diğer maddelere göre zorluk ve kolaylık bakımından çok farklılaştığını söylenebilir. Bireyler için 226,933 olarak kestirilen varyans ise toplam varyansın %7,874’ünü açıklamaktadır. Bu durum ölçmenin uygulandığı grubun ölçülen yetenekleri bakımından çok farklılık göstermediğinin bir göstergesidir. Puanlayıcı ana etkisi için 154,78206 olarak kestirilen varyans toplam varyansın %5,370’ini açıklamaktadır. Puanlayıcılar için kestirilen bu varyans bileşeni puanlayıcıların bütün bireyler için verdikleri puanlarda farklılıkların olduğunun bir göstergesi olarak yorumlanabilir. Diğer bir deyişle bu boyut için puanlayıcıların katılık ve cömertliğinin bütün bireyler için birbirine yakın olmadığı, farklılaştığı söylenebilir.

Birey x madde (*bm*) ortak etkisi varyansı toplam varyansın %1,382'sini açıklamaktadır. Bu varyans bileşeni, maddelerin güçlük düzeylerinin bireyden bireye çok az farklılık gösterdiği şeklinde yorumlanabilir. Birey x puanlayıcı (*bp*) ortak etkisi varyansı toplam varyansın 0,153'ünü, açıklamaktadır. Bu da puanlayıcıların verdikleri puanların bireyden bireye farklılık göstermediğini ifade eder. Madde x puanlayıcı (*mp*) ortak etkisi varyansı ise toplam varyansın 0,241'ini açıklamaktadır. Bu da puanlayıcıların verdikleri puanların maddeden maddeye farklılık göstermediği şeklinde yorumlanabilir. Yapılan analiz sonucunda 2,91847 olarak kestirilen birey x madde x puanlayıcı ortak etkisi varyansının toplam varyans içindeki payı %0,101'dir. Artık varyans ya da hata varyansı olarak da adlandırılan bu varyansın sıfıra yakın olması nedeniyle bu alt testte birey x madde x puanlayıcı ortak etkisinden kaynaklanan tesadüfi hataların çok az olduğu söylenebilir.

Alt Test 2: Testin ikinci alt boyutu olan Alt Test 2'de çok değişkenli analizle kestirilen en büyük varyans bileşeni 999,62727 ile yine madde (*m*) ana etkisi için kestirilen varyans bileşenidir ve toplam varyansın %55,246'sını açıklamaktadır. Bu sonuç bu alt testte madde güçlüklerinin oldukça farklılaştığını göstermektedir. Puanlayıcı ana etkisi için 527,17708 olarak kestirilen varyans toplam varyansın %29,135'ini açıklamaktadır. Bu alt testte Puanlayıcı (*p*) ana etkisinin toplam varyans açıklama yüzdesi Alt Test 1'e göre daha yüksek bulunmuştur. Bu durum puanlayıcıların puanlamaları arasında Alt Test 1'e göre ve genel olarak tutarsızlık olduğunun göstergesidir. Bireyler için 205,79511 olarak kestirilen varyans bileşeni ise toplam varyansın %11,374'ünü açıklamaktadır. Ölçmenin amacı olarak tanımlanan birey (*b*) varyansının küçük olması Alt Test 2'de de evren puanı ya da klasik test kuramındaki gerçek puan varyansının düşük olduğunu göstermektedir. Birey x madde (*bm*) ortak etkisi toplam varyansın %2,196'sını açıklamaktadır. Bu ortak etki varyansının düşük olması nedeniyle madde güçlüklerinin bireyden bireye farklılık göstermediği söylenebilir. Birey x puanlayıcı (*bp*) ortak etkisinin toplam varyans içindeki yüzdesinin %0,426 olması nedeniyle puanlayıcıların puanlamalarının bireyden bireye farklılık göstermediği söylenebilir. Madde x puanlayıcı (*mp*) ortak etkisi ise toplam varyansın %1,424'ünü açıklamaktadır. Bu nedenle puanlayıcıların puanlamaları maddeden maddeye farklılık göstermemiştir. Testin bu alt boyutunda birey x madde x puanlayıcı

ortak etkisi varyansı 3,59313 olarak kestirilmiştir. Artık ya da hata varyansı olarak adlandırılan bu varyans toplam varyansın %0,198'ini oluşturmaktadır. Bu sonuç; Alt Test 1'de olduğu gibi Alt Test 2'de de bireyler, maddeler ve puanlayıcılar arası etkileşim ile bu çalışmada ölçülemeyen sistematik ya da tesadüfi değişkenlik kaynaklarının oldukça düşük olduğunun bir göstergesidir.

Alt Test 3: Testin son alt boyutu olan Alt Test 3'te en büyük varyans bileşeni 545,292 olarak kestirilen madde(*m*) ana etkisi için kestirilen varyans bileşenidir. Toplam varyansı açıklama yüzdesi ise %50,322'dir. İkinci en büyük varyans bileşeni ise 274,62438 olarak kestirilen birey (*b*) ana etkisi için varyans bileşenidir ve toplam varyansın %25,344'ünü açıklamaktadır. Analizde 206,71779 olarak kestirilen puanlayıcı (*p*) ana etkisi ise toplam varyansın %19,077'sini açıklamaktadır.

Birey x madde (*bm*) ortak etkisi için 32,92678 olarak kestirilen varyans toplam varyansın %3,039'unu; birey x puanlayıcı (*bp*) ortak etkisi için 5,32112 olarak hesaplanan varyans toplam varyansın %0,491'ini; madde x puanlayıcı (*mp*) ortak etkisi için 15,80771 olarak kestirilen varyans toplam varyansın %1,459'unu açıklamaktadır. Artık varyans olarak tanımlanan birey x madde x puanlayıcı ortak etkisi için varyans büyüklüğü 2,90787 olarak kestirilmiştir. Bu varyans ise toplam varyansın sadece %0,268'ini açıklamaktadır.

Madde(*m*) ana etkisi için kestirilen varyans bileşeninin en büyük değeri alması bazı maddelerin bazı bireylere diğer bireylere göre daha zor ya da kolay geldiğinin bir göstergesidir. Birey(*b*) ana etkisi için kestirilen varyans bileşeninin toplam varyansın %25'ini açıklaması bireyler arasındaki farklılıkların ortaya konulabildiğinin göstergesidir. Puanlayıcı ana etkisinin toplam varyansı açıklama yüzdesinin bu boyutta yüksek olması puanlayıcıların Alt Test 3'ü puanlarken birbirlerine göre daha katı ya da cömert puanlama yaptıkları şeklinde yorumlanabilir. Birey x madde (*bm*) ortak etkisinin toplam varyansa etkisinin düşük olması bireylerin bir maddeden diğerine bağlı durumlarının çok farklılık göstermediği şeklinde açıklanabilir. Birey x puanlayıcı (*bp*) ortak etkisinin toplam varyansı açıklama yüzdesi madde x puanlayıcı (*mp*) ortak etkisinin toplam varyansı açıklama yüzdesine göre daha düşük olması puanlayıcıların yaptıkları puanlamaların bir maddeden diğerine nispeten daha tutarsız olduğunun göstergesidir. Birey x madde x puanlayıcı ortak etkisinin toplam varyansı açıklama

yüzdesinin çok düşük olması bu alt testte tesadüfî hataların çok az olduğunu göstergesidir.

3.1.4. Çok değişkenli K çalışması sonuçları nasıldır?

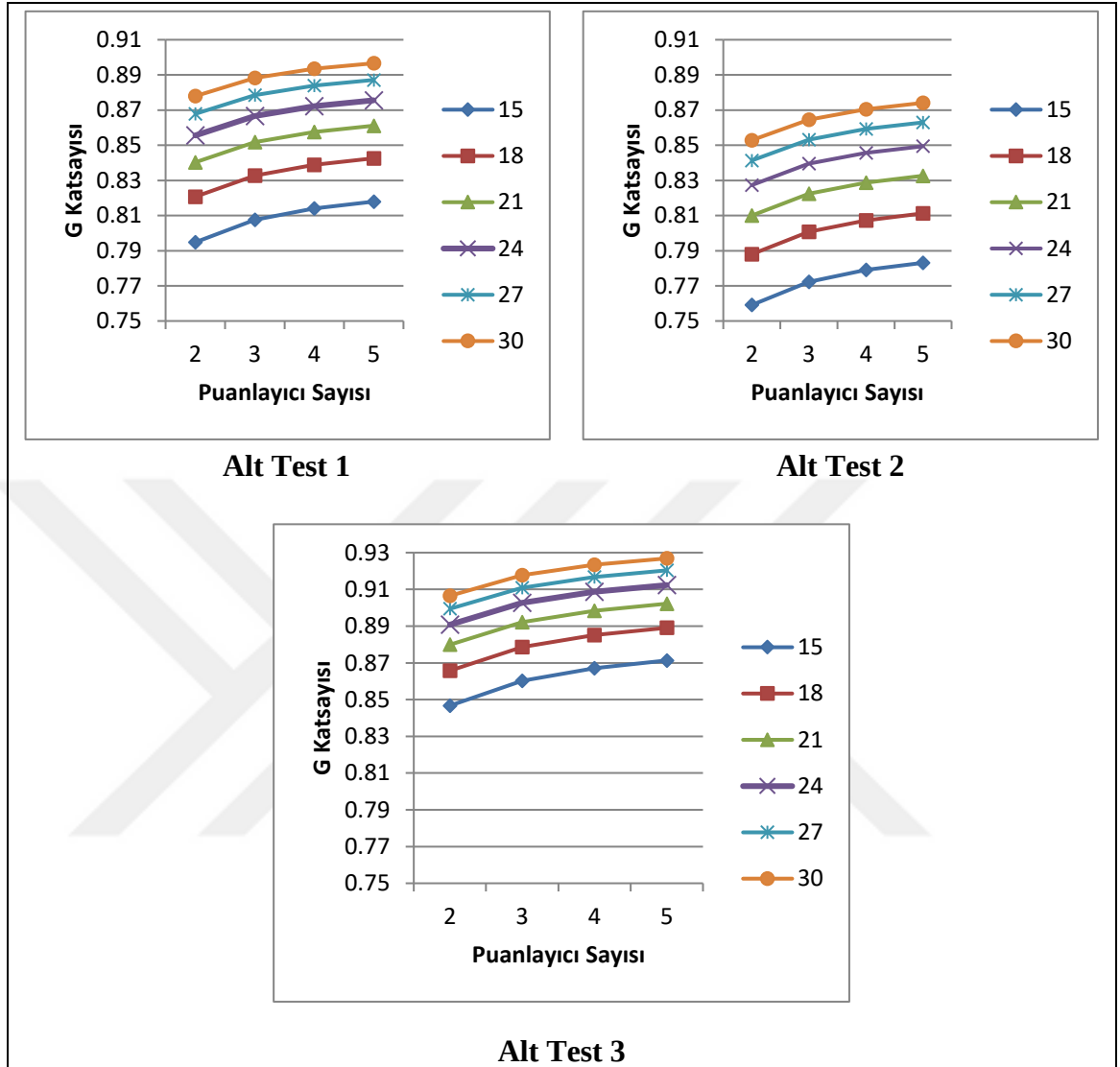
Açık uçlu matematik sınavının serbest puanlanmasında her bir alt testte bulunan madde sayısının sırasıyla 5, 6, 7, 8, 9, 10 yapılması ve puanlayıcı sayısının 5, 4, 3, 2 olarak alınması durumları için Çok Değişkenli K çalışması yapılmıştır. Bu K Çalışmasından elde edilen sonuçlar aşağıda Tablo 8’de verilmiştir.



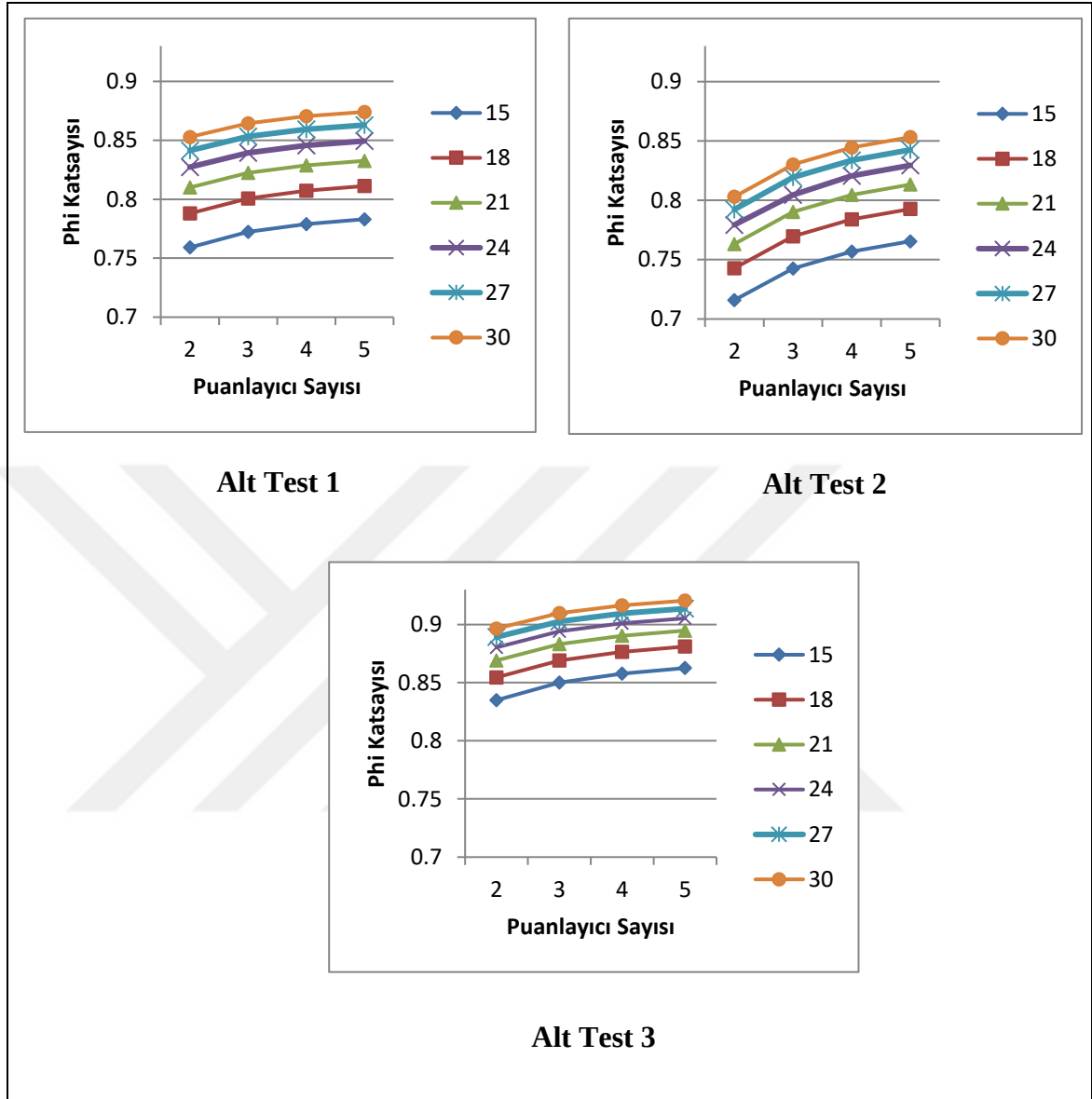
Tablo 8. Açık Uçlu Matematik Sınavının Serbest Puanlanmasında Alternatif Madde ve Puanlayıcı Sayıları Senaryoları ile Yapılan Çok Değişkenli K Çalışması Sonuçları

		G Katsayısı			Phi Katsayısı		
n_p	n_m	Alt Test 1	Alt Test 2	Alt Test 3	Alt Test 1	Alt Test 2	Alt Test 3
2	5	0,79481	0,75911	0,84671	0,75911	0,71594	0,83499
2	6	0,82071	0,78799	0,86574	0,78799	0,74278	0,85457
2	7	0,84027	0,80999	0,87988	0,80999	0,76322	0,86913
2	8	0,85556	0,82732	0,89078	0,82732	0,77930	0,88037
2	9	0,86784	0,84132	0,89945	0,84132	0,79229	0,88932
2	10	0,87793	0,85286	0,90651	0,85286	0,80299	0,89662
3	5	0,80749	0,77225	0,86020	0,77225	0,74266	0,85012
3	6	0,83271	0,80074	0,87856	0,80074	0,76971	0,86912
3	7	0,85172	0,82242	0,89215	0,82242	0,79026	0,88322
3	8	0,86656	0,83946	0,90263	0,83946	0,80462	0,89410
3	9	0,87846	0,85320	0,91095	0,85320	0,81945	0,90275
3	10	0,88822	0,86453	0,91772	0,86453	0,83018	0,90979
4	5	0,81397	0,77900	0,86711	0,77900	0,75678	0,85788
4	6	0,83885	0,80728	0,88511	0,80728	0,78391	0,87658
4	7	0,85757	0,82877	0,89842	0,82877	0,80452	0,89044
4	8	0,87216	0,84566	0,90867	0,84566	0,82070	0,90113
4	9	0,88386	0,85927	0,91681	0,85927	0,83374	0,90962
4	10	0,89345	0,87049	0,92342	0,87049	0,84447	0,91652
5	5	0,81792	0,78310	0,87132	0,78310	0,76551	0,86261
5	6	0,84257	0,81125	0,88908	0,81125	0,79269	0,88112
5	7	0,86111	0,83263	0,90222	0,83263	0,81332	0,89483
5	8	0,87556	0,84942	0,91234	0,84942	0,82951	0,90539
5	9	0,88714	0,86296	0,92036	0,86296	0,84255	0,91379
5	10	0,89662	0,87410	0,92688	0,87410	0,85329	0,92061

Not: Koyu yazılmış olanlar sınavda alt testlere göre kullanılan orijinal görev ve puanlayıcı sayısıdır.



Şekil 4. Açık Uçlu Matematik Sınavının Serbest Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan Çok Değişkenli K Çalışması Sonucunda Alt Testlere Göre G Katsayısındaki Değişim Grafiği



Şekil 5. Açık Uçlu Matematik Sınavının Serbest Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan Çok Değişkenli K Çalışması Sonucunda Alt Testlere Göre Phi Katsayısındaki Değişim Grafiği

Tablo 8 ve Şekil 4 incelendiğinde Alt Test 1 boyutunda orijinal puanlayıcı sayısı 5 ve orijinal madde sayıları 5 için çok değişkenli $b \times m \times p$ modeli ile yapılan K çalışmasında kestirilen G katsayısı 0,81792 olarak hesaplanmıştır. Madde sayısının sabit tutularak ($n_m=5$) puanlayıcı sayısının bir azaltılması ($n_p=4$) durumunda kestirilen G katsayısı 0,81397; puanlayıcı sayısının iki azaltılması ($n_p=3$) durumunda kestirilen G katsayısı 0,80749; puanlayıcı sayısının üç azaltılması ($n_p=2$) durumunda kestirilen G

katsayısı 0,80749 olarak hesaplanmıştır. Bu sonuçlara göre puanlayıcı sayısının 4, 3,2 olarak alınması durumunda G katsayısındaki azalma sırasıyla 0,003, 0,010 ve 0,023 olarak hesaplanmıştır. Tablo 8’de görüldüğü gibi Alt Test 1’de puanlayıcı ana etkisinin toplam varyansa katkısının %5,370 olması nedeniyle puanlayıcı sayısının azaltılması G katsayısını önemli ölçüde etkilememiştir. Buna karşılık puanlayıcı sayısının sabit tutulup madde sayısının artırılması durumunda G katsayısının daha fazla etkilendiği (arttığı) görülmüştür.

Puanlayıcı sayısının sabit tutulup ($n_p=5$) madde sayısının bir artırılması durumunda ($n_m=6$) kestirilen G katsayısı 0,84257 olarak hesaplanmıştır. Madde sayısının iki artırılması durumunda ($n_m=7$) kestirilen G katsayısı 0,86111; madde sayısının üç artırılması durumunda ($n_m=8$) G katsayısı 0,87556; madde sayısının dört artırılması durumunda kestirilen G katsayısı ($n_m=9$) ve madde sayısının beş artırılması ($n_m=9$) durumunda kestirilen G katsayısı 0,87410 olarak hesaplanmıştır. Bu sonuçlar dikkate alındığında puanlayıcı sayısı sabitken madde sayısının bir artırılması G katsayısını 0,024 arttırmıştır. Yine puanlayıcı sayısı sabitken madde sayısının iki artırılması G katsayısını 0.043 arttırmıştır. Puanlayıcı sayısı sabitken madde sayısının beş artırılması ise G katsayısı üzerinde 0,078’e kadar artış sağlamıştır. Puanlayıcı sayısının artırılması G katsayısına madde sayısının artırılması kadar katkı sağlamamıştır. Örneğin orijinal puanlayıcı sayısının bir azaltılıp ($n_p=4$) madde sayısının bir artırılması ile ($n_m=6$) yapılan K çalışması sonucu kestirilen G katsayısı 0,83885 olarak hesaplanmıştır. Puanlayıcı sayısının iki azaltılıp ($n_p=3$) madde sayısının iki artırılması ($n_m=7$) durumunda kestirilen G katsayısı 0,85172 olarak hesaplanmıştır. Puanlayıcı sayısının üç azaltılıp ($n_p=3$) madde sayısının beş artırılması durumunda ($n_m=10$)kestirilen G katsayısı 0,87793 olarak hesaplanmıştır. Çalışmada kestirilen en yüksek G katsayısı olan 0,89662; puanlayıcı sayısının beş, madde sayısının on olduğu senaryoda kestirilen G katsayısıdır. Görüldüğü gibi puanlayıcı sayısının artırılmasındansa madde sayısının artırılması G katsayısını arttırmakta hem de zaman, işgücü ve ekonomi açısından kolaylık sağlamaktadır.

Alt test 2 boyutunda orijinal puanlayıcı sayısının beş($n_p=5$) ve orijinal madde sayısının beş ($n_m=5$) olduğu durum için çok değişkenli $b \cdot x^m \cdot p$ modelle yapılan K çalışması ile kestirilen G katsayısı 0,78694 olarak hesaplanmıştır. Bu boyut için madde

sayısının sabit tutularak ($n_m=5$) puanlayıcı sayısının bir azaltılması sonucu ($n_p=4$) elde edilen G katsayısı 0,77963 olarak hesaplanmıştır. Yine madde sayısı sabitken ($n_m=5$) puanlayıcı sayısının iki azaltılması ($n_p=3$) sonucu kestirilen G katsayısı 0,76776; puanlayıcı sayısının üç azaltılması ($n_p=2$) sonucu kestirilen G katsayısı ise 0,74506 olarak hesaplanmıştır. Madde sayısının sabit tutulması durumunda puanlayıcı sayısının bir azaltılması sonucu kestirilen G katsayısındaki azalma 0,007 iken puanlayıcı sayısının iki azaltılması ile G katsayısındaki azalma 0,019 olarak hesaplanmıştır. Puanlayıcı sayısı ikiye düşürüldüğünde ise G katsayısındaki azalma 0,041 olmuştur. Madde sayısı sayısının sabit tutulup puanlayıcı sayısının azaltılması durumunda alt test 2 boyutunda alt test 1 boyutuna göre G katsayısında daha fazla azalma olmuştur. Bunun nedeni Tablo 8’de görüldüğü gibi Alt Test 1’de puanlayıcı ana etkisinin varyans açıklama yüzdesinin %5,370 iken Alt Test 2’de puanlayıcı ana etkisinin varyans açıklama yüzdesinin %29,135 olmasıdır.

Puanlayıcı sayısının sabit tutulup ($n_p=5$)madde sayısının bir arttırıldığı ($n_m=6$) durumda hesaplanan G katsayısı 0,81310 iken madde sayısının iki arttırıldığı ($n_m=7$) durumda hesaplanan G katsayısı 0,83288 olmuştur. Puanlayıcı sayısı sabit tutulup ($n_p=5$) madde sayısının beş arttırıldığı durumda hesaplanan G katsayısı 0,87101 olmuştur. Puanlayıcı sayısı sabit tutulup madde sayısının bir arttırılması ile G katsayısındaki artış 0,026 olmuş, madde sayısının iki arttırılması ile kestirilen G katsayısındaki artış 0,045 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken madde sayısı birer birer arttırılmıştır. Çalışmanın Alt Test 2 boyutunda kestirilen en yüksek G katsayısı puanlayıcı sayısının beş ($n_p=5$), madde sayısının on ($n_m=10$) alınması durumunda 0,87101 olarak hesaplanmıştır. Tablo 8 ve Şekil 4 incelendiğinde madde ana etkisinin varyans açıklama yüzdesinin %55,246 puanlayıcı ana etkisinin varyans açıklama yüzdesine %29,135göre daha yüksek olduğu görülmektedir. Bundan dolayı madde sayısının arttırılmasının G katsayısı üzerinde puanlayıcı sayısının arttırılmasına göre daha fazla etkisi vardır. Örneğin puanlayıcı sayısının üç ($n_p=3$) madde sayısının on ($n_m=10$) olarak alınması durumunda kestirilen G katsayısı 0,85292 iken puanlayıcı sayısının dört ($n_p=4$) madde sayısının on ($n_m=10$) olarak hesaplanması durumunda kestirilen G katsayısı 0,86414 olarak hesaplanmıştır. Madde sayısı on ($n_m=10$) puanlayıcı sayısı üç ($n_p=3$) iken puanlayıcı sayısının bir arttırılması durumunda

kestirilen G katsayısında 0,011 artış gözlenmiştir. Puanlayıcı sayısının arttırılması fazla işgücü gerektirdiği, ekonomik olmadığı ve G katsayısına fazlaca katkı sağlamadığı için testin bu alt boyutunda G katsayısını arttırmak için madde sayısının arttırılması önerilebilir.

Alt test 3 boyutunda orijinal puanlayıcı sayısı beş ($n_p=5$) ve orijinal madde sayıları beş ($n_m=5$) için çok değişkenli modelle K çalışması için kestirilen G katsayısı 0,87132 olarak hesaplanmıştır. Tablo 8 incelendiğinde bu boyut için madde sayısı sabit tutularak ($n_m=5$) puanlayıcı sayısının bir azaltılması ($n_p=4$) durumunda kestirilen G katsayısı 0,86711 olarak hesaplanmıştır. Madde sayısının sabit tutularak ($n_m=5$) puanlayıcı sayısının iki azaltılması ($n_p=3$) durumunda kestirilen G katsayısı 0,86020 olarak hesaplanmıştır. Madde sayısının sabit tutularak ($n_m=5$) puanlayıcı sayısının üç azaltılması durumunda ($n_p=2$) kestirilen G katsayısı 0,84671 olarak hesaplanmıştır. Görev sayısının sabit tutulması koşuluyla puanlayıcı sayısının bir azaltılması durumunda G katsayısındaki azalma 0,004 olarak hesaplanmıştır. Puanlayıcı sayısının sabit tutulup ($n_p=5$) madde sayısının bir arttırılması ($n_m=6$) sonucu kestirilen G katsayısı 0,88908; madde sayısının iki arttırılması ($n_m=7$) sonucu kestirilen G katsayısı 0,90222 ve madde sayısının üç arttırılması ($n_m=8$) ile kestirilen G katsayısı 0,91234 olarak hesaplanmıştır. Bu alt boyut için yapılan K çalışmasında yine puanlayıcı sayısının sabit tutulması koşuluyla madde sayısının dört arttırılması ($n_m=9$) durumunda kestirilen G katsayısı 0,92036 iken madde sayısının beş arttırılması durumunda ($n_m=10$) kestirilen G katsayısı 0,92688 olmuştur. Puanlayıcı sayısının sabit tutulup madde sayısının arttırılması bu boyutta da G katsayısını önemli ölçüde arttırmıştır. Örneğin orijinal madde sayısı 5 ($n_m=5$) ve puanlayıcı sayısına 5 ($n_p=5$) göre madde sayısının bir arttırılması durumunda G katsayısında 0,017 artış olmuştur. Madde sayısı iki arttırıldığında ise G katsayısındaki artış 0,030 olmuştur. Madde sayısı beş arttırıldığında G katsayısı 0,049 artmıştır. Bu boyut için de işgücü, zaman ve ekonomi düşünüldüğünde madde sayısının arttırılması uygun olacaktır.

Tablo 8 ve Şekil 5 incelendiğinde Alt Test 1 boyutunda orijinal puanlayıcı sayısı beş ($n_p=5$) ve orijinal madde sayısı beş ($n_m=5$) için çok değişkenli modelle K çalışması için kestirilen Phi katsayısı 0,78310 olarak hesaplanmıştır. Tablo 8’de görüldüğü bu

boyut için madde sayısı sabitken ($n_m=5$) puanlayıcı sayısının bir azaltılması ($n_p=4$) ile kestirilen Phi katsayısı 0,7790 olarak hesaplanmıştır. Madde sayısı sabit kalmak koşulu ile puanlayıcı sayısının iki azaltılması ($n_p=3$) durumunda hesaplanan Phi katsayısı 0,77225 olmuştur. Puanlayıcı sayısının üç azaltılması ($n_p=2$) durumunda yapılan K çalışmasında Phi katsayısı ise 0,75911 olmuştur. Madde sayısı sabitken puanlayıcı sayısının bir azaltılması durumunda Phi katsayısında 0,004 azalma olmuştur. Puanlayıcı sayısının iki azaltılması ile yapılan K çalışmasında Phi katsayısında 0,010; puanlayıcı sayısının üç azaltılması ile yapılan K çalışmasında ise hesaplanan Phi katsayısında 0,023 azalma olmuştur.

Puanlayıcı sayısı sabitken ($n_p=5$) madde sayısının bir arttırılması ($n_m=6$) ile yapılan K çalışmasında Phi katsayısı 0,81125; madde sayısının iki arttırılması ($n_m=7$) ile kestirilen Phi katsayısı 0,83263; madde sayısının üç arttırılması ($n_m=8$) ile hesaplanan Phi katsayısı ise 0,84942 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken madde sayısının dört arttırılması ($n_m=9$) ile kestirilen Phi katsayısı 0,86296; madde sayısının beş arttırılması ile kestirilen Phi katsayısı 0,87410 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken ($n_p=5$) madde sayısının iki arttırılması ($n_m=7$) ile yapılan K çalışmasında hesaplanan Phi katsayısı, orijinal madde ve puanlayıcı sayısına göre hesaplanan Phi katsayısından 0,04 fazladır. Buna karşılık madde sayısı sabitken ($n_m=5$) puanlayıcı sayısının iki azaltılması ($n_p=3$) ile yapılan K çalışmasında hesaplanan Phi katsayısında orijinal madde ve puanlayıcı sayısına göre hesaplanan Phi katsayısına göre 0,010 azalma olmuştur. Bu nedenle yapılan çalışmada G katsayısında olduğu gibi Phi katsayısını arttırmak için de madde sayısının arttırılması puanlayıcı sayısının arttırılmasından daha çok katkı sağlayacaktır.

Alt test 2 boyutunda orijinal puanlayıcı beş ($n_p=5$) ve orijinal madde sayıları beş ($n_m=5$) için çok değişkenli modelle K çalışması için kestirilen Phi katsayısı 0,76551 olarak hesaplanmıştır. Tablo 8 incelendiğinde madde sayısı sabitken ($n_p=5$) puanlayıcı sayısının bir azaltılması ($n_p=4$) ile kestirilen Phi katsayısı 0,75678; puanlayıcı sayısının iki azaltılması ile ($n_p=3$) kestirilen Phi katsayısı 0,74266 ve puanlayıcı sayısının üç azaltılması ($n_p=2$) ile kestirilen Phi katsayısı ise 0,71594 olarak hesaplanmıştır. Alt test 2’de bulunan madde sayısı değiştirilmeden puanlayıcı sayısı bir azaltıldığında ($n_p=4$)

kestirilen Phi katsayısı 0,008 azalmıştır. Madde sayısı değiştirilmeden puanlayıcı sayısı ikiye düşürüldüğünde ise kestirilen Phi katsayısında 0,04 azalma olmuştur.

Puanlayıcı sayısı sabit kalmak koşulu ile ($n_p=5$) madde sayısının bir arttırılması ile ($n_m=6$) kestirilen Phi katsayısı ise 0,80474 olarak hesaplanmıştır. Madde sayısının sırasıyla iki, üç dört ve beş arttırılması ile hesaplanan Phi katsayıları sırasıyla 0,81332; 0,82951; 0,84255 ve 0,85329 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken ($n_p=5$) madde sayısının bir arttırılması ile ($n_p=6$) kestirilen Phi katsayısı 0,008 artmıştır. Madde sayısının beş arttırılması ile hesaplanan Phi katsayısında 0,04 artma olmuştur.

Açık uçlu matematik sınavının serbest puanlanması ile oluşan verilerin çok değişkenli G kuramı ile analizinde Alt Test 2 boyutunda madde sayısının arttırılması veya puanlayıcı sayısının azaltılması Phi katsayısına aynı oranda katkı sağlamıştır. Fakat her bir alt testteki madde sayısının arttırılması puanlayıcı sayısının arttırılması işlemine göre daha az zaman gerektiren ve daha ekonomik bir yöntem olduğu için serbest puanlamada Alt test 2 boyutunda madde sayısının arttırılması Phi katsayısını arttırmak için tercih edilebilir.

Alt test 3 boyutunda orijinal puanlayıcı beş ($n_p=5$) ve orijinal madde sayıları beş ($n_m=5$) için çok değişkenli modelle K çalışması için kestirilen Phi katsayısı 0,86261 olarak hesaplanmıştır. Tablo 8 incelendiğinde madde sayısı sabit tutulup ($n_m=5$) puanlayıcı sayısı bir azaltıldığında ($n_p=4$) hesaplanan Phi katsayısı 0,85788 olarak hesaplanmıştır. Puanlayıcı sayısı iki azaltıldığında ($n_p=3$) hesaplanan Phi katsayısı 0,85012; puanlayıcı sayısı üç azaltıldığında ($n_p=2$) kestirilen Phi katsayısı 0,83499 olarak hesaplanmıştır. Puanlayıcı sayısı sabit kalmak şartıyla ($n_p=5$) madde sayısının bir arttırılması ile kestirilen Phi katsayısı 0,88112; madde sayısının beş arttırılması ile kestirilen Phi katsayısı ise 0,92061 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken Alt Test 3'te bulunan madde sayısının bir arttırılması ile Phi katsayısında 0,018; Alt Test 3'te bulunan madde sayısının beş arttırılması ile Phi katsayısında 0,058 artış hesaplanmıştır. Madde sayısının iki arttırılıp puanlayıcı sayısının bir azaltılması ile kestirilen Phi katsayısında orijinal madde ve puanlayıcı sayısına göre 0,027 artış hesaplanmıştır. Madde sayısının beş arttırılıp ($n_m=10$) puanlayıcı sayısının iki azaltılması ($n_p=3$) ile yapılan K çalışmasında Phi katsayısında orijinal madde ve puanlayıcı sayısına göre 0,04 artış hesaplanmıştır. Bu nedenle alt test 3 boyutunda Phi

katsayısını arttırmak için madde sayısının artırılması zaman ve işgücü açısından daha kolay olduğu için tercih edilebilir.

3.2.Alt Problem 2.Tüm bireyleri (b), tüm maddelerde (m) aynı puanlayıcıların (p) anahtarla puanladığı açık uçlu matematik sorularının kullanıldığı ölçme durumunda;

3.2.1. Tek değişkenli G çalışması sonuçları nasıldır?

Açık uçlu matematik sınavının puanlayıcılar tarafından anahtarla puanlanmasıyla elde edilen veriler üzerinde tek değişkenli G Kuramının $b \times m \times p$ deseni için G çalışması yapılmıştır. Bu çalışma ile varyans kaynakları için elde edilen varyansı büyüklüğü ile toplam varyansı açıklama oranları Tablo 9’da verilmiştir.

Tablo 9. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasıyla Elde Edilen Tek Değişkenli G Çalışması Sonuçları

Varyans Kaynağı	Varyans	Toplam Varyans İçindeki Payı (%)
Birey (b)	542,42269	22,6204
Madde (m)	1778,85613	74,1828
Puanlayıcı (p)	8,11080	0,3382
Birey-Madde ($b \times m$)	48,82684	2,0362
Birey-Puanlayıcı ($b \times p$)	2,13461	0,089
Madde-Puanlayıcı ($m \times p$)	15,95352	0,6653
Birey-Madde-Puanlayıcı ($b \times m \times p$,e)	1,62974	0,0679
Toplam	2397,9343	100

Tablo 9 incelendiğinde açık uçlu matematik sınavının anahtarla puanlanmasıyla elde edilen verilerle yapılan tek değişkenli G çalışması sonucunda kestirilen en büyük varyans 1778,85613 ile toplam varyansın %74,1828’ini açıklayan madde (m) ana etkisi varyansıdır. Tüm varyans bileşenleri içinde en büyük olan madde varyansı testin maddelerinin birbirinden oldukça farklılaştığını göstermektedir. Büyüklük bakımından ikinci sırada olan birey (b) ana etkisi için kestirilen varyans bileşeni 542,42269 olarak toplam varyansın %22,6204’ünü açıklamaktadır. Birey (b) ana etkisi için kestirilen varyans bileşeninin toplam varyansı açıklama yüzdesinin büyük olması beklenmektedir. Bu varyans bileşeninin büyük olması nedeniyle maddeleri cevaplayan bireylerin bağlı durumlarının maddeden maddeye değişmediği söylenebilir. Puanlayıcı (p) ana etkisi

için kestirilen varyans bileşeni 8,11080 olarak toplam varyansın %0,3382'sini açıklamaktadır. Puanlayıcı (p) ana etkisi için kestirilen varyans bileşeninin toplam varyansı açıklama yüzdesinin küçük olması puanlayıcıların puanlamalarında farklılık olmadığının göstergesidir.

Birey $\times$ puanlayıcı(bp) ortak etkisinin 2,13461 varyans oranına ve %0,089varyans açıklama yüzdesine sahip olması nedeniyle belli puanlayıcıların belli bireyler için yaptıkları puanlamalarının diğer puanlayıcılara göre fazlaca değişmediğini gösterir. Madde $\times$ puanlayıcı (mp) ortak etkisi 15,95352 varyans oranına ve %0,6653varyans açıklama yüzdesine sahiptir. Bu varyans bileşeninin payının toplam varyans içinde küçük olması nedeniyle puanlayıcıların bir maddeden diğer maddeye puanlamalarının farklılaşmadığı söylenebilir. Ölçülemeyen varyans kaynaklarını bileşik olarak gösteren birey $\times$ madde $\times$ puanlayıcı (bmp) ortak etkisi toplam varyansı açıklamada 1,62974 varyans oranı ve %0,0679varyans yüzdesi ile en düşük paya sahiptir. Bu varyansın yüzdesinin sıfıra yakın olması birey $\times$ madde $\times$ puanlayıcı ortak etkisinden kaynaklanan hataların az olduğunu göstermektedir.

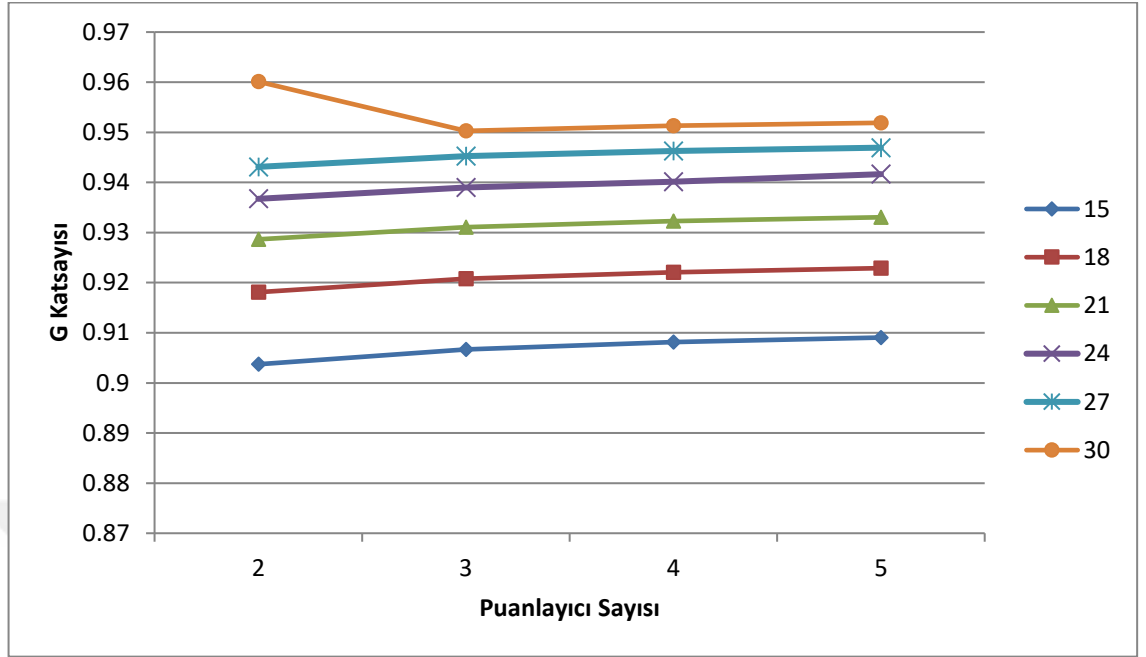
3.2.2. Tek değişkenli K çalışması sonuçları nasıldır?

Açık uçlu matematik sınavının anahtarla puanlanmasıyla elde edilen puanlarda tüm değişkenlerin çaprazlandığı $b \times m \times p$ deseninde birey (b) ölçmenin nesnesi (object of measurement), m , p ise yüzey (facet) olarak ele alınır. Elde edilen varyans bileşenlerinden hesaplanan mutlak ve bağıl hata varyanslarına ve görev, puanlayıcı sayılarının artırılıp azaltılmasına dayalı senaryolar için kestirilen genellenebilirlik katsayısı (G), ϕ (Φ) katsayısına ait değerler Tablo 10'da verilmiştir.

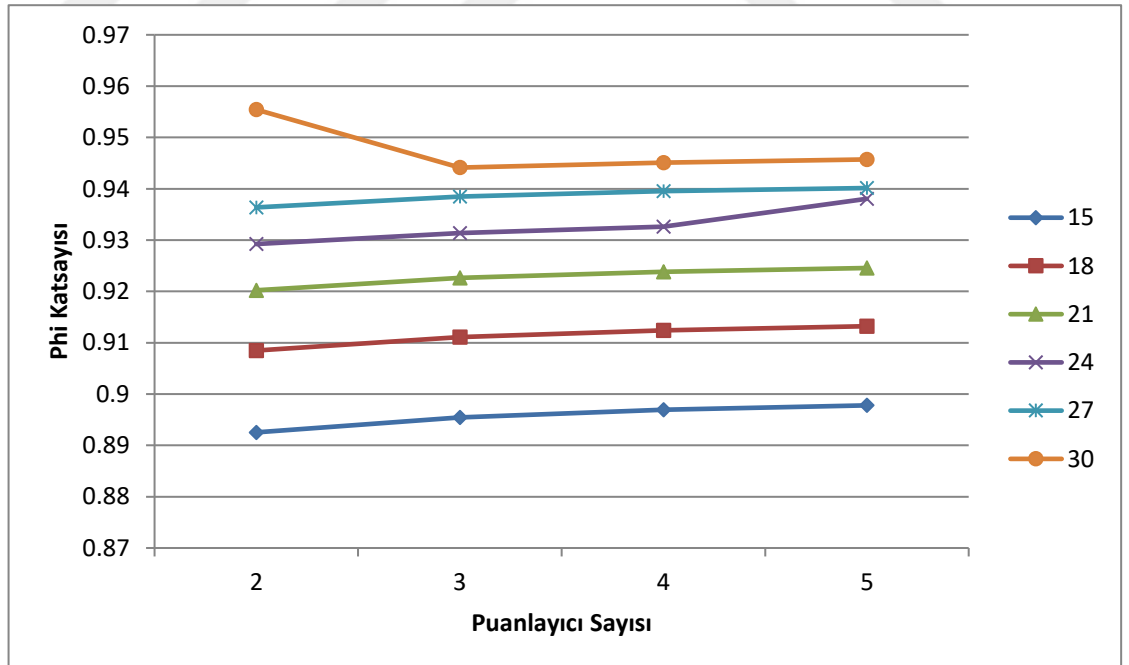
Tablo 10. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasında Bileşik Testte Bulunan Farklı Madde ve Puanlayıcı Sayısı Senaryoları ile Yapılan K Çalışmasının Sonuçları.

n_p	n_m	G Katsayısı	Phi Katsayısı
2	15	0,90372	0,89253
2	18	0,91810	0,90848
2	21	0,92865	0,92022
2	24	0,93673	0,92923
2	27	0,94311	0,93637
2	30	0,96011	0,95543
3	15	0,90667	0,89544
3	18	0,92076	0,91110
3	21	0,93109	0,92263
3	24	0,93899	0,93137
3	27	0,94523	0,93846
3	30	0,95028	0,94413
4	15	0,90816	0,89691
4	18	0,92210	0,91242
4	21	0,93231	0,92384
4	24	0,94013	0,93259
4	27	0,94630	0,93951
4	30	0,95129	0,94512
5	15	0,90905	0,89779
5	18	0,92290	0,91322
5	21	0,93305	0,92456
5	24	0,94165	0,93805
5	27	0,94694	0,94014
5	30	0,95190	0,94571

Not: Koyu yazılmış olanlar sınavda kullanılan orijinal görev ve puanlayıcı sayısıdır.



Şekil 6. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan K Çalışması Sonucunda G Katsayısındaki Değişim Grafiği



Şekil 7. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan K Çalışması Sonucunda Phi Katsayısındaki Değişim Grafiği

Tablo 10 ve Şekil 6 incelendiğinde 15 madde ve 5 puanlayıcıdan oluşan orijinal bileşik testin anahtarla puanlanması sonucu elde edilen puanlamaların tek değişkenli modelle kestirilen G katsayısının 0,90905 olduğu görülmektedir. Puanlayıcı sayısı sabit tutulup bileşik testte bulunan madde sayısı üç arttırıldığında ($n_m=18$) kestirilen G katsayısı 0,92290; bileşik testte bulunan madde sayısının altı arttırılması ($n_m=21$) ile kestirilen G katsayısı 0,93305 olarak, bileşik testteki toplam madde sayısının dokuz arttırılması ($n_m=24$) ile kestirilen G katsayısı ise 0,94165 olarak hesaplanmıştır. Puanlayıcı sayısı sabit kalmak koşulu ile ($n_p=5$) madde sayısının on iki arttırılması ($n_m=27$) ile kestirilen G katsayısı 0,95300 olarak, madde sayısının on beş arttırılması ile ($n_m=30$) kestirilen G katsayısı ise 0,95739 olarak hesaplanmıştır. Puanlayıcı sayısı sabit kalmak koşulu ile madde sayısı arttırıldığında G katsayısında artış gözlenmiştir. Örneğin puanlayıcı sayısı sabit kalmak koşuluyla bileşik testteki toplam madde sayısının üç arttırıldığında G katsayısında 0,01385 artmış, madde sayısı altı arttırıldığında G katsayısı 0,024 artmış ve madde sayısı on beş arttırıldığında G katsayısında 0,04285 artış olmuştur.

Çalışmada madde sayısının sabit tutulup ($n_m=15$) puanlayıcı sayısının bir azaltılması durumunda kestirilen G katsayısı 0,90816, puanlayıcı sayısının iki azaltılması ile kestirilen G katsayısı 0,90667 ve puanlayıcı sayısının üç azaltılması ile kestirilen G katsayısı 0,90372 olarak hesaplanmıştır. Puanlayıcı sayısının azaltılması durumunda G katsayısında beklendiği gibi düşüş olmuştur. Örneğin puanlayıcı sayısının bir azaltılması durumunda G katsayısı 0,00089, puanlayıcı sayısının iki azaltılması durumunda G katsayısında 0,00238 ve puanlayıcı sayısının üç azaltılması durumunda G katsayısında 0,00533 azalma olmuştur. Bu örneklerden de görüleceği gibi testin diğer alt boyutlarında olduğu gibi bileşik test açısından da kestirilen G katsayısının arttırmak için madde sayısını arttırmak, puanlayıcı sayısını arttırmaktan daha uygun olacaktır.

Açık uçlu matematik sınavının serbest puanlanması ile elde edilen verilerin Tek Değişkenli G çalışmasıyla analizinden elde edilen sonuçlarının gösterildiği Tablo 10 ve Şekil 7 incelendiğinde; orijinal madde ve puanlayıcı sayısı ile yapılan K çalışması sonucu kestirilen Phi katsayısı 0,89779 olarak hesaplanmıştır. Madde sayısı sabit kalmak şartı ile ($n_m=15$) puanlayıcı sayısının bir azaltılması ($n_p=4$) ile kestirilen Phi

katsayısında 0,00088 azalma olmuştur. Yine madde sayısı sabitken ($n_m=15$) puanlayıcı sayısının üç azaltılması ile hesaplana Phi katsayısı orijinal madde ve puanlayıcı sayısına göre yapılan K çalışmasında hesaplanan Phi katsayısına göre 0,00235 azalmıştır.

Puanlayıcı sayısı sabit kalmak şartıyla ($n_p=5$) toplam madde sayısının üç arttırılması ($n_m=18$) ile yapılan K çalışmasında kestirilen Phi katsayısı 0,01543; toplam madde sayısının iki arttırılması ($n_m=21$) ile yapılan K çalışmasında kestirilen Phi katsayısı 0,02677 olarak hesaplanmıştır. Puanlayıcı sayısı değiştirilmeden her bir alt testte bulunan madde sayısının beş, toplam madde sayısının on beş arttırılması ile yapılan K çalışması sonucu kestirilen Phi katsayısı 0,94571 olarak hesaplanmıştır. Puanlayıcı sayısını değiştirmeden açık uçlu matematik sınavında bulunan toplam madde sayısı on beş arttırıldığında Phi katsayısında orijinal madde ve puanlayıcı sayısı ile yapılan K çalışmasında kestirilen Phi katsayısına göre 0,04792 artış hesaplanmıştır.

Puanlayıcı sayısının bir azaltılıp ($n_p=4$) her bir alt testte bulunan madde sayısı bir arttırılmak üzere toplam madde sayısı üç arttırıldığında hesaplanan Phi katsayısı orijinal madde ve puanlayıcı sayısı ile yapılan K çalışmasına göre hesaplanan Phi katsayısından 0,01463 fazladır.

Puanlayıcı sayısı iki azaltılıp ($n_p=3$) toplam madde sayısının altı arttırılması ($n_m=21$) ile yapılan K çalışması sonucu kestirilen Phi katsayısı orijinal madde ve puanlayıcı sayısına göre hesaplanan Phi katsayısından 0,02484 fazla hesaplanmıştır. Açık uçlu matematik sınavının serbest puanlanmasında bileşik testte bulunan madde sayısının arttırılması puanlayıcı sayısının arttırılmasına göre işgücü ve zaman açısından daha uygun olacaktır.

3.2.3. Çok deęişkenli G alışması sonuçları nasıldır?

Tablo 11. Açık Uçlu Matematik Sınavının Serbest Puanlanmasıyla Elde Edilen Çok Deęişkenli G alışması Sonuçları

Varyans Kaynağı*	Çok Deęişkenli G alışması					
	Alt Test 1		Alt Test 2		Alt Test 3	
	Varyans	%	Varyans	%	Varyans	%
B	211,92287	7,234	227,51930	16,317	267,32975	38,562
M	2621,25043	89,480	1063,51897	76,273	362,24213	52,252
P	35,7557	1,221	53,83636	3,861	13,29154	1,917
bxm	42,92655	1,465	47,32842	3,394	39,55166	5,705
bxp	1,85629	0,063	2,15970	0,155	1,96694	0,284
m xp	13,99955	0,478	10,77184	0,773	7,37277	1,064
bxm xp,e	1,69927	0,580	1,54398	0,111	1,49876	0,216
Toplam	2929,4107	100	1394,36275	100	693,25355	100

* b: Birey, m: Madde, p: Puanlayıcı

Alt Test 1: Bu boyutta çok deęişkenli analizle kestirilen en büyük varyans bileşeni 2621,25043 ile madde (m) ana etkisi için kestirilen varyans bileşenidir ve toplam varyansın %89,480'ini açıklamaktadır. Büyük olan madde varyansına dayalı olarak bazı maddelerin dięer maddelere göre zorluk ve kolaylık bakımından çok farklılaştığını söylenebilir. Bireyler için 211,92287 olarak kestirilen varyans ise toplam varyansın %7,234'ünü açıklamaktadır. Bu nedenle ölçmenin uygulandığı grubun ölçülen yetenekleri bakımından çok farklılık göstermedięi söylenebilir. Puanlayıcı ana etkisi için 35,7557 olarak kestirilen varyans toplam varyansın %1,221'ini açıklamaktadır. Puanlayıcılar için kestirilen bu varyans bileşeni puanlayıcıların bütün bireyler için verdikleri puanlarda farklılıkların az olduęunun bir göstergesi olarak yorumlanabilir. Dięer bir deyişle bu boyut için puanlayıcıların katılık ve cömertlięinin bütün bireyler için birbirine yakın olduęu söylenebilir.

Birey x madde (*bm*) ortak etkisi varyansı toplam varyansın %1,465'ini açıklamaktadır. Bu varyans bileşeni, maddelerin güçlük düzeylerinin bireyden bireye az farklılık gösterdiği şeklinde yorumlanabilir. Birey x puanlayıcı (*bp*) ortak etkisi varyansı toplam varyansın %0,063'ünü, açıklamaktadır. Bu da puanlayıcıların verdikleri puanların bireyden bireye farklılık göstermediğini ifade eder. Madde x puanlayıcı (*mp*) ortak etkisi varyansı ise toplam varyansın %0,478'ini açıklamaktadır. Bu da puanlayıcıların verdikleri puanların maddeden maddeye farklılık göstermediği şeklinde yorumlanabilir. Yapılan analiz sonucunda 1,69927 olarak kestirilen birey x madde x puanlayıcı ortak etkisi varyansının toplam varyans içindeki payı %0,580'dir. Artık varyans ya da hata varyansı olarak da adlandırılan bu varyansın sıfıra yakın olması bu alt testte birey x madde x puanlayıcı ortak etkisinden kaynaklanan tesadüfi hataların çok az olduğunu göstergesidir.

Alt Test 2: Testin ikinci alt boyutu olan Alt Test 2'de çok değişkenli analizle kestirilen en büyük varyans bileşeni 1063,51897 ile yine madde (*m*) ana etkisi için kestirilen varyans bileşenidir ve toplam varyansın %76,273'ünü açıklamaktadır. Bu sonuç bu alt test için de madde güçlüklerinin oldukça farklılaştığını göstermektedir. Puanlayıcı ana etkisi için 53,83636 olarak kestirilen varyans toplam varyansın %3,861'ini açıklamaktadır. Bu alt testte puanlayıcı (*p*) ana etkisinin toplam varyansı açıklama yüzdesi Alt Test 1'e göre daha yüksek bulunmuştur. Bu durumda puanlayıcıların puanlamaları arasında Alt Test 1'e göre tutarsızlık olduğu söylenebilir. Bireyler için 227,51930 olarak kestirilen varyans bileşeni ise toplam varyansın %16,317'sini açıklamaktadır. Ölçmenin amacı olarak tanımlanan birey (*b*) varyansının küçük olması nedeniyle testi alan gruptaki bireylerin ölçüleri bakımından benzer özellikler taşıdığını söyleyebiliriz. Birey x madde (*bm*) ortak etkisi toplam varyansın %3,394'ünü açıklamaktadır. Bu ortak etki varyansının düşük olması nedeniyle madde güçlüklerinin bireyden bireye farklılık göstermediği söylenebilir. Birey x puanlayıcı (*bp*) ortak etkisinin toplam varyans içindeki yüzdesinin %0,155 olması nedeniyle puanlayıcıların puanlamalarının bireyden bireye farklılık göstermediği söylenebilir. Madde x puanlayıcı (*mp*) ortak etkisi ise toplam varyansın %0,773'ünü açıklamaktadır. Bu nedenle puanlayıcıların puanlamaları maddeden maddeye farklılık göstermemiştir. Testin bu alt boyutunda birey x madde x puanlayıcı ortak etkisi varyansı 1,54398 olarak

kestirilmiştir. Artık ya da hata varyansı olarak adlandırılan bu varyans toplam varyansın %0,111'ini oluşturmaktadır. Bu sonuç; Alt Test 1'de olduğu gibi Alt Test 2'de de bireyler, maddeler ve puanlayıcılar arası etkileşim ile bu çalışmada ölçülemeyen sistematik ya da tesadüfi değişkenlik kaynaklarının oldukça düşük olduğunun bir göstergesidir.

Alt Test 3: Testin son alt boyutu olan Alt Test 3'te en büyük varyans bileşeni 362,24213 olarak kestirilen madde (*m*) ana etkisi için kestirilen varyans bileşenidir. Toplam varyansı açıklama yüzdesi ise %52,252'dir. İkinci en büyük varyans bileşeni ise 267,32975 olarak kestirilen birey (*b*) ana etkisi için varyans bileşenidir ve toplam varyansın %38,562'sini açıklamaktadır. Analizde 13,29154 olarak kestirilen puanlayıcı (*p*) ana etkisi ise toplam varyansın %1,917'sini açıklamaktadır.

Birey x madde (*bm*) ortak etkisi için 39,55166 olarak kestirilen varyans toplam varyansın %5,705'ini; birey x puanlayıcı (*bp*) ortak etkisi için 1,96694 olarak hesaplanan varyans toplam varyansın %0,284'ünü; madde x puanlayıcı (*mp*) ortak etkisi için 7,37277 olarak kestirilen varyans toplam varyansın %1,064'ünü açıklamaktadır. Artık varyans olarak tanımlanan birey x madde x puanlayıcı ortak etkisi için varyans büyüklüğü 1,49876 olarak kestirilmiştir. Bu varyans ise toplam varyansın sadece %0,216'sını açıklamaktadır.

Madde(*m*) ana etkisi için kestirilen varyans bileşeninin en büyük değeri alması bazı maddelerin bazı bireylere diğer bireylere göre daha zor ya da kolay geldiğinin bir göstergesidir. Birey(*b*) ana etkisi için kestirilen varyans bileşeninin toplam varyansın %38'ini açıklaması bireyler arasındaki farklılıkların ortaya konulabildiğinin göstergesidir. Puanlayıcı ana etkisinin toplam varyansı açıklama yüzdesinin bu boyutta da düşük olması puanlayıcıların Alt Test 3'ü puanlarken birbirlerine göre tutarlı puanlama yaptıkları şeklinde yorumlanabilir. Birey x madde (*bm*) ortak etkisinin toplam varyansa etkisinin düşük olması bireylerin bir maddeden diğerine bağlı durumlarının çok farklılık göstermediği şeklinde açıklanabilir. Birey x puanlayıcı (*bp*) ortak etkisinin toplam varyansı açıklama yüzdesi madde x puanlayıcı (*mp*) ortak etkisinin toplam varyansı açıklama yüzdesine göre daha düşük olması puanlayıcıların yaptıkları puanlamaların bir maddeden diğerine nispeten daha tutarsız olduğunun göstergesidir. Birey x madde x puanlayıcı ortak etkisinin toplam varyansı açıklama yüzdesinin çok

düşük olması bu alt testte tesadüfî hataların çok az olduğunu göstergesidir.

3.2.4. Çok değişkenli K çalışması sonuçları nasıldır?

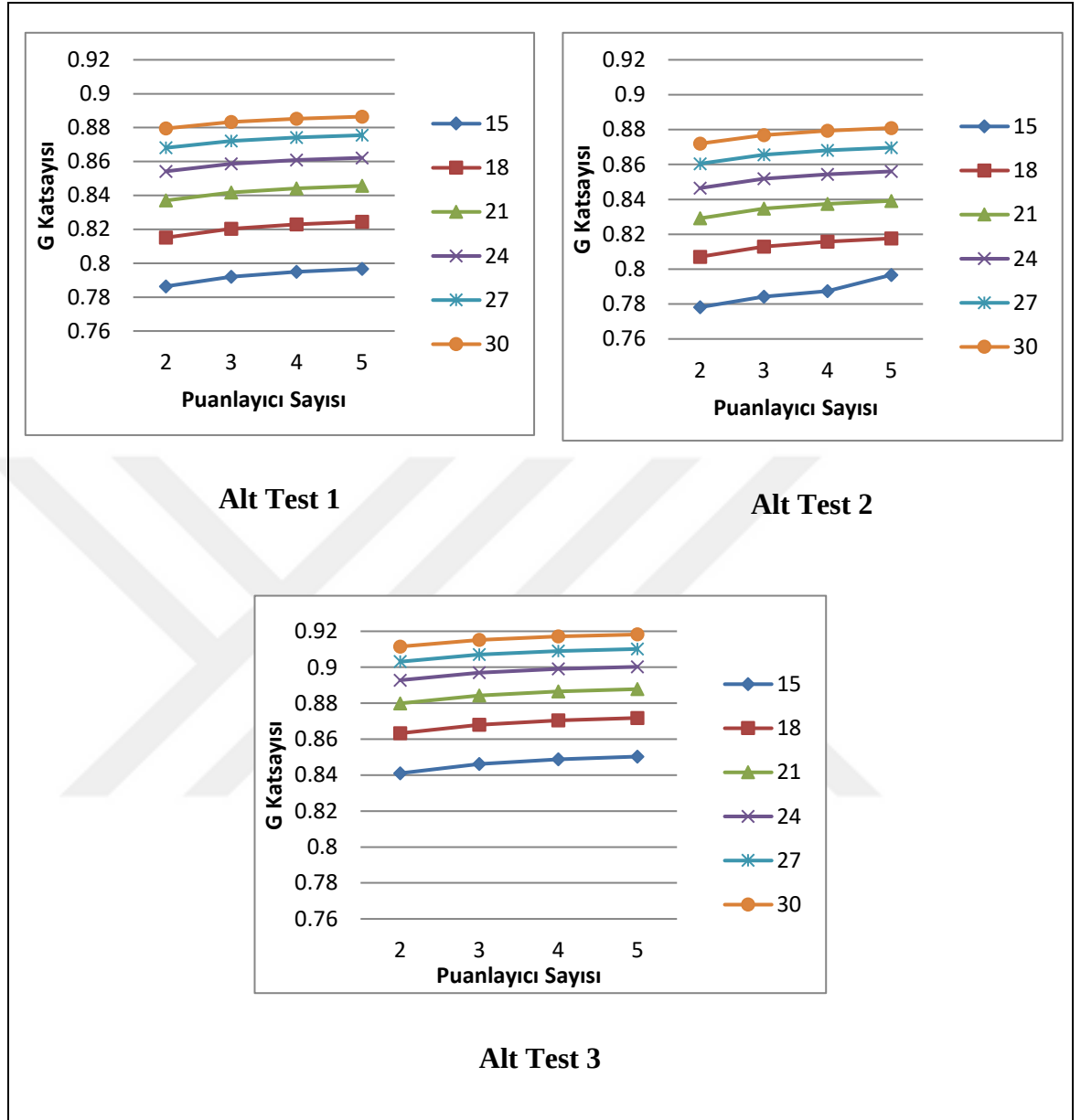
Açık uçlu matematik sınavının anahtarla puanlanmasında her bir alt testte bulunan madde sayısının sırasıyla 5, 6, 7, 8, 9, 10 yapılması ve puanlayıcı sayısının 5, 4, 3, 2 olarak alınması durumları için $b \cdot x m^{\circ} x p^{\circ}$ Çok Değişkenli K çalışması yapılmıştır. Bu K Çalışmasından elde edilen sonuçlar aşağıda Tablo 12’de verilmiştir.

Tablo 12 ve Şekil 8 incelendiğinde Alt Test 1 boyutunda orijinal puanlayıcı sayısı 5 madde sayıları 5 için çok değişkenli $b \cdot x m^{\circ} x p^{\circ}$ modeli ile yapılan K çalışmasında kestirilen G katsayısı 0,79670 olarak hesaplanmıştır. Madde sayısının sabit tutularak ($n_m=5$) şartıyla puanlayıcı sayısının bir azaltılması ($n_p=4$) durumunda kestirilen G katsayısı 0,79496; puanlayıcı sayısının iki azaltılması durumunda ($n_p=3$) kestirilen G katsayısı 0,79208; puanlayıcı sayısının üç azaltılması durumunda ($n_p=2$) ise kestirilen G katsayısı ise 0,78637 olarak hesaplanmıştır. Bu sonuçlara göre puanlayıcı sayısının 4, 3, 2 alınması durumunda G katsayısındaki azalma sırayla 0,001, 0,004 ve 0,010 olarak hesaplanmıştır. Tablo 11’de görüldüğü gibi AltTest1’de puanlayıcı ana etkisinin toplam varyansa katkısının %1,221 olması nedeniyle puanlayıcı sayısının azaltılması G katsayısını önemli ölçüde etkilememiştir. Buna karşılık puanlayıcı sayısının sabit tutulup madde sayısının artırılması durumunda G katsayısının daha fazla arttığı görülmüştür.

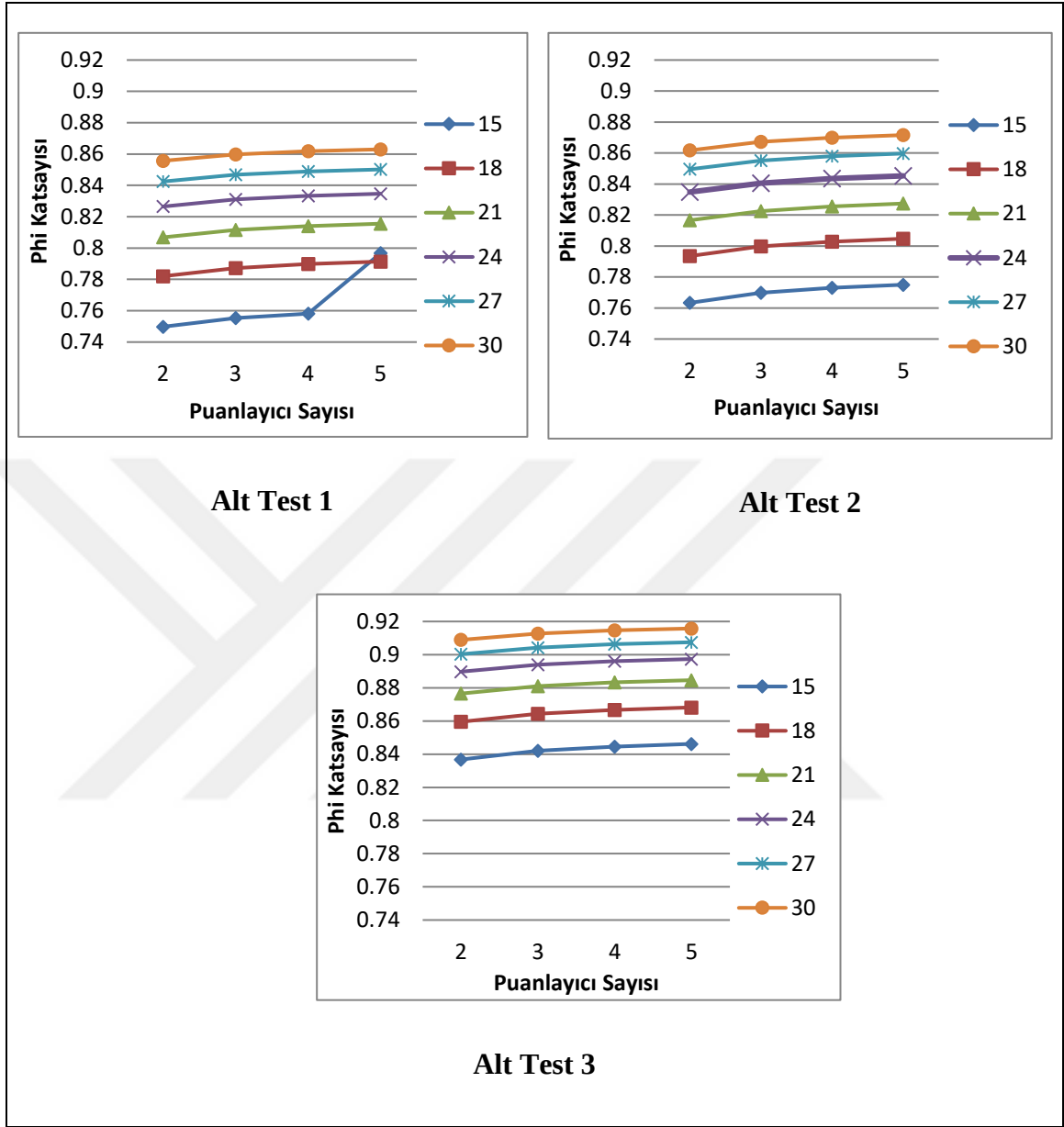
Tablo 12. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasında Alternatif Madde ve Puanlayıcı Sayıları Senaryoları ile Yapılan Çok Değişkenli K Çalışması Sonuçları

n_p	n_m	G Katsayısı			Phi Katsayısı		
		Alt Test 1	Alt Test 2	Alt Test 3	Alt Test 1	Alt Test 2	Alt Test 3
2	5	0,78637	0,77819	0,84102	0,74978	0,76334	0,83679
2	6	0,81514	0,80713	0,86327	0,78204	0,79354	0,85953
2	7	0,83702	0,82916	0,87990	0,80684	0,81661	0,87654
2	8	0,85422	0,84648	0,89279	0,82650	0,83482	0,88974
2	9	0,86809	0,86047	0,90309	0,84246	0,84955	0,90029
2	10	0,87951	0,87199	0,91150	0,85568	0,86172	0,90892
3	5	0,79208	0,78431	0,84615	0,75534	0,76979	0,84199
3	6	0,82034	0,81293	0,86798	0,78721	0,79972	0,86432
3	7	0,84179	0,83468	0,88428	0,81166	0,82257	0,88100
3	8	0,85863	0,85177	0,89692	0,83102	0,84059	0,89395
3	9	0,87220	0,86556	0,90700	0,84674	0,85516	0,90428
3	10	0,88337	0,87692	0,91522	0,85974	0,86718	0,91272
4	5	0,79496	0,78741	0,84874	0,75816	0,77305	0,84461
4	6	0,82296	0,81585	0,87036	0,78981	0,80285	0,86673
4	7	0,84419	0,83747	0,88650	0,81409	0,82559	0,88325
4	8	0,86085	0,85444	0,89899	0,83331	0,84350	0,89606
4	9	0,87427	0,86813	0,90896	0,84889	0,85798	0,90628
4	10	0,88531	0,87940	0,91710	0,86178	0,86993	0,91463
5	5	0,79670	0,79670	0,85030	0,79670	0,77502	0,84619
5	6	0,82454	0,81762	0,87179	0,79139	0,80474	0,86818
5	7	0,84564	0,83915	0,88783	0,81556	0,82741	0,88461
5	8	0,86219	0,85605	0,90024	0,83468	0,84526	0,89734
5	9	0,87552	0,86968	0,91015	0,85018	0,85969	0,90749
5	10	0,88648	0,88090	0,91822	0,86301	0,87159	0,91578

Not: Koyu yazılmış olanlar sınavda kullanılan orijinal madde ve puanlayıcı sayısıdır.



Şekil 8. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan Çok Değişkenli K Çalışması Sonucunda Alt Testlere Göre G Katsayısındaki Değişim Grafiği.



Şekil 9. Açık Uçlu Matematik Sınavının Anahtarla Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan Çok Değişkenli K Çalışması Sonucunda Alt Testlere Göre Phi Katsayısındaki Değişim Grafiği.

Puanlayıcı sayısının sabit tutulup ($n_p=5$) madde sayısının bir arttırılması durumunda ($n_m=6$) kestirilen G katsayısı 0,82454 olarak hesaplanmıştır. Madde sayısının iki arttırılması durumunda ($n_m=7$) kestirilen G katsayısı 0,84564; madde sayısının üç arttırılması durumunda ($n_m=8$) G katsayısı 0,86219; madde sayısının dört arttırılması durumunda kestirilen G katsayısı 0,87552 ($n_m=9$) ve madde sayısının beş

arttırılması ($n_m=10$) durumunda kestirilen G katsayısı 0,88648 olarak hesaplanmıştır. Bu sonuçlar dikkate alındığında puanlayıcı sayısı sabitken madde sayısının bir arttırılması G katsayısını 0,02784 arttırmıştır. Yine puanlayıcı sayısı sabitken madde sayısının iki arttırılması G katsayısını 0,04894 arttırmıştır. Puanlayıcı sayısı sabitken madde sayısının beş arttırılması ise G katsayısı üzerinde 0,08978'e kadar artış sağlamıştır. Puanlayıcı sayısının arttırılması G katsayısına madde sayısının arttırılması kadar katkı sağlamamıştır. Örneğin orijinal puanlayıcı sayısının bir azaltılıp ($n_p=4$) madde sayısının bir arttırılması ile ($n_m=6$) yapılan K çalışması sonucu kestirilen G katsayısı 0,82296 olarak hesaplanmıştır. Puanlayıcı sayısının iki azaltılıp ($n_p=3$) madde sayısının iki arttırılması ($n_m=7$) durumunda kestirilen G katsayısı 0,84179 olarak hesaplanmıştır. Puanlayıcı sayısının üç azaltılıp ($n_p=2$) madde sayısının beş arttırılması durumunda ($n_m=10$) kestirilen G katsayısı 0,88337 olarak hesaplanmıştır. Çalışmada kestirilen en yüksek G katsayısı olan 0,88648; puanlayıcı sayısının beş, madde sayısının on olduğu senaryoda kestirilen G katsayısıdır. Görüldüğü gibi puanlayıcı sayısının arttırılmasındansa madde sayısının arttırılması G katsayısını arttırmakta hem de zaman, işgücü ve ekonomi açısından kolaylık sağlamaktadır.

Alt test 2 boyutunda orijinal puanlayıcı sayısının beş ($n_p=5$) ve orijinal madde sayısının beş ($n_m=5$) olduğu durum için çok değişkenli $b \cdot x m^{\circ} x p^{\circ}$ modelle yapılan K çalışması ile kestirilen G katsayısı 0,79670 olarak hesaplanmıştır. Bu boyut için madde sayısının sabit tutularak ($n_m=5$) puanlayıcı sayısının bir azaltılması sonucu ($n_p=4$) elde edilen G katsayısı 0,78741 olarak hesaplanmıştır. Yine madde sayısı sabitken ($n_m=5$) puanlayıcı sayısının iki azaltılması ($n_p=3$) sonucu kestirilen G katsayısı 0,78431; puanlayıcı sayısının üç azaltılması ($n_p=2$) sonucu kestirilen G katsayısı ise 0,77819 olarak hesaplanmıştır. Madde sayısının sabit tutulması durumunda puanlayıcı sayısının bir azaltılması sonucu kestirilen G katsayısındaki azalma 0,00929 iken puanlayıcı sayısının iki azaltılması ile G katsayısındaki azalma 0,01239 olarak hesaplanmıştır. Puanlayıcı sayısı ikiye düşürüldüğünde ise G katsayısındaki azalma 0,01851 olmuştur. Madde sayısı sayısının sabit tutulup puanlayıcı sayısının azaltılması durumunda alt test 2 boyutunda alt test 1 boyutuna göre G katsayısında daha fazla azalma olmuştur. Bunun nedeni Tablo 11'de görüldüğü gibi Alt Test 1'de puanlayıcı ana etkisinin varyans açıklama yüzdesinin %1,221 iken Alt Test 2'de puanlayıcı ana etkisinin varyans

açıklama yüzdesinin %3,861 olmasıdır.

Puanlayıcı sayısının sabit tutulup ($n_p=5$) madde sayısının bir arttırıldığı ($n_m=6$) durumda hesaplanan G katsayısı 0,81762 iken madde sayısının iki arttırıldığı ($n_m=7$) durumda hesaplanan G katsayısı 0,83915 olmuştur. Puanlayıcı sayısı sabit tutulup ($n_p=5$) madde sayısının beş arttırıldığı durumda hesaplanan G katsayısı 0,88090 olmuştur. Puanlayıcı sayısı sabit tutulup madde sayısının bir arttırılması ile G katsayısındaki artış 0,02092 olmuş, madde sayısının iki arttırılması ile kestirilen G katsayısındaki artış 0,04245 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken madde sayısının birer birer arttırılmıştır. Çalışmanın Alt Test 2 boyutunda kestirilen en yüksek G katsayısı puanlayıcı sayısının beş ($n_p=5$), madde sayısının on ($n_m=10$) alınması durumunda 0,88090 olarak hesaplanmıştır. Tablo 11 incelendiğinde madde ana etkisinin varyans açıklama yüzdesinin %76,273 ve puanlayıcı ana etkisinin varyans açıklama yüzdesine %3,861 göre daha yüksek olmasından dolayı madde sayısının arttırılmasının G katsayısı üzerinde puanlayıcı sayısının arttırılmasına göre daha fazla etkisi vardır. Örneğin puanlayıcı sayısının üç ($n_p=3$) madde sayısının on ($n_m=10$) olarak alınması durumunda kestirilen G katsayısı 0,87692 iken puanlayıcı sayısının dört ($n_p=4$) madde sayısının on ($n_m=10$) olarak hesaplanması durumunda kestirilen G katsayısı 0,87940 olarak hesaplanmıştır. Madde sayısı on ($n_m=10$) puanlayıcı sayısı üç ($n_p=3$) iken puanlayıcı sayısının bir arttırılması durumunda kestirilen G katsayısında 0,00248 artış gözlenmiştir. Puanlayıcı sayısının arttırılması fazla işgücü gerektirdiği ve ekonomik olmadığı ve G katsayısına fazlaca katkı sağlamadığı için testin bu alt boyutunda G katsayısını arttırmak için madde sayısının arttırılması önerilebilir.

Alt test 3 boyutunda orijinal puanlayıcı sayısı beş ($n_p=5$) ve orijinal madde sayıları beş ($n_m=5$) için çok değişkenli modelle K çalışması için kestirilen G katsayısı 0,85030 olarak hesaplanmıştır. Tablo 12 incelendiğinde bu boyut için madde sayısı sabit tutularak ($n_m=5$) puanlayıcı sayısının bir azaltılması ($n_p=4$) durumunda kestirilen G katsayısı 0,84874 olarak hesaplanmıştır. Madde sayısının sabit tutularak ($n_m=5$) puanlayıcı sayısının iki azaltılması ($n_p=3$) durumunda kestirilen G katsayısı 0,84615 olarak hesaplanmıştır. Madde sayısının sabit tutularak ($n_m=5$) puanlayıcı sayısının üç azaltılması durumunda ($n_p=2$) kestirilen G katsayısı 0,84102 olarak hesaplanmıştır. Görev sayısının sabit tutulması koşuluyla puanlayıcı sayısının bir azaltılması

durumunda G katsayısındaki azalma 0,00156 olarak hesaplanmıştır. Puanlayıcı sayısının sabit tutulup ($n_p=5$) madde sayısının bir arttırılması ($n_m=6$) sonucu kestirilen G katsayısı 0,87179; madde sayısının iki arttırılması ($n_m=7$) sonucu kestirilen G katsayısı 0,88783 ve madde sayısının üç arttırılması ($n_m=8$) ile kestirilen G katsayısı 0,90024 olarak hesaplanmıştır. Bu alt boyut için yapılan K çalışmasında yine puanlayıcı sayısının sabit tutulması koşuluyla madde sayısının dört arttırılması ($n_m=9$) durumunda kestirilen G katsayısı 0,91015 iken madde sayısının beş arttırılması durumunda ($n_m=10$) kestirilen G katsayısı 0,91822 olmuştur. Puanlayıcı sayısının sabit tutulup madde sayısının arttırılması bu boyutta da G katsayısını önemli ölçüde arttırmıştır. Örneğin orijinal madde sayısı 5 ($n_m=5$) ve puanlayıcı sayısına 5 ($n_p=5$) göre madde sayısının bir arttırılması durumunda G katsayısında 0,2149 artış olmuştur. Madde sayısı iki arttırıldığında ise G katsayısındaki artış 0,03753 olmuştur. Madde sayısı beş arttırıldığında G katsayısı 0,06792 artmıştır. Bu boyut için de işgücü, zaman ve ekonomi düşünüldüğünde madde sayısının arttırılması uygun olacaktır.

Tablo 12 ve Şekil 9 incelendiğinde Alt Test 1 boyutunda orijinal puanlayıcı sayısı beş ($n_p=5$) ve orijinal madde sayısı beş ($n_m=5$) için çok değişkenli $b \cdot x_m \cdot x_p$ modelle K çalışması için kestirilen Phi katsayısı 0,79670 olarak hesaplanmıştır. Şekil 9'da görüldüğü bu boyut için madde sayısı sabitken ($n_m=5$) puanlayıcı sayısının bir azaltılması ($n_p=4$) ile kestirilen Phi katsayısı 0,75816 olarak hesaplanmıştır. Madde sayısı sabit kalmak koşulu ile puanlayıcı sayısının iki azaltılması ($n_p=3$) durumunda hesaplanan Phi katsayısı 0,75534 olmuştur. Puanlayıcı sayısının üç azaltılması ($n_p=2$) durumunda yapılan K çalışmasında Phi katsayısı ise 0,74978 olmuştur. Madde sayısı sabitken puanlayıcı sayısının bir azaltılması durumunda Phi katsayısında 0,03854 azalma olmuştur. Puanlayıcı sayısının iki azaltılması ile yapılan K çalışmasında Phi katsayısında 0,04136; puanlayıcı sayısının üç azaltılması ile yapılan K çalışmasında ise hesaplanan Phi katsayısında 0,04692 azalma olmuştur.

Puanlayıcı sayısı sabitken ($n_p=5$) madde sayısının bir arttırılması ($n_m=6$) ile yapılan K çalışmasında Phi katsayısı 0,79139; madde sayısının iki arttırılması ($n_m=7$) ile kestirilen Phi katsayısı 0,81556; madde sayısının üç arttırılması ($n_m=8$) ile hesaplanan Phi katsayısı ise 0,84526 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken madde

sayısının dört arttırılması ($n_m=9$) ile kestirilen Phi katsayısı 0,85018; madde sayısının beş arttırılması ile kestirilen Phi katsayısı 0,86301 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken ($n_p=5$) madde sayısının iki arttırılması ($n_m=7$) ile yapılan K çalışmasında hesaplanan Phi katsayısı, orijinal madde ve puanlayıcı sayısına göre hesaplanan Phi katsayısından 0,01886 fazladır. Buna karşılık madde sayısı sabitken ($n_m=5$) puanlayıcı sayısının iki azaltılması ($n_p=3$) ile yapılan K çalışmasında hesaplanan Phi katsayısında orijinal madde ve puanlayıcı sayısına göre hesaplanan Phi katsayısına göre 0,04136 azalma olmuştur. Bu nedenle yapılan çalışmada G katsayısında olduğu gibi Phi katsayısını arttırmak için de madde sayısının arttırılması puanlayıcı sayısının arttırılmasından daha çok katkı sağlayacaktır.

Alt test 2 boyutunda orijinal puanlayıcı beş ($n_p=5$) ve orijinal madde sayıları beş ($n_m=5$) için çok değişkenli modelle K çalışması için kestirilen Phi katsayısı 0,77502 olarak hesaplanmıştır. Tablo 12 incelendiğinde madde sayısı sabitken ($n_m=5$) puanlayıcı sayısının bir azaltılması ($n_p=4$) ile kestirilen Phi katsayısı 0,77305; puanlayıcı sayısının iki azaltılması ile ($n_p=3$) kestirilen Phi katsayısı 0,76979 ve puanlayıcı sayısının üç azaltılması ($n_p=2$) ile kestirilen Phi katsayısı ise 0,76334 olarak hesaplanmıştır. Alt test 2’de bulunan madde sayısı değiştirilmeden puanlayıcı sayısı bir azaltıldığında ($n_p=4$) kestirilen Phi katsayısı 0,00197 azalmıştır. Madde sayısı değiştirilmeden puanlayıcı sayısı ikiye düşürüldüğünde ise kestirilen Phi katsayısında 0,00523 azalma olmuştur.

Puanlayıcı sayısı sabit kalmak koşulu ile ($n_p=5$) madde sayısının bir arttırılması ile ($n_m=6$) kestirilen Phi katsayısı ise 0,80474 olarak hesaplanmıştır. Madde sayısının sırasıyla iki, üç dört ve beş arttırılması ile hesaplanan Phi katsayıları sırasıyla 0,82741;0,84526; 0,85969 ve 0,87159 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken ($n_p=5$) madde sayısının bir arttırılması ile ($n_p=6$) kestirilen Phi katsayısı 0,02972 artmıştır. Madde sayısının beş arttırılması ile hesaplanan Phi katsayısında 0,096570 artma olmuştur.

Açık uçlu matematik sınavının anahtarla puanlanması ile oluşan verilerin çok değişkenli G kuramı ile analizinde Alt Test 2 boyutunda madde sayısının arttırılması veya puanlayıcı sayısının azaltılması Phi katsayısına aynı oranda katkı sağlamıştır. Fakat her bir alt testteki madde sayısının arttırılması puanlayıcı sayısının arttırılması işlemine göre daha az zaman gerektiren ve daha ekonomik bir yöntem olduğu için

anahtarla puanlamada Alt test 2 boyutunda madde sayısının arttırılması Phi katsayısını arttırmak için tercih edilebilir.

Alt test 3 boyutunda orijinal puanlayıcı beş ($n_p=5$) ve orijinal madde sayıları beş ($n_m=5$) için çok değişkenli modelle K çalışması için kestirilen Phi katsayısı 0,84619 olarak hesaplanmıştır. Tablo 12 incelendiğinde bu boyut için madde sayısı sabit tutularak ($n_m=5$) puanlayıcı ($n_p= 4$) kestirilen Phi katsayısı 0,84461; puanlayıcı ($n_p= 3$) kestirilen Phi katsayısı 0,84199 olarak hesaplanmıştır. Puanlayıcı ($n_p= 2$) Phi katsayısı 0,83679 olarak hesaplanmıştır. Puanlayıcı sayısı sabit kalmak şartıyla ($n_p=5$) madde sayısının bir arttırılması ile kestirilen Phi katsayısı 0,86818; madde sayısının beş arttırılması ile kestirilen Phi katsayısı ise 0,91578 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken Alt Test 3'te bulunan madde sayısının bir arttırılması ile Phi katsayısında 0,02199; Alt Test 3'te bulunan madde sayısının beş arttırılması ile Phi katsayısında 0,06959 artış hesaplanmıştır. Madde sayısının iki arttırılıp puanlayıcı sayısının bir azaltılması ile kestirilen Phi katsayısında orijinal madde ve puanlayıcı sayısına göre 0,03706 artış hesaplanmıştır. Madde sayısının beş arttırılıp ($n_m=10$) puanlayıcı sayısının iki azaltılması ($n_p=3$) ile yapılan K çalışmasında Phi katsayısında orijinal madde ve puanlayıcı sayısına göre 0,06653 artış hesaplanmıştır. Bu nedenle alt test 3 boyutunda Phi katsayısını arttırmak için madde sayısının arttırılması zaman ve işgücü açısından daha kolay olduğu için tercih edilebilir.

3.3.Alt Problem 3.Tüm bireyleri (b), tüm maddelerde(m) aynı puanlayıcıların (p) anahtarla ve serbest puanladığı ölçme durumunda;

3.3.1. Tek değişkenli G çalışması sonuçları nasıldır?

Açık uçlu matematik sınavının puanlayıcılar tarafından anahtarla ve serbest puanlanmasıyla elde edilen veriler üzerinde tek değişkenli G Kuramının $b \times m \times p$ deseni için G çalışması yapılmıştır. Bu çalışma ile her iki puanlama türü için de varyans kaynakları için elde edilen varyansı büyükleri ile toplam varyansı açıklama oranları Tablo13'te verilmiştir.

Tablo 13. Açık Uçlu Matematik Sınavının Anahtarla ve Serbest Puanlamasıyla Elde Edilen Tek Değişkenli G Çalışması Sonuçları

Varyans Kaynağı	Anahtarla Puanlama		Serbest Puanlama	
	Varyans	Toplam Varyans İçindeki Payı (%)	Varyans	Toplam Varyans İçindeki Payı (%)
Birey (b)	542,42269	22,6204	561,008	18,720
Madde (m)	1778,85613	74,1828	1574,682	52,543
Puanlayıcı (p)	8,11080	0,3382	783,837	26,155
Birey-Madde ($b \times m$)	48,82684	2,0362	42,595	1,421
Birey-Puanlayıcı ($b \times p$)	2,13461	0,089	10,313	0,344
Madde-Puanlayıcı ($m \times p$)	15,95352	0,6653	21,356	0,713
Birey-Madde-Puanlayıcı ($b \times m \times p$)	1,62974	0,0679	3,120	0,104
Toplam:	2397,9343	100	2996,911	100

Tablo 13 incelendiğinde açık uçlu matematik sınavının serbest ve anahtarla puanlanması ile elde edilen verilerle yapılan tek değişkenli G çalışması sonucunda kestirilen en büyük varyans bileşeni her iki puanlama için de madde (m) ana etkisi varyansıdır. Bu durum hem serbest hem de anahtarla puanlama için testin maddelerin birbirinden oldukça farklılaştığının göstergesidir. Serbest puanlamada madde (m) ana

etkisi 1574,682 ile toplam varyansın %52,543 'ünü oluştururken; anahtarla puanlamada madde (*m*) ana etkisi 1778,85613 ile toplam varyansın %74,1828'ini oluşturmaktadır. Bu durumda anahtarla puanlamada testin maddelerinin serbest puanlamaya göre daha fazla farklılaştığı söylenebilir.

Birey (*b*) ana etkisi için kestirilen varyans bileşeni serbest puanlamada 561,008 ile toplam varyansın %18,72'sini açıklamaktadır. Anahtarla puanlamada ise birey (*b*) ana etkisi için kestirilen varyans bileşeni 542,42269 ile toplam varyansın %22,6204'ünü açıklamaktadır. Bu durumda anahtarla puanlamada birey (*b*) ana etkisi için kestirilen varyans yüzdesinin daha büyük olması nedeni ile testi alan bireylerin bağıl durumlarının maddeden maddeye serbest puanlamaya göre daha fazla değiştiği söylenebilir. Puanlayıcı (*p*) ana etkisi için kestirilen varyans bileşeni serbest puanlamada 783,837 ile toplam varyansın %26,155'ini açıklarken anahtarla puanlamada 8,11080 ile toplam varyansın %0,3382'sini açıklamaktadır. Anahtarla puanlamada beklendiği gibi puanlayıcı ana etkisi için kestirilen varyans yüzdesi serbest puanlamaya göre oldukça düşük çıkmıştır. Serbest puanlamada anahtarla puanlamaya göre puanlayıcıların puanlamalarının oldukça farklılık gösterdiği söylenebilir.

Birey x madde ortak etkisi (*bm*) serbest puanlamada 42,595 ile toplam varyansın %1,421'ini açıklarken anahtarla puanlamada 48,82684 ile toplam varyansın %2,0362'sini açıklamaktadır. Serbest ve anahtarla puanlamada birey x madde (*bm*) ortak etkisi için kestirilen varyans bileşeninin toplam varyansın içindeki yüzdesinin küçük olması nedeniyle her iki puanlama türü için de bireylerin bağıl durumlarının maddeden maddeye çok farklılık göstermediği söylenebilir. Birey x puanlayıcı(*bp*) ortak etkisi serbest puanlamada 10,313 ile toplam varyansın %0,344'ünü açıklarken; anahtarla puanlamada 2,13461 ile toplam varyansın %0,089'unu açıklamaktadır. Anahtarla puanlamada birey x puanlayıcı(*bp*) ortak etkisinin toplam varyans içindeki payının daha düşük olduğu görülmektedir. Bu nedenle puanlayıcıların puanlamalarının bireyden bireye serbest puanlamaya göre anahtarla puanlamada daha az farklılık gösterdiği söylenebilir. Madde x puanlayıcı(*mp*) ortak etkisi serbest puanlamada 21,356 ile toplam varyansın %0,713'ünü açıklarken; anahtarla puanlamada 15,95352 ile toplam varyansın %0,6653'ünü açıklamaktadır. Her iki puanlama türü için de puanlayıcıların puanlamalarının maddeden maddeye farklılık göstermediği görülmektedir. Son olarak

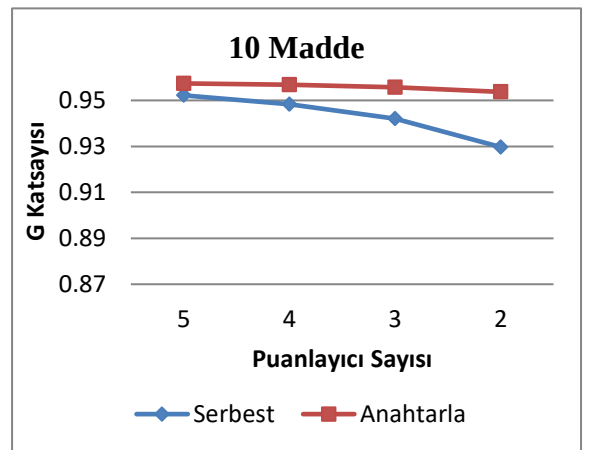
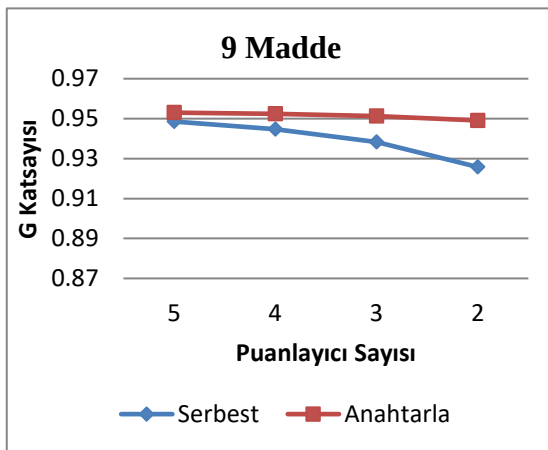
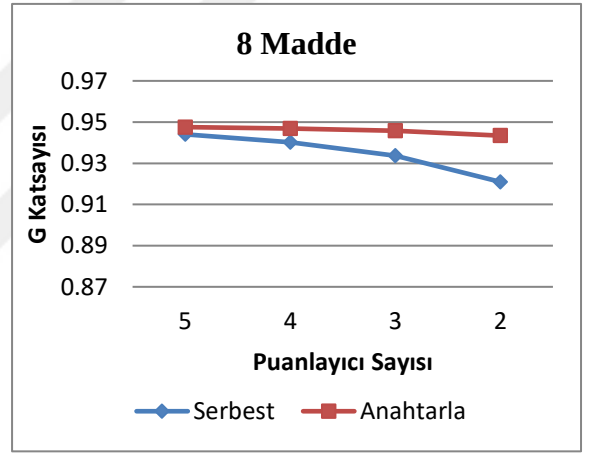
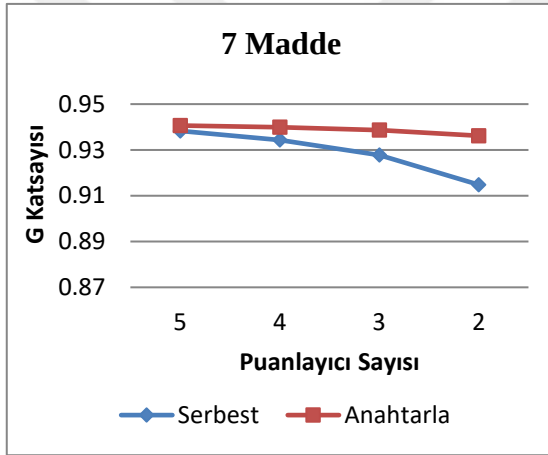
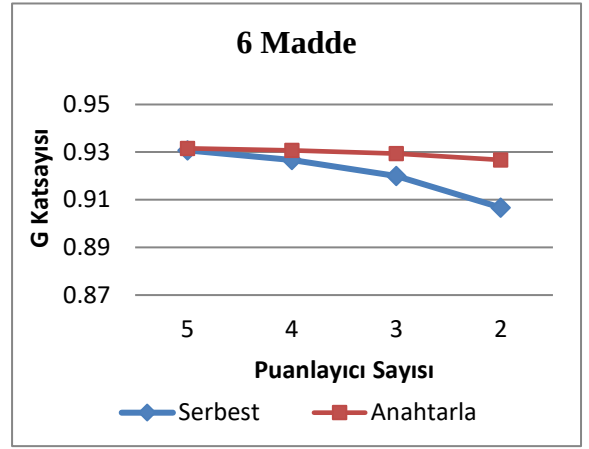
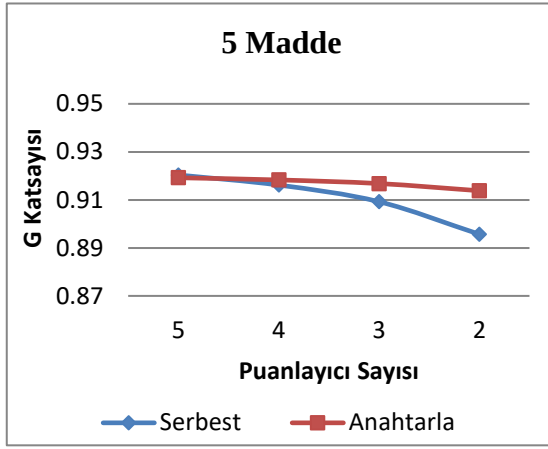
birey x madde x puanlayıcı ortak etkisi serbest puanlamada 3,120 ile toplam varyansın %0,104'ünü açıklarken anahtarla puanlamada 1,62974 ile toplam varyansın %0,0679'unu açıklamaktadır. Bu varyansın serbest ve anahtarla puanlamada yüzdesinin sıfıra yakın olması birey x madde x puanlayıcı ortak etkisinden kaynaklanan hataların her iki puanlama türü için de oldukça az olduğunu göstermektedir. Ayrıca anahtarla puanlamada bu ortak etki varyansının serbest puanlamaya göre daha da düşük olduğu görülmektedir.

3.3.2. Tek değişkenli K çalışması sonuçları nasıldır?

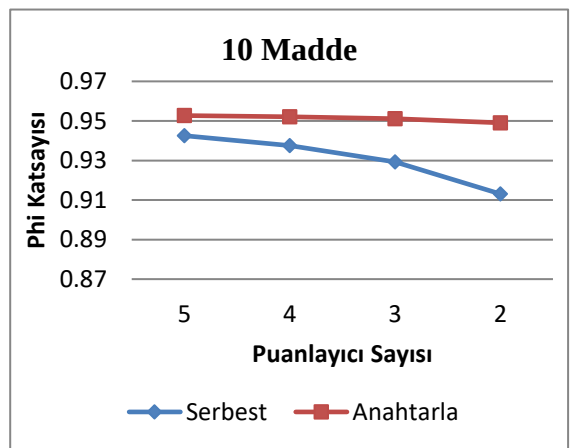
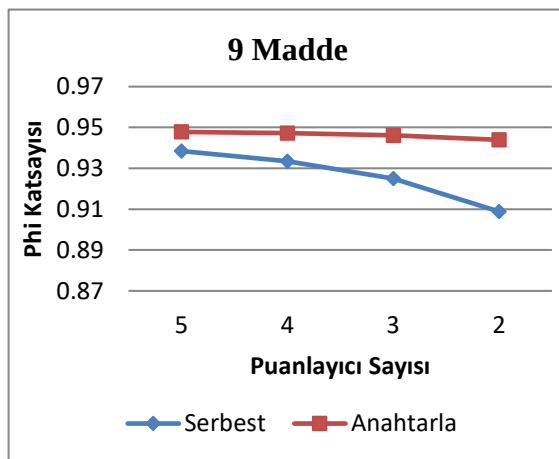
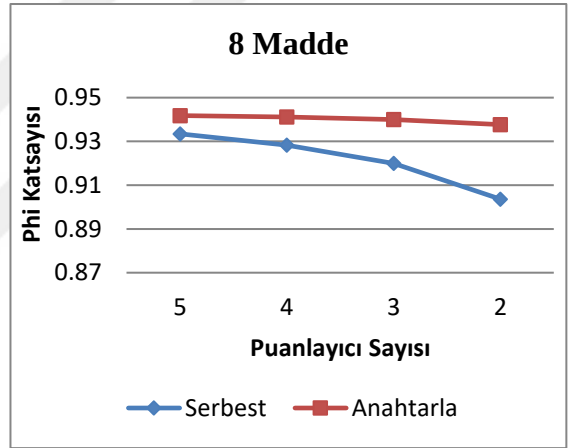
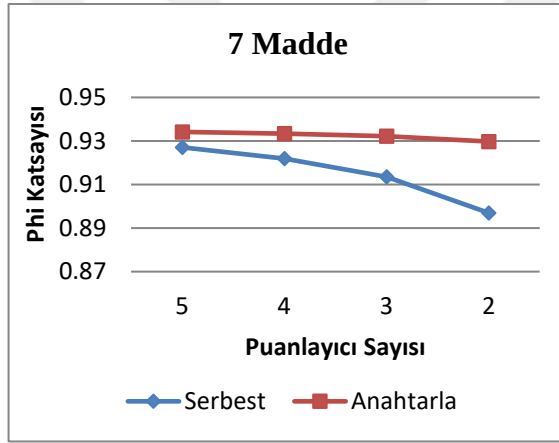
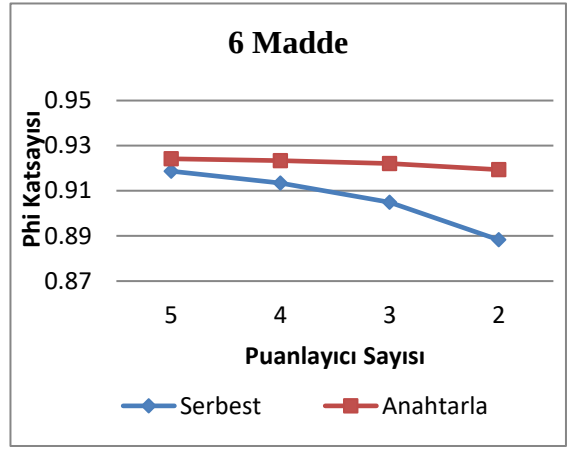
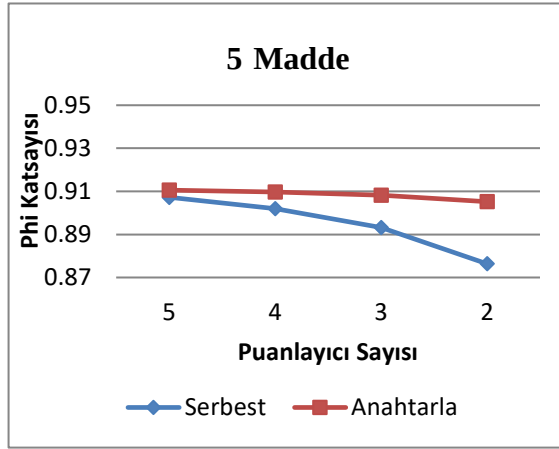
Açık uçlu matematik sınavının anahtarla ve serbest puanlanmasıyla elde edilen puanlarda tüm değişkenlerin çaprazlandığı $b \times m \times p$ deseninde birey (b) ölçmenin nesnesi (object of measurement), m , p ise yüzey (facet) olarak ele alınır. Elde edilen varyans bileşenlerinden hesaplanan mutlak ve bağıl hata varyanslarına ve görev, puanlayıcı sayılarının artırılıp azaltılmasına dayalı senaryolar için kestirilen genellenebilirlik katsayısı (G), phi (ϕ) katsayısına ait değerler Tablo 14'te verilmiştir.

Tablo 14. Açık Uçlu Matematik Sınavının Anahtarla ve Serbest Puanlanmasında Bileşik Testte Bulunan Farklı Madde ve Puanlayıcı Sayısı Senaryoları ile Yapılan K Çalışmasının Sonuçları

n_p	n_m	<i>Serbest Puanlama</i>		<i>Anahtarla Puanlama</i>	
		G Katsayısı	Phi Katsayısı	G Katsayısı	Phi Katsayısı
2	15	0,88694	0,86646	0,90372	0,89253
2	18	0,89926	0,87976	0,91810	0,90848
2	21	0,90827	0,88952	0,92865	0,92022
2	24	0,91515	0,89699	0,93673	0,92923
2	27	0,92058	0,90288	0,94311	0,93637
2	30	0,93498	0,91856	0,96011	0,95543
3	15	0,90036	0,88313	0,90667	0,89544
3	18	0,91234	0,89625	0,92076	0,91110
3	21	0,92109	0,90585	0,93109	0,92263
3	24	0,92776	0,91320	0,93899	0,93137
3	27	0,93302	0,91899	0,94523	0,93846
3	30	0,93727	0,92368	0,95028	0,94413
4	15	0,90723	0,89171	0,90816	0,89691
4	18	0,91902	0,90472	0,92210	0,91242
4	21	0,92763	0,91425	0,93231	0,92384
4	24	0,93420	0,92153	0,94013	0,93259
4	27	0,93937	0,92727	0,94630	0,93951
4	30	0,94355	0,93191	0,95129	0,94512
5	15	0,91140	0,89694	0,90905	0,89779
5	18	0,92308	0,90988	0,92290	0,91322
5	21	0,93160	0,91936	0,93305	0,92456
5	24	0,93811	0,92660	0,94165	0,93805
5	27	0,94322	0,93230	0,94694	0,94014
5	30	0,94736	0,93692	0,95190	0,94571



Şekil 10. Açık Uçlu Matematik Sınavının Anahtarla ve Serbest Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan K Çalışması Sonucunda Hesaplanan G Katsayısındaki Değişim Grafiği



Şekil 11. Açık Uçlu Matematik Sınavının Anahtarla ve Serbest Puanlanmasında Farklı Madde ve Puanlayıcı Sayılarına Göre Yapılan K Çalışması Sonucunda Hesaplanan Phi Katsayısındaki Değişim Grafiği

Tablo 14 ve Şekil 10 incelendiğinde 15 madde ve 5 puanlayıcıdan oluşan orijinal bileşik testin serbest puanlanması sonucu elde edilen puanlamaların tek değişkenli modelle kestirilen G katsayısının 0,91140 olduğu görülürken serbest puanlamada bu katsayı 0,90905 olmuştur. Her iki puanlama için G katsayılarının birbirine yakın olduğu görülmektedir. Serbest puanlamada puanlayıcı sayısı sabit tutulup bileşik testte bulunan madde sayısı üç arttırıldığında ($n_m=18$) kestirilen G katsayısı 0,92308; bileşik testte bulunan madde sayısı altı arttırıldığında ($n_m=21$) kestirilen G katsayısı 0,93160 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken ($n_p=5$) her bir alt testte bulunan madde sayısı üç; bileşik testte bulunan madde sayısı dokuz ($n_m=24$) arttırıldığında kestirilen G katsayısı 0,93811; bileşik testte bulunan madde sayısı on beş arttırıldığında ($n_m=30$) kestirilen G katsayısı 0,94736 olarak hesaplanmıştır. Anahtarla puanlamada ise puanlayıcı sayısı sabit tutulup bileşik testte bulunan madde sayısı üç arttırıldığında ($n_m=18$) kestirilen G katsayısı 0,92290; bileşik testte bulunan madde sayısının altı arttırılması ($n_m=21$) ile kestirilen G katsayısı 0,93305 olarak, bileşik testteki toplam madde sayısının dokuz arttırılması ($n_m=24$) ile kestirilen G katsayısı ise 0,94165 olarak hesaplanmıştır. Puanlayıcı sayısı sabit kalmak koşulu ile ($n_m=5$) madde sayısının on iki arttırılması ($n_m=27$) ile kestirilen G katsayısı 0,95300 olarak, madde sayısının on beş arttırılması ile ($n_m=30$) kestirilen G katsayısı ise 0,95739 olarak hesaplanmıştır. Bu durumda serbest puanlamada puanlayıcı sayısı sabit kalmak koşuluyla toplam testte bulunan madde sayısı üç arttırıldığında G katsayısı 0,01168 artmış; madde sayısı altı arttırıldığında G katsayısı 0,0202 artmış ve madde sayısı on beş arttırıldığında G katsayısında 0,03596 artış olmuştur. Aynı senaryo anahtarla puanlama için yapılan K çalışmasında uygulandığında bileşik testteki toplam madde sayısının üç arttırılması durumunda G katsayısı 0,01385 artarken; madde sayısı altı arttırıldığında G katsayısı 0,024 artmış ve madde sayısı on beş arttırıldığında G katsayısında 0,04285 artış olmuştur. Açık uçlu matematik sınavının anahtarla puanlaması ile elde edilen verilerle yapılan K çalışması sonuçlarına göre puanlayıcı sayısı sabitken madde sayısının arttırılması G katsayısını serbest puanlamaya göre daha fazla arttırmıştır.

Puanlayıcı sayısının azaltılması durumunda hem anahtarla hem serbest puanlamada G katsayısında beklendiği gibi düşüş olmuştur. Örneğin serbest

puanlamada puanlayıcı sayısının bir azaltılması durumunda G katsayısı 0,00417; puanlayıcı sayısının iki azaltılması durumunda G katsayısı 0,01104; puanlayıcı sayısının üç azaltılması durumunda G katsayısı 0,02446 azalmıştır. Anahtarla puanlamada ise puanlayıcı sayısının bir azaltılması durumunda G katsayısı 0,00089, puanlayıcı sayısının iki azaltılması durumunda G katsayısında 0,00238 ve puanlayıcı sayısının üç azaltılması durumunda G katsayısında 0,00533 azalma olmuştur. Görüldüğü gibi madde sayısı sabitken puanlayıcı sayısının azaltılmasından anahtarla puanlanan açık uçlu matematik sınavının verileri serbest puanlanan verilere göre daha az etkilenmiştir. Bu örneklerden de görüleceği gibi hem serbest hem anahtarla puanlamada bileşik test için kestirilen G katsayısının arttırmak için madde sayısını arttırmak, puanlayıcı sayısını arttırmaktan daha uygun olacaktır.

Tablo 14 ve Şekil 11 incelendiğinde açık uçlu matematik sınavının sırasıyla serbest ve anahtarla puanlanması sonucu elde edilen verilerle orijinal madde ve puanlayıcı sayısı ile yapılan K çalışması sonucu kestirilen Phi katsayısı 0,89694 ve 0,91057 olarak hesaplanmıştır. Anahtarla puanlama ile yapılan çalışmada hesaplanan Phi katsayısının serbest puanlamaya göre daha yüksek olduğu görülmektedir. Buna göre anahtarla puanlamanın serbest puanlamaya göre daha güvenilir sonuçlar verdiği söylenebilir. Tablo 14 incelendiğinde serbest puanlamada madde sayısı sabit kalmak şartı ile ($n_m=15$) puanlayıcı sayısının bir azaltılması ($n_p=4$) ile kestirilen Phi katsayısında 0,00523 azalma görülürken anahtarla puanlamada madde sayısı sabit kalmak şartı ile ($n_m=15$) puanlayıcı sayısının bir azaltılması ($n_p=4$) ile kestirilen Phi katsayısında 0,00088 azalma olmuştur. Yine serbest puanlamada madde sayısı sabit kalmak şartı ile ($n_m=15$) puanlayıcı sayısının üç azaltılması ($n_p=2$) ile kestirilen Phi katsayısında orijinal madde ve puanlayıcı sayısı ile yapılan K çalışmasına göre 0,0304 azalma olurken anahtarla puanlamada aynı durum için bu azalma 0,00235 olarak hesaplanmıştır. Bu nedenle anahtarla puanlama ile elde edilen verilerin puanlayıcı sayısının azaltılmasından serbest puanlamaya göre daha az etkilendiği söylenebilir.

Serbest puanlamada puanlayıcı sayısının bir azaltılıp ($n_p=4$) her bir alt testte bulunan madde sayısı bir arttırılmak üzere toplam madde sayısı üç arttırıldığında hesaplanan Phi katsayısı orijinal madde ve puanlayıcı sayısı ile yapılan K çalışmasına göre hesaplanan Phi katsayısından 0,00516 fazladır. Anahtarla puanlamada ise

puanlayıcı sayısının bir azaltılıp ($n_p=4$) her bir alt testte bulunan madde sayısı bir arttırılmak üzere toplam madde sayısı üç arttırıldığında hesaplanan Phi katsayısı orijinal madde ve puanlayıcı sayısı ile yapılan K çalışmasına göre hesaplanan Phi katsayısından 0,01463 fazladır. Puanlayıcı sayısının bir azaltılıp toplam madde sayısının üç arttırıldığı durumda her iki puanlama türü için de Phi katsayısında artma olduğu; anahtarla puanlamada serbest puanlamaya göre daha fazla artış olduğu görülmektedir. Tüm bu durumlar incelendiğinde anahtarla puanlamanın işgücü ve zaman açısından serbest puanlamaya göre daha uygun olacağı için tercih edilebileceği söylenebilir.

3.3.3. Çok değişkenli G çalışması sonuçları nasıldır?

Açık uçlu matematik sınavının puanlayıcılar tarafından serbest ve anahtarla puanlanmasıyla elde edilen veriler üzerinde çok değişkenli G Kuramının $b \cdot x m \cdot x p$ deseni için G çalışması yapılmıştır. Bu çalışma ile varyans kaynakları için elde edilen varyans büyükleri ile toplam varyansı açıklama oranları karşılaştırmalı olarak Tablo 15'te verilmiştir.

Tablo 15. Açık Uçlu Matematik Sınavının Serbest ve Anahtarla Puanlanmasıyla Elde Edilen Çok Değişkenli G Çalışması Sonuçları

<i>Varyans</i> <i>Kaynağı*</i>	<i>Serbest Puanlama ile Yapılan Çok Değişkenli G Çalışması</i>						<i>Anahtarla Puanlama ile Yapılan Çok Değişkenli G Çalışması</i>					
	<i>Alt Test 1</i>		<i>Alt Test2</i>		<i>Alt Test 3</i>		<i>Alt Test1</i>		<i>Alt Test 2</i>		<i>Alt Test 3</i>	
	Varyans	%	Varyans	%	Varyans	%	Varyans	%	Varyans	%	Varyans	%
<i>b</i>	226,933	7,874	205,795	11,374	274,624	25,344	211,92287	7,234	227,51930	16,317	267,32975	38,562
<i>m</i>	2446,319	84,878	999,627	55,246	545,292	50,322	2621,25043	89,480	1063,51897	76,273	362,24213	52,252
<i>p</i>	154,782	5,370	527,177	29,135	206,718	19,077	35,7557	1,221	53,83636	3,861	13,29154	1,917
<i>bxm</i>	39,841	1,382	39,729	2,196	32,927	3,039	42,92655	1,465	47,32842	3,394	39,55166	5,705
<i>bxp</i>	4,398	0,153	7,712	0,426	5,321	0,491	1,85629	0,063	2,15970	0,155	1,96694	0,284
<i>m xp</i>	6,959	0,241	25,770	1,424	15,808	1,459	13,99955	0,478	10,77184	0,773	7,37277	1,064
<i>b x m xp, e</i>	2,919	0,101	3,593	0,198	2,9079	0,268	1,69927	0,580	1,54398	0,111	1,49876	0,216
Toplam	2882,151	100	1809,403	100	1083,598	100	2929,4107	100	1394,36275	100	693,25355	100

Alt Test 1: Bu boyutta çok deęişkenli analizle serbest puanlama için kestirilen en büyük varyans bileşeni 2446,31921 ile madde (m) ana etkisi için kestirilen varyans bileşenidir ve toplam varyansın %84,878'ünü açıklamaktadır. Anahtarla puanlama için kestirilen en büyük varyans bileşeni de 2621,25043 ile madde (m) ana etkisi için kestirilen varyans bileşenidir ve toplam varyansın %89,480'ünü açıklamaktadır. Bu boyutta serbest ve anahtarla puanlama için kestirilen en büyük varyans bileşeni aynıdır ve toplam varyansı açıklama oranları yakındır. Bu nedenle her iki puanlama türü için de bazı maddelerin diğer maddelere göre zorluk ve kolaylık bakımından çok farklılaştığını söylenebilir. Serbest puanlamada bu boyutta bireyler için 226,933 olarak kestirilen varyans ise toplam varyansın %7,874'ünü açıklamaktadır. Anahtarla puanlamada bu varyans bileşeni ise 211,92287 olarak kestirilerek toplam varyansın %7,234'ünü açıklamaktadır. Serbest ve anahtarla puanlama için birey ana etkisinin toplam varyans içindeki payları birbirine çok yakındır. Bu nedenle her iki puanlama türü için de ölçmenin uygulandığı grubun ölçülen yetenekleri bakımından çok farklılık göstermediği söylenebilir. Serbest puanlamada puanlayıcı ana etkisi için 154,78206 olarak kestirilen varyans toplam varyansın %5,370'ini açıklarken; anahtarla puanlamada puanlayıcı ana etkisi için 35,7557 olarak kestirilen varyans toplam varyansın %1,221'ini açıklamaktadır. Puanlayıcıların puanlamalarındaki farklılıkları gösteren puanlayıcı ana etkisinin toplam varyans içindeki payının anahtarla puanlamada düştüğü görülmektedir. Bu nedenle anahtarla puanlamanın serbest puanlamaya göre puanlayıcı yanlılığından daha arınık olduğu söylenebilir.

Serbest puanlama için birey x madde (*bm*) ortak etkisi varyansı toplam varyansın %1,382'sini; anahtarla puanlama için ise toplam varyansın %1,465'ini açıklamaktadır. Her iki puanlama türü için de birey x madde (*bm*) ortak etkisi için kestirilen varyans bileşeni küçük olmasından dolayı maddelerin güçlük düzeylerinin bireyden bireye çok az farklılık gösterdiği söylenebilir. Serbest puanlama için birey x puanlayıcı(*bp*) ortak etkisi varyansı toplam varyansın 0,153'ünü; anahtarla puanlama için birey x puanlayıcı(*bp*) ortak etkisi varyansı toplam varyansın %0,063'ünü, açıklamaktadır. Her iki puanlama türü için de birey x puanlayıcı(*bp*) ortak etkisi için kestirilen varyans bileşeni küçük olmakla birlikte anahtarla puanlamada daha küçüktür. Bu da puanlayıcıların verdikleri puanların anahtarla puanlamada serbest puanlamaya göre

bireyden bireye daha az farklılık gösterdiğini ifade eder. Serbest puanlama için madde x puanlayıcı(*mp*) ortak etkisi varyansı ise toplam varyansın 0,241'ini açıklamaktayken anahtarla puanlama için madde x puanlayıcı(*mp*) ortak etkisi varyansı toplam varyansın %0,478'ini açıklamaktadır. Bu durum her iki puanlama türü için de puanlayıcıların verdikleri puanların maddeden maddeye farklılık göstermeği şeklinde yorumlanabilir. Serbest puanlamada 2,91847 olarak kestirilen birey x madde x puanlayıcı ortak etkisi varyansının toplam varyans içindeki payı %0,101'dir. Anahtarla puanlamada ise bu ortak etki varyansının toplam varyans içindeki payı 1,69927 ile %0,580 olarak hesaplanmıştır. Artık varyans olarak da adlandırılan bu varyansın serbest ve anahtarla puanlama için sifıra yakın olması bu alt testte birey x madde x puanlayıcı ortak etkisinden kaynaklanan tesadüfî hataların her iki puanlama türü için de çok az olduğunun göstergesidir.

Alt Test 2: Serbest puanlama için Alt Test 2'de çok değişkenli analizle kestirilen en büyük varyans bileşeni 999,62727 ile madde (m) ana etkisi için kestirilen varyans bileşenidir ve toplam varyansın %55,246'sını açıklamaktayken; anahtarla puanlamada en büyük varyans bileşeni 1063,51897 ile yine madde (m) ana etkisi için kestirilen varyans bileşenidir ve toplam varyansın %76,273'ünü açıklamaktadır. Bu sonuca göre her iki puanlama türü için de testin bu alt boyutunda madde güçlüklerinin oldukça farklılaştığı söylenebilir. Serbest puanlamada puanlayıcı ana etkisi için 527,17708 olarak kestirilen varyans toplam varyansın %29,135'ini; anahtarla puanlamada puanlayıcı ana etkisi için 53,83636 olarak kestirilen varyans toplam varyansın %3,861'ini açıklamaktadır. Buradan görüldüğü gibi anahtarla puanlamada puanlayıcıların puanlamaları testin bu alt boyutu için de serbest puanlamaya göre oldukça tutarlıdır. Bu nedenle puanlayıcı yanlılığından arınık daha güvenilir bir ölçme aracı için anahtarla puanlamanın daha avantajlı olduğu söylenebilir. Bu boyutta serbest puanlamada bireyler için 205,79511 olarak kestirilen varyans bileşeni ise toplam varyansın %11,374'ünü açıklamaktadır. Anahtarla puanlamada ise birey varyansı 227,51930 ile toplam varyansın %16,317'sini açıklamaktadır. Anahtarla puanlama durumunda birey varyansının daha büyük çıkması nedeni ile testin bu alt boyutunda bireylerin ölçülen özellikler bakımında serbest puanlamaya göre daha çok farklılaştığı söylenebilir. Serbest puanlamada birey x madde (*bm*) ortak etkisi toplam varyansın

%2,196'sini; anahtarla puanlamada birey x madde (*bm*) ortak etkisi toplam varyansın %3,394'ünü açıklamaktadır. Testin bu alt boyutunda her iki puanlama türünde birey x madde (*bm*) ortak etkisi için kestirilen varyans bileşenleri birbirine oldukça yakındır. Bu nedenle bu alt boyut için madde güçlüklerinin bireyden bireye puanlama türüne göre farklılık göstermediği söylenebilir. Serbest puanlamada birey x puanlayıcı(*bp*) ortak etkisinin toplam varyans içindeki yüzdesinin %0,426 iken anahtarla puanlamada aynı varyansın toplam varyans içindeki yüzdesi %0,155'tir. Bu nedenle puanlayıcıların puanlamalarının bireyden bireye puanlama türüne göre farklılık göstermediği söylenebilir.

Serbest puanlamada madde x puanlayıcı (*mp*) ortak etkisi ise toplam varyansın %1,424'ünü açıklamaktadır. Anahtarla puanlama durumunda bu ortak etkinin toplam varyans içindeki payı %0,773'tür. Her iki puanlama türü için de bu ortak etki varyansının toplam varyans içindeki payı oldukça düşüktür. Bu nedenle puanlayıcıların puanlamalarının maddeden maddeye puanlama türüne göre fazla değişkenlik göstermediği söylenebilir. Testin bu alt boyutunda hata varyansı olarak adlandırılan birey x madde x puanlayıcı(*bmp*) ortak etkisi serbest puanlamada toplam varyansın %0,198'ini; anahtarla puanlamada ise toplam varyansın %0,111'ini oluşturmaktadır. Bu nedenle her iki puanlama türü için bireyler, maddeler ve puanlayıcılar arası etkileşim ile bu çalışmada ölçülemeyen sistematik ya da tesadüfî değişkenlik kaynakları oldukça düşükken anahtarla puanlamada göre bu varyans bileşeninin toplam varyans içindeki yüzdesinin serbest puanlamaya göre daha da düşük olduğu görülmektedir.

Alt Test 3: Serbest puanlama için Alt Test 3'te en büyük varyans bileşeni 545,292 olarak kestirilen madde(m)ana etkisi için kestirile varyans bileşeni toplam varyansın %50,322'sini oluşturmaktadır. Serbest puanlamada olduğu gibi anahtarla puanlamada da en büyük varyans bileşeni madde (m) ana etkisi için kestirilen varyans bileşenidir. 362,24213 ile toplam varyansın %52,252'sini açıklamaktadır. Testin bu alt boyutu için her iki puanlama türü için de en büyük varyans bileşeninin madde (m) ana etkisi için kestirilen varyans bileşeni olduğu ve toplam varyans içindeki paylarının oldukça yakın olduğu görülmektedir. Bu nedenle maddelerin bazı bireylere diğer bireylere göre daha zor ya da kolay geldiğinin bir göstergesi olan bu varyans bileşenin puanlama türünden fazla etkilenmediği söylenebilir. Testin bu alt boyutunda her iki

puanlama türü içinde ikinci en büyük varyans bileşeni birey(b) ana etkisi için kestirilen varyans bileşenidir. Serbest puanlamada birey ana etkisi (b) için kestirilen varyans bileşeni 274,62438 ile toplam varyansın %25,344'ünü açıklamaktadır. Anahtarla puanlamada ise bu varyans bileşeni 267,32975 ile toplam varyansın %38,562'sini açıklamaktadır. Bu nedenle anahtarla puanlamada bireylerin ölçülen özellikler bakımından serbest puanlamaya göre daha çok değişkenlik gösterdiği söylenebilir. Serbest puanlamada 206,71779 olarak kestirilen puanlayıcı (p) ana etkisi toplam varyansın %19,077'sini açıklarken; anahtarla puanlamada 35,7557 olarak kestirilen puanlayıcı (p) ana etkisi toplam varyansın %1,221'ini açıklamaktadır. Anahtarla puanlama ve serbest puanlama puanlayıcı ana etkisi (p) için kıyaslandığında bu ana etkinin toplam varyans içindeki yüzdesinin puanlama anahtarla yapıldığında oldukça düştüğü söylenebilir.

Testin bu alt boyutu için serbest puanlamada birey x madde (*bm*) ortak etkisi için 32,92678 olarak kestirilen varyans toplam varyansın %3,039'ünü; anahtarla puanlamada ise 39,55166 ile toplam varyansın %5,705'ini açıklamaktadır. Her iki puanlama türü için de birey x madde (*bm*) ortak etkisinin toplam varyans içindeki payının düşük olması nedeniyle bireylerin bir maddeden diğerine bağlı durumlarının çok farklılık göstermediği söylenebilir. Serbest puanlamada birey x puanlayıcı(*bp*) ortak etkisi için 5,32112 olarak hesaplanan varyans toplam varyansın %0,491'ini; anahtarla puanlamada 1,96694 olarak hesaplanan varyans toplam varyansın %0,284'ünü açıklamaktadır. Her iki puanlama türünde de puanlayıcıların puanlamalarının bireyden bireye fazlaca farklılık göstermediği görülmektedir. Madde x puanlayıcı(*mp*) ortak etkisi için serbest puanlamada 15,80771 olarak kestirilen varyans toplam varyansın %1,459'unu açıklamaktayken; anahtarla puanlamada bu ortak etki varyansı 7,37277 ile toplam varyansın %1,064'ünü açıklamaktadır. Her iki puanlama türünde de bu ortak etki varyansının düşük olması nedeniyle puanlayıcıların puanlamalarının maddeden maddeye puanlama türüne göre farklılık göstermediği görülmektedir. Birey x madde x puanlayıcı ortak etkisi için serbest puanlamada kestirilen varyans büyüklüğü 2,90787 ile toplam varyansın %0,268'ini açıklamaktadır. Anahtarla puanlamada ise artık varyansın toplam varyans içindeki payı 1,49876 olarak toplam varyansın sadece %0,216'sını açıklamaktadır. Serbest ve anahtarla puanlama için artık hata varyansının düşük olması

nedeniyle ölçülemeyen sistematik ya da tesadüfi değişkenlik kaynaklarının her iki puanlama türü için de oldukça düşük olduğunu söyleyebilir.

3.3.4. Çok değişkenli K çalışması sonuçları nasıldır?

Açık uçlu matematik sınavının serbest ve anahtarla puanlanmasında her bir alt testte bulunan madde sayısının sırasıyla 5, 6, 7, 8, 9, 10 yapılması ve puanlayıcı sayısının 5, 4, 3, 2 olarak alınması durumları için Çok Değişkenli K çalışmaları yapılmıştır. Bu K çalışmalarından elde edilen sonuçlar aşağıda Tablo 16’de verilmiştir.



Tablo 16. Açık Uçlu Matematik Sınavının Serbest ve Anahtarla Puanlanmasında Alternatif Madde ve Puanlayıcı Sayıları ile Yapılan Çok Değişkenli K Çalışması Sonuçları

Serbest Puanlama ile Yapılan Çok Değişkenli K Çalışması							Anahtarla Puanlama ile Yapılan Çok Değişkenli K Çalışması						
Alt Test 1			Alt Test 2		Alt Test 3		Alt Test 1			Alt Test 2		Alt Test 3	
n _p	n _m	G Katsayısı			Phi Katsayısı			G Katsayısı			Phi Katsayısı		
2	5	0,79481	0,75911	0,84671	0,75911	0,71594	0,83499	0,78637	0,77819	0,84102	0,74978	0,76334	0,83679
2	6	0,82071	0,78799	0,86574	0,78799	0,74278	0,85457	0,81514	0,80713	0,86327	0,78204	0,79354	0,85953
2	7	0,84027	0,80999	0,87988	0,80999	0,76322	0,86913	0,83702	0,82916	0,87990	0,80684	0,81661	0,87654
2	8	0,85556	0,82732	0,89078	0,82732	0,77930	0,88037	0,85422	0,84648	0,89279	0,82650	0,83482	0,88974
2	9	0,86784	0,84132	0,89945	0,84132	0,79229	0,88932	0,86809	0,86047	0,90309	0,84246	0,84955	0,90029
2	10	0,87793	0,85286	0,90651	0,85286	0,80299	0,89662	0,87951	0,87199	0,91150	0,85568	0,86172	0,90892
3	5	0,80749	0,77225	0,86020	0,77225	0,74266	0,85012	0,79208	0,78431	0,84615	0,75534	0,76979	0,84199
3	6	0,83271	0,80074	0,87856	0,80074	0,76971	0,86912	0,82034	0,81293	0,86798	0,78721	0,79972	0,86432
3	7	0,85172	0,82242	0,89215	0,82242	0,79026	0,88322	0,84179	0,83468	0,88428	0,81166	0,82257	0,88100
3	8	0,86656	0,83946	0,90263	0,83946	0,80462	0,89410	0,85863	0,85177	0,89692	0,83102	0,84059	0,89395
3	9	0,87846	0,85320	0,91095	0,85320	0,81945	0,90275	0,87220	0,86556	0,90700	0,84674	0,85516	0,90428
3	10	0,88822	0,86453	0,91772	0,86453	0,83018	0,90979	0,88337	0,87692	0,91522	0,85974	0,86718	0,91272

Tablo 16 devam

		Alt Test	Alt Test	Alt Test	Alt Test	Alt Test	Alt Test	Alt Test	Alt Test	Alt Test	Alt Test	Alt Test	Alt Test
		1	2	3	1	2	3	1	2	3	1	2	3
n_p	n_m	G Katsayısı			Phi Katsayısı			G Katsayısı			Phi Katsayısı		
4	5	0,81397	0,77900	0,86711	0,77900	0,75678	0,85788	0,79496	0,78741	0,84874	0,75816	0,77305	0,84461
4	6	0,83885	0,80728	0,88511	0,80728	0,78391	0,87658	0,82296	0,81585	0,87036	0,78981	0,80285	0,86673
4	7	0,85757	0,82877	0,89842	0,82877	0,80452	0,89044	0,84419	0,83747	0,88650	0,81409	0,82559	0,88325
4	8	0,87216	0,84566	0,90867	0,84566	0,82070	0,90113	0,86085	0,85444	0,89899	0,83331	0,84350	0,89606
4	9	0,88386	0,85927	0,91681	0,85927	0,83374	0,90962	0,87427	0,86813	0,90896	0,84889	0,85798	0,90628
4	10	0,89345	0,87049	0,92342	0,87049	0,84447	0,91652	0,88531	0,87940	0,91710	0,86178	0,86993	0,91463
5	5	0,81792	0,78310	0,87132	0,78310	0,76551	0,86261	0,79670	0,79670	0,85030	0,79670	0,77502	0,84619
5	6	0,84257	0,81125	0,88908	0,81125	0,79269	0,88112	0,82454	0,81762	0,87179	0,79139	0,80474	0,86818
5	7	0,86111	0,83263	0,90222	0,83263	0,81332	0,89483	0,84564	0,83915	0,88783	0,81556	0,82741	0,88461
5	8	0,87556	0,84942	0,91234	0,84942	0,82951	0,90539	0,86219	0,85605	0,90024	0,83468	0,84526	0,89734
5	9	0,88714	0,86296	0,92036	0,86296	0,84255	0,91379	0,87552	0,86968	0,91015	0,85018	0,85969	0,90749
5	10	0,89662	0,87410	0,92688	0,87410	0,85329	0,92061	0,88648	0,88090	0,91822	0,86301	0,87159	0,91578

Tablo 16 incelendiğinde Alt Test 1 boyutunda serbest puanlamada orijinal puanlayıcı sayısı 5 ve orijinal madde sayıları 5 için çok değişkenli modelle yapılan K çalışmasında kestirilen G katsayısı 0,81792 olarak hesaplanırken; anahtarla puanlamada 0,79670 olarak hesaplanmıştır. Serbest puanlamada madde sayısı sabitken ($n_m=5$) puanlayıcı sayısının 4, 3, 2 olarak alınması durumunda G katsayısındaki azalma sırasıyla 0,003, 0,010 ve 0,023 olmuştur. Anahtarla puanlamada ise yine madde sayısı sabitken puanlayıcı sayısının 4, 3, 2 alınması durumunda G katsayısındaki azalma sırasıyla 0,001, 0,004 ve 0,010 olarak hesaplanmıştır. Madde sayısı sabitken puanlayıcı sayısının azaltılması ile yapılan senaryolarda her iki puanlama türü için de G katsayısını fazlaca etkilemediği görülmektedir. Bu nedenle madde sayısı sabitken puanlayıcı sayısının azaltılması durumunda G katsayısının puanlama türünden etkilenmediği söylenebilir. Serbest puanlamada puanlayıcı sayısı sabitken madde sayısının bir artırılması G katsayısını 0,024 arttırmıştır. Yine puanlayıcı sayısı sabitken madde sayısının iki artırılması G katsayısını 0,043 arttırmıştır. Puanlayıcı sayısı sabitken madde sayısının beş artırılması ise G katsayısı üzerinde 0,078'e kadar artış sağlamıştır. Anahtarla puanlamada puanlayıcı sayısı sabitken madde sayısının bir artırılması G katsayısını 0,02784 arttırmıştır. Yine puanlayıcı sayısı sabitken madde sayısının iki artırılması G katsayısını 0,04894 arttırmıştır. Puanlayıcı sayısı sabitken madde sayısının beş artırılması ise G katsayısı üzerinde 0,08978'e kadar artış sağlamıştır. Bu sonuçlara göre Alt test 1 boyutunda serbest ve anahtarla puanlama ile yapılan K çalışması sonuçlarına göre her iki puanlama türü için de madde sayısının artırılması puanlayıcı sayısının artırılmasına göre G katsayısına daha fazla katkı sağlamıştır. Bu nedenle hem serbest hem de anahtarla puanlama için puanlayıcı sayısının artırılmasından da madde sayısının artırılması G katsayısını arttırmakta hem de zaman, işgücü ve ekonomi açısından kolaylık sağlamaktadır.

Alt test 2 boyutunda serbest puanlamada orijinal puanlayıcı sayısının beş ($n_p=5$) ve orijinal madde sayısının beş ($n_m=5$) olduğu durum için çok değişkenli modelle yapılan K çalışması ile kestirilen G katsayısı 0,78694 olarak hesaplanmıştır. Anahtarla puanlamada ise orijinal puanlayıcı sayısının beş ($n_p=5$) ve orijinal madde sayısının beş ($n_m=5$) olduğu durum için çok değişkenli modelle yapılan K çalışması ile kestirilen G

katsayısı 0,79670 olarak hesaplanmıştır. Bu boyut için serbest puanlamada madde sayısının sabit tutulması durumunda puanlayıcı sayısının bir azaltılması sonucu kestirilen G katsayısındaki azalma 0,007 iken puanlayıcı sayısının iki azaltılması ile G katsayısındaki azalma 0,019 olarak hesaplanmıştır. Puanlayıcı sayısı ikiye düşürüldüğünde ise G katsayısındaki azalma 0,041 olmuştur. Anahtarla puanlamada madde sayısının sabit tutulması durumunda puanlayıcı sayısının bir azaltılması sonucu kestirilen G katsayısındaki azalma 0,009 iken puanlayıcı sayısının iki azaltılması ile G katsayısındaki azalma 0,012 olarak hesaplanmıştır. Puanlayıcı sayısı ikiye düşürüldüğünde ise G katsayısındaki azalma 0,019 olmuştur. Serbest ve anahtarla puanlamada madde sayısının sabit tutulup puanlayıcı sayısının düşürülmesi ile yapılan senaryolarda G sayısında meydana gelen düşüşler birbirine çok yakındır. Bu nedenle madde sayısı sabitken puanlayıcı sayısının azaltılması durumunda kestirilen G katsayılarının puanlama türünden fazlaca etkilenmediği söylenebilir.

Alt Test 3 boyutunda serbest puanlamada orijinal puanlayıcı sayısı beş ($n_p=5$) ve orijinal madde sayıları beş ($n_m=5$) için çok değişkenli modelle K çalışması için kestirilen G katsayısı 0,87132 olarak hesaplanmıştır. Anahtarla puanlamada ise bu katsayı 0,85030 olarak hesaplanmıştır. Tablo 16 incelendiğinde serbest puanlamada madde sayısının sabit tutulması koşuluyla puanlayıcı sayısının bir azaltılması durumunda G katsayısındaki azalma 0,004 olarak hesaplanırken; anahtarla puanlamada aynı senaryo için G katsayısındaki azalma 0,00156 olarak hesaplanmıştır. Bu nedenle her iki puanlama türü için de madde sayısı sabitken puanlayıcı sayısının azaltılması ile G katsayısının çok az oranda azaldığı; anahtarla puanlamada G katsayısının serbest puanlamaya oranla daha az etkilendiği söylenebilir. Serbest puanlamada puanlayıcı sayısı sabitken madde sayısının bir artırılması durumunda G katsayısında 0,017 artış olmuştur. Madde sayısı iki arttırıldığında ise G katsayısındaki artış 0,030; madde sayısı beş arttırıldığında G katsayısı 0,049 artmıştır. Anahtarla puanlama için puanlayıcı sayısı sabitken madde sayısı bir arttırıldığında G katsayısındaki artış 0,2149; madde sayısı iki arttırıldığında ise G katsayısındaki artış 0,03753; ve madde sayısı beş arttırıldığında G katsayısı 0,06792 olarak hesaplanmıştır. Her iki puanlama türü için de Alt Test 3 boyutunda işgücü, zaman ve ekonomi düşünüldüğünde madde sayısının arttırılması puanlayıcı sayısının arttırılmasından daha uygun olacaktır.

Tablo 16 incelendiğinde serbest puanlama için Alt Test 1 boyutunda orijinal puanlayıcı sayısı beş ($n_p=5$) ve orijinal madde sayısı beş ($n_m=5$) için çok değişkenli modelle K çalışması için kestirilen Phi katsayısı 0,78310 olarak hesaplanmıştır. Anahtarla puanlama için orijinal puanlayıcı sayısı beş ($n_p=5$) ve orijinal madde sayısı beş ($n_m=5$) için çok değişkenli modelle K çalışması için kestirilen Phi katsayısı ise 0,79670 olarak hesaplanmıştır. Tablo 16 incelendiğinde madde sayısı sabitken puanlayıcı sayısının bir azaltılması durumunda Phi katsayısında 0,004; puanlayıcı sayısının iki azaltılması ile yapılan K çalışmasında Phi katsayısında 0,010; puanlayıcı sayısının üç azaltılması ile yapılan K çalışmasında ise hesaplanan Phi katsayısında 0,023 azalma olmuştur. Anahtarla puanlamada aynı senaryolar için hesaplanan Phi katsayılarındaki azalma ise sırasıyla 0,03854; 0,04136; 0,04692 olarak hesaplanmıştır.

Tablo 16 incelendiğinde serbest puanlamada için puanlayıcı sayısı sabitken ($n_p=5$) madde sayısının iki arttırılması ($n_m=7$) ile yapılan K çalışmasında hesaplanan Phi katsayısı, orijinal madde ve puanlayıcı sayısına göre hesaplanan Phi katsayısından 0,04 fazladır. Buna karşılık madde sayısı sabitken ($n_m=5$) puanlayıcı sayısının iki azaltılması ($n_p=3$) ile yapılan K çalışmasında hesaplanan Phi katsayısında orijinal madde ve puanlayıcı sayısına göre hesaplanan Phi katsayısına göre 0,010 azalma olmuştur. Anahtarla puanlamada ise puanlayıcı sayısı sabitken ($n_p=5$) madde sayısının iki arttırılması ($n_m=7$) ile yapılan K çalışmasında hesaplanan Phi katsayısı, orijinal madde ve puanlayıcı sayısına göre hesaplanan Phi katsayısından 0,019 fazladır. Buna karşılık madde sayısı sabitken ($n_m=5$) puanlayıcı sayısının iki azaltılması ($n_p=3$) ile yapılan K çalışmasında hesaplanan Phi katsayısında orijinal madde ve puanlayıcı sayısına göre hesaplanan Phi katsayısına göre 0,041 azalma olmuştur. Bu sonuçlara göre serbest puanlamada madde sayısının arttırılması sonucu hesaplanan Phi katsayısı anahtarla puanlamaya göre daha fazla etkilenmiştir. Madde sayısı sabitken puanlayıcı sayısının arttırılması durumunda yapılan K çalışmasında Phi katsayısı anahtarla puanlamada serbest puanlamaya göre daha fazla etkilenmiştir. Yapılan çalışmada her iki puanlama türü için de Phi katsayısını arttırmak için madde sayısının arttırılmasını yeterli olacağı söylenebilir.

Alt Test 2 boyutunda serbest puanlama için orijinal puanlayıcı beş ($n_p=5$) ve

orijinal madde sayıları beş ($n_m=5$) için çok değişkenli modelle K çalışması için kestirilen Phi katsayısı 0,76551; anahtarla puanlama için ise 0,79670 olarak hesaplanmıştır. Serbest puanlamada madde sayısının sabit tutulması durumunda madde sayısı değiştirilmeden puanlayıcı sayısı bir azaltıldığında ($n_p=4$) kestirilen Phi katsayısı 0,008 azalmıştır. Madde sayısı değiştirilmeden puanlayıcı sayısı üçe düşürüldüğünde ise kestirilen Phi katsayısında 0,022 azalma olmuştur. Yine madde sayısı değiştirilmeden puanlayıcı sayısı ikiye düşürüldüğünde ise kestirilen Phi katsayısında ise 0,04 azalma olmuştur. Anahtarla puanlamada madde sayısı değiştirilmeden puanlayıcı sayısı bir azaltıldığında ($n_p=4$) kestirilen Phi katsayısı 0,002; puanlayıcı sayısı üçe düşürüldüğünde kestirilen Phi katsayısında 0,005; puanlayıcı sayısı ikiye düşürüldüğünde ise 0,006 azalma olmuştur. Bu sonuçlara göre puanlayıcı sayısının azaltılmasından serbest puanlama sonucu hesaplanan Phi katsayısının anahtarla puanlamaya göre fazlaca etkilendiği söylenebilir.

Serbest puanlamada puanlayıcı sayısı sabitken ($n_p=5$) madde sayısının bir arttırılması ile ($n_p=6$) kestirilen Phi katsayısı 0,008 artmıştır. Madde sayısının beş arttırılması ile hesaplanan Phi katsayısında 0,04 artma olmuştur. Anahtarla puanlamada ise puanlayıcı sayısı sabitken ($n_p=5$) madde sayısının bir arttırılması ile ($n_p=6$) kestirilen Phi katsayısı 0,03 artmıştır. Madde sayısının beş arttırılması ile hesaplanan Phi katsayısında 0,097 artma olmuştur. Serbest ve anahtarla puanlama için alt test 2 boyutunda puanlayıcı sayısının azaltılması veya madde sayısının arttırılması Phi katsayısını aynı oranda etkilemiştir. Fakat madde sayısının arttırılmasının zaman ve işgücü bakımından daha uygun olması nedeniyle alt test 2 boyutunda her iki puanlama türü için de madde sayısının arttırılması madde sayısının arttırılması için yeterli olacaktır.

Alt test 3 boyutunda serbest puanlama için orijinal puanlayıcı beş ($n_p=5$) ve orijinal madde sayıları beş ($n_m=5$) için çok değişkenli modelle K çalışması için kestirilen Phi katsayısı 0,86261 olarak hesaplanmıştır. Anahtarla puanlama için orijinal puanlayıcı beş ($n_p=5$) ve orijinal madde sayıları beş ($n_m=5$) için çok değişkenli modelle K çalışması için kestirilen Phi katsayısı 0,84619 olarak hesaplanmıştır. Tablo16 incelendiğinde puanlayıcı sayısı sabitken Alt Test 3'te bulunan madde sayısının bir

arttırılması ile Phi katsayısında 0,018; Alt Test 3'te bulunan madde sayısının beş arttırılması ile Phi katsayısında 0,058 artış hesaplanmıştır. Aynı senaryo anahtarla puanlama için uygulandığında puanlayıcı sayısı sabitken Alt Test 3'te bulunan madde sayısının bir arttırılması ile Phi katsayısında 0,022; Alt Test 3'te bulunan madde sayısının beş arttırılması ile Phi katsayısında 0,07 artış hesaplanmıştır. Her iki puanlama türü için de puanlayıcı sayısı sabitken madde sayısının arttırılması ile hesaplanan Phi katsayısında artış gözlenirken; anahtarla puanlamaya göre hesaplanan Phi katsayısı madde sayısının arttırılmasından serbest puanlamaya göre daha fazla etkilenmiştir.

Serbest puanlama için Alt Test 3 boyutunda madde sayısının beş arttırılıp($n_m=10$) puanlayıcı sayısının iki azaltılması($n_p=3$) ile yapılan Kçalışmasında Phi katsayısında orijinal madde ve puanlayıcı sayısına göre 0,04 artış hesaplanmıştır. Anahtarla puanlama için madde sayısının beşarttırılıp ($n_m=10$) puanlayıcı sayısının iki azaltılması ($n_p=3$) ile yapılan K çalışmasında Phi katsayısında orijinal madde ve puanlayıcı sayısına göre 0,067 artış hesaplanmıştır. Bu nedenle alt test 3 boyutunda hem serbest hem de anahtarla puanlama için Phi katsayısını arttırmak için madde sayısının arttırılması zaman ve işgücü açısından daha kullanışlı olduğu için tercih edilebilir.

BÖLÜM IV

SONUÇ VE ÖNERİLER

Bu araştırmada açık uçlu matematik sınavının serbest ve anahtarla farklı puanlayıcılar tarafından puanlanması sonucu elde edilen veriler G kuramına göre analiz edilerek her iki puanlama türü için de en uygun madde ve puanlayıcı araştırılmıştır. Araştırmada yapılan G çalışması ve K çalışmasının sonuçları aşağıda özetlenmiştir. Elde edilen sonuçlara göre açık uçlu sınav çalışmaları, puanlayıcı güvenirliği çalışmaları, genellenebilirlik kuramına dayalı çalışmalar ve benzer çalışmalar için önerilere yer verilmiştir.

4.1. Sonuçlar

Araştırma ile elde edilen bulgulardan ulaşılan sonuçlar aşağıda alt problemlerin sırasına göre verilmiştir.

Tüm bireyleri (b), tüm maddelerde (m) aynı puanlayıcıların (p) **serbest** puanladığı açık uçlu matematik sorularının kullanıldığı ölçme durumunda tek değişkenli G çalışmasının sonuçları aşağıda özetlenmiştir.

- 1) Atılgan (2005) yaptığı çalışmada çok değişkenli ve tek değişkenli modelle kestirilen varyans bileşenlerinin toplam varyansı açıklama oranlarının farklılık gösterdiğini sonucunu bulmuştur. Bu çalışmada da benzer sonuçlar elde edilmiştir. Açık uçlu matematik sınavının serbest puanlanması sonucu elde edilen verilerle G kuramının tek değişkenli $b \times m \times p$ modeli ile kestirilen birleşik testin kestirilen varyans yüzdeleri ile G kuramının çok değişkenli $b \times m \times p$ modeli ile alt testler için kestirilen varyans yüzdelerinin farklılık gösterdiği görülmektedir.
- 2) Yapılan tek değişkenli G çalışması sonucunda birleşik test için kestirilen en büyük varyans değeri, madde ana etkisi için kestirilen varyans değerine aittir. Bu durum açık uçlu matematik sınavında kullanılan maddelerin zorluk seviyelerinin farklı olduğunu, bazı maddelerin bazı öğrencilere diğerlerine göre daha kolay ya da zor geldiğini gösterir. Deliceoğlu (2009) yaptığı çalışmada tüm alt boyutlar için en büyük varyans bileşeni birey temel etkisi için kestirilen varyans bileşeni olarak hesaplamıştır. Bu çalışmada en büyük varyans değerinin birey ana etkisi için kestirilen varyans değeri olması beklenirken üçüncü en büyük varyans olması

nedeniyle sınavın uygulandığı grubun ölçülen özellikler bakımından benzerlik gösterdiği söylenebilir.

Tüm bireyleri (b), tüm maddelerde (m) aynı puanlayıcıların (p) serbest puanladığı açık uçlu matematik sorularının kullanıldığı ölçme durumunda tek değişkenli K çalışmasının sonuçları aşağıda özetlenmiştir.

- 1) Yapılan tek değişkenli K çalışması sonucu, farklı senaryolara göre kestirilen G katsayıları Phi katsayılarına göre daha yüksek çıkma eğilimindedir. Hem Phi katsayıları hem de G katsayıları 0,8 'in üzerindedir. Bu nedenle uygulanan ölçme aracının güvenilir olduğu söylenebilir.
- 2) Puanlayıcı sayısının sabit tutulup madde sayısının arttırıldığı senaryolarda beklendiği gibi G ve Phi katsayıları artmıştır. Bunun yanı sıra madde sayısı sabitken puanlayıcı sayısının azaltılması güvenilirliği fazla düşürmemiştir. Bu nedenle madde sayısının arttırılması puanlayıcı sayısının arttırılmasından zaman, işgücü ve ekonomiklik açısından daha kolay olduğu için yapılan çalışmada ölçme aracına yeni maddeler eklenerek güvenilirlik arttırılabilir.

Tüm bireyleri (b), tüm maddelerde (m) aynı puanlayıcıların (p) serbest puanladığı açık uçlu matematik sorularının kullanıldığı ölçme durumunda çok değişkenli G çalışmasının sonuçları aşağıda özetlenmiştir.

- 1) Açık uçlu matematik sınavının serbest puanlanması sonucu elde edilen verilerin çok değişkenli G kuramı ile analiz edilmesi sonucu kestirilen varyans bileşenlerine göre tüm alt testler için madde ana etkisi en yüksek varyans açıklama oranına sahiptir. Maddelerin güçlük düzeyleri, alt testlere göre farklılık göstermemektedir.
- 2) Puanlayıcı ana etkisinin varyans açıklama yüzdesi alt test 2'de diğer alt testlere göre daha yüksektir. Bu nedenle alt test 2 için yapılan puanlamalarda puanlayıcılardan kaynaklanan hatanın diğer alt testlere göre daha yüksek olduğu sonucuna varılabilir.
- 3) *Bmp* (birey x madde x puanlayıcı) ortak etkisinden kaynaklanan artık hata varyansı tüm alt testler için düşük çıkmıştır. Bu nedenle açık uçlu matematik sınavının serbest puanlanması sonucu elde edilen verilerle yapılan G çalışmasının sonuçlarına tesadüfî hatanın az karıştığı söylenebilir.

Tüm bireyleri (b), tüm maddelerde (m) aynı puanlayıcıların (p) serbest puanladığı açık uçlu matematik sorularının kullanıldığı ölçme durumunda çok

değişkenli K çalışmasının sonuçları aşağıda özetlenmiştir.

- 1) Tüm alt testler için puanlayıcı sayılarının azaltılması ve madde sayılarının arttırılması ile yapılan K çalışmaları sonucunda, G katsayıları madde sayısının arttırılmasından puanlayıcı sayısının azaltılmasına göre daha çok etkilenmiştir. Madde sayısının arttırılması daha büyük G katsayısı elde etmek için yeterli olmuştur.
- 2) Bu çalışmada serbest puanlama ile yapılan çok değişkenli K çalışması sonucunda puanlayıcı sayısının arttırılmasındansa madde sayısının arttırılması testin tüm alt boyutları için daha yüksek Phi katsayıları elde etmek için yeterlidir. Anadol (2007) yaptığı çalışmada G ve Phi katsayılarının, mutlak ve bağıl hata varyanslarının birbiriyle paralellik gösterdiğini söylemiştir. Bu çalışmada da G katsayısı ile Phi katsayısının paralellik gösterdiği görülmektedir.

Tüm bireyleri (b), tüm maddelerde (m) aynı puanlayıcıların (p) anahtarla puanladığı açık uçlu matematik sorularının kullanıldığı ölçme durumunda; tek değişkenli G çalışması sonuçları aşağıda özetlenmiştir.

- 1) Açık uçlu matematik sınavının anahtarla puanlanması sonucu elde edilen verilerin tek değişkenli G çalışması ile analizi sonucu madde ana etkisi en büyük varyans açıklama yüzdesine sahiptir. Serbest puanlamada olduğu gibi anahtarla puanlama için de bazı maddelerin bazı öğrencilere daha kolay ya da daha zor geldiği sonucu çıkarılabilir.
- 2) Madde ana etkisinden sonra ikinci en büyük varyans bileşeni birey ana etkisi için kestirilen varyans bileşeni olmuştur. Bu nedenle bireylerin bağıl durumlarının maddeden maddeye değişmediği görülmüştür.
- 3) Puanlayıcı (p) ana etkisi için kestirilen varyans bileşeni 8,11080 olarak toplam varyansın %0,3382'sini açıklamaktadır. Bu nedenle anahtarla puanlamada puanlayıcıların puanlamalarında pek farklılık olmadığı söylenebilir. Ayrıca anahtarla puanlamada birey x madde x puanlayıcı (bmp) ortak etkisi toplam varyansı açıklamada en düşük paya sahiptir. Buna dayanarak bu ortak etkiden kaynaklanan hataların oldukça az olduğu söylenebilir.

Tüm bireyleri (b), tüm maddelerde (m) aynı puanlayıcıların (p) anahtarla puanladığı açık uçlu matematik sorularının kullanıldığı ölçme durumunda tek değişkenli K çalışmasının sonuçları aşağıda özetlenmiştir.

1) Açık uçlu matematik sınavının anahtarla puanlanması sonucu elde edilen verilerin tek değişkenli K çalışması ile analizi sonucunda kestirilen G ve Phi katsayılarının 0.80'in üstünde olduğu görülmüştür. Bir ölçme durumunda yapılan çalışmada puanlayıcı sayısının arttırılması, G Kuramı çerçevesinde güvenilirlik katsayısını arttırmaktadır (Alharby, 2006). Yapılan araştırmada puanlayıcı sayısı ve madde sayılarının değiştirilmesi ile yapılan alternatif K çalışması sonucunda kestirilen G katsayıları puanlayıcı sayısının değiştirilmesinden madde sayısının değiştirilmesine göre daha az etkilenmiştir. Bu nedenle açık uçlu matematik sınavının anahtarla puanlanması sonucu elde edilen verilerle yapılan K çalışması sonucunda G ve Phi katsayılarının arttırılması için alt testlerde kullanılan madde sayısının arttırılması yeterli olacaktır.

2) Araştırmada yapılan K çalışması sonucu en yüksek güvenilirliğe beş puanlayıcı ve her bir alt testte on madde olmak üzere toplam otuz madde ile ulaşıldığı görülmektedir. Bunun yanında dört puanlayıcı ve yirmi dört madde ile de yüksek G ve Phi katsayılarına ulaşılabileceği de görülmektedir.

Tüm bireyleri (b), tüm maddelerde (m) aynı puanlayıcıların (p) anahtarla puanladığı açık uçlu matematik sorularının kullanıldığı ölçme durumunda çok değişkenli G çalışmasının sonuçları aşağıda özetlenmiştir.

1) Taşdelen (2009) yaptığı çalışmada Nedelsky ve Angoff standart belirleme yönteminin birlikte ele alındığı tümüyle çaprazlanmış model kullanarak yaptığı G çalışması sonucunda her iki yöntemde de maddeler arası farklılıkların ortaya çıkarılabildiği ve puanlayıcılar arasında bir tutarlılığın olduğunu söylemiştir. Bu çalışmada da madde ana etkisi anahtarla puanlamada da tüm alt testler için en büyük varyans açıklama yüzdesine sahiptir. Bu nedenle hem serbest hem de anahtarla puanlama ile yapılan çok değişkenli G çalışması sonuçları düşünüldüğünde maddeler arası farklılıkların ortaya çıkarılabildiği söylenebilir.

2) Yapılan G çalışması sonucu alt testlere göre kestirilen birey ana etkisi %38,562 ile alt test 3 için en büyük değere sahiptir. Bu durumda alt test 3'ün anahtarla puanlamada diğer alt testlere göre bireysel öğrenme farklılıklarını daha iyi ortaya koyabildiği sonucuna varılabilir.

3) Anahtarla puanlama verileri ile yapılan çok değişkenli G çalışması sonucunda

puanlayıcı ana etkisi için alt testlere göre kestirilen varyans yüzdeleri serbest puanlamaya göre daha düşük çıkmıştır. Bu durumda anahtarla puanlamanın serbest puanlamaya göre puanlayıcı güvenilirliği açısından beklendiği gibi daha güvenilir sonuçlar verdiği söylenebilir.

Tüm bireyleri (b), tüm maddelerde (m) aynı puanlayıcıların (p) anahtarla puanladığı açık uçlu matematik sorularının kullanıldığı ölçme durumunda çok değişkenli K çalışmasının sonuçları aşağıda özetlenmiştir.

- 1) Yapılan çalışmada tüm alt testler için puanlayıcı sayısının azaltılması G katsayısını fazlaca etkilememiştir. Madde varyansının toplam varyansa oranının fazla olması nedeniyle madde sayısının arttırılmasının güvenilirlik üzerindeki etkisinin puanlayıcı sayısının arttırılmasından daha fazla olduğu görülmüştür.
- 2) Tüm alt testler için en yüksek Phi katsayısına puanlayıcı sayısının beş, her bir alt testte kullanılan madde sayısı on olmak üzere toplam madde sayısının otuz olduğu senaryo ile ulaşıldığı görülmektedir. Fakat puanlayıcı sayısı iki azaltılıp madde sayısının otuz olarak alındığı senaryoda Phi katsayısının alt testlere göre fazla değişmediği görülmüştür. Bu nedenle yapılan araştırmada yüksek G ve Phi katsayıları elde edebilmek için her iki puanlama türünde de madde sayısının arttırılması zaman, işgücü ve ekonomiklik açısından yeterli olacaktır.

Tüm bireyleri (b), tüm maddelerde(m) aynı puanlayıcıların (p) anahtarla ve serbest puanladığı ölçme durumunda tek değişkenli G çalışmasının sonuçları aşağıda özetlenmiştir.

- 1) Açık uçlu matematik sınavının serbest ve anahtarla puanlanması ile elde edilen verilerle yapılan tek değişkenli G çalışması sonucunda kestirilen en büyük varyans bileşeni her iki puanlama için de madde (m) ana etkisi varyansıdır. Bu durum hem serbest hem de anahtarla puanlama için testin maddelerin birbirinden oldukça farklılaştığının göstergesidir. Ayrıca anahtarla puanlamada madde varyansının daha büyük çıkması nedeniyle testin maddelerinin bu puanlama türünde daha çok farklılaştığı söylenebilir.
- 2) Anahtarla puanlamada birey (b) ana etkisi için kestirilen varyans yüzdesinin daha büyük olması nedeni ile testi alan bireylerin bağıl durumlarının maddeden maddeye serbest puanlamaya göre daha az değiştiği görülmüştür.

- 3) Anahtarla puanlamada puanlayıcı ana etkisinin toplam varyansa katkısının serbest puanlamaya göre düşük olması nedeni ile beklendiği gibi anahtarla puanlamada puanlayıcı yanlılığından kaynaklanan hataların oldukça azaldığı söylenebilir.

Tüm bireyleri (b), tüm maddelerde(m) aynı puanlayıcıların (p) anahtarla ve serbest puanladığı ölçme durumunda tek değişkenli K çalışmasının sonuçları aşağıda özetlenmiştir.

- 1) Açık uçlu matematik sınavının anahtarla ve serbest puanlanmasında farklı madde ve puanlayıcı sayılarına göre yapılan çok değişkenli K çalışması sonucunda her iki puanlama için G katsayılarının birbirine yakın olduğu görülmektedir. Açık uçlu matematik sınavının anahtarla puanlaması ile elde edilen verilerle yapılan K çalışması sonuçlarına göre puanlayıcı sayısı sabitken madde sayısının arttırılmasının, G katsayısını serbest puanlamaya göre daha fazla arttırdığı gözlenmektedir. Puanlayıcı sayısının azaltılması durumunda da hem anahtarla hem serbest puanlamada G katsayısında beklendiği gibi düşüş olmuştur.
- 2) Madde sayısı sabitken puanlayıcı sayısının azaltılmasından anahtarla puanlanan açık uçlu matematik sınavının verilerinin serbest puanlanan verilere göre daha az etkilendiği gözlenmiştir. Anahtarla puanlamada iki puanlayıcı ve on maddeyle ulaşılabilen G katsayısına serbest puanlama yapıldığında en az dört puanlayıcı ile ulaşılabilmektedir. Bu nedenle anahtarla puanlama yaparken güvenilir sonuçlar elde etmek için iki puanlayıcı yeterli olurken serbest puanlamada en az dört puanlayıcı gerekmektedir.
- 3) Yapılan çalışmada anahtarla puanlama ile hesaplanan Phi katsayısının serbest puanlamaya göre daha yüksek olması nedeni ile anahtarla puanlamanın serbest puanlamaya göre daha güvenilir sonuçlar vereceği söylenebilir. Bu nedenle anahtarla puanlamanın işgücü ve zaman açısından serbest puanlamaya göre daha uygun olacağı sonucuna varılabilir.

Tüm bireyleri (b), tüm maddelerde(m) aynı puanlayıcıların (p) anahtarla ve serbest puanladığı ölçme durumunda çok değişkenli G çalışmasının sonuçları aşağıda özetlenmiştir.

- 1) Birey (b) ana etkisi hem anahtarla puanlamada hem serbest puanlamada varyans

açıklama yüzdeleri arasında ikinci sıradadır. Her iki yöntemle yapılan G çalışmasının sonuçları paralellik göstermektedir.

- 2) Puanlayıcı (p) ana etkisi serbest puanlamada tüm alt testlerde yüksekken, anahtarla puanlamada puanlayıcıların yaptığı puanlamalardaki farklılıktan kaynaklanan puanlayıcı ana etkisi tüm alt testlerde ve birleşik testte serbest puanlamaya göre oldukça düşmüştür. Bu durumda anahtarla puanlamada puanlayıcıların katılık ve cömertliklerinin serbest puanlamaya göre daha düşük olduğu söylenebilir.

Tüm bireyleri (b), tüm maddelerde(m) aynı puanlayıcıların (p) anahtarla ve serbest puanladığı ölçme durumunda çok değişkenli K çalışmasının sonuçları aşağıda özetlenmiştir.

- 1) Testin tüm alt boyutları için orijinal puanlayıcı sayısı 5 ve orijinal madde sayıları 5 için çok değişkenli model ile yapılan K çalışmasında kestirilen G katsayısı her iki puanlama türü için de oldukça yakın olarak hesaplanmıştır. Puanlayıcı ve madde sayılarının arttırılıp azaltılması ile yapılan K çalışması sonuçlarına göre her iki puanlama türü için de madde sayısının arttırılması puanlayıcı sayısının arttırılmasına göre G katsayısına daha fazla katkı sağlamıştır.
- 2) Testin tüm alt boyutları için orijinal puanlayıcı sayısı 5 ve orijinal madde sayıları 5 için çok değişkenli model ile yapılan K çalışmasında serbest puanlama için hesaplanan Phi katsayısı anahtarla puanlama için hesaplanan Phi katsayısına yakın çıkmıştır. G katsayısında olduğu gibi Phi katsayısını arttırmak için de madde sayısının arttırılması yeterli olmuştur. Madde sayısının arttırılması Phi katsayısına puanlayıcı sayısının arttırılması kadar katkı sağlamıştır. Bu nedenle yapılan çalışmada G ve Phi katsayılarını arttırmak için puanlayıcı sayısının arttırılmasına göre madde sayısının arttırılmasının daha avantajlı olduğu söylenebilir.

4.2. Öneriler

1. Açık uçlu matematik sınavının uygulandığı grup homojen bir grup olduğu için bireysel farklılıklar yeteri kadar ortaya konulamamıştır. Yapılacak olan benzer araştırmalarda ölçme aracı daha heterojen bir gruba uygulanabilir.
2. Madde sayısının arttırılmasının güvenilirliğe etkisinin puanlayıcı sayısının arttırılmasından daha fazla olduğu görülmüştür. Benzer çalışmalarda daha az puanlayıcı sayısı ve daha fazla madde kullanılabilir.
3. Araştırmada kullanılan açık uçlu matematik sınavı daha büyük yaş gruplarına da uygulanarak araştırmanın sonuçları incelenebilir.
4. Benzer bir araştırma puanlayıcılar öğrenim düzeylerine göre sınıflandırılarak (lisans, yüksek lisans) yapılabilir.
5. Bu çalışmada açık uçlu matematik sınavının değerlendirilmesinin güvenilirliği genellenebilirlik kuramının $b \times m \times p$ modeli ile analiz edilmiştir. Benzer çalışmalar genellenebilirlik kuramının farklı modelleri ve farklı değişkenlik kaynakları ele alınarak yapılabilir.

KAYNAKÇA

- Alharby, E. (2006). A Comparasion Between Two Scoring Methods, Holistic vs Analytic, Using Two Measurement Models, The Generalizability Theory and Many-Faced Rasch Measurement Within the Context of Performance Assesment . *Yayınlanmamış Doktora Tezi* . The Pennysylvania State University.
- Anadol, H. (2017). Dereceli Puanlama Anahtarının Güvenirliğinin Farklı Deneyim Yıllarına Sahip Puanlayıcıların Kullanıldığı Durumda İncelenmesi. *Yüksek Lisans Tezi* . Ankara: Ankara Üniversitesi.
- Arsan, N. (2012). Buz Pateninde Hakem Değerlendirmelerinin Genellenebilirlik Kuramı ve Rasch Modeli ile İncelenmesi. *Yayınlanmamış Yüksek Lisans Tezi* . Ankara: Hacettepe Üniveristesi.
- Atılğan, H. (2004). Genellenebilirlik Kuramı ve Çok Değişkenlik Kaynaklı Rasch Modelinin Karşılaştırılmasına İlişkin Bir Araştırma. *Yayınlanmamış Doktora Tezi* . Ankara: Hacettepe Üniversitesi.
- Atılğan, H. (2005). Genellenebilirlik Kuramı ve Puanlayıcılar Arası Güvenirlik İçin Örnek Bir Uygulama. *Eğitim Bilimleri ve Uygulama* , 4 (7), 95-108.
- Atılğan, H. K. (2007). *Eğitimde Ölçme ve Değerlendirme*. Ankara: Anı Yayınları.
- Bachman, L. F. (1990). *Fundamental Considerations in Language Testing* . Oxford: Oxford University Press.
- Badger, E., & Thomas, B. (1992). Open-ended questions in reading. Practical Assesment. *Resarch and Evaluation* , 3(4), 03.
- Bağcan Büyükturan, E., & Çıkırcı Demirtaşlı, N. (2013). Çoktan seçmeli testler ile yapılandırılmış gridlerin psikometrik özellikleri bakımından karşılaştırılması. *Ankara Üniveristesi Eğitim Bilimleri Fakültesi Dergisi* , 46(1),395-415.
- Baker, F. B. (2016). *Madde Tepki Kuramının Temelleri*. Ankara: Pegem Akademi Yayınları.

- Başat, B. (2015). 6. Sınıf Fen Bilimleri Dersi Elektriği Üretelim Ünitesinin Değerlendirilmesinde Kullanılan Araçların Güvenirliinin Genellenebilirlik Kuramına Göre Analizi. *Yüksek Lisans Tezi* . Tokat: Gaziosmanpaşa Üniversitesi.
- Baykul, Y. (2000). *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması*. Ankara: ÖSYM Yayınları.
- Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement* , 27(2), 93-108.
- Brennan, R. L. (1992). *Elements of Generalizability Theory* . New York, NY: Springer Verlag.
- Brennan, R. L. (2001a). *Generalizability Theory*. New York, NY: Springer Verlag.
- Brennan, R. L. (2011). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education* , 24, 1-21.
- Brennan, R. L. (2001b). *Manual for mGENOVA Version 2.1.:Iowa Testing Programs Occasional Paper Number 50*. Iowa: The University of Iowa.
- Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement* , 29(3), 253-271.
- Crocker, L., & Algina, J. (1986). *Introduction to Classical And Modern Test Theory*. New York: Holt, Rinehart and Winston.
- Çelen, N. (2008). Klasik Test Kuramı ve Madde Tepki Kuramı'na Dayalı Olarak Geliştirilen İki Testin Psikometrik Özelliklerinin Karşılaştırılması. *Yayınlanmamış Doktora Tezi* . Ankara: Ankara Üniversitesi.
- Deliceoğlu, G. (2009). Futbol Yetilerine ilişkin Dereceleme Ölçeğinin Genellenebilirlik ve Klasik Test Kuramına Dayalı Güvenirliklerinin Karşılaştırılması. *Doktora tezi* . Ankara: Ankara Üniversitesi.

- Demirel, Ö., & Ün, K. (1987). *Eğitim terimleri*. Ankara: Şafak Matbaacılık.
- Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Psychological Education and Exercises Science* , 5 (1), 13-14.
- Graham, S., Hebert, M., Sandbank, M. P., & Harris, K. R. (2016). Assessing the Writing Achievement of Young Struggling Writers: Application of Generalizability Theory. *Learning Disability Quarterly* , 39(2), 72–82.
- Gronlund, N. E. (1998). *Assessment Of Student Achievement*. Boston: Allyn and Bacon.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Güler, N. (2014). Analysis of open-ended statistics questions with many facet Rasch model. *Eurasian Journal of Educational Research* , 55, 73-90.
- Güler, N. (2008). Klasik Test Kuramı Genellenebilirlik Kuramı ve Rasch Modeli Üzerine Bir Araştırma. *Yayınlanmamış Doktora Tezi* . Ankara: Hacettepe Üniversitesi.
- Güler, N. (2011). Rastgele Veriler Üzerinde Genellenebilirlik Kuramı ve Klasik Test Kuramı'na Göre Güvenirliğin Karşılaştırılması. *Eğitim ve Bilim* , 225-23.
- Güler, N., & S, G. (2007). Açık Uçlu Matematik Sorularının Güvenirliğinin Klasik Test Kuramı ve Genellenebilirlik Kuramına göre İncelenmesi. *KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ* , 991-1019.
- Güler, N., Eroğlu, Y., & Akbaba, S. (2014). Reliability of criterion-dependent measurement tools according to Generalizability Theory: Application in the case of eating skills. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi* , 14(2), 217-232.
- Güler, N., Uyanık, G. K., & Teker, G. T. (2012). *Genellenebilirlik Kuramı*. Ankara: Pegem Akademi.
- Gültekin, S., & Çıkrıkçı Demirtaşlı, N. (2012). Comparing the test information obtained through multiple-choice, open-ended and mixed item tests based on item response theory. *Elementary Education Online* , 11(1), 251-263.

- Haiyang, S. (2010). An Application of Classical Test Theory and Manyfacet Rasch Measurement in Analyzing the Reliability of an English Test for Non-English Major Graduates. *Chinese Journal of Applied Linguistics* , 88.
- Haiyang, S. (2010). An Application of Classical Test Theory and Manyfacet Rasch Measurement in Analyzing the Reliability of an English Test for Non-English Major Graduates. *Chinese Journal of Applied Linguistics* , 88.
- Hambelton, R. K., & Swaminathan, H. (1985). *Item Response Theory. Principles and Application*: Kluwer: Nijhoff Publishing.
- Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. 17(3), 123-139.
- İlhan, M. (2016). Açık Uçlu Sorularla Yapılan Ölçmelerde Klasik Test Kuramı ve Çok Yüzeyle Rasch Modeline Göre Hesaplanan Yetenek Kestirimlerinin Karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi* , 31(2): 346-368.
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review* , 2(2), 130-144.
- Kan, A. (2007). *An Alternative Method in the New Educational Program from the Point of Performance-based Assessment: Rubric Scoring Scales*. Mersin: Educational Sciences: Theory & Practice.
- Kantor, R., & Lee, Y. W. (2007). Evaluating Prototype Tasks and Alternative Rating Schemes for a New ESL Writing Test through G Theory. *International Journal of Testing* , 7(4), 353-385.
- Kieffer, K. M. (1998). *Why Generalizability Theory is Essential and Classical Test Theory is Often Inadequate?* . New Orleans, LA. USA.: Paper presented at the annual meeting of the Southwestern Psychological Association.
- Kline, T. (2005). *Psychological Testing: A Practical Approach to Design and Evaluation*. Thousand Oaks: Sage Publications.

- Kraiger, K. (1989). Generalizability Theory: An Assessment Of Its Relevance To The Air Force Job Performance Measurement Project.
- Lee, G., & Frisbie, D. A. (1999). Estimating Reliability Under a Generalizability Theory Model for Test Scores Composed of Testlets. *Applied Measurement in Education* , 12(3), 237-255.
- Lord, F. M., & Novick, R. M. (1968). *Statistical Theories of Mental Test Scores*. California: Addison-Wesley Publishing Company.
- McMillian, J. H. (1997). *Classroom assesment: Principles and practice for effective instruction*. Needham Heights: MA: Allyn and Bacon.
- Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analysis. *Behavior Research Methods* , 38, 542-547.
- Nalbantoğlu, F. (2009). Performans Ölçümlerinde Genellenebilirlik Kuramıyla Farklı Desenlerin Karşılaştırılması. *Yüksek Lisans Tezi* . Ankara: Hacettepe Üniversitesi.
- Nalbantoğlu, Y. F. (2012). Genellenebilirlik Kuramında Dengelenmiş ve Dengelenmemiş Desenlerin Karşılaştırılması-İntramuskuler Enjeksiyon Yapma İstasyonu Verileri Üzerinde Bir Uygulama. *Doktora Tezi* . Ankara: Ankara Üniversitesi.
- Öğrenci Seçme ve Yerleştirme Merkezi. (2015). Yazılı sınav (Açık uçlu sorularla sınav): <http://www.osym.gov.tr/belge/1-23308/yazili-sinav-acik-uclu-sorularla-sinav-04022015.html> adresinden alınmıştır.
- Özberk, H. E. (2012). Genellenebilirlik Kuramı Karar Çalışmalarında Kullanılan Farklı Katsayıların Karşılaştırılması. *Yayınlanmamış Yüksek Lisans Tezi* . Ankara: Hacettepe Üniversitesi.
- Özçelik, D. A. (2011). *Ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayıncılık.
- Popham, W. J. (1997). What's wrong-and what's right-with rubrics. *Educational Leadership* , 55(2), 72-75.

- Popping, R. (2008). *Analyzing open-ended questions by means of text analysis procedures*. Michigan: University of Michigan.
- Romagnano, L. (2011). The myth of objectivity in mathematics assessment. *Mathematics Teacher* , 94(1), 31-37.
- Shavelson, J. R., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. Newbury Park: CA: Sage Publications.
- Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology* , 34, 133–166.
- Taşdelen, G. (2009). Nedelsky ve Angoff Standart Belirleme Yöntemlerinin Genellenebilirlik Kuramı ile Karşılaştırılmasına İlişkin Bir Araştırma. *Yüksek Lisans Tezi* . Ankara: Hacettepe Üniversitesi.
- Tekindal, S. (1997). *Klasik Yazılı Sınavla ve Çok Sorulu Testle Elde Edilen Ölçümlerin Güvenirlik ve Geçerliği*. Samsun: Cem Ofset.
- Tekindal, S. (1997). *KLASİK YAZILI SINAVLA VE ÇOK SORULU TESTLE ELDE EDİLEN ÖLÇÜMLERİN GÜVENİRLİK VE GEÇERLİĞİ*. Samsun : Cem Ofset.
- 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.
- Turgut, M. F. (1997). *Eğitimde Ölçme ve Değerlendirme Metotları*. Ankara: Saydam.
- Turgut, M. F., & Baykul, Y. (2013). *Eğitimde Ölçme ve Değerlendirme*. Ankara: Pegem Akademi.
- Valenza, J. K. (2000). *Students and teachers alike can benefit from rubrics*. The Philadelphia Inquirer: Interview with Dr.Andi Stix. .
- Wang, Z. (2005). *Estimating Reliability Under A Generalizability Theory Model For Writing Scores in C-Base*. Master Thesis. Columbia: University of Missouri.

- Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). *Reliability coefficients and generalizability theory*. Amsterdam: Elsevier.
- Yelboğa, A. (2007). Klasik Test Kuramı ve Genellenebilirlik Kuramına Göre Güvenirliğin Bir İş Performansı Ölçeği Üzerinde İncelenmesi. *Yayınlanmamış Doktora Tezi* . Ankara: Ankara Üniversitesi.
- Yurdugöl, H. (2006). Paralel, Eşdeğer ve Konjenerik Ölçmelerde Güvenirlik Katsayılarının Karşılaştırılması. *A.Ü. Eğitim Bilimleri Fakültesi Dergisi* , 39(1). 15-37.

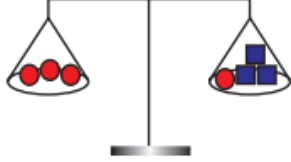


EKLER

EK - 1

Araştırmada Kullanılan Ölçme Aracı

1)



● → Bilinmeyen ■ → 10 Gram

Yukarıdaki terazi dengededir. Terazide verilen bilinmeyeni bulmak için gerekli olan denklemi kurarak çözünüz.

2)

Bir işçi maaşının $\frac{1}{4}$ ünü ev kirasına, kalan maaşının $\frac{1}{3}$ ünü yakıt parasına verdiğinde geriye 400 TL parası kalıyor.

Buna göre işçinin ev kirasının kaç TL olduğunu işlem yaparak bulunuz.

3) Süleyman bir şişe içindeki sütün $\frac{3}{4}$ ünü içtikten

sonra, kalan sütün $\frac{4}{5}$ ini daha içiyor.

Buna göre Süleyman işe içindeki sütün kaçta kaçını içmemiştir bulunuz.

4)

$1 - \frac{1}{1 - \frac{1}{1 - \frac{1}{2}}}$ işleminin sonucunu bulunuz.

5)

$$\frac{6}{5 - \frac{12}{\frac{x}{2} - 2}} = 3$$

Yukarıda verilen eşitliğe göre x 'in kaç olduğunu bulunuz.

6) $\left(4 - \frac{1}{4}\right) : \left(\frac{17}{4} - 4\right)$ işleminin sonucunu işlem basamaklarını yazarak bulunuz.

7) $(-4,3)$ noktasının $3x+4y=0$ doğrusunun üzerinde olup olmadığını işlem adımlarını göstererek belirleyiniz.

8) A $(2x-5; 3y+12)$ noktasının apsisi 7, ordinatı 15 olduğuna göre $x+y$ değerini gerekli işlemleri yaparak bulunuz.

9) $3x-y=3$ doğrusunun grafiğini eksenleri kestiği noktaları bularak koordinat sisteminde gösteriniz.

10) $ax-2y+3=0$ doğrusunun $(-1 ; 3)$ noktasından geçmesi için a 'nın ne olması gerektiğini bulunuz.

11) $4x-3y=24$ doğrusunun eksenleri kestiği noktaların koordinatları toplamını bulunuz.

12)

$$\frac{x-1}{x+1} = \frac{3}{5} \quad \text{olduğuna göre } x \text{ değerini işlem yaparak bulunuz.}$$

13) $21+ 5(a+1)= a+10$ denkleminde a yerine gelmesi gereksayıyı işlem yaparak bulunuz.

14) Bir bilgi yarışmasında doğru cevapların her birine +10 puan, yanlış cevaplara da -5 puan verilmektedir. Doğru cevap sayısı yanlış cevap sayısının iki katı olan bir öğrenci 120 puan alıyorsa sınav kaç sorudan oluşmaktadır?

15) Adem her gün, bir önceki güne göre 150 m. fazla koşarak üç günde toplam 1650 m. koşmuştur. Buna göre Adem ilk gün kaç m. koşmuştur?

EK – 2

Araştırmada Kullanılan Puanlama Anahtarı

Aşağıda Ek.1’de verilen ölçme aracında yer alan açık uçlu matematik sorularına ait puanlama anahtarı verilmiştir.

1) $3x = x + 30$ (4 puan)
 $3x - x = 30$ (2 puan)
 $2x = 30$ (2 puan)
 $x = 15$ (2 puan)

2) $1 - \frac{1}{4} = \frac{3}{4}$ (2 puan)
 $\frac{3}{4} \cdot \frac{1}{3} = \frac{1}{4}$ (2 puan)

$\frac{1}{4} + \frac{1}{4} = \frac{2}{4}$ (2 puan)

$1 - \frac{2}{4} = \frac{2}{4}$ (2 puan)

$400 \div 2 = 200$ (2 puan)

3) $1 - \frac{3}{4} = \frac{1}{4}$ (2 puan)

$\frac{1}{4} \cdot \frac{4}{5} = \frac{1}{5}$ (4 puan)

$\frac{1}{4} - \frac{1}{5} = \frac{5}{20} - \frac{4}{20} = \frac{1}{20}$ (4 puan)
(5) (4)

4) $1 - \frac{1}{2} = \frac{1}{2}$ (2 puan)

$1 \cdot \frac{2}{1} = 2$ (4 puan)

$1 - 2 = -1$ (2 puan)

$1 - (-1) = 2$ (2 puan)

5) $5 - \frac{12}{\frac{x}{2} - 2} = 2$ (2 puan)

$\frac{12}{\frac{x}{2} - 2} = 3$ (2 puan)

$\frac{x}{2} - 2 = 4$ (2 puan)

$\frac{x}{2} = 6$ (2 puan)

$x = 12$ (2 puan)

6) $4 - \frac{1}{4} = \frac{16}{4} - \frac{1}{4} = \frac{15}{4}$ (4 puan)

$\frac{17}{4} - 4 = \frac{17}{4} - \frac{16}{4} = \frac{1}{4}$ (4 puan)

$\frac{15}{4} \cdot \frac{4}{1} = 15$ (2 puan)

- 7) $3 \cdot (-4) + 4 \cdot (-3) = 0$ olduğunu kontrol edelim. (4 puan)

$$-12 + 12 = 0 \text{ (2 puan)}$$

$$0 = 0 \text{ (2 puan)}$$

Denklem sağlandığı için nokta doğrunun üzerindedir. (2 puan)

- 8) $2x - 5 = 7$ (2 puan)

$$2x = 7 + 5$$

$$2x = 12$$

$$x = 6 \text{ (2 puan)}$$

$$3y + 12 = 15 \text{ (2 puan)}$$

$$3y = 15 - 12$$

$$3y = 3$$

$$y = 1 \text{ (2 puan)}$$

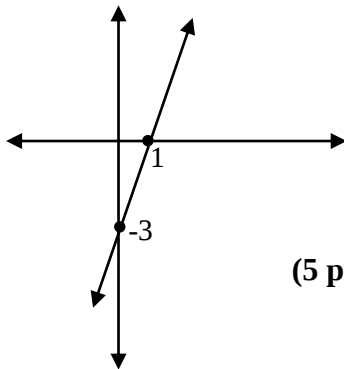
$$6 + 1 = 7 \text{ (2 puan)}$$

9)

x	y
0	-3
1	0

$$x = 0 \text{ ise } -y = 3 \text{ ve } y = -3$$

$$y = 0 \text{ ise } 3x = 3 \text{ ve } x = 1 \text{ (5 puan)}$$



(5 puan)

- 10) $x = -1$ ve $y = 3$ yazalım.

$$-a - 6 + 3 = 0 \text{ (4 puan)}$$

$$-a - 3 = 0 \text{ (2 puan)}$$

$$-a = 3 \text{ (2 puan)}$$

$$a = -3 \text{ (2 puan)}$$

- 11) $4x - 3y = 24$

x	y
0	-8
6	0

$$-3y = 24 \quad y = -8 \text{ (4 puan)}$$

$$4x = 24 \quad x = 6 \text{ (4 puan)}$$

$$-8 + 6 = -2 \text{ (2 puan)}$$

$$12) \frac{x-1}{x+3} = \frac{3}{5}$$

$$5(x - 1) = 3(x + 3) \text{ (2 puan)}$$

$$5x - 5 = 3x + 9 \text{ (2 puan)}$$

$$5x - 3x = 9 + 5 \text{ (2 puan)}$$

$$2x = 14 \text{ (2 puan)}$$

$$x = 7 \text{ (2 puan)}$$

- 13) $21 + 5a + 5 = a + 10$ (2 puan)

$$26 + 5a = a + 10 \text{ (2 puan)}$$

$$5a - a = 10 - 26 \text{ (2 puan)}$$

$$4a = -16 \text{ (2 puan)}$$

$$a = -4 \text{ (2 puan)}$$

14) Doğru cevap= 2x

Yanlış cevap= x

$$2x \cdot (10) + x \cdot (-5) = 120 \text{ (4 puan)}$$

$$20x - 5x = 120$$

$$15x = 120$$

$$x = 8 \text{ (4 puan)}$$

$$2x + x = 3x$$

$$3 \cdot 8 = 24 \text{ (2 puan)}$$

15) 1. Gün= x

$$2. \text{Gün} = x + 150$$

$$3. \text{Gün} = x + 300$$

$$x + x + 150 + x + 300 = 1650$$

(4 puan)

$$3x + 450 = 1650 \text{ (2 puan)}$$

$$3x = 1200 \text{ (2 puan)}$$

$$x = 400 \text{ (2 puan)}$$

EK - 3

Araştırmaya Katılan Puanlayıcılara Ait Bilgiler

Puanlayıcı Adı	Mezun Olduğu Üniversite	Öğrenim Düzeyi	Çalıştığı Kurum
Duygu Savuran	Balıkesir Necatibey Eğitim	Yüksek Lisans	50. Yıl
Albayrak	Fakültesi	(Matematik Eğitimi)	Orta Okulu
Feride Yeter	Balıkesir Necatibey Eğitim	Lisans	50. Yıl
	Fakültesi		Orta Okulu
Bilge Öztürk	ODTÜ Eğitim Fakültesi	Lisans	Çiğli Akiş Öğütçü Orta Okulu
Ayşegül Kırca	Bolu Abant İzzet Baysal	Lisans	Birgi
	Eğitim Fakültesi		Kazım Paşa Orta Okulu
Berrin Köse	Balıkesir Necatibey Eğitim	Lisans	50. Yıl
	Fakültesi		Orta Okulu

TEŞEKKÜR

Öğretmenlik mesleğine başladığım ilk günden beri yapmayı çok istediğim yüksek lisans eğitimimi tamamlayarak bu bölümü yazdığım için çok büyük mutluluk ve gurur duyuyorum.

Bu uzun, yorucu bir o kadar da keyifli süreçte Ölçme ve Değerlendirme alanında engin bilgi birikimi, sakinliği, yapıcı geri bildirimleriyle bana her zaman destek olan; bir eğitimci olarak ders anlatım becerilerini örnek aldığım tez danışmanım ve hocam Doç. Dr. Hakan ATILGAN' a sonsuz teşekkürlerimi sunuyorum. Kendisi ile çalıştığım için şanslı hissediyorum.

Yüksek lisans eğitimim boyunca derslerine severek girdiğim, bana katkı sağladıklarına inandığım bölüm hocalarım Prof. Dr. Tuncay ÖĞRETMEN' e ve Doç. Dr. Tahsin Oğuz BAŞOKÇU' ya teşekkürlerimi sunuyorum.

Tezimi yazım aşamasında okuyarak değerli görüşlerini benimle paylaşan Arş. Gör. Elif Kübra DEMİR' e katkılarından dolayı çok teşekkür ederim.

En özel teşekkürlerden biri de yüksek lisans eğitimi sırasında tanıştığım ve dostluğumuzun baki kalacağına inandığım arkadaşım Hayriye Candan ÇİÇEKÇİ' ye. Tezime ve bana yaptığın katkılardan dolayı ne kadar teşekkür etsem az. Ayrıca araştırmamı yürütürken hep görüş ve önerilerini aldığım desteklerini benden hiç esirgemeyen, tüm arkadaşlarıma da teşekkür ederim.

Bu zorlu süreçte her kahrımı çeken, hep beni cesaretlendiren çalışmalarına anlam katan sevgili aileme; hayatın her alanında rol model aldığım canım annem Nejla DAMLAR' a, hep arkamda duran canım babam Nazif DAMLAR' a, iyi ki varsın dediğim canım kardeşim Gökberk DAMLAR' a sonsuz sevgi ve minnetlerimi sunuyorum.

Hayatın her alanında olduğu gibi eğitimim sırasında da yükümü paylaşan, benim için her desteği sağlayan, varlığından huzur duyduğum canım eşim Tayfun DEMİRCİ' ye çok teşekkür ederim.

Tezimi yüksek lisans eğitimim sırasında dünyaya gelen, doğumu ile hayatıma büyük anlam katan, yorucu çalışma tempomda bir gülüşü ile yorgunluğumu alan canım oğlum Sarp DEMİRCİ' ye ithaf ediyorum.

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı: Pelin DAMLAR DEMİRCİ

Doğum Yeri ve Tarihi: İzmir 31/10/1987

Eğitim Durumu

Yüksek Lisans Öğrenimi: Ege Üniversitesi Sosyal Bilimler Enstitüsü, Eğitim Bilimleri
Anabilim Dalı Ölçme ve Değerlendirme Yüksek Lisans Programı

Lisans Öğrenimi: Dokuz Eylül Üniversitesi, Eğitim Fakültesi İlköğretim Matematik
Öğretmenliği

İş Deneyimi

Çalıştığı Kurumlar: Ovakent Merkez Ortaokulu, Mustafa Ayşe Yanbastı Ortaokulu,
Şehit Er Kamil Alkan Ortaokulu, 50. Yıl Ortaokulu

İletişim

E-Posta Adresi: pelindamlar@hotmail.com