

**PISA MATEMATİK OKURYAZARLIĞI MADDELERİNE İLİŞKİN
YETERLİLİK ŞEMASIYLA ELDE EDİLEN ÖLÇÜMLERİN
GÜVENİRLİĞİNİN GENELLENEBİLİRLİK VE KLASİK TEST
KURAMINA DAYALI OLARAK KARŞILAŞTIRILMASI**

Fatma Kızıltoprak

YÜKSEK LİSANS TEZİ

EĞİTİM BİLİMLERİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

EKİM, 2016

TELİF HAKKI VE TEZ FOTOKOPİ İZİN FORMU

Bu tezin tüm hakları saklıdır. Kaynak göstermek koşuluyla tezin teslim tarihinden itibaren tezden fotokopi çekilebilir.

YAZARIN

Adı: Fatma

Soyadı: KIZILTOPRAK

Bölümü: Eğitim Bilimleri

İmza:

Teslim tarihi: 14.11.2016

TEZİN

Türkçe Adı: PISA Matematik Okuryazarlığı Maddelerine İlişkin Yeterlilik Şemasıyla Elde Edilen Ölçümlerin Güvenirliğinin Genellenebilirlik ve Klasik Test Kuramına Dayalı Olarak Karşılaştırılması

İngilizce Adı: The Comparision of Reliability of Pisa Mathematical Literacy Items' Competency Needs Scores Obtained With Comptency Scheme Based on Generalizability Theory And Classical Test Theory

ETİK İLKELERE UYGUNLUK BEYANI

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

Yazar Adı Soyadı: Fatma KIZILTOPRAK

İmza:

JÜRİ ONAY SAYFASI

Fatma KIZILTOPRAK tarafından hazırlanan “PISA Matematik Okuryazarlığı Maddelerine İlişkin Yeterlilik Şemasıyla Elde Edilen Ölçümlerin Güvenirliğinin Genellenebilirlik ve Klasik Test Kuramına Dayalı Olarak Karşılaştırılması” adlı tez çalışması aşağıdaki jüri tarafından oy birliği ile Gazi Üniversitesi Eğitim Bilimleri Anabilim Dalı’nda Yüksek Lisans tezi olarak kabul edilmiştir.

Danışman: Prof. Dr. Mehtap ÇAKAN

(Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Gazi Üniversitesi)

Başkan: Doç. Dr. Şeref TAN

(Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Gazi Üniversitesi)

Üye: Doç. Dr. Burcu ATAR

(Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Hacettepe Üniversitesi)

Tez Savunma Tarihi:/....../2016

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

Prof. Dr. Ülkü ESER ÜNALDI

Eğitim Bilimleri Enstitüsü Müdürü

.....



Tüm sevdiklerime...

TEŞEKKÜR

Tez çalışmam, akademik çalışmalarım süresince güler yüzü ve sınırsız desteği ile her zaman yanımda olan değerli danışmanım Prof. Dr. Mehtap Çakan'a sonsuz teşekkürlerimi sunarım. Çalışma süresince önerileri ve görüşleri ile çalışmamın zenginleşmesi, fikirlerimin netleşmesine değerli görüşleri ve engin bilgisi ile katkı sağlayan değerli hocam Doç. Dr. Şeref TAN'a teşekkürü bir borç bilirim. Çalışmamın daha açık ve anlaşılır bir hale gelmesinde değerli görüşleri ile katkı sağlayan sayın hocam Doç. Dr. Burcu ATAR'a teşekkür ederim.

Çalışmamda ölçme aracının uyarlama sürecinde uzman görüşü ile destek veren değerli hocam Doç.Dr. Bahadır YANIK'a teşekkürü bir borç bilirim. Akademik çalışmalarım konusunda desteğini her zaman yanımda hissettiğim değerli hocam Doç. Dr. Tangül KABAEL'e çok teşekkür ederim. Çalışmama önemli ölçüde zaman ve emek harcayarak puanlayıcı olarak ve uzman görüşü ile destek veren arkadaşlarım, Ergün Cihat ÇORBACI, Derya ÇAVDAR, Esra TAŞ ve Sibel ADA, Ayça AKIN, Başak BARAK, Deniz EROĞLU ve Osman BAĞDAT'a değerli katkılarından dolayı çok teşekkür ederim. Yazım sürecimde benden yardımını esirgemeyen arkadaşlarım Yıldız YILDIRIM, Elif SEZER ve Fuat ELKONCA'ya çok teşekkür ederim.

Tez sürecimde bana evini açan ve beni kendi evimde gibi hissettiren çok sevgili arkadaşlarıma, kardeşlerim Mehmet-Beylül DUMAN çiftine ve KIZILTOPRAK ailesine teşekkürü bir borç bilirim. Maddi manevi olarak her koşulda yanımda olan ve beni destekleyen sevgili eşim Ayhan KIZILTOPRAK başta olmak üzere beni benden çok düşünen tüm sevdiklerime ve aileme çok teşekkür ederim.

Fatma KIZILTOPRAK

Ankara, 2016

**PISA MATEMATİK OKURYAZARLIĞI MADDELERİNE İLİŞKİN
YETERLİLİK ŞEMASIYLA ELDE EDİLEN ÖLÇÜMLERİN
GÜVENİRLİĞİNİN GENELLENEBİLİRLİK VE KLASİK TEST
KURAMINA DAYALI OLARAK KARŞILAŞTIRILMASI**

(Yüksek Lisans Tezi)

**FATMA KIZILTOPRAK
GAZİ ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ**

EYLÜL, 2016

ÖZ

Bu araştırmada PISA matematik okuryazarlığı maddelerinin yeterlilik gereksinimlerini PISA matematik okuryazarlığı çerçevesinde yer verilen temel matematiksel yeterlilikler bağlamında belirleyebilmek amacıyla geliştirilmiş olan Yeterlilik Şeması isimli puanlama anahtarı ile elde edilen ölçümlerin ölçüt ve yapı geçerliği ile güvenirliliğinin belirlenmesi amaçlanmıştır. Bu amaçla ilk olarak Yeterlilik Şeması puanlama anahtarının Türkçe'ye uyarlanma çalışmaları yürütülmüştür. Bu bağlamda alınan uzman görüşleri incelendiğinde çeviri formun dile kültüre ve matematik eğitimi alanı terminolojisine uygun olduğu görülmüştür. Yayınlanmış tüm PISA matematik okuryazarlığı maddeleri matematik eğitimi alanında doktora yeterliğini almış beş bağımsız uzman tarafından bir ay arayla iki kez puanlanmıştır. Elde edilen puanlamaların geçerliğini belirlemek amacıyla ölçüt geçerliği ve yapı geçerliği incelenmiştir. Ölçüt geçerliği bağlamında PISA matematik okuryazarlığı maddelerinin OECD tarafından açıklanan deneysel güçlük puanları yordanan değişken

olarak ele alınmıştır. Elde edilen puanlamaların güvenilirliğini belirlemek amacıyla ise Klasik Test Kuramı ve Genellenebilirlik Kuramı'ndan yararlanılmıştır. KTK çerçevesinde puanlayıcı içi ve puanlayıcılar arası güvenilirlik katsayıları incelenmiştir. G Kuramı çerçevesinde ise tümüyle çaprazlanmış desen tek değişkenli ve çok değişkenli model G ve K çalışmaları yapılmıştır. Araştırma sonuçlarına göre yeterlilik Şeması ile elde edilen puanlamaların maddelerin deneysel güçlük puanlarındaki değişkenliğin %74,7'sini anlamlı olarak açıkladığı görülmüştür. Ayrıca yapılan faktör analizi ile de teorik model doğrulanmıştır. KTK çerçevesinde puanlayıcı içi güvenilirlik için hesaplanan korelasyon katsayılarının puanlayıcılar ve kategoriler bazında farklılık gösterdiği, puanlayıcılar arası güvenilirlik bağlamında puanlayıcılar arasında güvenirliliğin güçlü olduğu sonucuna ulaşılmıştır. Çok değişkenli ve tek değişkenli G ve K çalışmaları sonucunda beş puanlayıcı ile iki zaman da yapılan puanlamaların güvenilirlik katsayılarının kabul edilebilir düzeyin üstünde olduğu görülmüştür. Ayrıca çok değişkenli model G ve K çalışmalarıyla kestirilen G ve Phi katsayılarının tek değişkenli modelle kestirilenlerden daha yüksek olduğu görülmektedir. KTK ve G kuramı çerçevesinde puanlamalara ilişkin elde edilen güvenilirlik düzeylerinin birbirine paralel olduğu bulunmuştur.

Anahtar Kelimeler: Klasik Test Kuramı, Genellenebilirlik Kuramı, Güvenirlilik, Puanlama anahtarı güvenirliliği, PISA

Sayfa Adedi: 160

Danışman: Prof. Dr. Mehtap ÇAKAN

**THE COMPARISON OF RELIABILITY OF PISA MATHEMATICAL
LITERACY ITEMS' COMPETENCY NEEDS SCORES OBTAINED
WITH COMPTENCY SCHEME BASED ON GENERALIZABILITY
THEORY AND CLASSICAL TEST THEORY**

(M. S. THESIS)

FATMA KIZILTOPRAK

GAZI UNIVERSITY

GRADUATE SCHOOL OF EDUCATIONAL SCIENCES

OCTOBER, 2016

ABSTRACT

The purpose of this study is to investigate the criterion and construct validity and reliability of the scores obtained by using Competency Scheme, which is improved for determining PISA mathematical literacy items' competency needs according to PISA mathematical literacy framework, based on generalizability theory and classical test theory. With this purpose initially the adaptation study of competency scheme was conducted. When experts' views in this context was examined, it was seen that the translated form was compatible with the language, culture and mathematical education jargon. All of the published PISA mathematical literacy items were scored at two different times with a month of interval by five independent raters who were PhD candidates in mathematical education field. With the purpose of determining the validity of scores, criterion validity and structural validity were examined. In the context of criterion validity, PISA mathematical literacy items' empirical difficulty scores were chosen as the predicted variable. Classical Test Theory and Generalizability Theory were utilized with the purpose of determining the reliability of

scores. In CTT framework, intra-rater and inter rater reliability were examined. In the framework of G theory, completely crossed designs via univariate and multivariate G and D studies were conducted. According to results, scores which were obtained via Competency Scheme explain %74,7 of the empirical difficulty scores' variability significantly. Besides, the conducted factor analysis confirmed the theoretical model. In the framework of CTT, it was seen that correlation coefficients which were calculated for intra-rater reliability differed from each other in terms of raters and categories and there is also a high reliability among the raters in the context of inter-rater reliability. At the end of the univariate and multivariate G and D studies, it was seen that the reliability coefficients which were predicted for scores obtained from five raters at two different times are higher than the acceptable levels. Furthermore, G and Phi coefficients which were estimated by multivariate model G and D studies are higher than G and Phi coefficients which were estimated by univariate G and D studies. It was found that in CTT and G theory frameworks the reliability levels are compatible with each other.

Key Words: classical test theory, generalizability theory, reliability, reliability of rubrics, PISA.

Page Number: 160

Supervisor: Prof. Dr. Mehtap ÇAKAN

İÇİNDEKİLER

TELİF HAKKI VE TEZ FOTOKOPİ İZİN FORMU	i
ETİK İLKELERE UYGUNLUK BEYANI.....	ii
JÜRİ ONAY SAYFASI.....	iii
TEŞEKKÜR.....	v
TABLolar LİSTESİ.....	xiv
ŞEKİLLER LİSTESİ.....	xvi
SİMGELER VE KISALTMALAR LİSTESİ.....	xvii
BÖLÜM I	1
GİRİŞ.....	1
Problem Durumu	1
Problem İfadesi	10
Araştırmanın Amacı	11
Alt Problemler	11
Araştırmanın Önemi.....	12
Sayıtlar	13
Sınırlılıklar.....	13
Tanımlar	13
BÖLÜM II.....	15
KAVRAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR.....	15
PISA- Uluslararası Öğrenci Değerlendirme Programı	15

PISA Maddelerinin Puanlanması ve Güçlük Düzeyleri	16
PISA 2012 Matematik Okuryazarlığı Çerçevesi	18
Yeterlilik Şeması.....	19
Ölçme Araçlarında Geçerlik ve Güvenirlik.....	21
Geçerlik	21
Güvenirlik	22
Klasik Test Kuramı.....	23
Puanlayıcı İçi Güvenirlik.....	26
Puanlayıcılar Arası Güvenirlik.....	26
Genellenebilirlik Kuramı	29
Temel Kavramlar	30
Genellenebilirlik ve Karar Çalışmaları.....	31
Çok Değişkenli G Çalışmaları.....	33
Davranışsal Ölçümlerin Simetrisi.....	33
İlgili Araştırmalar	34
Yeterlilik Şeması ile İlgili Araştırmalar	34
Puanlama Anahtarı Kullanılarak Elde Edilen Ölçümlerin Güvenirliğinin Genellenebilirlik ve Klasik Test Kuramına Dayalı Olarak İncelendiği Çalışmalar	36
BÖLÜM III	47
YÖNTEM.....	47
Araştırma Modeli.....	47
Evren ve Örneklem	47
Veri Toplama Araçları	49
Yeterlilik Şeması.....	50
Uzman Görüş Formu	52
Verilerin Toplanması.....	53

Verilerin Analizi	54
Birinci Araştırma Sorusuna İlişkin Veri Analizi	55
İkinci Araştırma Sorusuna İlişkin Veri Analizi	55
Üçüncü Araştırma Sorusuna İlişkin Veri Analizi	56
Dördüncü Araştırma Sorusuna İlişkin Veri Analizi	57
Beşinci Araştırma Sorusuna İlişkin Veri Analizi	58
Altıncı Araştırma Sorusuna İlişkin Veri Analizi	58
BÖLÜM IV	59
BULGULAR VE YORUMLAR.....	59
Birinci Araştırma Sorusuna İlişkin Bulgular	60
1.1.Araştırma Sorusuna İlişkin Bulgular	60
2.1.Araştırma Sorusuna İlişkin Bulgular	65
İkinci Araştırma Sorusuna İlişkin Bulgular	71
2.1.Araştırma Sorusuna İlişkin Bulgular	71
2.2.Araştırma Sorusuna İlişkin Bulgular	74
Üçüncü Araştırma Sorusuna İlişkin Bulgular	77
3.1.Araştırma Sorusuna İlişkin Bulgular	77
3.2.Araştırma Sorusuna İlişkin Bulgular	90
3.3.Araştırma Sorusuna İlişkin Bulgular	100
Dördüncü Araştırma Sorusuna İlişkin Bulgular	102
Beşinci Araştırma Sorusuna İlişkin Bulgular	104
Altıncı Araştırma Sorusuna İlişkin Bulgular	106
BÖLÜM V	109
SONUÇ, TARTIŞMA VE ÖNERİLER	109
Birinci Araştırma Sorusuna İlişkin Sonuç ve Tartışma	109
1.1.Araştırma Sorusuna İlişkin Sonuç ve Tartışma	110

1.2.Araştırma Sorusuna İlişkin Sonuç ve Tartışma	111
İkinci Araştırma Sorusuna İlişkin Sonuç ve Tartışma	111
2.1.Araştırma Sorusuna İlişkin Sonuç ve Tartışma	112
2.2.Araştırma Sorusuna İlişkin Sonuç ve Tartışma	113
Üçüncü Araştırma Sorusuna İlişkin Sonuç ve Tartışma	114
3.1.Araştırma Sorusuna İlişkin Sonuç ve Tartışma	114
3.2.Araştırma Sorusuna İlişkin Sonuç ve Tartışma	115
3.3.Araştırma Sorusuna İlişkin Sonuç ve Tartışma	116
Dördüncü Araştırma Sorusuna İlişkin Sonuç ve Tartışma	117
Beşinci Araştırma Sorusuna İlişkin Sonuç ve Tartışma	118
Altıncı Araştırma Sorusuna İlişkin Sonuç ve Tartışma	118
Öneriler	119
Araştırma Sonuçlarına Yönelik Öneriler	120
İleride Yapılacak Araştırmalara Yönelik Öneriler	120
KAYNAKLAR.....	122
EKLER	129
EK 1. Yeterlilik Şeması.....	130
EK 2: Kolmogorov-Smirnov Normal Dağılım Testi Sonuçları.....	135
EK 3: Betimsel İstatistikler	137
EK 4: Uzman Görüş Formu	140

TABLÖLAR LİSTESİ

Tablo 1. Maddelerin Uygulama Yıllarına ve Uygulama Türlerine Göre Dağılımı	48
Tablo 2. İlk ve İkinci Puanlamadan Elde Edilen Toplam Puanlara İlişkin Betimsel İstatistikler	60
Tablo 3. Güçlük Puanı ve Madde Puanına İlişkin Kolmogorov-Smirnov ve Shapiro Wilk Testi Sonuçları	61
Tablo 4. Güçlük Puanları ve Madde Puanları Arasındaki Pearson Korelasyon Katsayısı	62
Tablo 5. Yeterlilik Puanları ve Güçlük Puanlarına İlişkin Betimsel İstatistikler	64
Tablo 6. Yeterlilik Puanları ile Güçlük Puanları Arasındaki Yordama İlişkisi için Model Özeti	64
Tablo 7. Basit Doğrusal Regresyon Eşitliğine İlişkin Katsayılar ve Manidarlıkları	65
Tablo 8. Temel Bileşenler Analizi ile Yapılan Açıklayıcı Faktör Analizi Sonucunda Göstergelerin Faktörlere Göre Dağılımı ve Faktör Yük Değerleri.....	67
Tablo 9. Yeterlilik Şeması ile Beş Puanlayıcıdan Elde Edilen Puanların Ortalamalarına ait Açıklayıcı Faktör Analizi Sonuçları.....	68
Tablo 10. Varimax Döndürme Yöntemi ve Temel Bileşenler Analizi ile Yapılan Açıklayıcı Faktör Analizi Sonucunda Göstergelerin Faktörlere Göre Dağılımı ve Faktör Yük Değerleri	69
Tablo 11. Yeterlilik Şeması Puanlama Anahtarına İlişkin Puanlayıcı İçi Güvenirlik Katsayıları.....	72
Tablo 12. Puanlayıcıların Verdikleri Puanların Ortalamaları Arasındaki Farklara İlişkin İlişkili Örneklem T Testi Sonuçları.....	74
Tablo 13. İki Farklı Puanlama Zamanı için Yeterlilik Şeması Puanlama Anahtarına İlişkin Puanlayıcılar Arası Güvenirlik (Kendall'ın W) Katsayıları.....	75

Tablo 14. Puanlayıcıların İki Farklı Zamanda Her Bir Kategoriye Yönelik Verdiği Puanların Ortalaması Üzerinden Hesaplanan Yeterlilik Şeması Puanlama Anahtarına İlişkin Puanlayıcılar Arası Güvenirlik (Kendall'ın W) Katsayıları.....	76
Tablo 15. Yeterlilik Şeması İle Elde Edilen Ölçümlere İlişkin Tek Değişkenli ve Çok Değişkenli G Çalışması Sonucunda Kestirilen Varyanslar ve Toplam Varyansı Açıklama Yüzdeleri	79
Tablo 16. Yeterlilik Şeması ile Elde Edilen Ölçümlere İlişkin Çok Değişkenli K Çalışması Sonucunda Farklı Senaryolar için Kestirilen G ve Φ Katsayıları	92
Tablo 17. Toplam Puanlara için Zaman ve Puanlayıcı Sayıları Senaryolarına Göre Çok Değişkenli ve Tek Değişkenli K Çalışması ile Kestirilen G ve Φ Katsayıları.....	100
Tablo 18. Toplam Puanlar için Çok Değişkenli ve Tek Değişkenli K Çalışmalarıyla Kestirilen G ve Φ Katsayıları.....	105
Tablo 19. Yeterlilik Şeması ile Elde Edilen Toplam Puanlar ve Her Bir Kategoriye Ait Puanlar İçin KTK ve G Kuramı Çerçevesinde Hesaplanan Güvenirlik Katsayıları	107

ŞEKİLLER LİSTESİ

Şekil 1. PISA Maddeleri ve öğrenci yeterlikleri arasındaki ilişki Milli Eğitim Bakanlığı, 2005. PISA 2003 Uluslararası öğrenci değerlendirme projesi, Ulusal nihai rapor. Ankara Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü.	17
Şekil 2. PISA 2012 Matematik okuryazarlığı çerçevesi.....	18
Şekil 3. Toplam puanlar ortalaması ve güçlük puanları için serpm diyagramı ve standardize edilmiş yordanan değer ve standardize edilmiş hata grafiği.....	63
Şekil 4. Doğrulamalı faktör analizi ile test edilen teorik model.....	70
Şekil 5. Çok değişkenli K çalışmaları sonucunda alt kategoriler için kestirilen G katsayıları	93
Şekil 6. Çok değişkenli K çalışmaları sonucunda alt kategoriler için kestirilen Phi katsayıları	94
Şekil 7. Toplam puanlara için zaman ve puanlayıcı sayıları senaryolarına göre çok değişkenli ve tek değişkenli K çalışması ile kestirilen G ve Phi katsayıları	101

SİMGELER VE KISALTMALAR LİSTESİ

OECD	Organization for Economic Co-operation and Development - Ekonomik İşbirliği ve Kalkınma Örgütü
PISA	Programme for International Student Assessment- Uluslararası Öğrenci Değerlendirme Programı
G Kuramı	Genellenebilirlik Kuramı
KTK	Klasik Test Kuramı
G Çalışması	Genellenebilirlik Çalışması
K Çalışması	Karar Çalışması
G Katsayısı	Genellenebilirlik Katsayısı
Phi Katsayısı	Güvenirlilik Katsayısı
m	Madde
p	Puanlayıcı
z	Zaman
m x p	Madde x puanlayıcı ortak etkisi
m x z	Madde x zaman ortak etkisi
p x z	Puanlayıcı x zaman ortak etkisi
m x p x z	Madde x puanlayıcı x zaman karşılıklı etkisi (artık)

BÖLÜM I

GİRİŞ

Bu bölümde problem durumu, problem ve alt problemler, araştırmanın amacı, önemi, sayıltıları ve tanımlar kısmı yer almaktadır.

Problem Durumu

21. yüzyılda ülkelerin gelişmişlik düzeyi; ekonomik yapıları, bilgiyi üretme ve kullanma hızları, eğitim olanakları ve nitelikli birey sayısı gibi birçok etmenle ilişkilendirilmektedir. Bu etmenler arasındaki temel unsurun birey olması göz önünde bulundurulduğunda, ulusların küresel rekabette var olabilmeleri için sorgulayan ve üreten okuryazar vatandaşlara ihtiyaçları vardır (Jablonka, 2003, s.89). Okuryazarlık terimi son yıllarda farklı alanlarda sıklıkla kullanılmakta ve bilimsel okuryazarlık, matematik okuryazarlığı, teknoloji okuryazarlığı, finansal okuryazarlık, dijital okuryazarlık, görsel okuryazarlık... vb. gibi çok çeşitli okuryazarlık türlerinden bahsedildiği görülmektedir. Okuryazarlık teriminin bu türden kullanımlarının fen (science) ve eğitim ile ilgili tartışmalarda ve yayınlarda devamlı olarak kullanıldığı ve bilimsel okuryazarlığın iyi tanımlanmış bir terim olduğu belirtilmektedir (Turner, 2012). Okuryazarlığın ölçülmesi ve değerlendirmesine yönelik olarak Ekonomik İşbirliği ve Kalkınma Örgütü (OECD- Organization for Economic Co-operation and Development) tarafından yürütülen Uluslararası Öğrenci Değerlendirme Programı (PISA- Programme for International Student Assessment), yapılan geniş ölçekli ve güvenilir tarama araştırmaları arasında gösterilmektedir.

Öğrencilerin başarı düzeylerini artırmak, eğitim politikalarının öğrenci üzerindeki etkisini görmek, eğitim sistemini daha işlevsel hale getirebilmek, eğitim kalitesini yükseltmek amacıyla OECD tarafından yürütülen PISA, 15 yaş grubu öğrencilerinin okuma, matematik, problem çözme ve fen alanlarındaki okuryazarlık düzeylerinin değerlendirilmesi için yapılan bir tarama araştırmasıdır (OECD, 2004a, s.24). PISA çalışmalarının bir sonucu olarak matematik okuryazarlığına verilen önem de artmaktadır.

Ülkemizin matematik okuryazarlığını ölçmeyi amaçlayan PISA uygulamalarında almış olduğu sonuçlar öğrencilerimizin ortalamanın altında bir performans gösterdiklerini ortaya koymaktadır (OECD, 2004a, s.57; OECD, 2007, s.318; OECD, 2010a, s.47). PISA verilerinin yıllar arası analizini yapan Schleicher (2010), Kore, Finlandiya ve İsviçre'nin matematik okuryazarlıklarının yüksek, Meksika, Şili ve Türkiye'nin okuryazarlıklarının ise düşük olduğunu belirlemiştir.

2003'ten beri katıldığımız dört PISA sınavının (2003, 2006, 2009, 2012) da sonuçlarına göre ülkemiz son sıralarda yer almaktadır. Öğrencilerimizin performanslarındaki az miktarda artış olsa da ülkemiz son on ülke arasında yer almıştır. 2003'ten bu yana ülkemizin matematikte düzey 1 ve altındaki öğrenci oranı azalmış olmasına rağmen bu oranın hâlâ OECD ülkelerinin ortalamasındaki düzey 1 ve altı öğrenci oranının neredeyse iki katı olduğu görülmektedir (Milli Eğitim Bakanlığı [MEB], 2013, s.28). Berberoğlu (2007) matematik okuryazarlığında en üst düzeydeki öğrenci oranları açısından birçok ülkeden farklı olmayan ya da daha iyi durumda olan ülkemizin en büyük sorununun, en alt yeterlik düzeyinin de altındaki öğrenci sayılarının fazlalığı olduğunu ileri sürmüştür.

Alan yazında matematik okuryazarlığının bireyin matematiği gerçek hayat durumlarında kullanılabilir hale getirmesini etkileyen çok değerli bir beceri olduğu ortak savunudur (Höfer ve Beckmann, 2009; Jablonka, 2003; Kaiser ve Willander, 2005). Matematik okuryazarlığı bireylere matematiğin dünyada oynadığı rolü fark etme ve yapıcı, ilgili ve yansıtıcı vatandaşlar olarak iyi temellendirilmiş yargılarda bulunmada ve gerekli kararları vermede yardımcı olmaktadır (MEB, 2012, s.12; OECD, 2010b, s.19). OECD her ülkenin, karmaşık ve hızla değişen toplumlara ayak uydurabilmek için matematik okuryazar bir ulusa ihtiyacı olduğunu vurgulamaktadır (OECD, 2004a, s.37). OECD' ye göre çağa ayak uyduran ulusların tekrarlayan fiziksel işgücü yerine matematiksel düşünebilen vatandaşlara ihtiyaçları vardır. Özgen ve Bindak (2008) matematik okuryazarlığının öneminin kişinin temel bilgi ve becerileri kazanmasının yanında matematik ile ilgili düşünmesinin, problem

özmesinin, matematięe karşı olumlu tutum içinde olmasının ve matematięin gerçek yařamdaki önemini takdir etmesinin hedeflenmesinden kaynaklandığını belirtmiştir.

Uluslararası alanda giderek daha da önem kazanan matematik okuryazarlığına ilişkin PISA uygulamaları ile kapsamlı ve güvenilir değerlendirmeler edilmiş olmasına karşın değerlendirme sonuçlarının ileri sürülen kavramsal çerçeve ile tam olarak örtüşmüyor olması tartışma konusu olmuştur. PISA sonuçlarında matematik okuryazarlığı çerçevesinde yer verilen temel matematiksel yeterliliklere ilişkin değerlendirmeler yerine ülkeler bazında öğrencilerin yeterlilik düzeyleri ve ortalama puanlarına göre sıralamaları sunulmaktadır. Bu bağlamda Turner, Dossey, Blum ve Niss (2013) “PISA matematik okuryazarlığı maddelerinin güçlüğini etkileyen bilişsel faktörler nelerdir?” sorusundan yola çıkmış ve PISA matematik okuryazarlığı çerçevesinde tanımlanan temel matematiksel yeterlilikleri referans almışlardır. Madde güçlüğüne ilişkin bu bakış açısının PISA matematik ölçeğinin her düzeyini daha iyi değerlendirmeyi sağlayacak maddelerin üretilmesinde olduğu kadar, PISA değerlendirmelerine ilişkin öğrencilerin performanslarının raporlaştırılmasını da zenginleştirme potansiyeli olduğu belirtilmiştir. Ünlü, Kasper, Trendtel ve Schurig (2014), PISA çalışmaları kapsamında yapılan analizlerin, madde güçlüğü ile ilgili bilgilerle birlikte okuryazarlık düzeylerinin tanımlarını ve PISA maddelerinin bilişsel gereksinimlerinin (cognitive demands of items) ayrıntılı bir özetini ürettiğini belirtmişlerdir (s.422).

Embretson (1983, 1998) maddelerin çözümündeki bilişsel süreçlerle ilgili hipotezler test edilirken, maddelerin bilişsel gereksinimlerinin dikkate alınmasını önermiştir (Hartig, Frey, Nold ve Klieme, 2011). Daniel ve Embretson (2010) bilişsel karmaşıklığın matematik maddelerinin güçlüğündeki farklılıkları nasıl etkilediğine yönelik yapılan deneysel çalışmalardan bahsetmişlerdir. Daniel ve Embretson (2010)’un aktardığına göre, bu çalışmalarda Fisher (1973) türev alma kurallarının madde güçlüğini nasıl etkilediğini modellemiş; Medina-Diaz (1993) dokuzuncu sınıf cebir konusundaki maddelerin güçlüğünün, her bir madde için gerekli matematiksel işlemlerin güçlüğini ifade eden çarpma kuralları ile yordanabildiğini ortaya koymuş; matematiksel muhakeme testine yönelik bir bilişsel modelin (Embretson, 2006; Embretson & Daniel, 2008) madde güçlüğini ve madde yanıtlama süresini güçlü bir şekilde yordadığı görülmüştür.

Bu çalışmada ölçme aracı olarak kullanılan yeterliliklere dayalı madde puanlama şeması, ilk olarak Niss ve Danimarkalı meslektaşları tarafından tanımlanan bir takım yeterliliklerin, (Niss, 1999) PISA Matematik Uzmanları Grubu (MEG) tarafından kullanılmasından doğmuştur (Turner, Blum ve Niss, 2015, s.86). Bu bağlamda PISA matematik okuryazarlığı

maddelerinin bilişsel gereksinimleri boyutuna odaklanılmıştır. Maddelerin bilişsel gereksinimleri bir bireyin bir maddeyi doğru bir şekilde yanıtlayabilmesi için etkinleştirmesi gereken matematiksel yeterlilikleri (demand of activation of this competency) ve bu yeterliliklerin düzeyini içermektedir. MEG, PISA matematik okuryazarlığı maddelerinin bilişsel gereksinimlerini değerlendirmek ve öğrencilerin bu yeterliliklere sahip olma durumları ile matematiksel görevleri başarılı bir şekilde tamamlayabilme becerileri arasındaki ilişkiyi daha iyi anlayabilmek amacıyla kullanılacak bir şema ortaya çıkarmaya çalışmıştır. Kısa ve öz olması açısından bu çalışmada bu şemadan Yeterlilik Şeması olarak söz edilecektir.

Yeterlilik şeması altı yeterlilik alanını içermekte olup matematiksel problemlerin matematiksel yeterlilikler açısından niteliklerinin ölçülmesinde kullanılan her bir yeterliliğe ilişkin operasyonel tanımları, açıklamaları, yeterlilik düzeylerini ve bu düzeylere ilişkin tanımları içerdiğinden bir ölçme aracı olarak dereceli puanlama anahtarı türündedir.

Puanlama anahtarı kullanımının alanyazında daha çok performans değerlendirme ile birlikte ele alındığı görülmektedir. Palm (2008), performans değerlendirmenin gözlem ve kanıya dayandığını savunmaktadır. Gözlem ve kanıya dayanan ölçmelerde ise ölçümlerin nesnel bir şekilde elde edilmesi için çeşitli önlemler alınması gereklidir. Bu amaçla performans değerlendirme gibi alternatif ölçmeler için puanlama anahtarı geliştirilmesi ve kullanılması yoluna gidilmiştir. Puanlama anahtarı kullanımı puanlamanın öznel olduğu performans ölçümlerinde ilgilenilen değişkenlerin daha nesnel bir şekilde puanlamasını sağlamaktadır.

Puanlama anahtarının yaygın kullanılan tanımı “ölçütlerin listelenerek ve niteliklerin mükemmelden zayıfa doğru düzeyleri tanımlanarak bir göreve ilişkin beklentileri düzenleyen bir doküman” şeklinde ifade edilmektedir (Reddy ve Andrade, 2010). Popham (1997) bir puanlama anahtarının; değerlendirme ölçütleri, ölçüt tanımları ve bir puanlama stratejisi olmak üzere üç önemli özelliği olduğunu belirtmiştir. Parke (2001), değerlendirme ölçütünü bir nesnenin niteliğinin belirlenmesinde puanlayıcının göz önünde bulundurması gereken faktörler, birtakım göstergeler ya da rehber bir liste olarak tanımlamış olup, ölçütün ise süreci ve önemli görülen içeriği yansıttığı belirtmiştir (Reddy ve Andrade, 2010). Ölçüt tanımları bir öğrencinin bir beceriyi göstermede yapması gerekenlerin ayrıntılı bir açıklamasını vermektedir. Ölçüt tanımları hem puanlama amacı hem de öğrencilere sağlanacak geri bildirim için iyi ve zayıf yanıtları birbirinden ayırt etme gereğini

belirlemektedir (Reddy ve Andrade, 2010). Puanlama stratejileri ise bir süreç ya da ürüne ilişkin yargıların yorumlanması için ölçek kullanımını içerir.

Puanlama anahtarlarını çeşitli şekillerde sınıflandırmak mümkün olsa da genellikle bütüncül ve analitik olarak ikiye ayrıldığı görülmektedir. Bütüncül puanlama anahtarı kullanımıyla puanlayıcılar performansın niteliği hakkında genel bir karara varırken analitik puanlama anahtarı kullanımında görev ile değerlendirilen boyutların her biri için ayrı bir puan verirler.

Büyük ölçekli değerlendirmelerin yapıldığı günümüz öğrenme ortamında eğitimciler öğrenci çalışmalarını değerlendirmede puanlama anahtarı kullanmaktadır. Bu durum puanlama anahtarlarının değerlendirmelerin geçerliliğini ve güvenilirliğini arttıran araçlar olarak görüldüğünün bir göstergesidir (Rezaei ve Lovorn, 2010). Ancak yapılan bazı çalışmalar puanlama anahtarlarının doğrudan kullanılmasının puanlayıcılar arası güvenilirliği, geçerliliği ve genel olarak değerlendirmenin niteliğini ve geçerliliğini arttırdığı yönündeki görüşü zorlamaktadır (Örneğin; Rezai ve Lovorn, 2010). Analitik puanlama anahtarına dayalı değerlendirmelerin güvenilirliği ve geçerliliğine yönelik yeterli deneysel çalışmalar yürütülmemiştir. Ayrıca performans değerlendirme gibi alternatif ölçmeler için puanlamaların güvenilirliğini belgelemede zorluklar olduğu bilinmektedir (Taylor ve Pastor, 2013). Bununla birlikte yapılan her türlü ölçmeye bir miktar hata karıştığı kabul edildiğinden ve ölçme aracının gerçekte ölçmek istediği özelliği başka değişkenlerle karıştırmadan ölçebilme derecesinin belirlenmesi açısından ölçme araçları ile elde edilen verilerin güvenilirliğinin ve geçerliğinin incelenmesi gerekmektedir.

Güvenirlik ve geçerlik ölçme araçlarında bulunması istenen temel özelliklerdir. Çeşitli ölçme araçlarıyla elde edilen verilerin geçerli ve güvenilir olması, bu ölçme sonuçlarına dayalı olarak verilecek kararların isabetli olması için ön koşuldur. Geçerlik, ölçülen niteliğin başka değişkenlerle karıştırılmadan ölçülmesi ile ilgilidir. Buna bağlı olarak bir ölçme aracı ile elde edilen ölçümlerin geçerliği, aracın ölçme amacına hizmet etme derecesi olarak ifade edilebilir. Ölçme sonuçlarının geçerliği, ölçme yapılan öğrenci grubunun niteliklerine, ölçme ortamına, ölçme sonuçlarının kullanılacağı amaca yani değerlendirme ve karar verme işleminin gereklerine bağlı olarak değişir (Turgut ve Baykul, 2013, s.133). Bu çalışmada ölçme aracı olarak kullanılan Yeterlilik Şeması ile elde edilen ölçümlerin geçerliği, maddelerin gerektirdiği matematiksel yeterlilik düzeylerinin başka değişkenlerle karıştırılmadan ölçülmesi anlamındadır. Çalışma kapsamında ele alınan geçerlilik diğer bir

deyişle madde yeterlilik gereksinimi puanlarının, maddelerin gerektirdiği matematiksel yeterlilikleri ölçebilme derecesi olarak ifade edilebilir.

Güvenirlik bir ölçme aracının ölçülmek istenilen özelliği doğru olarak ölçüp ölçemediğini ifade eder. Bununla birlikte, sosyal bilimlerde yapılan ölçme çalışmalarının yüzde yüz doğru sonuçlar vermesi mümkün değildir. Bunun nedeni ölçme sonuçlarına farklı kaynaklardan gelen hataların karışmasıdır. Her ölçmeye mutlaka hata karışır. Bir ölçmeye sabit, sistematik ve tesadüfi hatalar olmak üzere üç türlü hata karışabilir. Güvenirlik ölçülen bir niteliğin hatasız ve doğru bir şekilde ölçülebilme derecesi olarak ölçme sonuçlarının tesadüfi hatalardan arınık olması demektir (Tan, 2013, s.133). Ölçme sonuçlarına dayalı olarak verilen kararların isabetli olabilmesi, ölçme sonuçlarının tesadüfi hatalardan arınık olabilmeye derecesine bağlıdır. Bu bağlamda çalışma kapsamında güvenilirlik, madde yeterlilik puanlarının tesadüfi hatalardan arınık olarak belirlenmesi, hatasız ve doğru bir şekilde ölçülebilmesi olarak ele alınmıştır.

Ölçme sonuçlarına karışan hata kaynaklarından hangisine odaklanıldığına da bağlı olarak farklı güvenilirlik kestirme yöntemleri uygulanabilmektedir. Ölçmelerde ele alınan hata kaynaklarının genellikle madde, zaman, test formu ve puanlayıcı olduğu görülmektedir. Puanlayıcı hata kaynağının önemli bir yere sahip olduğu ve performansın ölçülmesinde nesnellik sağlaması bakımından puanlama anahtarlarının kullanılması önerilmektedir. Puanlama anahtarlarıyla elde edilen ölçümlerin güvenilirliğinin incelenmesinde ise farklı kuramlar kapsamında farklı yöntemler kullanılabilmektedir. Klasik Test Kuramı, Genellenebilirlik Kuramı ve Madde Tepki Kuramı hata kaynaklarını farklı yaklaşımlarla ele almakta ve güvenilirlik hesaplama yöntemleri değişiklik göstermektedir.

Ölçme araçlarının güvenilirliklerinin belirlenmesinde en çok kullanılan yaklaşım Klasik Test Kuramı'na dayalı güvenilirlik katsayılarının incelenmesidir. Klasik test kuramına göre bir gözleme ilişkin olarak iç tutarlılık güvenilirliği, test tekrar test güvenilirliği, paralel formlar güvenilirliği, puanlayıcılar arası güvenilirlik gibi birden çok güvenilirlik katsayısı hesaplanabilir. Bu güvenilirlik katsayılarından biri yüksek çıkarken diğeri düşük çıkabilir. Bu durum klasik test kuramının sınırlılıklarından biri olup, her bir kaynaktan gelen hatanın ayrı ayrı ele alınmasından kaynaklanmaktadır (Crocker ve Algina, 2008, s.157). Alternatif ölçümlerin özelliği olan çoklu hata kaynaklarının modellenmesinde klasik test kuramına dayalı güvenilirlik ölçümleri yetersiz kalmaktadır (Taylor ve Pastor, 2013). Ancak alternatif ölçümlerin elde edildiği puanlama anahtarlarının etkililiğine ilişkin önemli eğilimlerden biri,

puanlayıcılar arası ve puanlayıcı içi güvenilirliği merkeze almaktadır (Alharby, 2006'dan aktaran Güler, 2008; Rezaei ve Lovorn, 2010).

Puanlayıcı içi güvenilirlik, bir puanlayıcının farklı zamanlarda aynı ölçme objesinin niteliklerine ilişkin verdiği puanlardaki tutarlılığın, kararlılığının göstergesidir. Puanlayıcıdan ya da puanlama anahtarından kaynaklanan bazı koşullar, puanlamalar arasında uyumsuzluk oluşmasına neden olabilir. Puanlayıcı içi güvenilirlik, test tekrar test güvenilirliğinde olduğu gibi, iki farklı zamanda verilen puanlar arasındaki korelasyonla hesaplanabilir. Test tekrar test güvenilirliğinde testin zamana karşı kararlılığı anlamında güvenilirlik kestirilirken, puanlayıcı içi güvenilirlikte puanlayıcının puanlamadaki tutarlılığı anlamındaki güvenilirlik kestirilir (Güler, 2008, s.17).

Puanlayıcılar arası güvenilirlik, birden fazla puanlayıcının birbirinden bağımsız olarak aynı ölçme objesinin niteliklerine ilişkin verdiği puanlardaki uyumun göstergesidir. Puanlayıcılar arası güvenilirlik, genellikle iki puanlayıcının bir grup öğrencinin performansına verdikleri puanlar arasındaki korelasyon ile ölçülmektedir (Güler, 2008, s.18; Rezai ve Lovorn, 2010). Ancak puanlayıcılar arası güvenilirliğin belirlenmesinde korelasyon katsayısının tek başına kullanımı yeterli görülmemektedir (Goodwin, 2001). Bu nedenle korelasyon katsayısı ile birlikte ortalamalar arasındaki farkların da test edilmesi ve birlikte yorumlanması önerilmiştir.

Ölçme araçlarının güvenilirliklerinin belirlenmesinde en çok kullanılan yaklaşım klasik test kuramına dayalı güvenilirlik katsayılarının incelenmesi olmasına karşın, ölçmeye karışan her bir hata için farklı güvenilirlik katsayılarının hesaplanması gerektiği ve bu katsayılar arasında tutarsızlıklar olabileceği bilinmektedir. Performans değerlendirmenin giderek artan bir şekilde kullanımı ile farklı hata kaynaklarının birlikte ele alınmasına fırsat veren Genellenebilirlik Kuramı'na dayalı güvenilirlik katsayılarının temel alındığı çalışmalara sıklıkla rastlanmaktadır (Atılgan, 2004; Deliceoğlu ve Demirtaşlı, 2012; Eser ve Gelbal, 2013; Smith ve Kulikowich, 2004; Taylor ve Pastor, 2013; Yelboğa ve Tavşancıl, 2010; Yılmaz ve Gelbal, 2011).

Klasik test kuramının farklı yöntemleri ile farklı anlamlarda elde edilen güvenilirlik katsayılarının birbirinden farklı olmasından hareketle Cronbach ve arkadaşları Genellenebilirlik (G) kuramını ortaya atmışlardır (Atılgan, 2005). G kuramı ölçümlerin birden fazla hata kaynağına bağlı olduğu durumlarda ölçme duyarlılığını kestirmeye yönelik bir yaklaşımdır (Cardinet, Johnson ve Pini, 2010, s.1). Bu kuram araştırmacılara, hata

kaynaklarını tanımlama ve her hata kaynağının göreceli katkısını analiz etme imkânı vermektedir (Taylor ve Pastor, 2013). Webb ve Shavelson (2005, s.717)' e göre bu kuramla hata kaynaklarının gerçek nedenleri tanımlanır, hata kaynakları ayırt edilir ve her bir hata kaynağı tahmin edilir. G kuramı puanlayıcı, zaman, test formu... vs. gibi bütün potansiyel hata kaynaklarını eş zamanlı olarak değerlendirir (Atılğan, 2005).

G kuramı kapsamında hata kaynaklarının her biri bir yüzey ya da değişkenlik kaynağı olarak alınmaktadır. Tek bir değişkenlik kaynağına odaklanılarak yapılan çalışmalar tek yüzeyli desenler, birden çok değişkenlik kaynağının birlikte ele alındığı çalışmalar çok yüzeyli desenler olarak isimlendirilmektedir. Tek yüzeyli desenler için genellenebilirlik çalışması yürütülmeden varyans analizi ile çalışmalar yürütülebilir. Çok yüzeyli desenlerde G çalışmaları yürütülerek ölçmeye karışan hata kaynaklarının açıkladığı varyans oranları kestirilebilmekte ve tüm hata kaynakları birlikte ele alınarak ölçümlerin güvenilirliği için tek bir katsayısı hesaplanabilmektedir. G çalışmalarının sonuçlarına dayalı olarak yürütülen K çalışmalarıyla ölçümlerin güvenilirliği arttırmak için tasarlanabilecek farklı senaryolar test edilerek ileride yapılacak ölçmelere ilişkin önerilerde bulunmaktadır.

G kuramı aynı zamanda kullanılan ölçme aracının tek değişkenli ya da çok değişkenli bir yapı göstermesine bağlı olarak farklı modellerin kullanımına izin vermektedir. Ölçümlerin elde edildiği ölçme aracının çok değişkenli bir yapı gösterdiği ya da bir testin alt boyutlardan oluştuğu durumlarda çok değişkenli G çalışmalarının yürütüldüğü görülmektedir (Atılğan, 2004; Deliceoğlu, 2009; Wilhelm ve Kim, 2015; Yelboğa, 2007). Çok değişkenli G çalışmaları ile hem ölçme aracının tamamına hem de alt boyutlarına göre varyans kestirimlerinin yapılması ve güvenilirlik katsayılarının hesaplanmasına ek olarak ölçme aracının geneline ilişkin açıklanan varyans oranlarına ve güvenilirlik kestirimine de tek çalışma ile ulaşılabilir.

Son yıllarda ülkemizde ve dünyada genellenebilirlik kuramına dayalı güvenilirlik hesaplamaya ve karşılaştırmaya yönelik çeşitli araştırmalar yapılmıştır (Atılğan, 2004; Deliceoğlu ve Demirtaşlı, 2012; Eser ve Gelbal, 2013; Güler ve Gelbal, 2010; Güler, 2011; Smith ve Kulikowich, 2004; Taylor ve Pastor, 2013; Yelboğa ve Tavşancıl, 2010; Yılmaz ve Gelbal, 2011). Puanlayıcılar arası güvenilirlik katsayısının hesaplanması, genellenebilirlik kuramı, klasik test kuramı, madde tepki kuramı, lojistik regresyon analizi ile kestirilen güvenilirlik katsayılarının karşılaştırılması, bir ölçeğin güvenilirliğini belirleme amacıyla genellenebilirlik kuramının kullanımı, alternatif ölçmelerin güvenilirliğini belirleme

amacıyla genellenebilirlik kuramının kullanımı, açık uçlu maddelere verilen cevapların puanlanmasında puanlayıcı güvenilirliği gibi pek çok konuda çalışmalar mevcuttur. Genellenebilirlik kuramıyla yapılan bu çalışmaların özel yetenek seçme sınavlarından futbol yetilerinin ölçümüne, fen okuryazarlığı testinden PISA uygulamalarında kullanılan maddelerden uyarlanan maddelerle hazırlanmış ölçeklere, alternatif ölçmelerin teknik kalitesinden, üst düzey problem çözme becerilerinin ölçümüne kadar geniş bir yelpazeye yayıldığı görülmüştür.

Ülkemizde genellenebilirlik kuramına dayalı güvenirlik hesaplamaya ve karşılaştırmaya yönelik çeşitli araştırmaların tümünde bireyler arası farklılaşmaya odaklanıldığı görülmüştür. Diğer bir deyişle, bu çalışmaların tamamında bireyler ölçme objesi olarak ele alınmıştır. Bireylerin ölçme objesi olarak ele alınması durumunda bireyler arası farklılaşmanın maksimum olması istenir ve birey bir hata kaynağı olarak görülmez. Bununla birlikte bireylerin ölçme objesi olarak ele alınması her çalışmanın amacına uygun olmayabilmektedir. Örneğin program değerlendirme gibi öğrenme durumlarının, öğretim durumlarının ya da gelişim basamaklarının farklılaşmasının beklendiği çalışmalarda bireyler arası farklılaşmanın sonuçları açıklamaya yardımcı olmadığı bilinmektedir (Cardinet, Tourneour ve Allal, 1976). Bununla birlikte bireylerin ölçme durumunda yer almaması, sadece puanlayıcı olarak yer alması ya da çalışmanın amacının farklı objeler arasındaki farklılaşmanın maksimum olmasını gerektirdiği durumlar (bu çalışmada olduğu gibi) için de G kuramından faydalanılabilmektedir. Cardinet vd. (1976), bireylerin çalışmanın merkezi objesi olmadığı problem durumları için de G kuramının güçlü bir tanımlayıcı ve analitik araç olduğunu belirtmektedir. Bu bağlamda, bu çalışmada maddeler ölçme objesi olarak ele alınmakta ve maddelerin yeterlilik gereksinimleri arasındaki farklılığın maksimum olması beklenmektedir.

Yeterlilik Şeması puanlama anahtarı ile PISA matematik okuryazarlığı maddelerin yeterlilik gereksinimlerine ilişkin değerlendirilmelerin geçerliği ve güvenirliğinin incelendiği çalışmalar bulunmaktadır (Turner ve Adams, 2012; Turner vd.,2013). Yeterlilik Şeması puanlama anahtarının geliştirildiği ilk çalışmada (Turner vd.,2013) değerlendirmelerin ölçüt ve yapı geçerliği çoklu regresyon yöntemi ve açıklayıcı faktör analiziyle, güvenirliği ise puanlayıcılar arası ve puanlayıcı içi güvenirlik yöntemleriyle incelenmiştir. Yeterlilik Şemasının puanlama anahtarının revize edilmiş versiyonunun kullanıldığı diğer çalışmada ise (Turner ve Adams, 2012), maddelerin yeterlilik gereksinimlerine ilişkin değerlendirmelerin ölçüt geçerliği çoklu regresyon yöntemiyle, güvenirliği klasik test

kuramı kapsamında puanlayıcılar arası güvenilirlik yöntemiyle ve çok yüzeyli Rasch modeli ile incelenmiştir.

Yeterlilik şeması puanlama anahtarının alt kategorilerden oluştuğu ve yeterlilik şeması ile yapılan ilk çalışmada (Turner vd., 2013) açımlayıcı faktör analizi sonucunda puanlama anahtarının tek boyutlu bir yapı göstermediği belirtilmiştir. Ayrıca PISA maddelerinin yeterlilik gereksinimlerinin belirlenmesine puanlayıcılardan ve uygulama zamanından kaynaklı hata karışması da mümkündür.

Ölçmelere karışan hata kaynaklarının birden fazla olması, güvenilirlik kestirimlerinin G kuramına dayalı olarak çok değişkenli modellerle incelenmesinin daha kapsamlı bir yaklaşım olacağını göstermektedir. Bununla birlikte puanlama anahtarları ile toplanan verilerin güvenilirliğinin incelenmesinde temel eğilimlerden biri olan ve klasik test kuramı kapsamında ele alınan puanlayıcı içi ve puanlayıcılar arası güvenilirliğinde göz ardı edilmemesi gerektiği düşünülmektedir. Bununla birlikte yeterlilik şeması kullanımının, PISA değerlendirmelerine ilişkin öğrenci performanslarının raporlaştırılmasını zenginleştirme potansiyeli olduğundan PISA maddelerinin yeterlilik gereksinimlerinin güvenilir bir şekilde değerlendirilmesinin önemli olduğu görülmektedir. Bu doğrultuda bu çalışmada diğer araştırmalardan farklı olarak PISA matematik okuryazarlığı maddelerinin yeterlik gereksinimlerine ilişkin değerlendirmelerin geçerliğinin yapı ve ölçüt geçerliğiyle, güvenilirliğinin ise genellenebilirlik kuramı ve KTK'ya dayalı puanlayıcı içi ve puanlayıcılar arası güvenilirlik yöntemleriyle test edilmesi amaçlanmıştır.

Problem İfadesi

2000 ile 2012 yılları arasındaki esas ve pilot uygulamalarda kullanılan yayınlanmış PISA matematik okuryazarlığı maddelerinin, Yeterlilik Şeması ile puanlamasıyla elde edilen madde yeterlilik gereksinimi puanlarının;

- Ölçüt ve yapı geçerliği ne düzeydedir?
- KTK ve G kuramı çerçevesinde belirlenen güvenilirlik kestirimleri nelerdir?
- Güvenirlik kestirimleri arasındaki ilişkiler nasıldır?

Araştırmanın Amacı

Bu araştırmanın genel amacı, yayınlanmış PISA matematik okuryazarlığı maddelerinin madde yeterlilik gereksinimi puanlarının ölçüt ve yapı geçerliğinin incelenmesi, KTK ve G Kuramı çerçevesinde güvenilirlik kestirimlerinin incelenmesi ve karşılaştırılmasıdır. Bu genel amaç doğrultusunda Yeterlilik Şeması'nın Türkçe'ye uyarlanması alt amaç olarak belirlenmiştir. Bu bağlamda bu araştırmada, 2000 ile 2012 yılları arasındaki esas ve pilot uygulamalarda kullanılan ve OECD tarafından yayınlanmış PISA matematik okuryazarlığı maddelerinin iki farklı zamanda beş puanlayıcı tarafından Yeterlilik Şeması ile puanlamasıyla elde edilen ölçümlerin ölçüt ve yapı geçerliği ile KTK ve G Kuramı çerçevesinde güvenilirliğinin incelenerek karşılaştırılması amaçlanmıştır.

Alt Problemler

1. Yeterlilik Şeması ile elde edilen ölçümlerin ölçüt ve yapı geçerliği ne düzeydedir?
 - 1.1. Yeterlilik Şeması ile elde edilen ölçümlerin ölçüt geçerliği ne düzeydedir?
 - 1.2. Yeterlilik Şeması puanlama anahtarının yapı geçerliği ne düzeydedir?
2. Yeterlilik Şeması ile elde edilen toplam puanlar ve her bir kategoriye ait puanlar için Klasik Test Kuramı çerçevesinde kestirilen puanlayıcılara ilişkin güvenilirlik katsayıları ne düzeydedir?
 - 2.1. Yeterlilik Şeması ile elde edilen toplam puanlar ve her bir kategoriye ait puanlar için puanlayıcı içi tutarlılık derecesi nedir?
 - 2.2. Yeterlilik Şeması ile elde edilen toplam puanlar ve her bir kategoriye ait puanlar için puanlayıcılar arası tutarlılık derecesi nedir?
3. Yeterlilik Şeması ile elde edilen ölçümler için kestirilen Genellenebilirlik Kuramı parametreleri nelerdir?
 - 3.1. Çok değişkenli ve tek değişkenli G çalışmaları sonucunda her bir kategoriye ve toplam puanlara ilişkin kestirilen varyanslar ve toplam varyansı açıklama yüzdeleri nelerdir?
 - 3.2. Orijinal puanlayıcı ve zaman sayıları ile puanlayıcı ve zaman sayılarının bir arttırılıp bir azaltılması senaryolarına göre çok değişkenli Karar (K) çalışmalarıyla kestirilen G ve Phi katsayıları alt kategorilere göre nasıl değişmektedir?
 - 3.3. Orijinal puanlayıcı ve zaman sayıları ile puanlayıcı ve zaman sayılarının bir arttırılıp bir azaltılması senaryolarına göre çok değişkenli ve tek değişkenli Karar (K)

çalışmalarıyla toplam puanlara ilişkin kestirilen G ve Phi katsayıları modellere göre nasıl değişmektedir?

4. Yeterlilik Şeması ile elde edilen her bir kategoriye ait puanlar için çok değişkenli model K çalışmalarıyla kabul edilebilir düzeyde bir G katsayısı elde etmek için gerekli minimum puanlayıcı ve zaman sayısı nedir?
5. Yeterlilik Şeması ile elde edilen toplam puanlar için çok değişkenli ve tek değişkenli model K çalışmalarıyla kabul edilebilir düzeyde bir G katsayısı elde etmek için gerekli minimum puanlayıcı ve zaman sayısı modellere göre nasıl değişmektedir?
6. Yeterlilik Şeması ile elde edilen toplam puanlar ve her bir kategoriye ait puanlar için KTK ve G Kuramı çerçevesinde hesaplanan güvenilirlik katsayıları iç tutarlılık ölçütlerine göre tutarlı mıdır?

Araştırmanın Önemi

Puanlama anahtarı kullanımının maddelerin yeterlilik gereksinimlerine ilişkin yapılacak ölçümlerin güvenilirliğinin sağlanmasında gerekli fakat yeterli olmadığı bilinmektedir. Genellikle performans değerlendirme çalışmalarında kullanılan puanlama anahtarlarının öğrenci performansının ölçülmesi, uygulanan programın etkililiği ve öğrenme çıktılarının değerlendirilmesi gibi farklı amaçlarla kullanımına rastlanmaktadır (Jonsson ve Svingby, 2007). Ancak puanlama anahtarlarının ölçme araçlarını meydana getiren maddelerin yeterlilik gereksinimlerinin belirlenmesinde kullanımı çok rastlanan bir durum değildir. Bununla birlikte maddelerin yeterlilik gereksinimlerine ilişkin yapılan değerlendirmelerin geçerli ve güvenilir olması gerekmektedir. Bu çalışmanın, PISA matematik okuryazarlığı maddelerinin bilişsel gereksinimlerini değerlendirmek için geliştirilen Yeterlilik Şemasının güvenilirlik ve geçerliğini incelemesi yönüyle alana katkı sağlayacağı düşünülmektedir.

Bu çalışma ile Yeterlilik Şeması puanlama anahtarının güvenilirlik kestirimlerinin iki farklı kurama göre belirlenmesi ve karşılaştırılması amaçlanmaktadır. Puanlama anahtarlarının güvenilirliğinin belirlenmesinin hem kategoriler bazında hem de toplam puanlar üzerinden araştırılacak olmasının, ölçme aracının revize çalışmalarına katkı sağlayacağı düşünülmektedir. Ayrıca Yeterlilik Şeması'nın yalnızca PISA maddeleri için değil, tüm matematiksel problemlerin analizinde de kullanılabilecek olması, matematik eğitimcilerinin yapacakları çalışmalarda madde örneklemini seçerken faydalanabilecekleri bir kaynak olması açısından önem taşıdığı düşünülmektedir. Bu çalışmanın, Yeterlilik Şeması

puanlanlama anahtarının Türkçe'ye uyarlanmış olması ve puanlanlama anahtarının ulusal alanyazına kazandırılmış olması açısından da alana katkı sağlayabileceği düşünülmüştür.

Sayıtlar

1. Puanlayıcılar maddeleri ciddiyeyle puanlamıştır.
2. Puanlayıcılar ikinci puanlamayı ilk puanlamadan ve her iki puanlamayı birbirlerinden bağımsız olarak yapmışlardır.

Sınırlılıklar

Bu araştırma;

1. OECD tarafından yayınlanmış PISA matematik okuryazarlığı maddeleri ile sınırlıdır.
2. PISA matematik okuryazarlığı maddeleri 2000 ile 2012 arasındaki uygulamalar ile sınırlıdır.
3. Araştırma gönüllü puanlayıcı olan beş matematik eğitimsi ile sınırlıdır.

Tanımlar

- a. PISA Matematik Maddelerinin Deneysel Güçlük Puanı: PISA uygulamaları sonucunda, bir maddenin güçlük düzeyi ile bir öğrencinin yetenek düzeyinin ilişkilendirildiği, maddelerin öğrencilerden toplanan verilerle aynı ölçek üzerinden ölçeklendirilmesi sonucu maddelere atanan puanlardır. Bu puanların ortalaması 500, standart sapması 100 olacak şekilde ölçeklendirme yapılmaktadır.
- b. Maddelerin Bilişsel Gereksinimi: Bireylerin bir maddeyi çözebilmesi için gerekli olan matematiksel yeterliliklerin neler olduğunu ve bunların düzeylerini ifade etmektedir.
- c. Ölçme Objesi: Bu çalışmada maddeler ölçme objesi olarak alınmıştır ve ana etki madde etkisidir.



BÖLÜM II

KAVRAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

Bu bölümde PISA matematik okuryazarlığı çerçevesi, puanlayıcı güvenirliği ile klasik test kuramı ve genellenebilirlik kuramı kısaca tanıtılmıştır. Çalışma kapsamında kullanılacak olan klasik test kuramına dayalı puanlayıcı içi ve puanlayıcılar arası güvenirlik katsayıları hakkında bilgi verilmiştir. Bölüm sonunda ise Yeterlilik Şeması ve genellenebilirlik kuramı ile ilgili araştırma sonuçlarına yer verilmiştir.

PISA- Uluslararası Öğrenci Değerlendirme Programı

Öğrencilerin başarı düzeylerini artırmak, eğitim politikalarının öğrenci üzerindeki etkisini görmek, eğitim sistemini daha işlevsel hale getirebilmek, eğitim kalitesini yükseltmek amacıyla OECD tarafından yürütülen PISA, 15 yaş grubu öğrencilerinin okuma, matematik, problem çözme ve fen alanlarındaki okuryazarlık düzeylerinin değerlendirilmesi için yapılan bir tarama araştırmasıdır (OECD, 2004a, s.24). İlk PISA uygulaması 2000 yılında gerçekleştirilmiş ve okuma alanındaki okuryazarlık becerileri geniş kapsamlı olarak değerlendirilmiştir. Her üç yılda bir tekrarlanan PISA uygulamalarında daha sonraki yıllarda sırasıyla matematik okuryazarlığı ve problem çözme okuryazarlığı (2003) ve fen okuryazarlığı (2006) alanlarına ağırlık verilmiştir ve bu döngü devam etmektedir. Şimdiye kadar yapılan uygulamalar arasından PISA 2003 ve PISA 2012 uygulamalarında matematik okuryazarlığı ve problem çözme okuryazarlığı alanlarına ağırlık verilmiştir.

PISA deęerlendirmelerine verilen önemin bir sonucu olarak PISA uygulamalarına katılan ülke sayısı giderek artmaktadır. 2000 yılında yapılan ilk PISA uygulamasına katılan 32 ülkeden 28'i OECD üyesidir. PISA 2012 uygulamasına ise 34'ü OECD üyesi olmak üzere toplam 65 ülkenin katıldığı bilinmektedir. Ülkemizde OECD üyesi bir ülke olarak PISA 2003 uygulamasından bu yana PISA uygulamalara katılmaktadır.

Ülkemizin matematik okuryazarlığını ölçmeyi amaçlayan PISA uygulamalarında almış olduęu sonuçlar öğrencilerimizin ortalamanın altında bir performans gösterdiklerini ortaya koymaktadır (OECD, 2004a, s.57; OECD, 2007, s.318; OECD, 2010a, s.47). PISA verilerinin yıllar arası analizini yapan Schleicher (2010), Kore, Finlandiya ve İsviçre'nin matematik okuryazarlıklarının yüksek, Meksika, Şili ve Türkiye'nin okuryazarlıklarının ise düşük olduğunu belirlemiştir.

Matematik okuryazarlığı tanımlarının geniş bir aralıęa yayıldığı görölmektedir. Matematik okuryazarlığının alan yazında kullanılabilecek ortak tanımının yapılmasına yönelik çalışmalar matematik okuryazarlığının farklı yönlerini vurgulamaktadır. Bununla birlikte matematik okuryazarlığını açık bir şekilde tanımlamaya yönelik ilk girişimin ilk PISA çerçevesinde OECD (1999) tarafından yapıldığı belirtilmektedir (Niss& Jablonka, 2014, s.392; Turner, 2012).

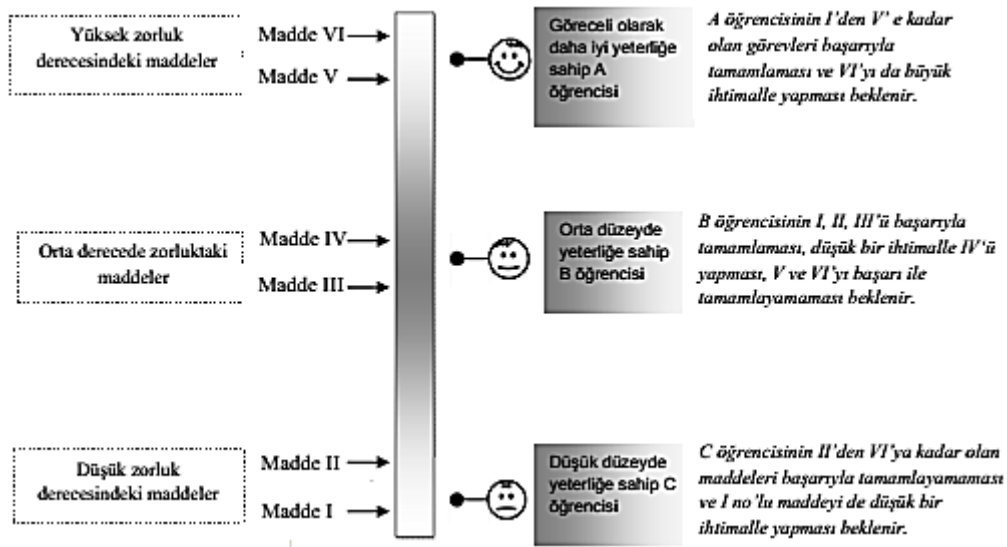
OECD (2004b), matematik okuryazarlığını, bireyin dünyada matematięin oynadığı rolü anlama ve belirleme, iyi temellendirilmiş yargılarda bulunma, yapıcı, ilgili ve düşünceli birer vatandaş olarak şimdiki ve gelecekteki hayatındaki ihtiyaçlarını karşılayacak şekilde matematikle meşgul olma kapasitesi olarak tanımlamıştır. Bu tanıma ilişkin gelen eleştiriler doğrultusunda PISA 2012' de matematik okuryazarlığı yeniden düzenlenerek, çeşitli bağlamlarda bireyin formüle etme, matematięi kullanma ve yorumlama kapasitesi olarak tanımlanmıştır. Bu kapasite matematiksel olarak akıl yürütmeyi, bir olguyu açıklamak ve tahmin edebilmek için matematiksel kavramları, işlemleri ve araçları kullanmayı içermektedir (MEB, 2012, s.12).

PISA Maddelerinin Puanlanması ve Güçlük Düzeyleri

PISA deęerlendirmelerinde farklı başarı puanları kestirilmekte olup, madde tepki kuramı modelleri kullanılarak her bir öğrenci için her konu alanında farklı puanlar tanımlanmaktadır. Öğrencilerin yetenek düzeyleri madde tepki kuramında önemli bir deęişkendir ve öğrencilerin yetenek düzeyleri doğrudan maddelerle ilişkilendirilerek

yorumlanmaktadır Madde ve öğrenci istatistikleri genellikle madde tepki kuramı çerçevesinde tek parametrelili lojistik regresyon yöntemi kullanılarak kestirilmektedir (MEB, 2012, s.5).

Matematik puanlama ölçekleri belirli bir yılda OECD ülkelerinin matematik alanındaki ortalama puanı 500 ve standart sapması 100 olacak şekilde 2003 yılı temel alınarak hazırlanmaktadır (MEB, 2013, s.11). Bu ölçeklerin hazırlanmasında aynı zamanda OECD ülkelerindeki öğrencilerin yaklaşık üçte ikisinin aldığı puanlar 400 ile 600 arasında yer alacak şekilde olmasına dikkat edilmektedir.



Şekil 1. PISA Maddeleri ve öğrenci yeterlikleri arasındaki ilişki Milli Eğitim Bakanlığı, 2005. *PISA 2003 Uluslararası öğrenci değerlendirme projesi, Ulusal nihai rapor*. Ankara Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü.

Şekil 1'de görüldüğü gibi oluşturulan puanların bir öğrencinin performansını ya da bir maddenin zorluk düzeyini belirtmek için kullanılabildiği ifade edilmektedir. Buna göre, bir maddenin zorluk düzeyi ile bir öğrencinin yetenek düzeyi ilişkilendirilmektedir. Bu doğrultuda bir öğrencinin yeterlilik puanına bakılarak aynı düzeydeki ya da altındaki güçlük puanına sahip maddeleri çözebileceği ileri sürülmektedir (MEB, 2005, s8. ; MEB, 2012, s.7). Örneğin, 550 puan alan bir öğrencinin 550 ya da daha düşük güçlük düzeyindeki maddeleri başarılı bir şekilde yanıtlayabilmesi beklenmektedir

PISA 2012 Matematik Okuryazarlığı Çerçevesi

PISA 2012’ de yeniden düzenlenen matematik okuryazarlığı çerçevesinde yer alan temel yapılar biri diğerinin içinde yer alan üç kutu ile Şekil 1’ de gösterildiği gibi modellenmiştir (OECD, 2014; p. 37). Bu modelde yer alan en büyük kutu matematik okuryazarlığının gerçek yaşamda ortaya çıkan bir problem ya da zorluk bağlamında değerlendirildiğini göstermektedir. Ortadaki kutu problemi çözmek için kullanılabilecek matematiksel düşünme ve eylemlerin doğasına ışık tutmakta ve en küçük kutu ise problem çözen bireyin bir çözüm oluşturmak için kullandığı süreçleri tanımlamaktadır. PISA 2012 matematik okuryazarlığı çerçevesinde yapılan değerlendirme üç farklı açıdan ele alınmaktadır. Bunlar; kullanılan matematiksel içerik, matematiksel süreçler ve bağlamlardır.

PISA 2012 Matematik Okuryazarlığı Çerçevesi Genel Özellikleri

Gerçek yaşam bağlamındaki zorluk

Matematiksel alan kategorileri:

Nicelik; belirsizlik ve veri; değişim ve ilişkiler; uzay ve şekil

Gerçek yaşam bağlamı kategorileri: Kişisel; Toplumsal; Mesleki; Bilimsel

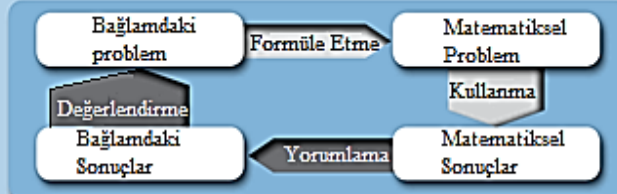
Matematiksel düşünce ve eylemler

Matematiksel kavramlar, bilgi ve beceriler

Temel matematiksel yeterlilikler:

İletişim; Temsil; Stratejiler üretme; Matematikleştirme; Muhakeme ve argüman; Sembolik, formal, teknik dilin ve işlemlerin kullanımı; matematiksel araçları kullanma

Süreçler: Formüle etme; Matematiği kullanma; Yorumlama/Değerlendirme



Şekil 2. PISA 2012 Matematik okuryazarlığı çerçevesi

PISA 2012 de kullanılan matematiksel içerik alanları; değişim ve ilişkiler, uzay ve şekil, nicelik ve belirsizlik ve veri konularıdır. PISA (2003) matematik okuryazarlığı çerçevesinde ele alınan içerik alanlarından cebir yerine nicelik, olasılık yerine belirsizlik ve veri alanlarının getirildiği görülmektedir. Matematiksel süreçler; bireyin problemi çözmek için problemin bulunduğu bağlamı matematikle ilişkilendirmesini ve aynı zamanda bu süreçleri oluşturan becerileri ifade etmektedir. Bu süreçler; formüle etme, matematiği kullanma ve yorumlama/ değerlendirme. Kullanılan bağlamlar ise, değerlendirme maddelerinin

problem durumlarının yer aldığı bağlamlardır. PISA 2012’de kullanılan bağlamlar; kişisel, mesleki, toplumsal ve bilimsel olmak üzere dört farklı kategoride ele alınmıştır. Matematiksel süreçlerde değişen düzeylerde kullanılan matematiksel yeterlilikler de PISA 2012 çerçevesinde düzenlenerek iletişim, matematik diline aktarma, temsil ile gösterim, akıl yürütme ve çıkarsama yapma, farklı stratejiler oluşturma ve kullanma, matematiksel dili ve işlemleri kullanma ve matematiksel araçları kullanma biçiminde belirlenmiştir (OECD, 2013, s.39). Bu yeterlilikler hem matematik okuryazarlığı için temel oluşturmakta, hem de matematiksel davranışı ifade etmede önemli bir araç olarak kullanılmaktadır (OECD, 2010b, s.31; MEB 2012, s.16).

Temel Matematiksel Yeterlilikler

Matematiksel yeterlilikler 1990’ların sonunda Danimarka KOM Projesi ile Niss ve meslektaşları tarafından ileri sürülmüştür (Niss & Højgaard, 2011, s.13). Matematiksel yeterlilik bir bireyin matematiğin gerçekten bir rolü olan ya da potansiyel olarak rol oynayabileceği bağlamlarda ya da durumlarda bilgiyi temel alan bir tutumla uygun bir şekilde davranma kapasitesi ve hazırbulunuşluğudur (Niss& Jablonka, 2014, s.392). Matematiksel yeterliliğin yapı taşları olan ve matematiksel yeterlilikler olarak adlandırılan bileşenleri en önemli özelliklerdir. Bu yeterlilikler; matematiksel düşünme, problem kurma ve çözme, matematiksel modelleme, matematiksel muhakeme, matematiksel temsilleri idare etme, sembolleştirme ve formelleştirme ile ilgilenme ve matematiksel araç gereçleri kullanma olmak üzere sekiz kategoride toplanmaktadır. Matematiksel yeterlilikler ne öğrenenlerin matematiğine ne de öğretilecek matematiğe özellikle odaklanmamaktadır. Ayrıca kapasite, eğilim, tutum gibi kişisel özellikler de vurgulanmamaktadır.

Yeterlilik Şeması

Yeterlilik şeması ilk olarak Niss ve Danimarkalı meslektaşları tarafından tanımlanan bir takım yeterliliklerin (Niss, 1999) PISA Matematik Uzmanları Grubu (MEG) tarafından kullanılmasından doğmuştur (Turner, Dossey, Niss ve Blum, 2015, s.86). PISA uygulamalarının başlangıcında bu yeterliliklerin matematik okuryazarlığı için bir temel teşkil ettiği kabul edildiğinden, her PISA uygulamasının altını dolduran bir matematik okuryazarlığı çerçevesi söz konusudur. MEG daha sonra PISA matematik okuryazarlığı

maddelerinin bilişsel gereksinimleri boyutunu değerlendirmek ve öğrencilerin bu yeterliliklere sahip olma durumları ile matematiksel değerlendirme görevlerini başarılı bir şekilde tamamlayabilme becerileri arasındaki ilişkiyi daha iyi anlayabilmek amacıyla kullanılacak bir şema ortaya çıkarmaya çalışmıştır.

Yeterlilik şemasının geliştirilmesi ve iyileştirilmesine yönelik çalışmalar halen devam etmekte olup ilk yayınlandığı 2009 tarihinden bugüne kadar bazı yenileme çalışmaları yapılmıştır (Turner ve Adams, 2012; Turner vd., 2013; Turner vd., 2015). Yeterlilik şeması, Niss ve arkadaşları tarafından belirlenen ve PISA matematik okuryazarlığı çerçevelerinde yer verilen sekiz matematiksel yeterlilikten matematiksel düşünme ile matematiksel muhakemenin birleştirilmesi ve matematiksel yardımlar ve araçlar yeterliliğinin PISA matematik okuryazarlığı maddeleri için uygun olmaması gerekçesi ile çıkarılması sonucunda altı yeterlilik alanını içermektedir (Turner vd. 2015, s.90). Şemanın ilk versiyonunda (Turner, 2012) bu yeterlikler muhakeme ve argüman, iletişim, modelleme, temsil, problem çözme ile sembolik, formal, teknik dilin ve işlemlerin kullanımı şeklinde adlandırılmış, her bir yeterliliğin operasyonel tanımları verilmiş ve düzeyleri belirtilmiştir. Şemanın revize edilmiş versiyonunda (Turner vd. 2015, s.90) ise bu yeterliliklerin iletişim, strateji planlama, matematikleştirme, temsil, sembollerin, işlemlerin ve formal dilin kullanımı ile muhakeme ve argümantasyon olarak yeniden adlandırıldığı, operasyonel tanımları ve düzeylere ilişkin tanımlamalarında çeşitli değişikliklerin yapıldığı görülmektedir.

Yeterlilik şeması matematiksel problemlerin matematiksel yeterlilikler açısından niteliklerinin ölçülmesinde kullanılan her bir yeterliliğe ilişkin operasyonel tanımları, açıklamaları, yeterlilik düzeylerini ve bu düzeylere ilişkin tanımları içerdiğinden bir ölçme aracı olarak dereceli puanlama anahtarı türündedir. Puanlama anahtarları kullanıldıkları her alanda etkili bir şekilde düzeylendirme yapılan araçlar olarak övülse de puanlama anahtarının doğrudan kullanımının değerlendirmelerin geçerliğini ve güvenilirliğini arttırdığı yönündeki görüşü zorlayan çalışmalar bulunmaktadır (Rezaei ve Lovorn, 2010). Bu nedenle puanlama anahtarlarının geçerliğinin ve güvenilirliğinin incelenmesi gerekmektedir.

Ölçme Araçlarında Geçerlik ve Güvenirlik

Güvenirlik ve geçerlik ölçme araçlarında bulunması istenen temel özelliklerdir. Çeşitli ölçme araçlarıyla elde edilen verilerin geçerli ve güvenilir olması bu ölçme sonuçlarına dayalı olarak verilecek kararların isabetli olması için ön koşuldur.

Geçerlik

Özberk (2012), Cronbach ve ark. (1972)'nin bir ölçme aracının gerçekte ölçülmek istenen özelliği ölçtüğüne ilişkin birçok kanıtı ihtiyaç duyulduğunu belirttiğini ve geçerliğin ölçme aracı ile elde edilen ölçümlerden çıkarılabilecek yordamaları desteklemek için toplanan kanıtlar olarak tanımladığını ifade etmiştir. Bu açıdan bakıldığında bir ölçme aracı ile elde edilen ölçümlerin geçerliği için tek bir tanım ya da katsayının kullanılması yerine o ölçme aracının kullanım amacına hizmet ettiği yönündeki kanıtların toplanmasının daha uygun olabileceği söylenebilir. Bir geçerlik çalışması planlarken istenen çıkarım açıkça ifade edilmeli daha sonra bu çıkarımlar için ölçümlerin kullanılabilirliğine ilişkin kanıtların toplanmasına geçilmelidir (Crocker ve Algina, 2008, s.217). Genel bir geçerlik tanımı çerçevesinde ölçme aracının kullanılış amacına göre farklı geçerlik türlerinden söz edilir. Bunlardan bazıları kapsam geçerliği, yapı geçerliği, ölçüt geçerliği ve görünüş geçerliğidir.

Kapsam geçerliği bir ölçme aracının bu araçla ölçülmek istenilen davranışları ne derecede kapsadığı ile ilgilidir. Ölçme konusu olacak davranışların neler olduğunun belirlenmiş olması gerekir (Turgut ve Baykul, 2013, s.134). Soyut kavramlarla tanımlanan yetenek, tutum, güdü gibi soyut değişkenlerin içeriğini ve sınırlarını açık şekilde belirlemek güç olduğundan kapsam geçerliği daha çok başarı testleri için önemlidir (Büyüköztürk, 2012, s.167). Kapsam geçerliliğini test etmede kullanılan yollardan biri uzman görüşüne başvurmaktır. Uzmanların her bir sorunun geçerli olduğu noktasında uyuşma düzeylerinin %90-100 olması beklenirken, %70-80 oranında uyuşma gösterdikleri maddeler eleştirilene göre düzeltmeler yapılarak ölçekte tutulabilmektedir (Büyüköztürk, 2012, s.167).

Yapı geçerliği ölçme aracı ile ölçülmek istenen hipotetik yapının pratikte ölçülebilme derecesi ile ilgidir (Güler, 2008, s.27). Testin ölçülmek istenen davranış bağlamında soyut bir kavramı (faktörü) doğru bir şekilde ölçebilme derecesini gösterir (Büyüköztürk, 2012, s.168). İlgilenilen niteliği ölçmek amacıyla çok sayıda ölçülebilir, gözlenebilir sorular oluşturularak bu soruların ilgilenilen nitelikleri ne derece doğru ölçtüğü yapı geçerliği ile

belirlenebilir. Faktör analizi, küme analizi, iç tutarlılık analizi ve hipotez testi gibi tekniklerden yararlanılabilir (Büyüköztürk, 2012, s.168).

Ölçüt geçerliği ya da ölçüte dayalı geçerlik ölçme sonuçlarının ikinci bir ölçütle arasındaki ilişkinin derecesi ile ilgilidir (Güler, 2008 s. 13). Bazı durumlarda bir ölçme aracının geçerliği yeterli güvenirlik ve geçerlikte olduğu bilinen başka bir ölçme aracıyla korelasyonuna bakılarak hesaplanmaktadır. Yeterli güvenirlik ve geçerlikte olduğu bilinen bir ölçme aracına ölçüt, geçerliği saptanacak olana yordayıcı denir. Ölçütün geçerlik ve güvenirliğinden emin isek, yordayıcının güvenirliği yüksek, ikisi arasındaki korelasyon düşük ise yordayıcının geçerliğinin düşük olduğu söylenebilir (Turgut ve Baykul, 2013, s.138). Crocker ve Algina (2008) ölçüte dayalı geçerlilik çalışmalarının genellikle beş aşamayı izlediğini belirtmişlerdir.

1. Uygun bir ölçüt davranış ve onu ölçecek bir yöntem tanımlama
2. Testin nihai olarak kullanılacağı kişileri temsil eden uygun bir örneklem tanımlama
3. Testi uygulama ve her bir bireyin puanlarını kaydetme
4. Ölçüt verisi elde edildiğinde her bir birey için ölçütle ilgili performans ölçümünü elde etme
5. Ölçüt performans ile test puanları arasındaki ilişkinin gücünü belirleme

Görünüş geçerliği ise testin ismi, açıklamalar, soruların ve testin düzeni gibi testin genel olarak geçerliğine ilişkin araştırmacının görüş sahibi olmasını sağlar (Büyüköztürk, 2012, s.169). Olumlu izlenimler testin sorularına güvenilir cevaplar almayı kolaylaştıracaktır. Testin görünümü ile ilgili olan bu geçerlik türü uzman görüşü ile değerlendirilir.

Güvenirlik

Güvenirlik, bir ölçme aracında bulunması gereken iki temel özellikten biridir. Sosyal bilimlerde yapılan ölçme çalışmalarında ölçme sonuçlarına farklı kaynaklardan gelen hataların karışması nedeniyle ölçme sonuçlarının tam olarak doğru bir şekilde elde edilememesi söz konusudur. Ölçmeye karışan tesadüfi hatalar ölçme araçlarının güvenirliğini düşürmektedir. Bu bağlamda güvenirlik, ölçülen bir özelliğin hatasız ve doğru bir şekilde ölçülebilme derecesi olarak ölçme sonuçlarının tesadüfi hatalardan arınık olması demektir (Tan, 2013, s.133).

Güvenirlik hassasiyet/duyarlılık, tutarlılık ve kararlılık olmak üzere üç farklı anlamda ele alınmaktadır (Tan, 2013, s.133). Duyarlılık ölçme aracının daha fazla ayrıntıyı ölçebilecek nitelikte olmasıyken, iç tutarlılık aynı özelliği ölçen birden çok test maddesinden elde edilen ölçümlerin tutarlı olmasıdır. Kararlılık ise aynı niteliğin tekrar tekrar ölçülmesi sonucunda elde edilen ölçümlerin kararlı olması anlamındadır. Uzun sürede değişmeyen yapıdaki değişkenlerin belli aralıklarla tekrar tekrar ölçülmesi sonucu elde edilen sonuçlar benzer olduğunda ölçümlerin kararlı olduğu yorumu yapılır.

Bir ölçme aracının uygulanış biçimine, puanlamasında kullanılan yönteme, ölçülen değişkenin tek veya çok boyutlu oluşuna ve uygulanan ölçme teorisine göre kullanılan güvenilirlik kestirme yöntemleri değişmektedir (Büyüköztürk, 2012, s.170). Bununla birlikte ölçme sonuçlarına karışan hata kaynaklarından hangisine odaklanıldığına da bağlı olarak farklı güvenilirlik kestirme yöntemleri uygulanabilmektedir. Ölçmelerde ele alınan hata kaynaklarının genellikle madde, zaman, test formu ve puanlayıcı olduğu görülmektedir. Puanlayıcı hata kaynağının önemli bir yere sahip olduğu ve performansın ölçülmesinde nesnellik sağlaması bakımından kullanılması önerilen puanlama anahtarlarıyla elde edilen ölçümlerin güvenilirliğinin incelenmesinde farklı kuramlar kapsamında farklı yöntemler kullanılabilmektedir.

Klasik Test Kuramı, Genellenebilirlik Kuramı ve Madde Tepki Kuramı hata kaynaklarını farklı yaklaşımlarla ele almakta ve güvenilirlik hesaplama yöntemleri değişiklik göstermektedir. Ölçme araçlarının güvenilirliklerinin belirlenmesinde en çok kullanılan yaklaşım Klasik Test Kuramı'na dayalı güvenilirlik katsayılarının incelenmesidir. Bununla birlikte performans değerlendirmenin giderek artan bir şekilde kullanımı ile farklı hata kaynaklarının birlikte ele alınmasına fırsat veren Genellenebilirlik Kuramı'na dayalı güvenilirlik katsayılarının temel alındığı çalışmalara da sıklıkla rastlanmaktadır.

Klasik Test Kuramı

Klasik test kuramında güvenilirliğin açıklanması genellikle gerçek puan modeline dayandırılmaktadır. Gerçek puan modeline göre ölçme sonuçlarında yanlışlığın, sistematik hatanın ya da sabit hatanın olmadığı varsayılarak gözlenen puanlar (X), gerçek puanlar (T) ile rastgele hataların (E) toplamına eşittir.

$$X = T + E \quad (1)$$

Modele göre bir bireyin gerçek puanı T , o bireyin aynı testi birçok defa bağımsız olarak alması sonucunda elde edildiği varsayılan puanların ortalaması olarak kabul edilir. Bireylerin gözlenen puanları bir durumdan diğerine değişiklik gösterse de gerçek puanları sabittir ve bu durum sistematik olmayan ölçme hatasından kaynaklanmaktadır (Webb, Shavelson ve Hartel, 2006, s.81). Bu model ölçme hatasının gerçek puanla korelasyonun sıfır olduğu ve hatanın tesadüfi olduğu varsayımına dayanmakta olup tesadüfi hataların toplamının beklenen değeri sıfırdır (Tan, 2013, s.134). Gerçek puan modelinden yola çıkılarak oluşturulan Eşitlik 2’de benzer şekilde gözlenen puanların varyansının gerçek puanların varyansı ile hata varyansının toplamına eşit olduğu kabul edilmiştir.

$$\sigma_x^2 = \sigma_T^2 + \sigma_E^2 \quad (2)$$

Crocker ve Algina (2008)’in gerçek puanların ve hata puanlarının temel özellikleri olarak ileri sürdüğü ve modelin varsayımları olarak nitelendirdiği üç temel bileşen söz konusudur (s.111). Bunlar;

1. Bireyler evreni için hata puanlarının ortalaması sıfırdır ($\mu_E = 0$).
2. Bireyler evreni için gerçek puanlar ile hata puanları arasındaki korelasyon sıfırdır ($\rho_{TE}=.00$).
3. Bireyler iki farklı test aldığı ve her bir bireyin iki testteki puanlarının olası gözlenen puanların bağımsız iki dağılımından random olarak seçildiği varsayıldığında, iki puanlamadaki hata puanları arasındaki korelasyon sıfırdır ($\rho_{E_1E_2}=.00$).

Bu varsayımlar gerçek puan modelini test puanlarının güvenilirliğine uygulanabilmesine olanak tanımaktadır. Bu bağlamda klasik test kuramında güvenilirlik, gerçek puanlarla gözlenen puanlar arasındaki korelasyonun karesi olan güvenilirlik katsayısının gerçek puan varyansının gözlenen puan varyansına oranı olarak tanımlanabilmektedir (Crocker ve Algina, 2008, s.115).

$$\rho_{xx} = \frac{\sigma_T^2}{\sigma_x^2} \quad (3)$$

Eşitlik 2’de görüldüğü gibi gözlenen puan varyansı, sistematik olduğu düşünülen gerçek puan varyansı ile random olduğu düşünülen hata puanı varyansından oluşmaktadır.

Güvenirlilik katsayısı hatanın bağlı olduğu kaynağa göre farklılık gösterebilir ve farklı isimlerle anılmaktadır.

Klasik test kuramına göre bir gözleme ilişkin olarak iç tutarlılık güvenilirliği, test tekrar test güvenilirliği, paralel formlar güvenilirliği, puanlayıcılar arası güvenirlilik gibi birden çok güvenirlilik katsayısı hesaplanabilir. İç tutarlılık anlamında güvenirlilik ölçme aracındaki maddelerden elde edilen puanların tutarlılığı olup, buradaki hata kaynağı maddelerdir. Test tekrar test güvenirliliğinde bir testin aynı koşullarda farklı zamanlarda uygulanmasından elde edilen sonuçların birbiriyle tutarlı olması beklenmektedir. Bu yöntem ile hesaplanan güvenirlilik ölçme aracının farklı zamanlarda kararlı ölçmeler yapabilme derecesi olarak ele alınmakta olup hata kaynağı zamandır. Paralel formlar güvenirliliğinde aynı ölçme için iki paralel formun uygulanması sonucu elde edilen puanların tutarlılığına bakılmakta olup buradaki hata kaynağı ölçme aracıdır. Puanlayıcılar arası güvenirlilikte ise birden fazla puanlayıcının aynı özelliğe verdikleri puanlar arasındaki tutarlılığa bakılmakta olup buradaki hata kaynağı puanlayıcılarıdır.

Bu güvenirlilik katsayılarından biri yüksek çıkarken bir diğeri düşük çıkabilir. Bu durum klasik test kuramının sınırlılıklarından biri olup, her bir kaynaktan gelen hatanın ayrı ayrı ele alınmasından kaynaklanmaktadır (Crocker ve Algina, 2008, s.157). Alternatif ölçümlerin özelliği olan çoklu hata kaynaklarının modellenmesinde klasik test kuramına dayalı güvenirlilik ölçümleri yetersiz kalmaktadır (Taylor ve Pastor, 2013).

Bir ölçme aracının uygulanış biçimine, puanlamasında kullanılan yöntem, ölçülen değişkenin tek veya çok boyutlu oluşuna ve uygulanan ölçme teorisine göre kullanılan güvenirlilik kestirme yöntemi değişmektedir. Test tekrar test güvenirliliği, paralel formlar güvenirliliği, iki yarı test güvenirliliği Kuder Richardson-20 (KR-20) ve Cronbach alfa güvenirliliği, madde toplam puan korelasyonu, madde ortalama puanları arasındaki farkların ilişkisiz grup t-testi ile sınanması ve ölçmenin standart hatası güvenirlilik kestirim yöntemlerinden bazılarıdır (Büyüköztürk, 2012, s.170). Bu çalışmada Klasik test kuramı kapsamında ele alınan güvenirlilik kestirim yöntemleri aşağıda açıklanmıştır. Puanlama anahtarlarının etkililiğine ilişkin önemli eğilimlerden biri puanlayıcılar arası ve puanlayıcı içi güvenirliliği merkeze almaktadır (Jonsson ve Svingby, 2007; Rezaei ve Lovorn, 2010).

Puanlayıcı İçi Güvenirlik

Bir puanlayıcının farklı zamanlarda aynı ölçme objesinin niteliklerine ilişkin verdiği puanlardaki tutarlılığı, kararlılığının göstergesidir. Puanlayıcıdan ya da puanlama anahtarından kaynaklanan bazı koşullar puanlamalar arasında uyumsuzluk oluşmasına neden olabilir. Puanlayıcı içi güvenirliliğin hesaplanması test tekrar test güvenirliliğinde olduğu gibi iki farklı zamanda verilen puanlar arasındaki korelasyonla hesaplanabilir. Test tekrar test güvenirliliğinde testin zamana karşı kararlılığı anlamında güvenirlilik kestirilirken, puanlayıcı içi güvenirlilikte puanlayıcının puanlamadaki tutarlılığı anlamındaki güvenirlilik kestirilir (Güler, 2008, s.17). Bununla birlikte puanlamalar arasındaki tam uyumu yansıtan basit yüzde hesaplama, farklı zamanlardaki puanlamaların ortalamaları arasındaki farkların test edilmesi gibi yöntemlerde kullanılabilir.

Basit yüzde puanlayıcının aynı puanları verdiği nesne sayısının puan verdiği tüm nesnelerin sayısına oranı ile kolayca hesaplanabilen bir teknik olup tutarlık anlamında güvenirliliğin bir ölçüsüdür. Bu yöntem sınıflama, sıralama, eşit aralıklı ve eşit oranlı ölçekler ile elde edilen verilerin tümüne uygulanabilmesine karşın şansa ya da tesadüfi olarak ortaya çıkan uyumu hesaplayamamaktadır (Goodwin, 2001). Puanlayıcı içi güvenirliliğin belirlenmesinde genellikle Pearson momentler çarpımı korelasyon katsayısından yararlanılmaktadır. Bununla birlikte Pearson korelasyon katsayısı aralarındaki ilişki belirlenecek değişkenlerin normal dağılım göstermesi varsayımına dayalıdır (Field, 2009, s.177).

Puanlayıcı içi güvenirliliğin belirlenmesinde normal dağılım varsayımının sağlanmadığı durumlarda nonparametrik bir istatistik olan (Field, 2009, s.179) Spearman rho korelasyon katsayısından yararlanılmaktadır. Ancak puanlayıcı içi güvenirliliğe ilişkin korelasyon katsayısı sunulurken, puanlayıcıların farklı zamanlardaki puanlamalarının ortalamaları arasındaki farkın anlamlılığı ilişkili örneklem t testi ya da varyans analizi ile test edilerek korelasyon katsayısı ile birlikte yorumlanması önerilmektedir (Goodwin, 2001).

Puanlayıcılar Arası Güvenirlik

Puanlayıcılar arası güvenirlilik birden fazla puanlayıcının birbirinden bağımsız olarak aynı ölçme objesinin niteliklerine ilişkin verdiği puanlardaki uyumun göstergesidir. Puanlayıcılar arası güvenirlilik genellikle iki puanlayıcının bir grup öğrencinin performansına verdikleri puanlar arasındaki korelasyon ile ölçülmektedir (Güler, 2008, s.18; Rezai ve Lovorn, 2010). Ancak puanlayıcılar arası güvenirliliğin belirlenmesinde korelasyon katsayısının tek başına

kullanımı yeterli görülmemektedir (Goodwin, 2001). Bu nedenle korelasyon katsayısı ile birlikte ortalamalar arasındaki farkların da test edilmesi ve birlikte yorumlanması önerilmiştir. Puanlayıcılar arası güvenilirliğin incelenmesinde iki puanlayıcı söz konusu olduğunda, puanlayıcılar arasındaki kesin uyumun yüzdesi, Cohen'in kappası, Scott'un pi katsayısı, korelasyon katsayısı, Kendall'ın uyuşum katsayısı ve Krippendorff α gibi güvenilirlik katsayıları kestirilebilir. Puanlayıcılar arası güvenilirliğin incelenmesinde ikiden fazla puanlayıcı söz konusu olduğundaysa puanlayıcılar arasındaki kesin uyumun yüzdesi, sınıf içi korelasyon katsayısı, Fleiss'in kappası, Kendall'ın uyuşum katsayısı ve Krippendorff α gibi güvenilirlik katsayıları kestirilebilir.

Puanlayıcılar arasındaki kesin uyumun yüzdesi basit yüzde yöntemi ile kolayca hesaplanabilmesine karşın şansla ortaya çıkabilecek uyumu göz ardı etmektedir. Cohen'in kappası basit yüzde yönteminin bir sınırlılığı olan şansla ortaya çıkabilecek uyumun düzeltilmesini içeren şans başarısı üzerindeki uyumun kestirimine olanak tanımaktadır. Ancak ikiden fazla puanlayıcının olması durumunda kullanılamamaktadır. Fleiss, Cohen's kappasını puanlayıcı sayısının ikiden fazla olduğu durumlar için genişletmiştir. Fleiss'in kappası tüm puanlayıcıların şansla uyumunun ötesinde gözlenen uyumlarının derecesini vermektedir. İkiden fazla puanlayıcı söz konusu iken yararlanılabilecek yöntemlerden biri de sınıf içi korelasyon katsayısıdır. Sınıf içi korelasyon katsayısı homojenliği ölçmek için kullanılan bir teknik olarak varyans analizine dayanmakta ve farklı modeller oluşturularak kestirilebilmektedir.

Puanlayıcılar arası güvenilirliğin kestirilmesinde parametrik olmayan bir istatistiksel teknik olarak sıklıkla kullanılan yöntemlerden biri de Kendall'ın uyuşum katsayısı olarak bilinen Kendall'ın W katsayısıdır. Kendall'ın W katsayısı, Spearman'ın sıra farkları korelasyon katsayısına (rank order correlation coefficient) oldukça benzemektedir (Mehta ve Patel, 2011, s.181). Friedman testi ile aynı sonuçları vermesine ek olarak bu testin Friedman testinden farkı bir uyum ölçüsü olarak yorumlanabilmesidir (Mehta ve Patel, 2011, s.181). Bu katsayı en az sıralama ölçeği ile elde edilmiş verilere uygulanabilmektedir (Karagöz, 2010; Kılıç, 2009, s.27;). Bu katsayısı N bağımsız puanlayıcının her birinin K adayın her birine atadığı sıralamalar üzerinden hesaplanmakta ve N puanlayıcının K adaya atadığı sıralamalar arasındaki uyumu ölçmektedir (Kendall ve Babington-Smith, 1939'dan aktaran Mehta ve Patel, 2011, s.108). Kendall'ın W katsayısı aşağıdaki formülle hesaplanır (Howell, 2013, s.320).

T_j ; her bir nesnenin farklı puanlayıcıdan aldığı sıralamaların toplamı

N ; sıralanacak nesne sayısı

k ; puanlayıcı sayısı olmak üzere

$$W = \frac{12 \sum T_j^2}{k^2 N(N^2 - 1)} - \frac{3(N + 1)}{N - 1}$$

Kendall'ın W katsayısı da çoğu güvenilirlik katsayısı gibi 0 ile 1 arasında değer almakta, katsayının değeri 1 iken puanlayıcıların sıralamaları arasında mükemmel uyumun var olduğu 0 iken hiç uyumun olamadığı şeklinde yorumlanmaktadır. Puanlayıcılar arasındaki uzlaşmazlık puanlayıcıların adayları farklı şekilde sıraladıklarını göstermektedir.

Alanyazında, Kendall'ın W katsayısının yorumlanmasında bazı çalışmalarda farklılıklar olduğu görülmüştür. Puanlayıcılar arası uyumun düzeyini belirlemede Kendall'ın W katsayısının alacağı 0,1; 0,3; 0,5; 0,7; 0,9 değerlerinin sırasıyla çok zayıf, zayıf, orta düzeyde, güçlü ve kesinlikle güçlü uyuma karşılık geldiği belirtilmiştir (Ergün, 1995'ten aktaran; Kurnaz ve Kutlu, 2016). Diğer yandan Deliceoğlu (2009)'un Szymanski ve Linkowski, (1993)'ten aktardığına göre 0,80 ve üstündeki değerlerin kabul edilebilir uyuma karşılık geldiği belirtilmiştir (s.40). Bununla birlikte puanlayıcılar arası uyumun Kendall'ın W katsayısı ile araştırıldığı araştırmaların büyük çoğunluğunda katsayının değerine göre yorum yapmak yerine asimptotik p değerine bakılarak puanlayıcıların sıralamaları arasında uyum var ya da uyum yok şeklinde yorum yapılmıştır.

Kendall'ın uyum katsayısı puanlayıcılar arası uyumun olmadığı varsayılan yokluk hipotezi ile test edilmektedir. Asimptotik p değerinin 0,05'den küçük olması durumunda yokluk hipotezi reddedilir ve puanlamalar arası uyum olduğu şeklinde yorumlanırken, 0,05'den büyük olması durumunda yokluk hipotezi kabul edilerek puanlayıcılar arasında uyum olmadığı yorumu yapılır (Mehta ve Patel, 2011, s.100). Bununla birlikte puanlayıcı sayısının küçük olması nedeniyle bu çıkarımın test edilmesi gerekmektedir. Bu amaçla gerçek p değeri için orta sıraların özgün çift yönlü düzeninin 10.000 random permütasyona dayalı olarak Monte Carlo kestirimine başvurulabilir. Puanlayıcılar arasındaki uyum gerçek p değerinin %99 güven aralığında alabileceği değerler incelenerek ve 0,05 düzeyinin aşılp aşılmadığına bakılarak yorumlanmalıdır.

Ölçme araçlarının güvenilirliklerinin belirlenmesinde en çok kullanılan yaklaşım Klasik Test Kuramı'na dayalı güvenilirlik katsayılarının incelenmesi olmasına karşın ölçmeye karışan her

bir hata için farklı güvenilirlik katsayılarının hesaplanması gerektiği ve bu katsayılar arasında tutarsızlıklar olabileceği bilinmektedir. Performans değerlendirmenin giderek artan bir şekilde kullanımı ile farklı hata kaynaklarının birlikte ele alınmasına fırsat veren Genellenebilirlik Kuramı'na dayalı güvenilirlik katsayılarının temel alındığı çalışmalara sıklıkla rastlanmaktadır.

Genellenebilirlik Kuramı

Klasik test kuramının farklı yöntemleri ile farklı anlamlarda elde edilen güvenilirlik katsayılarının birbirinden farklı olmasından hareketle Cronbach ve arkadaşları Genellenebilirlik (G) kuramını ortaya atmışlardır (Atılğan, 2005). G kuramı ölçümlerin birden fazla hata kaynağına bağlı olduğu durumlarda ölçme duyarlılığını kestirmeye yönelik bir yaklaşımdır (Cardinet, Johnson ve Pini, 2010, s.1). Bu kuram araştırmacılara, hata kaynaklarını tanımlama ve her bir hata kaynağının göreceli katkısını analiz etme imkânı vermektedir (Taylor ve Pastor, 2013). Webb ve Shavelson (2005, s.717)' e göre bu kuramla hata kaynaklarının gerçek nedenleri tanımlanır, hata kaynakları ayırt edilir ve her bir hata kaynağı tahmin edilir. G kuramı puanlayıcı, zaman, test formu... vs. gibi bütün potansiyel hata kaynaklarını eş zamanlı olarak değerlendirir (Atılğan, 2005).

Klasik test kuramı gözlenen puan varyansını gerçek puan varyansı olarak adlandırılan sistematik ve randum hata varyansı olarak adlandırılan iki parçaya ayırır. Hata tahminleri ve güvenilirlik kestirimleri veri toplama desenine bağlı olarak değişmektedir. Hata varyansı tek parçalı bir yapı değildir, hata birden fazla kaynaktan ortaya çıkar. Örneğin, testin uygulanma zamanı, puanlayıcılar ve test maddeleri gibi kaynakların her birinden hata karışır. Test tekrar test güvenirlğinde performansın günden güne değişimi hata olarak ele alınırken madde örneklemeden gelen değişim dikkate alınmaz. İç tutarlılık katsayısı madde örneklemesinden kaynaklanan değişimi hata olarak ele alırken günden güne değişimi dikkate almaz. G kuramı klasik test kuramının aksine, gerçek ve hata puanlarının birden çok tanımı olabileceğini kabul eder (Webb, Shavelson ve Shalvey, 1989). Bu kuram davranış ölçmede güvenilirliğin değerlendirilmesini, güvenilir gözlemlerin tasarlanmasını, araştırılmasını, kavramsallaştırılmasını sağlayan istatistiksel bir kuram olup varyans analizine (ANOVA) dayalıdır (Atılğan, 2005).

Genellenebilirlik kuramıyla yapılan çalışmalarda tasarlanan desenlerin açıklanmasında, değişkenlik kaynağı, yüzey, koşul, kabul edilebilir gözlemler evreni, genelleme evreni ve

evren puanı kavramları ortaya çıkmaktadır. Bu kavramlara dayalı olarak çeşitli desenler geliştirildiği için öncelikle bu kavramların tanımlanmasının uygun olduğu düşünülmüştür.

Temel Kavramlar

Daha öncede belirtildiği gibi G kuramı ile ölçümlere puanlayıcı, zaman, test formu gibi farklı kaynaklardan karışan hatalar eş zamanlı ve bir bütün olarak hesaplanabilmektedir. Değişkenlik kaynağı, ölçmede potansiyel hata kaynağı olan ölçme sürecinin karakteristikleridir (Eser ve Gelbal, 2013). Puanlayıcı, zaman, test formu gibi hata kaynaklarının her biri bir değişkenlik kaynağı olarak ele alınır. Koşul kavramı ise değişkenlik kaynağının düzeyini diğer bir deyişle içerdiği düzey sayısını ifade eder. Örneğin üç puanlayıcının yer alacağı bir çalışmada puanlayıcılar bir değişkenlik kaynağı olup, koşulu üçtür.

Kabul edilebilir gözlemler evreni, eldeki örneklemin yerine geçebilecek olası gözlemler evrenidir. Diğer bir deyişle değişkenlik kaynaklarının bütün olası koşulları kabul edilebilir gözlemler evreni olarak tanımlanır (Atılğan, 2005). Genelleme evreni ise genellenmek istenilen bir değişkenlik kaynağının koşulları olarak ele alınmaktadır. Başka bir ifadeyle genelleme evreni bir K çalışmasının planlanmasında araştırmacının sonuçlarını genellemek istediği evrendir (Webb ve Shavelson, 2005, s.718). Genellenebilirlik kuramında adayın evren puanı genelleme evrenindeki ölçmelerin ortalaması olarak tanımlanır (Crocker ve Algina, 2008, s.159).

Genellenebilirlik kuramıyla yapılan çalışmalarda farklı desenler kullanılabilmektedir. G ve K çalışmaları, değişkenlik kaynağının tesadüfi (random) ya da sabit (fixed) olması, verilerin toplanmasında desenin çapraz (cross) ya da yuvalanmış (nested) olması, sonuçların yorumlanmasının görelî ya da mutlak olması ve genellenebilirlik (G) katsayısının ya da güvenirlik (Phi) kat sayısının hesaplanması mümkündür (Güler ve Gelbal, 2010).

Bir çalışmada yer alan bir değişkenlik kaynağının tüm yüzeyleri diğer değişkenlik kaynaklarının tüm yüzeylerinde bulunuyorsa bu çalışma deseni tümüyle çaprazlanmış desendir. Çaprazlanmış desenler değişkenlik kaynaklarının aralarına “x” işareti konularak temsil edilmektedir. Çalışmada yer alan bir değişkenlik kaynağının bazı yüzeyleri diğer değişkenlik kaynaklarının bazı yüzeylerinde bulunuyorsa bu çalışma deseni yuvalanmış desendir (Güler, 2011). Yuvalanmış desenlerin temsilinde ise değişkenlik kaynaklarının aralarında “:” işareti bulunmaktadır. Örneğin 5 puanlayıcının 10 bireyin 20 göreve verdiği

yanıtları puanlamasıyla elde edilen veriler için tümüyle çaprazlanmış desen söz konusu olduğunda tüm puanlayıcıların tüm bireylerin tüm görevlere verdiği yanıtları puanlaması söz konusudur ve desen $b \times m \times p$ şeklinde ifade edilir. Aynı ölçme durumu için tasarlanabilecek yuvalanmış desen bireylerin farklı görevleri yanıtlaması ve farklı puanlayıcıların farklı öğrencilerin yanıtlarını puanlaması durumunda söz konusudur ve desen $b:p:m$ şeklinde ifade edilir. Bu desenlere ek olarak bazı değişkenlik kaynaklarının çaprazlanıp bazılarının yuvalandığı desenlerde kullanılabilmektedir. Bu desenler ise karışık desenler olarak adlandırılmıştır. Çalışmalarda olası tüm değişkenlik kaynaklarından gelen ve bunlar arasındaki etkileşimden kaynaklanan varyansların açıklanabilmesine olanak verdiği için tümüyle çaprazlanmış desenlerin kullanımı önerilmektedir (Briesch, Swaminathan, Welsh ve Chafouleas, 2014).

G kuramı bir örneklemden elde edilen ölçüm sonuçlarının daha geniş ölçümlere genellenme derecesini veren tekniklerle ilgilidir ve elde edilen ölçmeler, ait oldukları gözlemler evrenine genellenmek istenir (Crocker ve Algina, 2008, s.158). Çoklu hata kaynakları temel bir kavram olan genelleme evrenini tanımlar. Bir kişiden elde edilen bir ölçüm ilgili davranışın bir örneklemini temsil eder (Webb vd, 1989; Briesch vd., 2014). Bir ölçümün kullanışlı olması davranışın daha geniş bir durumlar setine (diğer bir deyişle genelleme evrenine) genellenebilmesine izin verme derecesine bağlıdır. Klasik test kuramındaki güvenilirlik kavramı ile daha kapsamlı ve esnek olan genellenebilirlik kavramı yer değiştirir. Gözlenen puanların gerçek puanları ne derece doğru yansıttığını sorgulamak yerine G kuramı gözlenen puanların tanımlanmış bir durumlar evreni içinde kişilerin davranışları hakkında ne derece doğru genelleme yapmaya izin verdiği ile ilgilidir. Böylelikle güvenilirlik sorusu genellemelerin doğruluğu ya da genellenebilirlik sorusuna dönüşmüştür.(Cronbach et al., 1972, p. 15'den aktaran Webb vd., 1989). Genellenebilirlik kuramında güvenilirlik iki tür çalışma ile incelenebilir. Bu çalışmalarla ilgili bilgilere aşağıda yer verilmiştir.

Genellenebilirlik ve Karar Çalışmaları

Genellenebilirlik kuramıyla iki tür çalışma yürütülebilir. Bunlar G çalışması ve K çalışmasıdır. Yapılacak çalışmanın G çalışması mı K çalışması mı olacağını araştırmacının amacı belirlemektedir. Eğer amaç, ölçüm sonuçlarının ölçme evrenine genellenmesi ise bu çalışma G çalışmasıdır. Eğer araştırmacı yapacağı çalışmanın sonucunda mutlak, ölçüte dayalı bir karar almak istiyorsa (işe alma, programın etkililiğine karar verme, seçme ve

yerleştirme gibi kararlar) K çalışması yürütmelidir. K çalışması yapılmadan önce K çalışmasında kullanılacak desenin belirlenmesi için öncelikle G çalışması yürütülmelidir (Crocker ve Algina, 2008, s.158).

G-çalışmaları, ölçmedeki çeşitli varyans kaynakları hakkında mümkün olduğunca bilgi vermeli, potansiyel varyans kaynaklarını tanıtmalı ve içermelidir (Shavelson ve Webb, 1991). G çalışmaları ölçme desenine karar verilmesine, ölçme araçlarının geliştirilmesine ve ilerideki kullanımları için değişkenlik kaynaklarından gelen hataların azaltılmasına imkân vermektedir (Atılğan, 2005). K çalışmaları ise G çalışmalarının üzerine kuruludur. G çalışmalarıyla elde edilen bilgilerden yararlanılarak ele alınan ölçme durumunda güvenilirliğin sağlanması için en uygun senaryoların belirlenmesi amacına hizmet etmektedir (Brennan, 2001; Shavelson ve Webb, 1991).

G ve K çalışmaları ile ölçümlerin güvenilirliğine ilişkin iki farklı güvenilirlik katsayısı elde edilebilir. Bunlar göreceli, norma dayalı olan G katsayısı ve mutlak, ölçüte dayalı Phi katsayısıdır (Webb ve Shavelson, 2005, s.717). Çalışma sonucunda verilecek kararın niteliğine göre bu güvenilirlik katsayılarından uygun olanı seçilerek yorumlanmaktadır. Göreceli karar (G), bireylerin sıralamadaki yerine odaklanırken, mutlak karar (Phi) sıralamadan bağımsız olarak performans düzeyine odaklanmaktadır (Webb ve Shavelson, 2005, s.717).

G kuramının iki değişkenlik kaynaklı çaprazlanmış tesadüfi etki deseni kapsamında G çalışması ile kestirilen varyans bileşenleri aşağıda bir örnek üzerinden incelenmiştir. G kuramı ile yapılan araştırmalar de genellikle bireylerin ölçme objesi olarak ele alındığı ve değişkenlik kaynağı olarak adlandırılmadığından birey madde ve puanlayıcılardan oluşan bir desen iki yüzeyli olarak adlandırılmaktadır. G kuramının bmxp desende bir bireyin (b), bir maddeden (m) ve bir puanlayıcıdan (p) elde edilen gözlenen puanı aşağıdaki gibidir.

$$X_{bmp} =$$

μ	genel ortalama
$+ \mu_b - \mu$	öğrenci etkisi
$+ \mu_m - \mu$	madde etkisi
$+ \mu_p - \mu$	puanlayıcı etkisi
$+ \mu_{bm} - \mu_b - \mu_m + \mu$	öğrenci x madde ortak etkisi
$+ \mu_{bp} - \mu_b - \mu_p + \mu$	öğrenci x puanlayıcı ortak etkisi

$$+ \mu_{mp} - \mu_m - \mu_p + \mu$$

.....madde x puanlayıcı ortak etkisi

$$+ X_{bmp} - \mu_b - \mu_m - \mu_p + \mu_{bm} + \mu_{bp} + \mu_{mp} - \mu$$

.....artık etkisi

Çok Değişkenli G Çalışmaları

Bir testin alt testlerden oluşması ya da bir ölçme aracının alt boyutlarının olduğu durumlarda ölçme araçlarının güvenirliklerini incelemeye kullanılan yaklaşımlardan biri çok değişkenli G çalışmalarıdır. Çok değişkenli G çalışmaları ile birden fazla boyuttan oluşan ölçme araçlarının alt boyutlarına ve tamamına (bileşik test) ilişkin G ve Phi katsayıları kestirilebilmektedir (Brennan, 2001). Bununla birlikte bir ölçme aracının alt boyutlardan oluştuğu durumlarda toplam puanlar ile ilgileniliyorsa tek değişkenli G çalışmalarının yürütülmesi de mümkündür. Ancak alt boyutlar göz ardı edilerek tek değişkenli G çalışması ile hesaplanacak bileşik teste ilişkin güvenirlik sınırlayıcı olabilmektedir.

Davranışsal Ölçümlerin Simetrisi

Davranışsal ölçümlerin amacı genellikle bireyleri farklılaştırmak ve bazı konu alanları ve becerilerdeki bireylerin puanlarını derecelendirmektir. KTK'da geleneksel bir şekilde bireyler ölçme objesi olarak ele alınır ve bir ölçümün bireyleri ne kadar iyi ayırt ettiğinin uygunluğu yorumlanır. Diğer tüm yüzeylere hata kaynakları olarak bakılır. Cardinet vd. (1976) ölçme objesinin değişebileceğini ileri sürmüştür. Özellikle eğitimsel araştırmalarda değişimin bir göstergesi olarak belli bir test maddesi üzerindeki başarı ile ilgilenilebilir. Alternatif olarak bir sınıftaki her bir öğrencinin test puanını elde etmiş olsak bile her bir öğrenci yerine şubelere odaklanılabilir. Bu durumlarda bireyler ölçme objesi olmadığı için birer hata kaynağı olarak ele alınır. Cardinet, Tourneur, & Allal (1976) ölçme objesinin değişebileceği durumlar için simetri ilkesini ileri sürmüştür. Bu ilke bir desende her hangi bir yüzeyin ölçme objesi olarak ele alınabileceğini varsaymaktadır. Veriler üzerinden okulların, sınıfların ya da öğretimsel amaçların ne kadar iyi ayırt edildiği araştırılmak istenebilir. Bu tür sorular G kuramı kapsamında ele alınabilir. Bu tür durumlarda bireyler arasındaki farklılıklar ölçme objesi olarak ele alınmak yerine araştırılan farklılığı etkilemektedir Bu nedenle hata olarak görülür. Bu yaklaşımın dayanağı bir veri setinde varyans bileşenlerinin ölçme deseninden bağımsız olarak hesaplanabilmesidir (Webb, vd., 1989).

İlgili Araştırmalar

Bu bölümde çalışma ile ilgili olduğu düşünülen alanyazındaki daha önceki çalışmalar özetlenmiş ve kronolojik olarak sunulmuştur.

Yeterlilik Şeması ile İlgili Araştırmalar

Turner vd. (2013) yaptıkları araştırmada Yeterlilik Şeması'nın ilk versiyonunu oluşturmuşlar ve geçerlik ile güvenilirliğini incelemişlerdir. Çalışmada ilk olarak yeterliliklere ilişkin operasyonel tanımlar ve her bir yeterliliğin etkinleşme gereksinimine ilişkin dört düzeye ait açıklamalar oluşturulmuştur. Daha sonra PISA matematik okuryazarlığı maddeleri gereksinim düzeyleri açısından değerlendirilmiştir. Çalışmada PISA 2003 ve 2006 uygulamalarında yer alan 48 maddeye ilişkin Matematik Uzmanları Grubu'nda yer alan sekiz puanlayıcıdan toplanan veriler yeterlilik puan ortalamaları arasındaki korelasyonlar, puanlayıcı güvenilirliği, açıklanan varyans oranı, puanlamalar arası güvenirlik ve faktör yapısı açısından analiz edilmiştir. Puanlayıcı içi güvenirlik iki yıl ara ile dört puanlayıcıdan toplanan puanlamalar üzerinden korelasyon ile hesaplandığında korelasyonların 0,37 ile 0,91 arasında değiştiği ve 0,05 düzeyinde anlamlı olduğu görülmüştür. Puanlayıcı içi güvenirlik bağlamında sekiz puanlayıcının 48 maddeye verdikleri puanlar üzerinden her bir yeterlilik için Cronbach α değerleri hesaplanmış ve bu değerlerin 0,62 ile 0,95 arasında değiştiği görülmüştür. Güvenirliğin en düşük olduğu yeterliliğin muhakeme ve argüman en yüksek olduğu yeterliliğin ise iletişim olduğu görülmüştür. Bu sonuçlara dayanarak muhakeme ve argüman yeterliliği dışındaki yeterlilikler için puanlayıcılar arası güvenirliğin yüksek bulunduğu belirtilmiştir.

Açıklanan varyans oranı için en iyi kümeler yöntemi seçilerek çoklu regresyon analizi yapılmıştır. En iyi alt kümeler analizi sonucunda en iyi yordayıcıların sırasıyla muhakeme ve argüman, semboller ve formelleştirme, problem çözme, iletişim, modelleme ve son olarak da temsil yeterlilikleri olduğu belirtilmiştir. PISA 2003 maddeleri üzerinden yapılan bu analiz sonucunda açıklayıcı gücü en yüksek olan yeterliliklerin muhakeme ve argüman, semboller ve formelleştirme ile problem çözme olduğunu ve madde güçlüğündeki değişkenliğin bu modelle %70,5'inin (Düzeltilmiş R^2) açıklanabildiğine işaret etmiştir. PISA 2006 maddeleri üzerinden yapılan analiz sonucunda da yine açıklayıcı gücü en yüksek olan yeterliliklerin muhakeme ve argüman, semboller ve formelleştirme ile problem çözme olduğunu ve madde güçlüğündeki değişkenliğin bu modelle %71,4'ünün (Düzeltilmiş R^2)

açıklanabildiğine işaret etmiştir. Her iki durum içinde Durbin-Watson istatistiği sonucunda artıkların ilişkisiz olduğu belirtilmiştir. 48 maddeye ilişkin veriler temel bileşenler analizi ile döndürme yapılmaksızın analiz edilmiştir. Özdeğeri 1'den büyük iki faktör olduğu ve bu iki faktörün puanlamalardaki varyansın toplamda %64'ünü açıkladığı belirtilmiştir.

Turner ve Adams (2012) yaptıkları araştırmada Yeterlilik Şeması'nın revize edilmiş formunu kullanmıştır. Bu çalışmada Yeterlilik Şeması'nın deneysel geçerliği incelenmiştir. Araştırmada Yeterlilik Şeması ile yapılan puanlamalar için puanlayıcılar arası uyumun düzeyi, her bir yeterliliğin görevlerdeki bilişsel karmaşıklığın farklı boyutlarını ortaya çıkarıp çıkaramadığı ve puanlamaların madde gücünü hangi ölçüde yordadığını belirlemek amaçlanmıştır. PISA 2012 için geliştirilen matematik okuryazarlığı maddelerini kullanarak PISA görevleri için 67 ülkede 2011 yılında kapsamlı bir alan çalışması yapılarak araştırmanın verileri toplanmıştır. Çalışmada PISA 2012'de kullanılmak üzere 2011 de alan uygulaması yapılan 196 matematik görevi ele alınmıştır. Bu maddeler yeterlilikleri düzeyleri açısından beş puanlayıcı tarafından Yeterlilik Şeması ile değerlendirilmiş ve puanlanmıştır. Puanlayıcılardan dördü PISA 2012 Matematik Uzmanları Grubu'nda görev yapmakta olup bir puanlayıcı da PISA matematik madde geliştirme takımının lideridir. Alan uygulaması yapılan her bir madde random olarak seçilmiş yaklaşık 50 öğrenciye uygulanmıştır. Her öğrenci yaklaşık 50 maddeyi yanıtlamıştır. Çalışmada çok boyutlu Rasch modelleme ve gizil regresyon yöntemi ile veriler analiz edilmiştir. Puanlayıcılar arası uyum tam uyuşma yüzdesi ve Pearson momentler çarpımı korelasyon katsayısı ile kategoriler bazında puanlayıcı çiftleri için hesaplanmıştır. Bazı kategorilerde uyuşmaların ve korelasyonların diğerlerine göre daha yüksek olduğu görülmüştür. Semboller ve formelleştirme kategorisindeki uyuşmanın diğerlerinden daha yüksek olduğu görülmüştür. Her bir yeterlilik için puanlayıcıların verdikleri puanların tutarlılığı için Cronbach α , toplam puanlarla korelasyonu ve diğer kalanlarla korelasyonu hesaplanmıştır. Buna göre güvenilirliklerin oldukça yüksek ve madde toplam korelasyonlarının yüksek olduğu görülmüştür. En düşük güvenilirliğin 0,77 ile matematikleştirme en yüksek güvenilirliğin 0,87 ile semboller ve formelleştirme kategorisi olduğu görülmüştür. Puanlayıcılardan birinin puanlamaları için hem madde-toplam hem de madde kalan korelasyonlarının düşük olduğu ve en düşük uyumun bu puanlayıcının puanlamalarında ortaya çıktığı belirtilmiştir. Çok yüzeyli Rasch modeli ile kategoriler bazında kestirilen güvenilirlik değerlerinin Cronbach α değerlerinden az da olsa büyük olduğu görülmüştür. Modelde değişken çıkarılarak yapılan çoklu regresyon çalışmaları ile strateji planlama, semboller ve formelleştirme ile muhakeme ve argüman yeterlilikleriyle kurulan

modelin madde güçlüğünün %74 ile önemli bir ölçüde yordanabildiği görülmüştür. Bu sonuçlar PISA maddelerinin yeterlilik gereksinimlerinin madde güçlüğünün önemli bir yordayıcısı olduğunu göstermektedir.

Puanlama Anahtarı Kullanılarak Elde Edilen Ölçümlerin Güvenirliğinin Genellenabilirlik ve Klasik Test Kuramına Dayalı Olarak İncelendiği Çalışmalar

Cardinet vd. (1976) çalışmalarında, eğitimsel ve psikolojik ölçmelerin yapıldığı çalışmalar için G kuramının klasik test kuramından daha kapsamlı ve daha anlamlı bir çerçeve olduğunu ancak geleneksel psikometrik anlayışı (test edilen bireyler arasındaki farklılaşmanın en yüksek olması gibi) sürdürdüğünü belirtmişlerdir. Bu bağlamda bireylerin çalışmanın merkezi objesi olmadığı problem durumları için de G kuramının güçlü bir tanımlayıcı ve analitik araç olduğunu ileri sürmüşlerdir. Örneğin program değerlendirme gibi öğrenme durumlarının, öğretim durumlarının ya da gelişim basamaklarının farklılaşmasının beklendiği çalışmalarda bireyler arası farklılaşmanın sonuçları açıklamaya yardımcı olmaktan çok zarar verici olduğu belirtilmiştir. Bir araştırmada araştırmacılar ölçümlerin elde edildiği bireylerden çok ölçme koşullarına odaklanıyorsa ölçme desenindeki koşulların transpoze edilmesinin gerekli olduğu ileri sürülmüştür. Bununla birlikte Cronbach ve arkadaşlarının G kuramı kapsamında ele alınan değişkenlik kaynaklarından birey ile diğerleri arasında bir ayrım yaptığı ve araştırmacıların bireylerin daima ölçme objesi olarak ele alınması gerektiği örtük varsayımını sürdürdükleri belirtilmiştir. Birey değişkenlik kaynağının yüzey olarak ele alınmadığını ve G kuramında birey madde ve puanlayıcı gibi üç değişkenlik kaynaklı bir desenin bu nedenle iki yüzeyli desenler olarak adlandırıldığını ifade etmişlerdir. Bu noktada araştırmacılar G kuramı kapsamında diğer değişkenlik kaynaklarının da ölçme objesi olarak ele alınabileceğini belirterek model simetrisi kavramını ileri sürmüşlerdir.

Smith ve Kulikowich (2004) yaptıkları araştırmada karmaşık problem çözme becerilerini ölçmek için oluşturulan bir değerlendirmeden elde edilen yanıtların psikometrik özelliklerini incelemek için genellenebilirlik kuramının ve çok yüzeyli Rasch modelin kullanımını betimlemeyi amaçlamışlardır. Bu doğrultuda kentteki dezavantajlı gençlerin karma problem çözme becerilerini geliştirmek için yapılan boylamsal bir araştırma projesinde yer alan 44 dördüncü sınıf öğrencilerini kontrol grubunu oluşturmak amacıyla seçmişlerdir ve bu

öğrenciler diğer öğrencilere verilen deneysel değişkenlerin hiç birine tabi tutulmamışlardır. Öğrencilere üç eğitimli kişiden biri tarafından "Administration of the Kickball Assessment" (KBA) uygulanmıştır ve uygulama kapak mektubunu ve sonra gelen her bir soruyu okumaktan ibarettir. Bu uygulama 50 dakika sürmektedir ve üç ay ara ile iki defa yapılmıştır. Sonrasında 44 öğrencinin 5 soruya verdiği her iki cevabını da iki puanlayıcı puanlamıştır. Veriler hem Genellenebilirlik Kuramı hem de Çok Yüzeyle Rasch Model ile analiz edilmiştir. Genellenebilirlik Kuramı için random etki modeli tanımlanarak (44 katılımcı, 5 madde, 2 puanlayıcı, 2 durum) BMDP programı 8V ile varyans bileşenleri kestirilerek G ve D çalışmaları yapmışlardır. G çalışması sonucunda en büyük varyans bileşeninin bireysel farklılıklar olduğu, en küçük varyans bileşeninin ise durumlar olduğunu bulmuşlardır. Ayrıca araştırmanın sonuçları doğrultusunda iki ölçme modelinin de yüzeyler arasındaki varyansların göreceli büyüklüklerinde benzeştiklerini ancak varyans kaynaklarını ele alma konusunda farklılaştıklarını görmüşlerdir.

Atılgan (2004) çalışmasında 2002 ve 2003 yıllarında yapılan müzik öğretmenliği özel yetenek seçme sınavlarına ilişkin verileri kullanarak Genellenebilirlik Kuramı ve Çok Değişkenli Rasch Modeli ile analizlerini uygulamıştır. Genellenebilirlik Kuramı kapsamında tek değişkenli ve çok değişkenli G çalışmalarından kestirilen güvenilirlik parametrelerini karşılaştırmış ve alternatif karar çalışmalarının gerçek durumla tutarlılıklarını incelemiştir. Ayrıca Genellenebilirlik Kuramı ve Çok Değişkenli Rasch Modeline ait istatistikler karşılaştırılmıştır. Çalışmada birey görev ve puanlayıcı değişkenlik kaynaklarının çaprazlandığı $b \times g \times p$ deseni kullanılmıştır. Araştırma sonuçlarına göre Genellenebilirlik Kuramının tek değişkenli ve çok değişkenli modellerinin aynı ölçme durumları için alt boyutlarda oluşan testlerde farklı sonuçlar ürettiği G ve Phi katsayılarının modele göre farklılık gösterdiği alternatif karar çalışmalarıyla kestirilen G ve Phi katsayılarının gerçek durumda kestirilenlerden farklı olduğu görülmüştür. Ayrıca Genellenebilirlik Kuramı ve Çok Değişkenli Rasch Modeli yaklaşımlarından elde edilen sonuçların kullanışlı ve kısmen tutarlı olduğu bulunmuştur. Çok değişkenli ve tek değişkenli modeller sonucunda değişkenlik kaynakları için hesaplanan varyanslar ve varyans oranları modellere göre farklıdır. Çok değişkenli ve tek değişkenli model K çalışmalarında alt testler ve birleşik testler için yapılan çok değişkenli model K çalışmaları ile kestirilen G ve Phi katsayıları tek değişkenli modellerle kestirilenlerden büyüktür. Alt boyutlar ve birleşik testlerin güvenilirlik parametrelerinin arttırılması için puanlayıcı sayılarının arttırılması zaman sayılarının arttırılmasına göre daha uygun bulunmuştur.

Atılğan (2005) çalışmasında G kuramının temellerini tanıtmış ve klasik test kuramına göre avantajlarını puanlayıcıların da yer aldığı çok hata kaynaklı hipotetik bir örnek üzerinden açıklamıştır. Bu çalışmada ölçüt dayanklı ölçmeler için Phi, norm dayanaklı ölçmeler için G katsayılarının hesaplanması ve kullanımı gösterilmiştir. Araştırma verileri 10 adaya uygulanan dört görevden oluşan bir testin dört bağımsız puanlayıcı tarafından puanlandığı varsayılarak üretilmiştir. Çalışmada G kuramının bmxpx çaprazlanmış tesadüfi etki deseni kullanılmıştır. Çalışma sonucunda G kuramının klasik test kuramına göre oldukça güçlü bir alternatif olduğu, puanlayıcılar arası güvenilirliğin hesaplanmasında korelasyonel tekniklere göre daha güçlü ve kapsamlı bir alternatif olabileceği belirtilmiştir.

Jonsson ve Svingby (2007) puanlama anahtarı kullanımını güvenilirlik, geçerlilik ve eğitimsel sonuçları bakımından deneysel uygulamaya dayanan 75 çalışma üzerinden incelemiştir. İncelenen çalışma sonuçlarına göre kullanılan puanlama anahtarları analitikse, belli bir konuya özgüyse ve örneklerle ve/veya puanlayıcı eğitimi ile tamamlanmışsa performans değerlendirme puanlarının güvenilirliğini zenginleştirebilmektedir. Ancak performans değerlendirmenin geçerli yorumlanmasını kolaylaştırmamaktadır. Bununla birlikte ölçütlerin ve beklentilerin açık bir şekilde ortaya konulması ile öz değerlendirme ve geri bildirim sağladığından öğretimi desteklediği görülmektedir. Performans değerlendirme göz önüne alındığında, karmaşık davranışların gözlemlerinin inanılabilir ve güvenilir bir şekilde gerçekleştirilip gerçekleştirilmediği problemi ortaya çıkmaktadır. Bu problem en çok geniş ölçekli değerlendirmelerde ve performans değerlendirme kullanılarak geniş ölçekli kararlar almada önemli görülmekte ve bu ölçümlerin hem geçerli ve güvenilir olmasına ilişkin kanıtların sunulmasını gerektirmektedir. Bununla birlikte öğrencinin öğrenmesine yardımcı olan sınıf içi değerlendirmeler için bu gereksinim aynı düzeyde olmasa da yine de değerlendirmenin geçerli olması gerekmektedir. Puanlama anahtarı kullanımı inanılabilirlik problemini çözmenin bir yolu olarak görülmektedir.

Güler (2008) çalışmasında matematik başarısının ölçülmesi ile elde edilen puanlara Klasik Test Kuramı, Genellenebilirlik Kuramı ve Çok Değişkenlik Kaynaklı Rasch Modeli'ni uygulayarak puanların güvenilirlik kestirimlerini hesaplamış ve bu üç kuramdan elde edilen sonuçları karşılaştırmıştır. Araştırmada TIMMS 1999'da kullanılan 24 açık uçlu matematik sorusu 203 öğrenciye uygulanmış ve bu öğrencilerin verdikleri yanıtlar 4 puanlayıcı tarafından bütüncül puanlama anahtarı ile puanlanmıştır. Elde edilen puanların güvenilirliğini kestirmek amacıyla Klasik Test Kuramı kapsamında Cronbach alfa güvenilirlik katsayısı puanlayıcılar arası güvenilirliğin belirlenmesinde Kendall'ın W katsayısı ile puanlayıcılar arası

korelasyon katsayısı hesaplanmış ve puanlayıcıların verdikleri puanların ortalamaları arasında fark olup olmadığını belirlemek için F testi yapılmıştır. Genellenebilirlik Kuramı kapsamında ise tek değişkenli model G ve K çalışmaları yürütülerek $b \times g \times p$ tümüyle çaprazlanmış desen kullanılarak G ve Phi katsayıları hesaplanmıştır. Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli ile birey puanlayıcı ve madde boyutuna ilişkin güvenilirlik katsayıları hesaplanmıştır. Araştırma sonuçları puanların iç tutarlılığın oldukça yüksek olduğunu göstermiştir. Puanlayıcılar arası uyumun Kendall'ın W katsayısı ile orta düzeyde, korelasyon katsayısı ile oldukça yüksek düzeyde olduğu ve F testi sonuçlarına göre puanların ortalamaları arasında farklılık olduğu belirtilmiştir. Genellenebilirlik Kuramı kapsamında G ve Phi katsayılarının 0,92 ve 0,90 olarak kestirildiği ve puanlayıcı değişkenlik kaynağının toplam varyansı açıklama oranının % 2,1 olduğu belirtilmiştir. Çok Değişkenlik Kaynaklı Rasch Ölçme Modeli'ne göre ise öğrenci boyutuna ilişkin güvenilirlik 0,95 puanlayıcılar arası güvenilirlik 0,99 olarak bulunmuştur. Araştırma sonuçlarına dayanarak kullanılan ölçme aracının güvenilir sonuçlar verdiği ve güvenilirliğin belirlenmesinde en az iki kuramdan yararlanmanın daha uygun olacağı ileri sürülmüştür.

Deliceoğlu (2009) çalışmasında Futbol Yetilerine İlişkin Dereceleme Ölçeği'nden elde edilen verilere ilişkin güvenlik katsayılarını Klasik Test Kuramı ve Genellenebilirlik Kuramı'na dayalı olarak belirlemiş ve karşılaştırmıştır. Araştırmaya katılan 72 futbolcunun teknik yetilerinin ölçülmesi amacıyla 56 maddeden oluşan "Futbol Yetilerine İlişkin Dereceleme Ölçeği" Tükçe'ye uyarlanarak kullanılmıştır. 4 puanlayıcı, futbolcuların yetilerini video kayıtları üzerinde 1 hafta arayla iki kez puanlamışlar ve ikinci puanlamaya ait veriler analizlerde kullanılmıştır. Verilerin analizinde Klasik Test Kuramı kapsamında ilk ve ikinci puanlama arasındaki tutarlılık için Pearson momentler çarpımı korelasyon katsayısı, iç tutarlılık güvenirliliği için Cronbach alfa katsayısı, puanlayıcılar arası güvenilirlik için Kendall'ın uyum katsayısı hesaplanmıştır. Genellenebilirlik Kuramı kapsamında ise çok değişkenli $b \times m \times p$ modeli ile G ve K çalışmaları yapılmıştır. Verilerin analizi sonucunda puanlayıcıların puanlamalarının birbirine benzer olduğu, çok değişkenli model K çalışmasında farklı puanlayıcı madde senaryolarına göre kestirilen G ve Phi katsayılarında artış ve azalışın birbirine paralel olduğu görülmüştür. Güvenirlik parametreleri iç ölçütlere göre incelendiğinde G katsayısı ile Cronbach alfa katsayılarının kabul edilebilir düzeyin üstünde Phi katsayısı ile Kendall W güvenirlik katsayılarının kabul edilebilir düzeyin altında olduğu belirtilmiştir. Ölçeğin tamamına ve alt boyutlarına ilişkin güvenilirlik katsayılarının yüksek olmasına dayandırılarak ölçeğin güvenilir bir ölçek olduğu ileri sürülmüştür.

Güler ve Gelbal (2010) matematik başarısına ilişkin elde edilen puanlamaların güvenilirliğini klasik test kuramı ve genellenebilirlik kuramı yaklaşımları çerçevesinde incelemişlerdir. Bu bağlamda TIMSS-1999 uygulamasında yer verilen 24 tane açık uçlu matematik sorusu 203 öğrenciye uygulanmış ve dört puanlayıcı tarafından öğrenci yanıtları puanlanmıştır. Yapılan açımlayıcı ve doğrulayıcı faktör analizi sonucunda 24 maddeden 6 tanesinin iki faktörde birden yüksek yük değerleri aldığı görülmüş ve tüm analizler bu maddelerin çıkarılması sonucu kalan 18 madde üzerinden yürütülmüştür. Verilerin analizinde klasik test kuramı kapsamında Cronbach alfa ve Kendall'in W katsayısı hesaplanmıştır. G kuramı kapsamında tüm değişkenlik kaynaklarının çaprazlandığı bmxpxp tamamen tesadüfi etki deseni tanımlanmış, G ve Phi katsayıları hesaplanmıştır. Puanların iç tutarlılığı için hesaplanan Cronbach alfa katsayısının (0,92) oldukça yüksek olduğu ve puanlayıcılar arası güvenilirlik için hesaplanan Kendall'in W katsayısının (0,52) puanlamalar arasında tutarlılık olduğunu gösterdiği belirtilmiştir. Puanlamalar arası tutarlılık için korelasyon ve ortalamalar arası farklar incelendiğinde korelasyonların oldukça yüksek olduğu bununla birlikte birinci puanlayıcı için ortalamalar arasındaki farkın anlamlı olduğu görülmüştür. Puanlayıcıların puanlamalarının katılık/cömertlik düzeylerinin farklı olduğu ancak puan sıralamalarının birbirine paralel olduğu ve puanlayıcıların birbirleriyle tutarlı puanlamalar yaptıkları sonucuna ulaşılmıştır. G kuramı kapsamında G katsayısı 0.92 ve Phi katsayısı 0.90 olarak bulunmuştur. Yapılan karar çalışmaları sonucunda puanlamaların güvenilirliğinin yükseltilmesi için madde sayısının üçte bir artırılmasının puanlayıcı sayısını bir arttırmaya göre daha etkili olduğu belirtilmiştir. Çalışma sonuçları kullanılan ölçme aracının öğrencilerin matematik başarısını belirlemede güvenilir sonuçlar verdiğini göstermiştir. G kuramı güvenilirlik ve geçerlik arasındaki geleneksel anlamdaki farkı ortadan kaldırdığı ve tek bir çalışma ile belirleyebilmesine olanak tanıdığı için daha ekonomik ve pratik bir kuram olarak önerilmiştir. Birden fazla puanlamanın yapıldığı durumlarda güvenilirliğin kestirilmesinde klasik test kuramı ve G kuramının birlikte kullanılmasının daha açıklayıcı ve aydınlatıcı olacağı belirtilmiştir.

Rezai ve Lovorn (2010) sosyal bilimlerden bir yazma görevine ilişkin öğrencilerin yazılı yanıtlarını değerlendirmede kullanılan bir puanlama anahtarının geçerlilik ve güvenilirliğini araştırmışlardır. Bu bağlamda araştırmacılar tarafından iki farklı rapor ve bir puanlama anahtarı hazırlanmıştır. Raporlardan biri puanlama anahtarındaki teknik kategorisi dışındaki tüm koşulları sağlarken, diğeri yalnızca teknik kategorisindeki koşulları sağlamaktadır. Araştırmacılar eğitim ve işletme alanından 326 lisans öğrencisinin bu raporları önce

puanlama anahtarı kullanmadan sonra puanlama anahtarı kullanarak değerlendirmelerini istemişlerdir. Verilerin analizinde puanlama anahtarının güvenilirliğini değerlendirmek için ölçmenin standart hatasını kullanmışlardır. Araştırma sonuçları puanlama anahtarı kullanmadan yapılan değerlendirmede açıklığın ve standart hatanın daha düşük olduğunu, puanlama anahtarı kullanarak yapılan değerlendirmenin güvenilirliği arttırmadığını göstermiştir.

Yelboğa ve Tavşancıl (2010) klasik test kuramı ve genellenebilirlik kuramı ile kestirilen güvenilirlik katsayılarının ilişkisini inceledikleri çalışmada bir işletmede çalışan 176 personelin iş performansı üç farklı puanlayıcı tarafından iş performans ölçeği kullanılarak belirlenmiştir. Ölçek dört boyut altında yer alan 32 maddeden oluşmakta olup ölçek ile elde edilen verilerle yapılan açıklayıcı ve doğrulayıcı faktör analizi sonuçlarının ölçeğin yapısını doğruladığı belirtilmiştir. Ayrıca modelin uyum iyiliği değerleri modelin kabul edilebilir olduğunu göstermiştir. Çalışmada puanlayıcılar bir yıl ara ile personelin iş performansını iki kez değerlendirmiş ve bu ölçümler üzerinden test tekrar test yöntemi kapsamında uygulamalar arası tutarlılık Pearson momentler çarpımı korelasyon katsayısı ile incelenmiştir. İç tutarlılık için Cronbach α güvenilirlik katsayısı ve puanlayıcılar arası güvenilirlik için Kendall'in uyum katsayısı hesaplanmıştır. Genellenebilirlik kuramı kapsamında G ve Phi katsayıları hesaplanmıştır. Pearson momentler çarpımı korelasyon katsayılarının alt boyutlar ve toplam test için 0,77 ile 0,89 arasında değiştiği görülmüş ve iki uygulamada yüksek kararlılık gösterdiği belirtilmiştir. Ölçeğin alt boyutları ve toplam test için kestirilen Cronbach α katsayılarının 0,66 ile 0,94 arasında değiştiği ve ölçeğin alt boyutlarından sadece birinin düşük değerler aldığı belirtilmiştir. Bu nedenle bu boyuttaki maddelerin ölçeğin diğer boyutlarındaki maddelere göre daha düşük iç tutarlılığa sahip olduğu belirtilmiştir. Puanlayıcılar arası güvenilirlik bağlamında her iki uygulamaya göre alt boyutlar ve toplam test için hesaplanan Kendall'in uyum katsayılarının 0,83 ile 0,98 arasında değiştiği görülmüştür. Buna dayanarak puanlayıcılar arası uyumun yüksek olduğu ve puanlayıcıların birbirine benzer puanlamalar yaptıkları ileri sürülmüştür. Ayrıca birinci uygulamadan elde edilen veriler üzerinden kestirilen Kendall'in W katsayıları ikinci uygulamada kestirilenlerden daha düşüktür. Toplam puanlar için kestirilen Kendall W katsayıları alt boyutlar için kestirilenlerden daha yüksektir. Ayrıca test tekrar test bağlamındaki Pearson momentler çarpımı korelasyon katsayıları puanlayıcılar ve alt boyutlar bazında göreceli olarak farklılık göstermektedir. G kuramı kapsamında $b \times m \times p$ deseni ile her iki uygulama için ayrı ayrı çok değişkenli G ve K çalışmaları yürütülmüştür.

Her iki uygulama için yapılan çok deęişkenli G ve K alıřmaları sonucunda alt boyutlardan biri için kestirilen G ve Phi katsayılarının dięerlerine gre kestirilenlerden olduka dřk olduęu, G katsayılarının Phi katsayılarından byk olduęu ve puanlayıcı-madde sayılarındaki artış ya da azalışa paralel olarak G ve Phi katsayılarının da arttığı ya da azaldığı grlmřtr. KTK ve G Kuramına gre kestirilen gvenirlik katsayılarının her iki uygulama için birbiriyle tutarlı sonular verdięi belirtilmiřtir.

Anıl ve Bykkıdık (2012) alıřmalarında performansa dayalı durum belirleme yoluyla elde edilen lmlerin gvenirliğini Genellenebilirlik kuramı kapsamında karışık desen kullanarak incelemiřlerdir. Bu amala birey, sınıf, puanlayıcı, grev, lt deęişkenlik kaynaklarının ele alındığı, “(b : s) x p x  x g” řeklinde ifade edilen drt yzeyli karışık desen ile G ve K alıřmaları yrtmřlerdir. Bu baęlamda 90 ęrencinin performansı drt puanlayıcı tarafından analitik puanlama anahtarı ve yanıt tanıma kodları kullanılarak puanlanmıştır. Arařtırma sonucunda G kuramı kapsamında toplam varyansı aıklama oranı en yksek olan etkinin lme objesi olarak ele alınan birey ana etkisi olduęu belirtilmiştir. Sınıf ana etkisinin aıkladığı varyansın nemsenmeyecek dzeyde olsa da farklı sınıflardaki bireylerin puanlarının puanlayıcıdan puanlayıcıya kısmen deęiřebildięi grlmřtr. Verilerin gvenirliğine iliřkin G katsayısının 0,911 ve Phi katsayısının 0,898 olduęu, karar alıřması sonucunda puanlayıcı sayısını arttırıp azaltmanın grev sayısını arttırıp azaltmaya gre daha etkili olduęu sonucuna ulařılmıştır.

Eser ve Gelbal (2012) G Kuramı ve lojistik regresyon analizinden yararlanarak gerekleřtirilen bir performans puanlamada ortaya ıkan puanlayıcı tutarlığını belirlemek ve karřılařtırmak amacıyla yaptıkları alıřmada 106 ęrenciye 15 maddelik bir lme aracı vermiřlerdir. ęrencilerin verdikleri cevaplar  puanlayıcı tarafından dereceli puanlama anahtarı kullanılarak puanlanmıştır. Puanlama ile elde edilen veriler, G kuramı ve lojistik regresyon analizi ile madde bazında ve testin tamamına dayalı olarak analiz edilmiştir. Analiz sonucunda G kuramı ile elde edilen puanlayıcı varyans bileřenleri ve toplam varyansı aıklama yzdeleri ile lojistik regresyon analizi ile elde edilen sınıflama yzdeleri yorumlanmıştır. Elde edilen bulgulara dayalı olarak G Kuramı ve lojistik regresyon analizinin puanlayıcılar arası tutarlığı belirlemede paralel sonular rettięi, ancak lojistik regresyon analizinin G kuramı kadar hassas ıktılar vermedięi ve G kuramına gre daha yzeyssel bir istatistik olduęu sonucuna varılmıştır.

Taylor ve Pastor (2013) çalışmalarında alternatif değerlendirmenin teknik niteliğini değerlendirmek amacıyla genellenebilirlik kuramını temel almışlardır ve gözlenen puan varyansına puanlayıcı türü, değerlendirme girişimleri ve görevlerden her birinin göreceli katkısını belirlemek için bir G çalışması yapmışlardır. Bu doğrultuda oluşturdukları desen görevlerden (t), puanlayıcı türünden (r), değerlendirme girişimlerinden (a) ve öğrencilerden (s) oluşmaktadır ve “(t:s) x r x a” şeklinde gösterilmiştir. Kullanılan veri 5., 8. ve 10. sınıf okuma ve matematik değerlendirmelerinden elde edilmiştir ve ağır bilişsel engelli öğrencilere okulun son altı haftası boyunca iki görev uygulanmıştır. Bu performans görevlerini iki farklı puanlayıcı türü gözlemlemiş ve her bir görev için üç ayrı girişimle öğrencilerin performanslarına göre bağımsız değerlendirme yapmışlardır. Araştırmacılar verilerin analizi aşamasında GENOVA programını kullanarak G çalışmasının yanı sıra D çalışması da yapmışlardır. G çalışması sonucunda tüm sınıf düzeylerinde ve hem okuma hem de matematik dersinde en büyük varyans bileşeninin (ta,sta) etkileşiminden kaynaklandığını gözlemlemişlerdir. Ayrıca okuma ve matematik puanları için Phi katsayısının 0,51 ve 0,62 arasında olduğunu ve okuma için hesaplanan katsayıların daha küçük olduğunu bulmuşlardır. Yaptıkları D çalışmasında ise puanlayıcı türü sayısını sabit tutarak değerlendirme girişimi sayısını ve görev sayısını bir ile beş arasında değiştirmişlerdir ve daha etkili sonuçlar elde etmek için girişim sayısını arttırmaktansa görev sayısının artmasının puanların güvenilirliği konusunda daha önemli olduğunu vurgulamışlardır.

Briesch vd. (2014), G kuramının değerlendirme araştırmalarında oldukça faydalı ve güçlü bir yöntem olmasına karşın yeterince kullanılmadığı fikrinden yola çıkarak yürüttükleri çalışmada G kuramının kavramsal temellerini, termonolojisini, G kuramı kapsamında yürütülecek çalışmalarda göz önünde tutulması gereken boyutları tartışmışlardır. Bu çalışmada, eğitim ve psikoloji alanındaki araştırmacılar tarafından G kuramının kullanımını ve anlaşılmasını sağlamak için metodolojik ve analitik konulara odaklanılarak G kuramının genel bir tanıtımının yapılması amaçlanmıştır. Bu bağlamda bir çalışmanın desenin planlanmasında modelde yer verilecek değişkenlik kaynaklarının iyi bir şekilde belirlenmesi elde edilecek ölçümlere etki edebilecek değişkenlik kaynaklarını belirlemenin önemli olduğu belirtilmiştir. Eğitimsel ve psikolojik ölçümlere en fazla etki edebilecek değişkenlik kaynaklarının genellikle, puanlayıcı, madde, düzenleme (setting), yöntem, boyut, zaman ve form olduğu ileri sürülmüştür. Ayrıca tüm hata kaynaklarının bağımsız yordamalarına olanak verdiği için tümüyle çaprazlanmış desenlerin istendik olduğu ve yüzey sayısının

arttırılmasının yüzeyler arasındaki etkileşimlerin yorumlanmasını karmaşıktırdığı belirtilmiştir.

Wilhelm ve Kim (2015) yaptıkları çalışmada Öğretimsel Kalite Değerlendirme adlı sınıf içi değerlendirmelerde yaygın olarak kullanılan ölçme araçları takımının güvenilirliğini incelemiş ve öğretmenlerin öğretim uygulamalarının güvenilir bir biçimde ölçülebilmesi için gerekli gözlem sayısını belirlemeyi amaçlamışlardır. Dört yıl süren bu araştırma kapsamında ölçme objesi olarak ele alınan 150 öğretmen iki yıl ara ile iki kez gözlemlenmiştir. Araştırmacılar tarafından dört beş gün boyunca puanlayıcılara eğitim verilmiş ve kodlama öncesi uyuşmanın %80 olduğu görülmüştür. Araştırma verileri çok değişkenli G çalışması ile analiz edilmiştir. Puanlama anahtarları ile toplanan verilere faktör analizi yapılmış, faktör analizi sonucunda 3 boyutun olduğu görülmüştür. Çok değişkenli K çalışması sonuçlarına göre öğretmen sayısının 150 (tüm öğretmenler örnekleme) ve 96 (sınıf içi tartışma yaptıran öğretmenler örnekleme) olduğu, gözlem sayısının 2 ile 10 arasında olduğu durumlarda Özgün Öğretimsel Kalite Değerlendirme ve Genişletilmiş Öğretimsel Kalite Değerlendirme araçlarına ilişkin güvenilirlik parametreleri kestirilmiştir. K çalışması sonuçlarına göre Özgün Öğretimsel Kalite Değerlendirme araçları için kestirilen G katsayılarının 0,519 ve 0,576, Genişletilmiş Öğretimsel Kalite Değerlendirme araçları için kestirilen G katsayılarının 0,554 ve 0,626 olduğu görülmüştür. Gerekli olan gözlem sayısı genelleme yapılmak istenen evrene, modelleme yaklaşımına, kullanılan ölçme aracına istenilen güvenilirlik düzeyine bağlı olarak değişmektedir. 0,65 güvenilirlik düzeyine ulaşmak için sınıf içi tartışma yapan öğretmenler örnekleminde 3 gözlemin yeterli olduğu belirtilmiştir. 0,80 güvenilirlik düzeyinde G katsayısı elde etmek için en az 9 gözleme ihtiyaç vardır. Araştırma sonuçları Öğretimsel Kalite Değerlendirme araçları kullanılarak bir öğretmenin öğretim uygulamalarının güvenilir bir biçimde ölçülebilmesi için bazı durumlarda en az üç gözleme gerek duyulduğunu göstermiştir.

Bağcı (2015) çalışmasında Matematiksel Muhakemeyi Değerlendirme Ölçeği'nden elde edilen puanların güvenilirliğini Klasik Test Kuramı ve Genellenebilirlik Kuramı kapsamında incelemiştir. Araştırmanın verileri 187 öğrenciye Matematiksel Muhakemeyi Değerlendirme Ölçeği uygulanmış ve öğrencilerin yanıtları üç puanlayıcı tarafından puanlanması sonucunda araştırmanın verileri elde edilmiştir. Genellenebilirlik kuramı kapsamında α x s x p tümüyle çaprazlanmış desen ile α x (s:p) yuvalanmış desen için ayrı ayrı G ve K çalışmaları yapılmış ölçümlerin güvenilirlik kestirimleri hesaplanmıştır. Klasik Test Kuramı kapsamında Cronbach alfa, Tabakalanmış alfa ve sınıf içi korelasyon katsayısı hesaplanmıştır. Araştırma sonuçları

her iki kuramla elde edilen güvenilirlik katsayılarının kabul edilebilir olduğunu göstermiştir. Yuvalanmış desende G çalışması sonucunda kestirilen G ve Phi katsayılarının çaprazlanmış desende kestirilenlerden daha yüksek olduğu görülmüştür. G katsayılarının birbirine yakın olduğu G kuramı ile yapılan analizin Klasik Test Kuramı ile yapılanlara göre daha ayrıntılı bilgi verdiği belirtilmiştir.

Büyükkıdık ve Anıl (2015) çalışmalarında performansa dayalı değerlendirmelerin güvenilirliğinin, G kuramında farklı desenler kullanıldığında yüzey sayısının değişiminden nasıl etkilendiğini incelemişlerdir. Araştırma kapsamında 132 öğrencinin rutin olmayan problem çözme performanslarını dört puanlayıcı tarafından bütünsel ve analitik puanlama anahtarı kullanılarak puanlanmıştır. G kuramı kapsamında bireylerin ölçme objesi olarak ele alındığı üç farklı desen oluşturulmuştur. İki türdeki puanlama anahtarı için ayrı ayrı birey, görev ve puanlayıcı değişkenlik kaynaklarının tümüyle çaprazlandığı bxgxp deseni ile birey, görev, puanlayıcı ve puanlama anahtarı değişkenlik kaynaklarının tümüyle çaprazlandığı bxgxpax deseni ile G ve K çalışmaları yürütülmüştür. Araştırma sonucunda farklı desenlerle elde edilen G ve Phi katsayılarının birbirinden farklı olduğu, değişkenlik kaynağı sayısı arttıkça ölçme objesinin açıkladığı varyans oranının azaldığı görülmüştür. Analitik puanlama anahtarı ile elde edilen puanların güvenirliliği bütünsel puanlama anahtarı ile elde edilen puanların güvenirliliğine göre daha yüksek bulunmuştur.

Güler & Teker (2015), puanlayıcılar arası güvenirliliğin belirlenmesinde korelasyon, ortalamaların karşılaştırılması, uyum yüzdesi ve genellenebilirlik kuramı yaklaşımlarını ele almışlardır. 43 öğrencinin on açık uçlu maddeye verdikleri yanıtlar iki puanlayıcı tarafından puanlanmasıyla elde edilen verilere ait güvenilirlik değerlerinin farklılaştığı görülmüştür. Ele alınan yaklaşımlara göre güvenilirlik kestirimleri genellenebilirlik kuramıyla 0,90 olarak bulunmuş, basit korelasyonla yüksek düzeyde ve pozitif yönlü (0,74) bir ilişki ortaya çıkmış, puanlayıcılar arası uyuma yüzdesiyle kestirilen tam uyum ise % 58,9 olarak belirlenmiştir. Maddeler tek tek ele alındığında puanlayıcıların ortalamaları arasında üç maddede anlamlı bir farklılık bulunmuş olmasına karşın genel ortalama da anlamlı bir farklılık bulunmamıştır. Araştırma sonucuna göre bu yaklaşımlar arasından genellenebilirlik kuramının pek çok hata kaynağını birlikte ele alabilmesinden dolayı özellikle puanlayıcı güvenirliliğinin belirlenmesinde kullanılması önerilmiştir.

İlgili araştırmalar incelendiğinde KTK çerçevesinde puanlayıcı içi güvenirlilik yaklaşımına göre puanlayıcılar arası güvenirlilik yaklaşımına daha fazla başvurulduğu görülmüştür. G

kuramı çerçevesinde ise genellikle tek deęişkenli modellerin kullanıldığı çalışmalara rastlanmıştır. Bununla birlikte çok deęişkenli G çalışmalarının yapıldığı araştırmalar genellikle performans deęerlendirmeye dayalı araştırmalardır. Her iki kurama dayalı olarak güvenilirlik katsayıları karşılaştırıldığında G kuramı ile elde edilen güvenilirlik parametrelerinin daha yüksek olduğu görülmüştür. Ayrıca G kuramı çerçevesinde kestirilen G katsayılarının Phi katsayılarından daha büyük olduğu görülmektedir. Puanlayıcılar arası güvenilirlik katsayıları kullanılan ölçme araçlarının alt boyutları için ayrı ayrı kestirildiğinde toplam punalar için kestirilenlere göre daha düşük çıkmaktadır. Ayrıca çok boyutlu bir ölçme aracının alt boyutları için kestirilen güvenilirlik katsayıları tutarlılık anlamında farklılık göstermemektedir.



BÖLÜM III

YÖNTEM

Bu bölümde araştırmanın modeli, çalışma grubu, veri toplama araçları, verilerin toplanması ve analizine ilişkin bilgiler verilmiştir.

Araştırma Modeli

Araştırma, matematik okuryazarlığı maddelerinin yeterlilik düzeylerinin değerlendirilmesi çerçevesinde Yeterlilik Şeması ile elde edilen ölçümlerin Genellenebilirlik Kuramı'na ve Klasik Test Kuramı'na dayalı olarak güvenilirlik kestirimlerinin incelenmesine ve karşılaştırılmasına dayalıdır. Bu yönüyle bu araştırma, güvenilirlik kestirimlerinin belirlenmesi açısından betimsel, elde edilen güvenilirlik kestirimlerinin iki farklı kurama göre karşılaştırılması açısından temel araştırmadır.

Evren ve Örneklem

Araştırmanın evreni 2000- 2012 yılları arasında PISA esas ve pilot uygulamalarında yer verilen matematik okuryazarlığı maddelerinden oluşmakta olup, PISA matematik okuryazarlığı madde havuzunda kaç madde olduğu açıklanmadığından evrendeki madde sayısı bilinmemektedir. OECD bir maddeye birden fazla uygulamada yer verdiği için madde havuzundaki maddelerin tamamını yayınlamamaktadır. Bugüne kadar PISA 2000, 2003, 2006, 2009 ve 2012 esas ve pilot uygulamalarında yer verilen ve OECD (2009a, s.310; 2013b, s.3) tarafından yayınlanmış 156 matematik okuryazarlığı maddesi mevcuttur. Bu

bağlamda bu araştırmanın örneklemini PISA 2000, 2003, 2006, 2009 ve 2012 esas ve pilot uygulamalarında yer verilen ve OECD (2009a, s.310; 2013b, s.3) tarafından yayınlanmış 156 matematik okuryazarlığı maddesi oluşturmaktadır.

Araştırmada elde edilen veriler üzerinden güvenilirlik kestirimleri yapılacağı ve elde edilen güvenilirlik katsayıları karşılaştırılacağı, Genellenebilirlik Kuramına göre oluşturulan desenlerde maddeler ölçme objesi olarak ele alınacağı için evreni temsil edecek şekilde bir örneklem seçilmesine gerek duyulmamıştır. Bununla birlikte araştırmada ele alınan yayınlanmış maddeler OECD tarafından madde örneklemini olarak sunulduğundan bu örneklemin evreni temsil edebileceği düşünülmektedir. Maddelerin uygulama yıllarına ve uygulama türlerine göre dağılımı Tablo 1’de sunulmuştur:

Tablo 1

Maddelerin Uygulama Yıllarına ve Uygulama Türlerine Göre Dağılımı

Uygulama Yılı	Uygulama Türü	Madde Sayısı Frekans (f)	Madde Sayısı Yüzde %
2000	Esas	19	12
2003	Esas	23	15
2003	Pilot	47	30
2003/ 2006	Esas	11	7
2012	Esas	26	17
2012	Pilot	30	19
Toplam		156	100

Araştırmada ele alınan maddelerden %51’inin yani 79 maddenin esas uygulamada kullanıldığı, %49’unun yani 77 maddenin ise pilot uygulamalarda kullanıldığı görülmektedir. Esas uygulamada kullanılan maddelere ait madde güçlük puanlarına OECD (2009a; 2013b) raporlarında yer verilirken pilot uygulamalarda kullanılan maddelere ait madde güçlük puanları belirtilmemiştir. Bu bağlamda örnekleme alınan 156 maddeden 79’unun (%51) yarısından fazlasına ait madde güçlük puanları bilinmektedir.

Maddelerin güçlük puanları ve öğrencilerin yeterlilik düzeyleri aynı ölçek üzerinden ölçeklendirilmekte olup, madde güçlük puanının madde güçlük katsayısından farklı olduğuna dikkat edilmelidir. Madde güçlük puanları kavramsal çerçevede de açıklandığı gibi, PISA uygulamaları sonucunda elde edilen veriler üzerinden, öğrencilerin yetenek düzeyleri doğrudan maddelerle ilişkilendirilerek, genellikle madde tepki kuramı

çerçevesinde tek parametrelili lojistik regresyon yöntemi kullanılarak kestirilmektedir. Matematik puanlama ölçekleri 2003 yılı temel alınarak belirli bir yılda OECD ülkelerinin matematik alanındaki ortalama puanı 500 ve standart sapması 100 olacak şekilde hazırlanmaktadır (MEB, 2013, s.11). Bu ölçeklerin hazırlanmasında aynı zamanda OECD ülkelerindeki öğrencilerin yaklaşık üçte ikisinin aldığı puanlar 400 ile 600 arasında yer alacak şekilde olmasına dikkat edilmektedir. Bu çalışmada kullanılan esas uygulamalarda yer verilen yayınlanmış 79 maddenin güçlük puanları 262 ile 840 arasında değişmekte olup, ranjı 578, ortalaması 545,09, ve standart sapması 104,92'dir.

Araştırmada Yeterlilik Şeması puanlama anahtarının uyarlama aşamasında altı farklı uzmandan görüş alınmıştır. Bu uzmanlardan biri İngilizce Öğretmenliği alanından, dördü Eğitimde Ölçme ve Değerlendirme alanından ve biri Matematik Eğitimi alanından seçilmiştir. Eğitim düzeyleri açısından incelendiğinde İngilizce Öğretmenliği alanındaki uzmanın yüksek lisans tez aşamasında, Eğitimde Ölçme ve Değerlendirme alanındaki üç uzmanın doktora ders aşamasında ve Matematik Eğitimi alanındaki bir uzmanın post doktrasını bitirmiş olup matematik eğitimi alanında doçentlik unvanına sahip olduğu görülmektedir.

Ayrıca verilerin toplanması aşamasında maddelerin yeterlilik gereksinim düzeylerinin değerlendirilmesinde beş farklı puanlayıcıdan yararlanılmıştır. Puanlayıcıların tamamı Matematik Eğitimi alanında doktora yapmakta ve tez aşamasında olan uzmanlardan gönüllülük esasına dayalı olarak seçilmiştir. Ayrıca bu puanlayıcıların seçiminde bir diğer ölçüt olarak matematik okuryazarlığı ve PISA ile ilgili daha önceden bilgi sahibi olmaları göz önünde tutulmuştur.

Veri Toplama Araçları

Veri toplama aracı olarak uyarlama sürecinde uzman görüşü almak amacıyla “Uzman Görüş Formu” ile puanlama sürecinde yayınlanmış PISA matematik okuryazarlığı maddelerinin puanlanmasında Ek 1’de verilen Yeterlilik Şeması kullanılmıştır. Aşağıda bu araçlara ilişkin ayrıntılı bilgi verilmiştir.

Yeterlilik Şeması

Yeterlilik şeması, Niss ve arkadaşları tarafından belirlenen ve PISA matematik okuryazarlığı çerçevelerinde yer verilen sekiz matematiksel yeterlilikten (matematiksel düşünme, problem kurma ve çözme, matematiksel modelleme, matematiksel muhakeme, matematiksel temsilleri idare etme, sembolleştirme ve formalleştirme ile ilgilenme ve matematiksel araç gereçleri kullanma) matematiksel düşünme ile matematiksel muhakemenin birleştirilmesi ve matematiksel yardımlar ve araçlar yeterliliğinin PISA matematik okuryazarlığı maddeleri için uygun olmaması gerekçesi ile çıkarılması sonucunda altı yeterlilik alanını içermektedir (Turner vd. 2015, s.90).

Şemanın ilk versiyonunda (Turner vd., 2013) bu yeterlikler *muhakeme ve argüman, iletişim, modelleme, temsil, problem çözme* ile *sembolik, formal, teknik dilin ve işlemlerin* kullanımı şeklinde adlandırılmış, her bir yeterliliğin operasyonel tanımları verilmiş ve düzeyleri belirtilmiştir. Şemanın revize edilmiş versiyonunda (Turner vd. 2015, s.90) ise bu yeterliliklerin *iletişim, strateji planlama, matematikleştirme, temsil, sembollerin, işlemlerin ve formal dilin kullanımı* ile *muhakeme ve argümantasyon* olarak yeniden adlandırıldığı, operasyonel tanımları ve düzeylere ilişkin tanımlamalarında çeşitli değişikliklerin yapıldığı görülmektedir.

Yeterlilik şemasında her bir yeterliliğe ilişkin dört düzey bulunmakta bu düzeyler 0 ile 3 arasında puanlanmaktadır. Düzey 0, o yeterliliğe ilişkin düşük bilişsel gereksinimi, Düzey 3 ise yüksek bilişsel gereksinimi ifade etmektedir. Bir maddenin alabileceği en düşük puan 0 iken en yüksek puan 18'dir. Yeterlilik şeması kullanılarak puanlanan bir maddenin tüm yeterlilik alanlarından yüksek ya da düşük puan alması diğer bir deyişle kategorilere verilen puanların birbirleriyle tutarlı olması gerekmemektedir. Örneğin bir maddenin iletişim yeterliliği gereksinimi yüksek olabilmesine karşın hiç temsil içermediğinden temsil yeterliliği gereksinimi düşük olabilmektedir (Turner vd., 2015, s.107). Burada yeterlilik gereksinimi (demand for activation of competencies) ile kast edilen, bir maddenin doğru bir şekilde yanıtlanabilmesi için öğrencinin etkinleştirmesi gereken bilişsel süreçlerdir. Diğer bir deyişle öğrencinin ilgili yeterliliği ne düzeyde etkinleştirmesi gerektiğini ifade etmektedir.

Yeterlilik şeması matematiksel problemlerin matematiksel yeterlilikler açısından niteliklerinin ölçülmesinde kullanılan altı yeterliliğe ilişkin operasyonel tanımları, açıklamaları, yeterlilik düzeylerini ve bu düzeylere ilişkin tanımları içerdiğinden bir ölçme

aracı olarak analitik dereceli puanlama anahtarı türündedir. Bu puanlama anahtarını geliştirilme çalışmaları PISA matematik okuryazarlığı maddelerinin deneysel güçlük düzeyleri ile matematik okuryazarlığı çerçevesinde yer verilen temel matematiksel yeterlilikler arasındaki ilişkiyi açığa çıkarabilmek amacıyla 2003 yılında başlamış olup halen devam etmektedir (Turner vd., 2015, s.87; Turner, Dossey, Blum ve Niss, 2013, s.24). Yeterlilik şemasıyla yapılan ilk çalışmada temel amaç madde güçlüğünün yordayıcılarını daha iyi anlamaktır (Turner vd., 2015, s.89).

Turner vd. (2013)'ün çalışmasında bu amaçla şemanın ilk versiyonu kullanılarak hem PISA 2003 hem de PISA 2006 yıllarındaki uygulamalarda yer verilen 48 matematik okuryazarlığı maddesi sekiz puanlayıcı tarafından Yeterlilik şeması kullanılarak iki yıl ara ile iki kez puanlanmıştır (s.28). Her bir maddeye ilişkin puanlayıcıların verdiği puanların ortalaması alınmış ve madde yeterlilik puanının PISA uygulamalarından elde edilen madde güçlüğünü yordayıp yordamadığı incelenmiştir. Maddelerin ortalama puanları temel alınarak hesaplanan kategoriler arasındaki korelasyonların 0,283 ile 0,721 arasında değiştiği, en yüksek korelasyonun problem çözme ile muhakeme ve argüman kategorileri arasında olduğu görülmüştür (Turner vd., 2013, s.28).

Puanlayıcı içi güvenilirliği hesaplamada dört puanlayıcının iki yıl ara ile aynı maddelere verdikleri puanlar üzerinden her bir kategori için iki puanlama arasındaki korelasyonlar hesaplanmıştır. Korelasyon değerlerinin 0,05 düzeyinde anlamlı olduğu ve 0,369 ile 0,906 arasında değiştiği görülmüştür. Araştırmacılar bu noktada şema ve kullanımı ile daha fazla deneyimi olan puanlayıcıların puanlamaları arasındaki korelasyonların daha yüksek olduğunu belirtmişlerdir.

Turner vd. (2013) puanlayıcılar arası güvenilirliği hesaplamada ise, her bir kategori için her bir maddeye sekiz puanlayıcının da aynı değeri verme derecesini belirlemek amacıyla Cronbach alfa değerlerini hesaplamışlardır. Bu değerlerin muhakeme ve argüman yeterliliği için 0,62 ile iletişim yeterliliği için 0,95 arasında değiştiği görülmüştür.

Maddelerin ortalama puanlarının madde güçlüğünü ne derecede yordadığını belirlemek amacıyla en iyi alt küme ve çoklu regresyon yönteminden yararlanılmıştır. Analiz sonuçları şemada yer alan altı yeterlilikten muhakeme ve argüman, semboller ve formalleştirme, problem çözme olmak üzere üçünün 48 maddenin güçlük düzeyindeki değişkenliğin %70'inden fazlasını açıkladığını göstermiştir. Puanlayıcılar arası uyumun önemli bir düzeyde olduğu fakat puanlayıcılar arası uzlaşma ve puanlayıcı içi tutarlığa ilişkin yeterli

kanıt bulunamadığı görülmüştür. Bu durum, puanlayıcıların madde yeterliliklerine verdikleri puanların birbirinin aynısı (puanlayıcılar arası uzlaşma) olmasa da, birlikte artma ve azalma eğiliminde (puanlayıcılar arası uyum) olduğunu göstermektedir. Ayrıca puanlayıcı içi tutarlığın düşük olması, puanlayıcıların farklı zamanlarda aynı maddelere farklı yeterlilik puanları atadığı anlamına gelmektedir. Madde puanlarının ortalamaları alınarak kurulan regresyon modeli madde güçlüğündeki değişkenliğin önemli bir bölümünü açıklasa da puanlayıcılar arası ve puanlayıcı içi güvenilirliğe ilişkin bulgulara dayanarak şemanın revize edilme çalışmaları yürütülmüştür.

Bir sonraki aşamada Turner Adams (2012)' nin çalışmasında PISA 2012 uygulamasında kullanılmak üzere geliştirilmiş 196 matematik okuryazarlığı maddesi şemanın revize edilmiş versiyonu ile beş puanlayıcı tarafından puanlanmıştır. Puanlayıcılar arası güvenilirlik için iki puanlayıcı arasındaki tam uzlaşma yüzdesi her bir kategori için hesaplanmış, bu değerlerin %32 ile %73 arasında değiştiği görülmüştür. Kategoriler bazında hesaplanan Cronbach alfa değerleri ise 0,77 ile 0,87 arasında değişmektedir. Madde güçlüğünün yordanmasına ilişkin olarak altı yeterlilik kategorisi ve madde güçlüğü arasındaki ilişkiler yedi boyutlu madde tepki modeli oluşturularak incelenmiştir. Analiz sonuçları şemada yer alan altı yeterlilikten strateji planlama, iletişim, semboller ve formelleştirme olmak üzere üçünün 196 maddenin güçlük düzeyindeki değişkenliğin %74'ünü açıkladığını göstermiştir. Bununla birlikte *strateji planlama, matematikleştirme ve muhakeme ve argüman* yeterliliklerinin önemli derecede örtüştüğü ve bu nedenle modelde bunlar arasından yalnızca birinin yer aldığı belirtilmiştir (s.91). Bu nedenle kategori tanımları ve düzey açıklamaları detaylandırılarak yeniden düzenlenmiş ve bu çalışmada kullanılmış olan şemanın son versiyonu oluşturulmuştur. Yeterlilik şemasının son versiyonuna ilişkin geçerlik ve güvenilirlik çalışmaları puanlama yapılacak yeni maddelerin üretilmesi gerektiği gerekçesi ile henüz yapılmamış ya da yayınlanmamıştır.

Uzman Görüş Formu

Araştırma kapsamında Yeterlilik Şeması'nın Türkçe' ye uyarlanması aşamasında çeviri formun dile, kültüre ve matematik eğitimi terminolojisine uygunluğunun değerlendirilmesine ve puanlama anahtarının kapsam geçerliğinin saptanmasına ilişkin olarak uzman görüşüne başvurulmuş ve bu amaçla araştırmacı tarafından hazırlanan uzman görüş formu kullanılmıştır. Ek 4'te bir bölümü verilen uzman görüş formu iki bölümden

oluşmaktadır. Bölüm I de çalışmanın amacı ve uzman görüşüne ihtiyaç duyulmasının nedeni belirtilmiş, uzmanların görüşlerini nasıl belirtecekleri açıklanmıştır.

Bölüm II de puanlama anahtarındaki her bir başlık, cümle, tanım ve düzey açıklaması bir madde gibi ele alınarak özgün forma karşılık gelen Türkçe ifadeler ile uygunluk seçenekleri bir tablo halinde birlikte verilmiştir. Uzman görüş formu bu şekilde ele alındığında uyarlama aşaması için uzmanlardan toplam 79 ifade için görüş alınmıştır. Uzmanlardan bu bölümde özgün formdaki ifadeler ile Türkçe karşılıklarını karşılaştırmaları, görüşlerini Uygun, Düzeltmeli ve Uygun Değil seçeneklerinden birini işaretleyerek belirtmeleri ve varsa önerilerini belirtmeleri istenmiştir.

Uzman görüş formu ile toplanan verilerin analizinde tam uzlaşma yüzdesi kullanılmıştır. İfadelerin uygun olduğu görüşünde uzmanlar arası tam uzlaşma oranı 0,81 iken düzeltmeli ve uygun değil görüşü için tam uzlaşma sağlanmadığı görülmüştür. Her bir uzmanın ifadelerin uygun olduğu görüşüne katılma oranları ise sırasıyla 0,91; 0,92; 0,94;0,97 olarak bulunmuştur.

Verilerin Toplanması

Verilerin toplanmasına ilişkin süreç aşağıdaki gibi gerçekleşmiştir:

1. Öncelikle Yeterlilik Şeması'nın uyarlanması için elektronik posta yoluyla puanlama anahtarının geliştiricilerinden izin alınmıştır. Daha sonra araştırmacının kendisi ve bir İngilizce Öğretmenliği alan uzmanı olmak üzere iki uzman tarafından Yeterlilik Şeması puanlama anahtarının Türkçe çevirisi yapılmıştır. Çeviri formlar karşılaştırılmış, farklılıkların olduğu ifadeler yeniden incelenmiştir. Yapılan düzeltmeler sonucu elde edilen form ile uzman görüş formu hazırlanmış ve İngilizce Öğretmenliği lisans mezunu ve Ölçme Değerlendirme alanında doktora yapan bir alan uzmanından özgün form ile çeviri formunun karşılaştırılması ve uygun olmayan çeviriler için öneride bulunulması istenmiştir. Gerekli düzeltmeler yapıldıktan uzman görüş formu revize edilerek üçü Matematik Öğretmenliği lisans mezunu olup Ölçme Değerlendirme alanında doktora yapan ve biri Matematik Eğitimi alanında doçentlik unvanına sahip dört uzmandan görüş alınmıştır. Bu uzmanlardan gelen görüşler analiz edilerek uzlaşma oranları belirlenmiştir. Uzmanlar arasında tam uzlaşma oranı 0,81, her bir uzmanın ifadelerin uygun olduğu görüşüne katılma oranları ise sırasıyla 0,91; 0,92; 0,94; 0,97 olarak bulunmuştur. Bu bağlamda

uyarlama formun dile, kültüre ve Matematik Eğitimi alanı terminolojisine uygun olduğu görülmüştür.

2. Puanlama anahtarının çevirisinden sonra, puanlama sürecinde ölçme objesi olarak ele alınacak maddelerin belirlenmesi için OECD tarafından açıklanan matematik okuryazarlığı maddelerine ulaşılmıştır. Tablo 1’ de uygulama türleri ve yılları belirtilen yayınlanmış maddelerin tümü olan 156 maddenin yeterlilik şeması ile puanlanmasına karar verilmiştir.
3. Puanlamayı yapacak uzmanların seçiminde çalışmaya gönüllü katılım, matematik eğitimi alanında doktora yeterliliğini almış olma ve matematik okuryazarlığına ilişkin ön bilgiye sahip olma ölçütleri göz önünde tutulmuştur. Bu ölçütleri karşılayan biri araştırmacının kendisi olmak üzere beş uzman belirlenmiştir.
4. Uzmanlara puanlama anahtarının amacı, geliştirilme süreci ve kullanımı ile ilgili araştırmacı tarafından eğitim verilmiş, maddelerin nasıl puanlandığına ilişkin örnekler incelenmiştir. Bu eğitimin içeriği, Yeterlilik Şeması’nın tanıtıldığı, örnek puanlamaların gösterilerek açıklandığı Turner vd. (2015)’in makalesinden yararlanarak hazırlanmıştır.
5. Puanlayıcılar iki hafta içinde 156 maddeyi puanlamışlardır. Puanlama süreci sona erdikten dört hafta sonra puanlayıcılardan aynı maddeleri ilk puanlamadan bağımsız olarak yeniden puanlamaları istenmiştir. İkinci puanlamanın tamamlanması için puanlayıcılara ilk puanlama da olduğu gibi iki hafta süre verilmiştir.

Verilerin Analizi

Verilerin analizinde öncelikle 156 PISA matematik okuryazarlığı maddesinin yeterlilik gereksinim düzeylerine ilişkin beş puanlayıcıdan iki farklı zamanda toplanan verilerin normal dağılıma sahip olup olmadıkları Kolmogorov-Simirov testi ile test edilmiş ve bu veriler için betimsel istatistikler hesaplanmıştır. Kolmogorov- Simirnov testi ve betimsel istatistikler SPSS 21.0 programı ile hesaplanmış olup ulaşılan sonuçlar Ek 2 ve Ek 3’de sunulmuştur. Puanlayıcıların maddelere verdikleri puanların toplamaları için hesaplanan betimsel istatistikler ise bulgulardan önce sunularak verilerin dağılımları incelenmiştir. Her bir araştırma sorusunun analizine ilişkin bilgiler aşağıda verilmiştir.

Birinci Araştırma Sorusuna İlişkin Veri Analizi

Birinci araştırma sorusuna ilişkin olarak Yeterlilik Şeması ile elde edilen ölçümler üzerinden geçerlik çalışmaları yürütülmüştür. Geçerlik çalışmaları bağlamında ölçüt geçerliği ve yapı geçerliği incelenmiştir. Alt araştırma sorularının analizine ilişkin bilgiler aşağıda verilmiştir.

1.1.Araştırma Sorusuna İlişkin Veri Analizi

Geçerlik çalışmaları bağlamında ilk olarak ölçüt geçerliğine bakılmış, Pearson Momentler Çarpımı Korelasyon Katsayısı hesaplanarak regresyon eşitliği kurumuştur. Bu bağlamda 156 madde içinde yalnızca esas uygulamalarda kullanılan 79 maddenin deneysel güçlük puanları bilindiğinden bu 79 madde üzerinden ölçüt geçerliği incelenmiş ve deneysel güçlük puanları geçerlik ölçütü olarak alınmıştır. Maddelerin puanlayıcıdan puanlayıcıya ve uygulamadan uygulamaya her bir kategori için farklı puanlar aldığı göz önünde tutularak her bir madde için her iki uygulamada beş farklı puanlayıcıdan aldıkları puanların ortalamaları kullanılmıştır. Hesaplanacak korelasyon katsayısına karar vermeden önce elde edilen bu ortalama puanlar için normal dağılım varsayımı ve regresyon uygulanmadan önce iki değişken arasındaki doğrusallık varsayımı SPSS programı ile analiz edilmiştir.

1.2.Araştırma Sorusuna İlişkin Veri Analizi

Geçerlik çalışmaları bağlamında ikinci olarak yapı geçerliği açımlayıcı ve doğrulayıcı faktör analizi ile incelenmiştir. Bu amaçla puanlayıcıların her bir kategoriye iki farklı zamanda verdikleri puanların ortalamaları birer gösterge olarak ele alınmış ve beş puanlayıcının altı kategoriye verdikleri puanlardan oluşan toplam 30 gösterge ve 156 madde için analizler gerçekleştirilmiştir. Açımlayıcı faktör analizi için SPSS 21.0, doğrulayıcı faktör analizi için LISREL 8.8 programı kullanılmıştır. Analizlerden önce verilerin faktör analizine uygun olup olmadıkları KMO ve Barlett testleri ile incelenmiş verilerin faktör analizine uygun olduğu sonucuna varılmıştır.

İkinci Araştırma Sorusuna İlişkin Veri Analizi

İkinci araştırma sorusu için Klasik Test Kuramı çerçevesinde Yeterlilik Şeması ile elde edilen ölçümlerin güvenirliği puanlayıcı içi ve puanlayıcılar arası güvenirlik bağlamında

incelenmiş ve analizlerde SPSS 21.0 programından yararlanılmıştır. Alt araştırma sorularının analizine ilişkin bilgiler aşağıda verilmiştir.

2.1.Araştırma Sorusuna İlişkin Veri Analizi

2.1. araştırma sorusuna yönelik olarak toplam puanlar ve alt kategoriler bazında Yeterlilik Şeması kullanılarak beş puanlayıcının iki farklı zamanda yaptığı puanlamalar üzerinden birinci ve ikinci puanlamanın her biri için puanlayıcı içi güvenilirliğin belirlenmesi amacıyla veriler normal dağılım göstermediğinden Spearman rho korelasyon katsayısı hesaplanmıştır.

2.2.Araştırma Sorusuna İlişkin Veri Analizi

2.2. araştırma sorusu için beş puanlayıcının Yeterlilik Şema'sını kullanarak iki farklı zamanda PISA matematik okuryazarlığı maddelerinin gereksinim düzeylerini belirlemesi sonucunda elde edilen verilerle puanlayıcılar arası güvenilirliğin hesaplanmasında nonparametrik bir istatistik olan Kendall'ın W katsayısı kullanılmıştır. Puanlayıcılar arası güvenilirlik katsayıları ilk olarak her bir alt kategori ve toplam puanlar için 156 maddenin birinci uygulamadan ve ikinci uygulamadan aldıkları puanlar üzerinden hesaplanmış olup son olarak maddelerin birinci ve ikinci uygulamadan aldıkları puanların ortalamaları üzerinden hesaplanmıştır.

Üçüncü Araştırma Sorusuna İlişkin Veri Analizi

Üçüncü araştırma sorusuna yönelik olarak Yeterlilik Şeması puanlama anahtarı ile beş puanlayıcıdan iki farklı zamanda elde edilen ölçümlerin güvenirliliği Genellenebilirlik Kuramı çerçevesinde çok değişkenli ve tek değişkenli model G ve K çalışmaları ile incelenmiştir. Yeterlilik Şeması puanlama anahtarı ile elde edilen puanlarla Genellenebilirlik Kuramı kapsamında toplam puanlar için madde (m), puanlayıcı (p) ve zaman (z) yüzeylerinin çaprazlandığı, $m \times p \times z$ deseni oluşturularak tek değişkenli G çalışması ve alt kategorilere ait puanlar için $m^* \times p^* \times z^*$ deseni oluşturularak çok değişkenli G çalışması yapılmıştır. Alt araştırma sorularının analizine ilişkin bilgiler aşağıda verilmiştir.

3.1.Araştırma Sorusuna İlişkin Veri Analizi

3.1. araştırma sorusu için tek değişkenli ve çok değişkenli G çalışmaları sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri m , p , z ana etkileri ile mp , mz , pz , mpz ortak etkileri bağlamında ayrı ayrı incelenmiştir. Çok değişkenli G çalışmaları sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri her bir alt kategori için ayrı ayrı açıklanmıştır.

3.2.Araştırma Sorusuna İlişkin Veri Analizi

3.2. araştırma sorusu için G kuramının çok değişkenli $m^* \times p^* \times z^*$ deseni ile puanlayıcı ve zaman sayılarının bir arttırılıp bir azaltılmasıyla oluşturulan senaryolara göre K çalışmaları yapılmıştır. K çalışmaları sonucunda alt kategoriler için kestirilen G ve Phi katsayıları ve bu katsayıların değişimi her bir kategori için ayrı ayrı incelenmiştir.

3.3.Araştırma Sorusuna İlişkin Veri Analizi

3.3. araştırma sorusuna yönelik olarak Yeterlilik Şeması ile elde edilen toplam puanlar için puanlayıcı ve zaman sayılarının bir arttırılıp bir azaltıldığı senaryolara göre tek değişkenli ve çok değişkenli Karar çalışmaları yapılmıştır. Tek değişkenli ve çok değişkenli Karar çalışmalarıyla kestirilen G ve Phi katsayıları ile bu katsayıların modellere göre değişimi incelenmiştir.

Dördüncü Araştırma Sorusuna İlişkin Veri Analizi

Araştırmanın dördüncü sorusuna yönelik olarak Yeterlilik Şeması ile elde edilen her bir kategoriye ait puanlar için zaman sayılarının bir ile üç, puanlayıcı sayılarının bir ile altı arasında değiştiği senaryolar için çok değişkenli model K çalışmaları yapılmıştır. K çalışmaları sonucunda farklı senaryolara göre kestirilen G ve Phi katsayıları incelenerek kabul edilebilir düzeyde bir G katsayısı elde etmek için gerekli minimum puanlayıcı ve zaman sayıları önerilmiştir.

Beşinci Araştırma Sorusuna İlişkin Veri Analizi

Araştırmanın beşinci sorusuna yönelik olarak Yeterlilik Şeması ile elde edilen toplam puanlar için zaman sayılarının bir ile üç, puanlayıcı sayılarının bir ile altı arasında değiştiği senaryolar için çok değişkenli ve tek değişkenli model K çalışmaları yapılmıştır. K çalışmaları sonucunda farklı senaryolara ve modellere göre kestirilen G katsayıları incelenerek kabul edilebilir düzeyde bir G katsayısı elde etmek için gerekli minimum puanlayıcı ve zaman sayıları önerilmiştir.

Altıncı Araştırma Sorusuna İlişkin Veri Analizi

Araştırmanın altıncı sorusuna yönelik olarak Yeterlilik Şeması ile elde edilen toplam puanlar ve her bir kategoriye ait puanlar için KTK ve G Kuramı çerçevesinde kestirilen güvenilirlik katsayıları karşılaştırılmıştır. KTK çerçevesinde puanlayıcı içi güvenilirlik için kestirilen Spearman rho korelasyon katsayısı ve puanlayıcılar arası güvenilirlik için kestirilen Kendall'ın W katsayısı ile G kuramı çerçevesinde kestirilen G ve Phi katsayıları iç tutarlılık ölçütlerine göre karşılaştırmıştır.

BÖLÜM IV

BULGULAR VE YORUMLAR

Bu bölümde araştırma soruları sırasıyla incelenerek bulgulara ve bulgulara ilişkin yapılan yorumlara yer verilmiştir. Bulgulara ve bulguların yorumlanmasına geçilmeden önce beş puanlayıcıya göre ayrı ayrı, 156 maddenin Yeterlilik Şeması ile iki farklı zamanda puanlanmasından elde edilen toplam puanlara ilişkin betimsel istatistikler Tablo 2’de verilmiştir.

Tablo 2 incelendiğinde P5 dışındaki diğer puanlayıcıların ilk puanlamada verdikleri puanların ortalamalarının ikinci puanlamada verdikleri puanların ortalamalarından daha düşük olduğu görülmektedir. Toplam puanlara ilişkin en düşük ortalama (5,79) P3’ün ilk puanlamasından, en yüksek ortalama (6,79) P1’in ikinci puanlamasından elde edilmiştir. Her iki puanlama incelendiğinde puanlamalara ilişkin dağılımların sağa çarpık olduğu görülmektedir. Çarpıklık katsayısının standart hatası ile bölümünden elde edilen değerlerden 1,96’nın üstünde kalan puanlamaların normal dağılım göstermediği 1,96’nın altında kalanların ise normal dağılıma benzediği söylenebilir. Buna bağlı olarak, P4 ve P5’in her iki puanlamasına ve P2’nin ilk puanlamasına ilişkin dağılımların normale yakın olduğu sonucuna varılmıştır. Bununla birlikte tüm puanlamalara ilişkin dağılımların normal dağılım gösterip göstermediğini incelemek için yapılan Kolmogorov-Smirnov uyum iyiliği testi sonuçları Ek 2’de verilmiştir. Bu test sonuçlarına göre yalnızca toplam puan ortalamalarına ilişkin dağılımın 0,05 anlamlılık düzeyinde manidar olmadığı için normal dağıldığı, diğer puanlamalara ilişkin dağılımların ise normal dağılım göstermediği gözlenmiştir. Basıklık

katsayıları incelendiğinde normal dağılıma göre P2, P3 ve P4'ün ikinci puanlamalarının daha sivri bunların dışında kalan diğer puanlamaların ise daha basık olduğu görülmektedir.

Tablo 2

İlk ve İkinci Puanlamadan Elde Edilen Toplam Puanlara İlişkin Betimsel İstatistikler

Puanlayıcı	Puanlama Zamanı	Ortalama	ss	Çarpıklık	Çarpıklık/ Çarpıklığın Standart Hatası (0,194)	Basıklık	Basıklık/ Basıklığın Standart Hatası (0,386)
P1	İlk						
	Puanlama	6,44	2,46	0,497	2,56	-0,302	-0,78
P2	İkinci						
	Puanlama	6,79	2,48	0,493	2,54	-0,142	-0,37
P3	İlk						
	Puanlama	6,18	2,49	0,265	1,37	-0,330	-0,85
P4	İkinci						
	Puanlama	6,38	2,47	0,407	2,10	0,438	1,13
P5	İlk						
	Puanlama	5,79	2,73	0,608	3,13	-0,077	-0,20
P6	İkinci						
	Puanlama	6,08	2,70	0,542	2,79	0,088	0,23
P7	İlk						
	Puanlama	6,59	2,62	0,374	1,93	-0,138	-0,36
P8	İkinci						
	puanlama	6,71	2,49	0,266	1,37	0,164	0,42
P9	İlk						
	Puanlama	6,57	2,61	0,157	0,81	-0,353	-0,91
P10	İkinci						
	Puanlama	6,57	2,42	0,204	1,05	-0,024	-0,06

Birinci Araştırma Sorusuna İlişkin Bulgular

Araştırmanın ilk sorusuna yönelik olarak PISA matematik okuryazarlığı maddelerinin gereksinim düzeylerine ilişkin Yeterlilik Şeması puanlama anahtarı ile beş puanlayıcıdan iki farklı zamanda toplanan verilerin ölçüt ve yapı geçerliği incelenmiştir.

1.1.Araştırma Sorusuna İlişkin Bulgular

“Yeterlilik Şeması ile elde edilen ölçümlerin ölçüt geçerliliği ne düzeydedir?”

Bu alt araştırma sorusu için Yeterlilik Şeması ile elde edilen ölçümlerin ölçüt geçerliği incelenmiştir. Bu bağlamda 156 madde içinde esas uygulamalarda kullanılan 79 maddenin deneysel güçlük puanlarına ulaşılmış ve bu maddelerin deneysel güçlük puanları geçerlik

ölçütü olarak alınmıştır. Bu amaçla, yayınlanmış ve esas uygulamalarda kullanılan 79 PISA matematik okuryazarlığı maddesinin deneysel uygulamasından elde edilen ve OECD (2013c) raporlarında açıklanan deneysel güçlük puanları ile maddelerin Yeterlilik Şeması ile her bir puanlayıcı tarafından iki farklı zamanda aldığı yeterlilik toplam puanların ortalamaları arasındaki korelasyona bakılmıştır. Beş puanlayıcıya ilişkin elde edilen sonuçlar farklılık gösterdiğinden tüm puanlayıcıların verdikleri puanların ortalamaları üzerinden korelasyon hesaplanmıştır. Öncelikle hangi korelasyon katsayısının hesaplanacağını belirlemek amacıyla maddelerin güçlük puanları ile toplam puan ortalamalarının normal dağılım gösterip göstermediğini belirlemek amacıyla Kolmogorov-Smirnov ve Shapiro-Wilk testine başvurulmuştur. Örneklem 50'den büyük olduğu için normal dağılım varsayımı Kolmogorov-Smirnov uyum iyiliği testi ile incelenmiştir. Test sonuçlarının anlamlı olması verinin normal dağılımdan farklı olduğu şeklinde yorumlanırken, p değerinin 0,05'den yüksek olması verilerin dağılımının normal dağılıma yakın olduğunu göstermektedir (Field, 2009, s. 144). Yapılan analiz sonucu Tablo 3'te verilmiştir.

Tablo 3.

Güçlük Puanı ve Madde Puanına İlişkin Kolmogorov-Smirnov ve Shapiro Wilk Testi Sonuçları

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	İstatistik	sd	p	İstatistik	sd	p
Toplam puanlar ortalaması	0,069	79	0,200*	0,984	79	0,453
Güçlük Puanı	0,060	79	0,200*	0,989	79	0,761

Maddelerin deneysel uygulamadan elde edilen güçlük puanlarının ve toplam puanların ortalamalarının normal dağılım gösterip göstermediğini test etmek amacıyla yapılan Kolmogorov-Smirnov uyum iyiliği istatistiği sonucunda her iki değişken içinde dağılımların normal dağıldığı gözlenmiştir. Veriler normal dağılım gösterdiği için iki değişken arasındaki korelasyon Pearson Momentler Çarpımı Korelasyon Katsayısı ile incelenmiştir. Yapılan analiz sonucu Tablo 4'te verilmiştir.

Tablo 4

Güçlük Puanları ve Madde Puanları Arasındaki Pearson Korelasyon Katsayısı

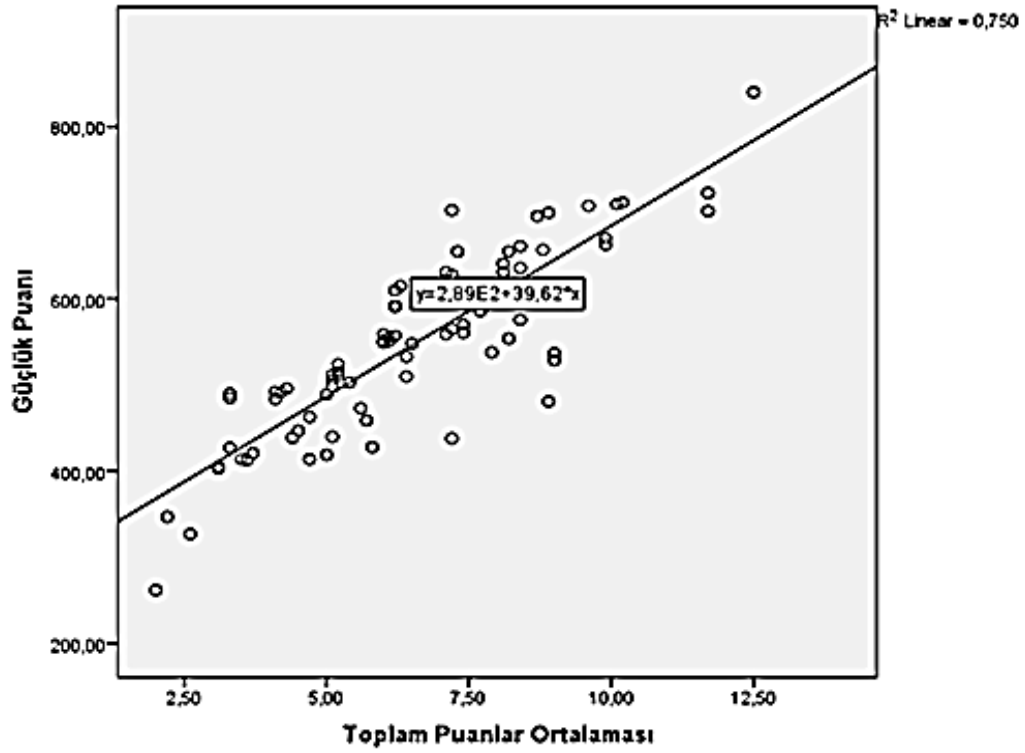
			Toplam Puanlar Ortalaması	Güçlük Puanı
Toplam Ortalaması	Puanlar	r	1	0,866**
		Sig. (çift yönlü)		,000
		N	79	79
Güçlük Puanı		r	0,866**	1
		Sig. (çift yönlü)	,000	
		N	79	79

**Korelasyon 0.01 düzeyinde (çift yönlü) anlamlıdır.

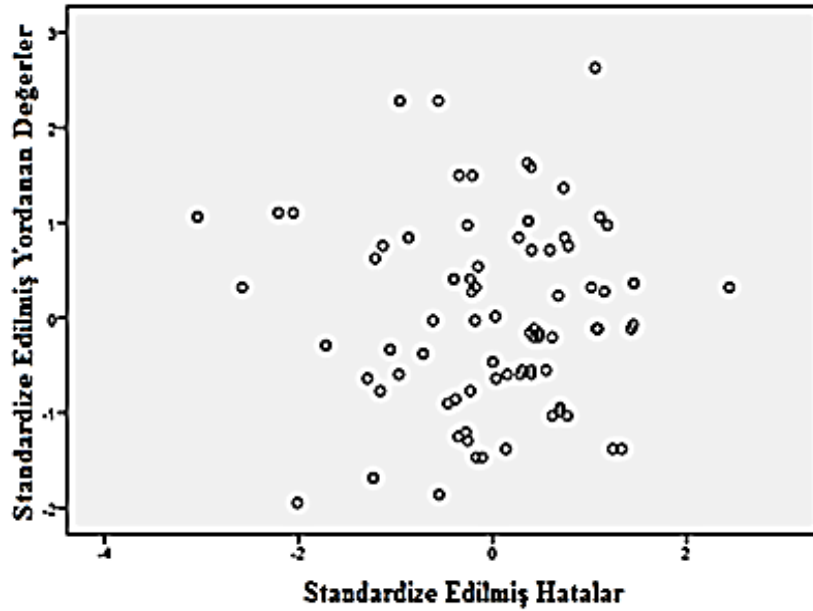
Tablo 4'te görüldüğü gibi değişkenler arasındaki Pearson korelasyon katsayısı 0,01 düzeyinde çift yönlü olarak manidar bulunmuştur. Ayrıca korelasyon katsayısının 0,866 olması maddelerin toplam puanlar ortalaması ile güçlük puanları arasında kuvvetli ve pozitif yönde bir ilişki olduğunu göstermektedir. Buna dayanarak ölçüt geçerliğinin sağlandığı söylenebilir. Ölçüt olarak seçilen madde güçlük puanları bilinmesine karşın madde yeterlilik puanlarının belirlenmesinin de önemli olduğu düşünülmektedir. Bir maddenin güçlük puanı deneysel uygulama yapıldıktan sonra belirlenebilmekte, uygulamada öğrencilerden toplanan veriler üzerinden kestirilmektedir. Buna karşın madde yeterlilik puanları daha maddenin yazım aşamasındayken belirlenebilir. Bu durum uygulamaya alınacak maddelerin güçlük puanlarının önceden kestirilebilmesine olanak sağlamaktadır. Ayrıca madde güçlük puanı bir maddenin gerektirdiği matematiksel yeterliliklerin düzeyini belirlemek için kullanılamazken, maddelerin yeterlilik gereksinimlerinin belirlenmesi bir maddenin hangi yeterliliği ne düzeyde gerektirdiğini ortaya çıkarmak için kullanılabilir.

Ölçüt geçerliğine ilişkin bu kanıtı ek olarak yordama geçerliği bağlamında maddelerin beş puanlayıcının her biri tarafından iki farklı zamanda puanlanmasıyla elde edilen toplam yeterlilik puanları ortalamasının madde güçlük puanını yordayıp yordamadığını belirlemek amacıyla basit doğrusal regresyon analizi yapılmıştır. Basit doğrusal regresyon analizi için değişkenlerin normal dağılım göstermesi ve aralarında pozitif yönlü kuvvetli bir ilişki bulunması basit doğrusal regresyon analizinin varsayımlarının karşılandığını göstermektedir. Ayrıca bağımlı değişken ve bağımsız değişkenlerin sürekli olması da basit doğrusal regresyon analizinin varsayımlarından bir diğerini karşılamaktadır. Bunun yanı sıra doğrusallık varsayımının test edilmesi için Şekil 3'te verilen serpm diyagramı ile

standardize edilmiş yordanan değer ve standardize edilmiş hata grafiği incelenmiştir (Tan, 2016, s.248).



Şekil 3.a. Toplam Puanlar Ortalaması ve Güçlük Puanları için Serpme Diyagramı



Şekil 3.b. Standardize Edilmiş Yordanan Değer ve Standardize Edilmiş Hata Grafiği

Şekil 3. Toplam puanlar ortalaması ve güçlük puanları için serpme diyagramı ve standardize edilmiş yordanan değer ve standardize edilmiş hata grafiği

Şekil 3'e bakıldığında madde toplam puanlar ortalaması ile madde güçlük puanları arasında doğrusal ilişki olduğu görülmektedir. Ayrıca standardize edilmiş yordanan değer ve standardize edilmiş hata grafiği noktalar sıfır ekseninde toplandığından doğrusal ilişkinin varlığını kanıtlamaktadır. Basit doğrusal regresyon analizinin varsayımlarının sağlandığı görülmüş olup basit doğrusal regresyon analizine geçmeden önce bu iki değişkene ait betimsel istatistikler aşağıda Tablo 5'te verilmiştir.

Tablo 5

Yeterlilik Puanları ve Güçlük Puanlarına İlişkin Betimsel İstatistikler

	Betimsel İstatistikler		
	$\bar{X}$	ss	n
Toplam Puanlar Ortalaması	6,47	2,29	79
Güçlük Puanı	545,09	104,92	79

Madde toplam puanlar ortalamasına ilişkin betimsel istatistikler incelendiğinde madde puanları ortalamasının 6,47, standart sapmasının 2,29 olduğu görülmektedir. Madde güçlük puanlarının ise 545,09 ortalamaya ve 104,92 standart sapmaya sahip olduğu görülmüştür. Bir sonraki aşamada yapılan basit doğrusal regresyon modelinin özeti Tablo 6'da verilmiştir.

Tablo 6

Yeterlilik Puanları ile Güçlük Puanları Arasındaki Yordama İlişkisi için Model Özeti

Bağımsız Değişken	r	R ²	Düzeltilmiş R ²	Tahminin Standart Hatası	Durbin-Watson
Toplam puanlar ortalaması	0,866	0,750	0,747	52,79307	2,007

Tablo 6 incelendiğinde toplam puanlar ortalaması ve güçlük puanları arasındaki korelasyonun (r) 0,866 olduğu ve açıklanan varyansın (Düzeltilmiş R²) 0,747 olduğu görülmüştür. Bu değerler güçlük puanlarındaki varyansın %74,7'sinin toplam puanlar ortalaması tarafından açıklandığına işaret etmektedir. Bununla birlikte tabloda yer verilen Durbin-Watson istatistiği artıkların ilişkisizliğinin belirlenmesinde kullanılmakta olup bu değer 0 ile 4 aralığında değiştiği ve 2 değerini aldığı anda artıkların ilişkisiz olduğu

belirtilmektedir (Field, 2009, s. 210). Bu bilgiye dayanarak toplam puanlar ortalamasının ve güçlük puanlarının artıkları arasında ilişki olmadığı sonucuna varılmıştır.

Tablo 7’de standartlaştırılmamış regresyon eşitliği için a ve b katsayılarına, standartlaştırılmış regresyon eşitliği için β katsayısına ve bu katsayıların manidarlıklarına yer verilmiştir.

Tablo 7

Basit Doğrusal Regresyon Eşitliğine İlişkin Katsayılar ve Manidarlıkları

Değişken	B	β	t	p
Sabit Sayı (a)	288,872		16,164	0,000
Toplam puanlar ortalaması (b)	39,618	0,866	15,201	0,000

F=231,078; p<0,01

Tablo 7 incelendiğinde yordamanın 0,01 düzeyinde manidar olduğu görülmüştür. Standartlaştırılmış regresyon eşitliği için β katsayısı 0,866 olarak bulunmuş ve regresyon eşitliği “Güçlük Puanları = $0,866 \times$ Toplam Puanlar Ortalaması” şeklinde elde edilmiştir. Standartlaştırılmamış regresyon eşitliği için a katsayısı 288,872 ve b katsayısı 39,618 olarak bulunmuştur. Buna bağlı olarak standartlaştırılmamış regresyon eşitliği “Güçlük Puanları = $288,872 + 39,618 \times$ Toplam Puanlar Ortalaması” şeklindedir. Bu modele göre yeterlilik puanlarındaki bir puanlık artışın madde güçlük puanında 39,618 puanlık bir artışa karşılık geldiği görülmektedir. Maddenin toplam yeterlilik puanı ortalaması sıfır olduğunda güçlük puanı 288,872 değerini almaktadır.

2.1.Araştırma Sorusuna İlişkin Bulgular

“Yeterlilik Şeması puanlama anahtarının yapı geçerliği ne düzeydedir?”

Bu alt problem için Yeterlilik Şeması ile elde edilen ölçümlerin yapı geçerliği açıklayıcı ve doğrulayıcı faktör analizi ile incelenmiştir. Yeterlilik Şeması puanlama anahtarında kategoriler altında göstergeler yer almadığı için her bir puanlayıcının her bir kategoriye vermiş olduğu puanların ortalamaları (ilk ve ikinci uygulama puanlarının ortalaması) göstergeler olarak ele alınmıştır. Böylelikle toplam altı kategori için beş puanlayıcının verdiği puanlar ile 30 gösterge olduğu kabul edilerek faktör analizi çalışmaları

yürütülmüştür. Açımlayıcı faktör analizi çalışmasıyla faktör yapıları belirlenmiş ve doğrulayıcı faktör analizi ile modelin veriye uyum sağlayıp sağlamadığı test edilmiştir.

Açımlayıcı faktör analizinde ilk olarak örneklem büyüklüğünün korelasyon güvenirliğini sağlayacak ölçüde olup olmadığının ve verilerin faktör analizi yapmaya uygun olup olmadığının bir göstergesi olan Kaiser-Meyer-Olkin (KMO) ve Barlett Küresellik testleri sonucunda KMO değerinin 0,89 olduğu ve Barlett Küresellik testinin χ^2 değeri 3661,209 ile 0,001 düzeyinde anlamlı olduğu bulunmuştur. Çokluk, Şekercioğlu ve Büyüköztürk (2014) KMO testi sonucunda değerin 0,50'den büyük olması durumunda analize devam edilebileceğini (s. 207) ve Barlett Küresellik testi anlamlı ise verilerin çok değişkenli normal dağılımdan geldiğini (s.208) belirtmişlerdir. Bu sonuçlar verilere faktör analizi uygulanabileceğini göstermiştir.

Açımlayıcı faktör analizi çalışmasında her bir puanlayıcının her bir kategoriye iki farklı zamanda verdiği puanların ortalamasından oluşan puanlamaların iki baskın faktörlü bir yapı sergilediği ve iki göstergenin (P3_MU, P5_SP) birden fazla faktörde yüksek faktör yüklerine sahip olduğu görülmüştür. Bu göstergelerin iki faktördeki yükleri arasındaki farkın 0,10'dan küçük olduğu için çıkarılmasına karar verilmiştir. 28 göstergeye ilişkin döndürme olmaksızın temel bileşenler analiziyle yapılan açımlayıcı faktör analizleri ile elde edilen faktör yük değerleri Tablo 8'de verilmiştir. Tablo 8 incelendiğinde 28 göstergeden P1_SM ile P3_ILT göstergelerinin birden fazla faktörde yüklendiği ve özdeğeri 1'den büyük altı faktör olmasına karşın genel olarak iki baskın faktörün olduğu görülmektedir. Bu sonuç doğrultusunda daha fazla gösterge analizden çıkarılmadan 28 gösterge Varimax döndürme yöntemi ve temel bileşenler analizi yöntemiyle analiz edilmiştir.

P3_MU ve P5_SP göstergeleri çıkartılarak Varimax döndürme yöntemi ve temel bileşenler analizi ile açımlayıcı faktör analizi tekrarlanmıştır. Varimax döndürme yöntemi ile yapılan analizde faktör yük değerleri 0,40 ve üzerinde olan göstergeler alındığında 28 göstergenin altı faktörlü bir yapı sergilediği ve faktörlerde yer alan göstergelerin Yeterlilik Şeması alt kategorilerine paralel olduğu ortaya çıkmıştır. Örneğin birinci faktörde puanlayıcıların temsil kategorisine ilişkin verdikleri puanları temsil eden göstergeler (P5_TM, P4_TM, P2_TM, P1_TM, P3_TM) toplanırken, ikinci faktörde matematikleştirme kategorisine ilişkin verdikleri puanları (P4_MT, P2_MT, P5_MT, P1_MT, P3_MT) temsil eden göstergelerin yer aldığı görülmüştür.

Tablo 8

Temel Bileşenler Analizi ile Yapılan Açımlayıcı Faktör Analizi Sonucunda Göstergelerin Faktörlere Göre Dağılımı ve Faktör Yük Değerleri

	Bileşenler Matrisi^a					
	Bileşenler					
	1	2	3	4	5	6
P1_SP	,817					
P4_SP	,807					-,449
P2_SP	,770					-,423
P5_SM	,754					
P2_ILT	,732					
P3_SP	,719					-,467
P4_MT	,700		-,412			
P2_MT	,692		-,414			
P3_MT	,688					
P5_MT	,674					
P5_ILT	,671					
P4_SM	,671					
P1_MT	,652		-,405			
P4_MU	,639					
P1_ILT	,635		,462			
P2_MU	,631					
P4_ILT	,631		,499			
P2_SM	,624					
P3_SM	,602				,482	
P5_MU	,549	,409				
P1_MU	,516					
P1_SM	,516			,444		
P4_TM		,899				
P5_TM		,886				
P3_TM		,843				
P1_TM		,842				
P2_TM		,697			,511	
P3_ILT	,535		,544			

Yeterlilik Şeması ile beş puanlayıcıdan elde edilen puanların ortalamalarına ait Varimax döndürme yöntemi ile temel bileşenler analizi sonucuna göre faktörlerin öz değerleri, faktörlerle açıklanan varyans yüzdeleri ve faktörlerle açıklanan yığılmalı varyans yüzdeleri aşağıdaki Tablo 9’da verilmiştir.

Temel bileşenler analizinde öz değeri 1,00’den büyük olan altı faktörden birinci ve ikinci faktörlerin özdeğerlerin diğer faktörlerin özdeğerlerinden oldukça büyük olduğu ve altı faktörün birlikte toplam varyansın % 77,22’sini açıkladığı görülmektedir. Bu sonuçlara

dayanarak Yeterlilik Şeması puanlama anahtarının çok değişkenli bir yapı sergilediği söylenebilir.

Tablo 9

Yeterlilik Şeması ile Beş Puanlayıcıdan Elde Edilen Puanların Ortalamalarına ait Açıklayıcı Faktör Analizi Sonuçları

Faktörler	Öz değerler	Açıklanan Varyans Yüzdeleri	Açıklanan Yıgmal Varyans Yüzdeleri
1	10,300	14,432	14,432
2	4,735	14,137	28,569
3	2,014	13,513	42,082
4	1,784	13,450	55,533
5	1,680	11,974	67,507
6	1,110	9,714	77,220

Varimax döndürme yöntemi sonucunda göstergelerin faktör yük değerleri ve hangi faktöre bağlandıkları aşağıdaki Tablo 10'da görülmektedir. Tablo 10 incelendiğinde ilişkili puanlamaların aynı faktörler altında toplandığı ve faktör yüklerinin 0,887 ile 0,678 arasında değiştiği görülmektedir.

Maddelerin matematik okuryazarlığı yeterlilik gereksinimlerine ilişkin 30 gösterge üzerinden doğrulayıcı faktör analizi ile test edilen teorik model Şekil 4'te verilmiştir. Verinin modele uyum sağlayıp sağlamadığını test etmek için yapılan analizde çok değişkenli normallik sağlanmasına (Relative Multivariate Kurtosis =1.099>1 olduğundan) karşın tek değişkenli normallik sağlanmadığı için Robust Maximum Likelihood kestirim yöntemi kullanılmıştır. Analiz sonucunda göstergelere ait 0,3'den düşük faktör yükü ve 0,9'dan büyük disturbance bulunmadığı görülmüştür. Bununla birlikte göstergelere ait faktör yük değerlerinin 0,71 ile 0,93 arasında değiştiği 0,30'dan yüksek olduğu ve t değerlerinin de anlamlı olduğu görülmektedir. Bu noktada açıklayıcı faktör analizi sonucuna göre birden fazla faktörde yüksek yük değerleri aldığından çıkarılması önerilen iki göstergenin (P3_MU, P5_SP) doğrulayıcı faktör analizinde faktör yük değerlerinin sırasıyla 0,67 ve 0,84 olduğu ve analizden çıkarılması gerekmediği görülmüştür. Bu nedenle hiçbir gösterge analizden çıkarılmamıştır.

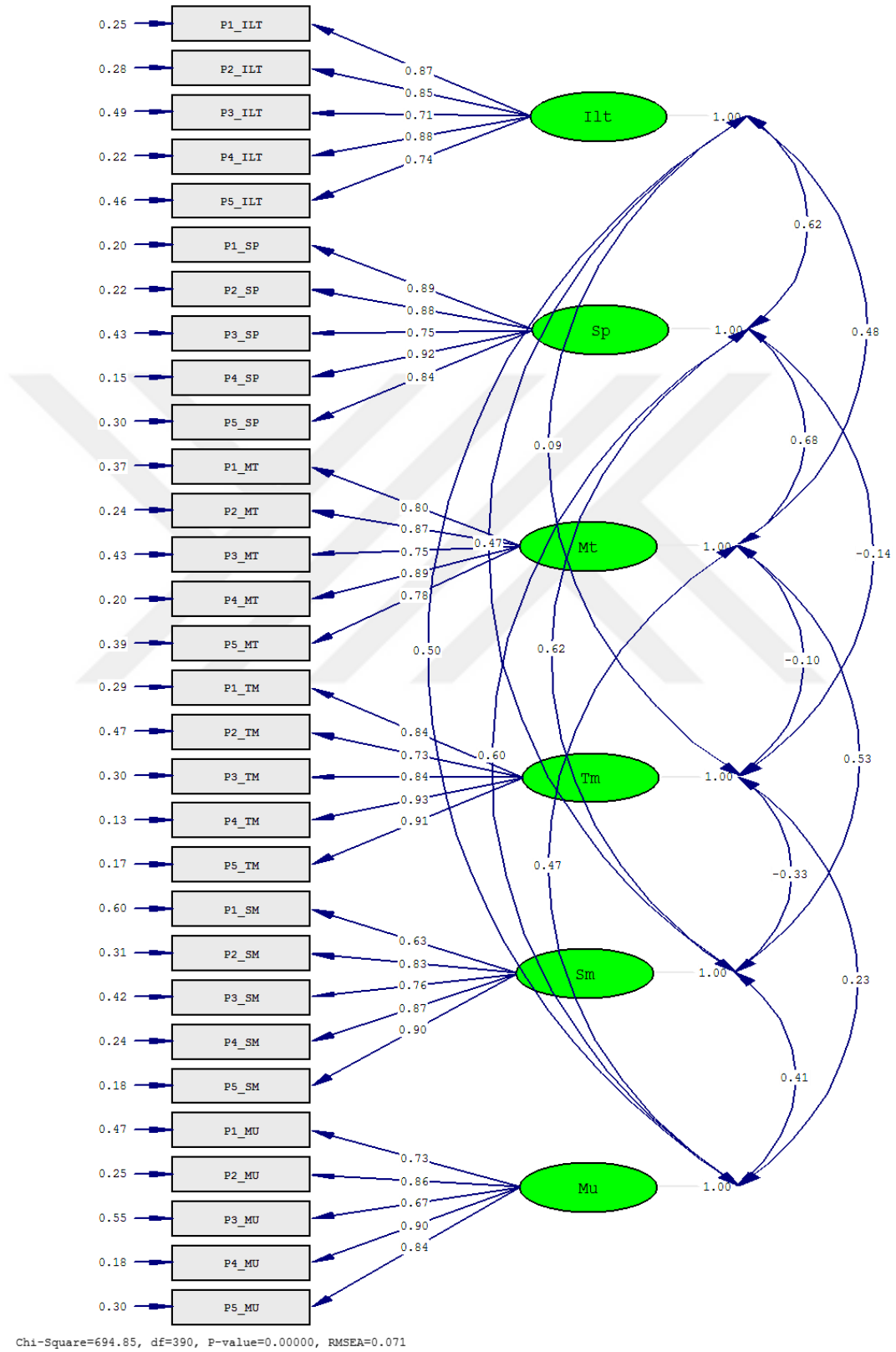
Tablo 10

Varimax Döndürme Yöntemi ve Temel Bileşenler Analizi ile Yapılan Açımlayıcı Faktör Analizi Sonucunda Göstergelerin Faktörlere Göre Dağılımı ve Faktör Yük Değerleri

	Döndürülmüş Bileşenler Matrisi ^a					
	1	2	3	4	5	6
P5_TM	,887					
P4_TM	,878					
P2_TM	,858					
P1_TM	,856					
P3_TM	,838					
P4_MT		,852				
P2_MT		,832				
P5_MT		,771				
P1_MT		,763				
P3_MT		,722				
P4_SM			,819			
P2_SM			,791			
P3_SM			,780			
P5_SM			,779			
P1_SM			,678			
P4_ILT				,859		
P1_ILT				,847		
P2_ILT				,759		
P3_ILT				,747		
P5_ILT				,707		
P5_MU					,844	
P4_MU					,800	
P1_MU					,791	
P2_MU					,782	
P4_SP						,758
P2_SP						,727
P3_SP						,720
P1_SP						,705
a. Döndürme 6 iterasyonla oluşturulmuştur.						

Uyum iyiliği indeksleri incelendiğinde $\chi^2/sd = 1,78$ ($p=0,0$) değeri 2'den küçük olduğu için mükemmel uyuma, RMSEA= 0,071 ile 0,08'den küçük olduğu için iyi uyuma işaret etmektedir (Hooper, Coughlan ve Mullen, 2008). NNFI = 0,97 ve CFI=0,97 olup 0,95'den büyük olduğu için mükemmel uyuma, Standartlaştırılmış RMR 0,065 olup 0,08'den küçük olduğu için iyi uyuma işaret etmektedir. GFI ve AGFI değerleri örneklem büyüklüğünden etkilendiği ve örneklem 156 madde ile sınırlı olduğu için yorumlanmamıştır. Uyum iyiliği

indeksi değerleri modelin veriye iyi uyum sağladığını ve modelin doğrulandığını göstermektedir.



Şekil 4. Doğrulayıcı faktör analizi ile test edilen teorik model

İkinci Araştırma Sorusuna İlişkin Bulgular

Araştırmanın ikinci sorusuna yönelik olarak PISA matematik okuryazarlığı maddelerinin gereksinim düzeylerine ilişkin Yeterlilik Şeması puanlama anahtarı ile beş puanlayıcıdan iki farklı zamanda toplanan verilerin güvenilirliği toplam puanlar ve her bir kategoriye ait puanlar için Klasik Test Kuramı çerçevesinde puanlayıcı içi ve puanlayıcılar arası güvenilirlik bağlamında incelenmiştir.

2.1.Araştırma Sorusuna İlişkin Bulgular

“Yeterlilik Şeması ile elde edilen toplam puanlar ve her bir kategoriye ait puanlar için puanlayıcı içi tutarlılık derecesi nedir?”

Bir testin iki farklı zamanda uygulanmasına dayalı olarak elde edilen ölçümlerin güvenilirliğinin belirlenmesinde test tekrar test güvenilirliği kapsamında genellikle Pearson Momentler Çarpımı Korelasyon Katsayısına başvurulmaktadır. Ancak Pearson korelasyon katsayısı aralarındaki ilişki belirlenecek değişkenlerin normal dağılım göstermesi varsayımına dayalıdır (Field, 2009, s.177). Bu çalışmada her bir puanlayıcıdan toplam puanlar ve alt kategoriler için elde edilen puanlamaların normal dağılım göstermediği EK 2’de verilen Kolmogorov-Simirnov ve Shapiro-Wilk testi sonuçlarında görülmektedir. Bu nedenle puanlayıcı içi güvenirliliğin belirlenmesinde normal dağılım varsayımının ihlal edilebildiği durumlarda kullanılabilen nonparametrik bir istatistik olan (Field, 2009, s.179) Spearman rho korelasyon katsayısından yararlanılmıştır. Yeterlilik Şeması kullanılarak beş puanlayıcının iki farklı zamanda yaptığı puanlamalar üzerinden birinci ve ikinci puanlama arasında puanlayıcı içi güvenirliliğin belirlenmesi amacıyla hesaplanan Spearman rho korelasyon katsayıları Tablo11’de sunulmuştur.

Tablo 11 incelendiğinde tüm alt kategoriler ve toplam puanlar için hesaplanan korelasyon katsayılarının tamamının 0,01 düzeyinde çift yönlü olarak anlamlı olduğu ve biri dışında tamamının 0,70’den büyük olduğu dolayısıyla Yeterlilik Şeması ile elde edilen ölçümler için bu ölçme durumunda puanlayıcı içi güvenirliliğin yüksek olduğu söylenebilir. Ancak Tablo 11’de P1’in sembollerin, işlemlerin ve formal dilin kullanımı kategorisine verdiği puanlara ilişkin puanlayıcı içi güvenirlilik katsayısının 0,421 ile diğer puanlayıcılara göre oldukça düşük olduğu göze çarpmaktadır. Bu durum P1’in maddelerin sembollerin işlemlerin ve formal dilin kullanımı kategorisine ilişkin gereksinimlerini belirlerken ilk uygulamada ve ikinci uygulamada maddelere verdiği puanlar arasında önemli ölçüde tutarsızlık olduğunu

göstermektedir. Bununla birlikte korelasyon katsayısının 0,421 olması ilk iki puanlama arasında düşük düzeyde bir ilişki olduğunu göstermektedir. Bu tutarsızlığın, puanlayıcının puanlama zamanında içinde bulunduğu koşullardan, sağlık durumundan, yorgunluk, dalgınlık gibi fiziksel nedenlerden ya da iş yoğunluğundan kaynaklanabileceği gibi iletişim kategorisine yönelik ilk puanlama esnasında algısı ile ikinci puanlama esnasındaki algısı arasındaki farklılıktan da kaynaklanmış olabileceği düşünülmüştür.

Tablo 11

Yeterlilik Şeması Puanlama Anahtarına İlişkin Puanlayıcılar İçeri Güvenirlik Katsayıları

Kategoriler	P1	P2	P3	P4	P5
İletişim	0,845	0,936	0,889	0,829	0,725
Strateji Planlama	0,934	0,907	0,919	0,852	0,699
Matematikleştirme	0,930	0,842	0,911	0,800	0,705
Temsil	0,951	0,954	0,904	0,936	0,756
Sembollerin, İşlemlerin ve Formal Dilin Kullanımı	0,421	0,892	0,928	0,877	0,698
Muhakeme ve Argüman	0,919	0,861	0,909	0,879	0,752
Toplam Puan	0,894	0,939	0,946	0,902	0,799

Tüm korelasyonlar 0,01 düzeyinde anlamlıdır.

Yeterlilik Şeması'nın alt kategorileri ve toplam puanlar için hesaplanan puanlayıcı içeri güvenirlik katsayılarının 0,421 ile 0,954 arasında değiştiği, puanlayıcı içeri güvenirliliğin en yüksek olduğu alt kategorinin temsil, en düşük olduğu alt kategorinin ise sembollerin işlemlerin ve formal dilin kullanımı kategorisi olduğu söylenebilir. Bununla birlikte P5 için hesaplanan güvenirlik katsayılarının diğer puanlayıcılar için hesaplananlara göre oldukça düşük olduğu görülmektedir. Bu durum P5'in madde gereksinimlerini belirlemede zorlandığı şeklinde yorumlanabilir. Güvenirlik katsayılarının genellikle 0,70'den yüksek olması puanlayıcıların maddelerin matematik okuryazarlığına ve alt kategori yeterliliklerine ilişkin gereksinimleri kolay bir şekilde belirleyebildiklerine ilişkin bir gösterge olarak ele alınabilir ve puanlama anahtarının kararlılığının yüksek olduğu söylenebilir.

Korelasyon katsayıları güvenirlik için bir ölçü olarak ele alınabilse de ortalamadan bağımsız hesaplandığı için puanlayıcıların verdikleri puanların ortalamaları arasındaki farklılıkları

hesaba katmamaktadır. Bu nedenle Goodwin (2001), korelasyon katsayısının hesaplanmasına ek olarak puanlayıcıların verdikleri puanların ortalaması arasındaki farklılıkların test edilmesini önermiştir.

Aşağıdaki Tablo 12’de hesaplanan her bir korelasyon katsayına paralel olarak puanların ortalamaları arasındaki farklılıkların ilişkili örneklem t testi ile test edilmesine ilişkin sonuçlar sunulmuştur.

Tablo 12 incelendiğinde puanlayıcıların farklı zamanlarda verdikleri puanların ortalamaları arasında farklılık olan ve olmayan durumların olduğu görülmektedir. Kategoriler bazında bakıldığında iletişim kategorisinde tüm puanlayıcıların farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı bir farklılık varken strateji planlama kategorisi için yalnızca P5’in, matematikleştirme kategorisinde yalnızca P3’ün farklı zamanlarda verdiği puanların ortalamaları arasında anlamlı farklılık olduğu görülmektedir. Temsil kategorisinde yalnızca P1 ve P5’in farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı farklılık görülürken, sembollerin, işlemlerin ve formal dilin kullanımı kategorisinde yalnızca P1 ve P5’in farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı bir farklılık olmadığı görülmüştür. Muhakeme kategorisi için tüm puanlayıcıların, toplam puanlar bazında ise yalnızca P4 ve P5’in farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı bir farklılık olmadığı sonucuna varılmıştır.

Genel olarak bakıldığında sembollerin, işlemlerin ve formal dilin kullanımı kategorisi ile toplam puanlar için puanlayıcıların çoğunun farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı bir farklılık olduğu, diğer kategoriler içinse puanlayıcıların çoğunun farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı bir farklılık olmadığı söylenebilir.

Tablo 12

Puanlayıcıların Verdikleri Puanların Ortalamaları Arasındaki Farklara İlişkin İlişkili Örneklem T Testi Sonuçları

Kategori	Puanlayıcılar	Fark	t	sd	p
İletişim	P1_ILT1 - P1_ILT2	-0,08333	2,768	155	,006
	P2_ILT1 - P2_ILT2	-0,03846	2,146	155	,033
	P3_ILT1 - P3_ILT2	-0,08974	3,634	155	,000
	P4_ILT1 - P4_ILT2	0,08333	2,768	155	,006
	P5_ILT1 - P5_ILT2	0,07051	2,141	155	,034
Strateji Planlama	P1_SP1 - P1_SP2	-0,03205	1,514	155	,132
	P2_SP1 - P2_SP2	-0,00641	0,257	155	,797
	P3_SP1 - P3_SP2	0,01282	0,533	155	,595
	P4_SP1 - P4_SP2	-0,05128	1,642	155	,103
	P5_SP1 - P5_SP2	0,12821	3,025	155	,003
Matematikleştirme	P1_MT1 - P1_MT2	-0,02564	1,267	155	,207
	P2_MT1 - P2_MT2	-0,03205	1,148	155	,253
	P3_MT1 - P3_MT2	-0,07051	2,907	155	,004
	P4_MT1 - P4_MT2	-0,00641	0,199	155	,842
	P5_MT1 - P5_MT2	0,00000	0,000	155	1,000
Temsil	P1_TM1 - P1_TM2	-0,05128	2,575	155	,011
	P2_TM1 - P2_TM2	-0,01923	0,904	155	,367
	P3_TM1 - P3_TM2	-0,04487	1,614	155	,109
	P4_TM1 - P4_TM2	-0,02564	1,070	155	,286
	P5_TM1 - P5_TM2	-0,18590	4,377	155	,000
Sembollerin, İşlemlerin ve Formal Dilin Kullanımı	P1_SM1 - P1_SM2	-0,11538	1,871	155	,063
	P2_SM1 - P2_SM2	-0,05769	2,357	155	,020
	P3_SM1 - P3_SM2	-0,05769	2,357	155	,020
	P4_SM1 - P4_SM2	-0,07692	2,739	155	,007
	P5_SM1 - P5_SM2	-0,03846	0,904	155	,367
Muhakeme ve Argüman	P1_MU1 - P1_MU2	-0,04487	1,959	155	,052
	P2_MU1 - P2_MU2	-0,05128	1,901	155	,059
	P3_MU1 - P3_MU2	-0,04487	1,821	155	,071
	P4_MU1 - P4_MU2	-0,04487	1,534	155	,127
	P5_MU1 - P5_MU2	0,02564	0,602	155	,548
Toplam Puan	P1_TOP1 - P1_TOP2	-0,35256	4,003	155	,000
	P2_TOP1 - P2_TOP2	-0,20513	2,895	155	,004
	P3_TOP1 - P3_TOP2	-0,29487	4,363	155	,000
	P4_TOP1 - P4_TOP2	-0,12179	1,458	155	,147
	P5_TOP1 - P5_TOP2	0,00000	0,000	155	1,000

2.2.Araştırma Sorusuna İlişkin Bulgular

“Yeterlilik Şeması ile elde edilen toplam puanlar ve her bir kategoriye ait puanlar için puanlayıcılar arası tutarlılık derecesi nedir?”

Bu alt araştırma sorusu için beş puanlayıcının Yeterlilik Şema'sını kullanarak iki farklı zamanda PISA matematik okuryazarlığı maddelerinin gereksinim düzeylerini belirlemesi sonucunda elde edilen verilerle, puanlayıcılar arası güvenilirliğin hesaplanmasında nonparametrik bir istatistik olan Kendall'ın W katsayısı kullanılmıştır. Puanlayıcılar arası güvenilirlik katsayıları ilk olarak her bir alt kategori ve toplam puanlar için 156 maddenin birinci uygulamadan ve ikinci uygulamadan aldıkları puanlar üzerinden hesaplanmış olup, bu katsayılar aşağıdaki Tablo 13'te sunulmuştur.

Tablo 13

İki Farklı Puanlama Zamanı için Yeterlilik Şeması Puanlama Anahtarına İlişkin Puanlayıcılar Arası Güvenirlik (Kendall'ın W) Katsayıları

Kategoriler	Uygulama Zamanı	Kendall W.	p
İletişim	1. Zaman	0,648	0,00
	2. Zaman	0,724	0,00
Strateji Planlama	1. Zaman	0,716	0,00
	2. Zaman	0,775	0,00
Matematikleştirme	1. Zaman	0,655	0,00
	2. Zaman	0,739	0,00
Temsil	1. Zaman	0,738	0,00
	2. Zaman	0,757	0,00
Sembollerin, İşlemlerin ve Formal Dilin Kullanımı	1. Zaman	0,593	0,00
	2. Zaman	0,723	0,00
Muhakeme ve Argüman	1. Zaman	0,675	0,00
	2. Zaman	0,701	0,00
Toplam Puan	1. Zaman	0,787	0,00
	2. Zaman	0,846	0,00

Tablo 13 incelendiğinde Yeterlilik Şeması'nın alt kategorileri ve toplam puanlar için hesaplanan puanlayıcılar arası güvenilirlik katsayılarının 0,648 ile 0,846 arasında değiştiği görülmektedir. Güvenirlik katsayıları anlamlı olduğundan, toplam puanlar ve alt kategoriler için birinci ve ikinci uygulamada puanlayıcılar arası güvenirliliğin yüksek ve 0,01 düzeyinde anlamlı olduğu, diğer bir deyişle puanlayıcılar arasındaki uyumun olduğu söylenebilir. Bununla birlikte her bir kategori ve toplam puanlar için birinci uygulama zamanındaki puanlayıcılar arası güvenilirlik katsayılarının ikinci uygulama zamanındakilere göre daha düşük olduğu görülmektedir. Bu durum puanlama anahtarının kullanımına aşina olmanın puanlayıcılar arası güvenirliliği arttırdığı şeklinde yorumlanabilir. Ayrıca birinci uygulama

zamanında puanlayıcılar arası uyumun en yüksek olduğu kategori temsil kategorisiyken, en düşük olduğu kategori sembollerin, işlemlerin ve formal dilin kullanımı kategorisidir. İkinci uygulama zamanında ise puanlayıcılar arası güvenilirliğin strateji planlama kategorisinde en yüksek, muhakeme ve argüman kategorisinde en düşük olduğu görülmektedir. Her bir puanlayıcının iki farklı zamanda her bir kategoriye yönelik verdiği puanların ortalaması alınarak puanlama anahtarının alt kategorileri ve toplam puanlar için puanlayıcılar arası güvenilirliğe ilişkin hesaplanan Kendall'ın W katsayıları aşağıdaki Tablo 14' te verilmiştir.

Tablo 14

Puanlayıcıların İki Farklı Zamanda Her Bir Kategoriye Yönelik Verdiği Puanların Ortalaması Üzerinden Hesaplanan Yeterlilik Şeması Puanlama Anahtarına İlişkin Puanlayıcılar Arası Güvenirlik (Kendall'ın W) Katsayıları

Kategoriler	Kendall W.	p
İletişim	0,711	0,00
Strateji Planlama	0,784	0,00
Matematikleştirme	0,721	0,00
Temsil	0,781	0,00
Sembollerin, İşlemlerin ve Formal Dilin Kullanımı	0,704	0,00
Muhakeme ve Argüman	0,709	0,00
Toplam Puan	0,833	0,00

Tablo 14 incelendiğinde her bir puanlayıcının iki farklı zamanda her bir kategoriye yönelik verdiği puanların ortalaması üzerinden Yeterlilik Şeması'nın alt kategorileri ve toplam puanlar için hesaplanan puanlayıcılar arası güvenirlilik katsayılarının 0,704 ile 0,833 arasında değiştiği görülmektedir. Güvenirlilik katsayılarına göre toplam puanlar için puanlayıcılar arası güvenilirliğin yüksek olduğu, alt kategoriler için ise her iki uygulamada da puanlayıcılar arası güvenilirliğin diğer bir deyişle puanlayıcılar arasındaki uyumun daha düşük olduğu söylenebilir. Bununla birlikte puanlayıcılar arası uyumun en yüksek olduğu kategori strateji planlama kategorisiyken, en düşük olduğu kategori yine sembollerin, işlemlerin ve formal dilin kullanımı kategorisidir.

Üçüncü Araştırma Sorusuna İlişkin Bulgular

Araştırmanın üçüncü sorusuna yönelik olarak PISA matematik okuryazarlığı maddelerinin gereksinim düzeylerine ilişkin Yeterlilik Şeması puanlama anahtarı ile beş puanlayıcıdan iki farklı zamanda elde edilen ölçümlerin güvenilirliği Genellenebilirlik Kuramı çerçevesinde çok değişkenli ve tek değişkenli model G ve K çalışmaları ile incelenmiştir.

3.1.Araştırma Sorusuna İlişkin Bulgular

“Çok değişkenli ve tek değişkenli G çalışmaları sonucunda her biri kategoriye ve toplam puanlara ilişkin kestirilen varyanslar ve toplam varyansı açıklama yüzdeleri nelerdir?”

Yayınlanmış PISA matematik okuryazarlığı maddelerinin tümü olan 156 maddenin yeterlilik gereksinim düzeylerini belirlemek amacıyla bu maddeler beş puanlayıcı tarafından iki farklı zamanda Yeterlilik Şeması kullanılarak puanlanmıştır. Elde edilen puanlarla Genellenebilirlik Kuramı kapsamında toplam puanlar için madde (m), puanlayıcı (p) ve zaman (z) yüzeylerinin çaprazlandığı, $m \times p \times z$ deseni oluşturularak tek değişkenli G çalışması ve alt kategorilere ait puanlar için $m^* \times p^* \times z^*$ deseni oluşturularak çok değişkenli G çalışması yapılmıştır. Tek değişkenli ve çok değişkenli G çalışmaları sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri m , p , z ana etkileri ile mp , mz , pz , mpz ortak etkileri Tablo 15’te sunulmuştur. Tablo 15’te verilen çok değişkenli G çalışmaları sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri her bir alt kategori için ayrı ayrı açıklanmıştır.

İletişim: Çok değişkenli G çalışması sonucunda İletişim alt kategorisi için kestirilen varyanslar ve toplam varyansı açıklama yüzdeleri incelendiğinde ölçme objesi olarak ele alınan madde ana etkisi (m) için kestirilen varyans bileşeninin (0,24267) toplam varyansın % 59,1’ini açıkladığı görülmektedir. Ölçme objesine ilişkin ana etki evren puan varyansı olarak değerlendirilir ve ölçülen özellik açısından maddeler arası farklılaşmayı ifade eder (Güler, 2008; Shavelson ve Webb, 1991; Brennan, 2001). Ölçme objesi olan maddeler için kestirilen varyansın toplam varyans içindeki oranının en büyük olması istenilen bir durumdur ve maddeler arası farklılıkların ortaya çıkarılabildiğinin bir göstergesidir (Güler, 2008, s.77). Bu durum iletişim yeterliliği gereksinimleri açısından maddelerin birbirinden farklılaştığını göstermektedir. Puanlayıcı (p) ana etkisi için kestirilen varyans bileşeni (0,00001) toplam varyansın % 0’ını, zaman (z) ana etkisi için kestirilen varyans bileşeni (0,00065) toplam varyansın % 0’ını açıklamaktadır. Puanlayıcı değişkenlik kaynağı için

kestirilen varyans bileşeni ölçülen özellik açısından bütün maddeler için puanlayıcılar arasındaki farklılaşmayı gösterirken zaman değişkenlik kaynağı için kestirilen varyans bileşeni ise ölçülen özellik açısından bütün maddeler için uygulamalar arasındaki farklılaşmayı ifade etmektedir. Bu varyans bileşenlerinin toplam varyansı açıklama oranının sıfır olması, puanlayıcıların bütün maddeler için puanlamadaki katılıklarının/cömertliklerinin birbirlerine çok yakın olduğuna ve bütün maddeler için uygulamalarda verilen puanlar arasında farklılığın olmadığına işaret etmektedir.

Ayrıca zaman değişkenlik kaynağı için kestirilen varyans bileşeninin negatif değer aldığı görülmektedir. Kestirilen varyansın negatif çıkması durumunda negatif varyans değerinin sıfır olarak ele alınmasının uygun olduğu belirtilmektedir (Atılğan, 2004; Brennan, 2001; Shavelson ve Webb, 1991). Bununla birlikte negatif varyans değeri araştırmada büyük bir evreni temsil eden oldukça küçük bir örneklemin kullanmış olabileceğine de işaret edebilmektedir (Güler, Uyanık ve Teker, 2012, s.69)

Madde x puanlayıcı (mp) ortak etkisi için kestirilen varyans bileşeni (0,10544) toplam varyansın % 25,7'sini açıklamaktadır. Bu varyans bileşenin toplam varyans içerisinde en yüksek ikinci orana sahip olduğu görülmektedir. m x p ortak etkisi varyansı belli bir maddenin bir puanlayıcıdan diğerine bağlı durumunun farklı olup olmadığını göstermektedir. Diğer bir deyişle, bazı puanlayıcıların bazı maddeleri puanlamalarındaki tutarsızlıkları ifade etmektedir. Bir puanlayıcı bir maddeyi puanlarken çok katı bir diğerini puanlarken çok cömert davranabilmektedir. Farklı puanlayıcılara göre farklı madde sıralamaları, maddeler ile puanlayıcılar arasındaki etkileşimden kaynaklanmaktadır (Güler, Uyanık ve Teker, 2012, s. 67). m x p ortak etkisi varyans oranının yüksek olması madde x puanlayıcı ortak etkisinden kaynaklanan farklılığın büyük olduğunu ve belli maddelerin bağlı durumlarının bir puanlayıcıdan diğerine çok farklılaştığını göstermektedir. Diğer bir deyişle puanlayıcıların maddelerin bağlı durumları hakkında uzlaşmazlıkları olduğu söylenebilir.

Madde x zaman (mz) ortak etkisi için kestirilen varyans bileşeni (0,00418) toplam varyansın % 1'ini açıklamaktadır. m x z ortak etkisi varyansı belli bir maddenin bir uygulamadan diğerine bağlı durumunun farklı olup olmadığını göstermektedir. Diğer bir deyişle, bazı uygulamalarda bazı maddelerin puanlarındaki tutarsızlıkları ifade etmektedir.

Tablo 15

Yeterlilik Şeması İle Elde Edilen Ölçümlere İlişkin Tek Değişkenli ve Çok Değişkenli G Çalışması Sonucunda Kestirilen Varyanslar ve Toplam Varyansı Açıklama Yüzdeleri

Etki	Çok Değişkenli G										Tek Değişkenli G			
	İletişim		Strateji Planlama		Matematikleştirme		Temsil		Sembollerin, İşlemlerin ve Formal Dilin Kullanımı		Muhakeme ve Argüman		Tüm Test	
	Varyans	%	Varyans	%	Varyans	%	Varyans	%	Varyans	%	Varyans	%	Varyans	%
m	0,24267	59,1	0,35911	67,0	0,26611	59,4	0,35088	66,0	0,26085	52,9	0,31209	57,5	4,78919	72,6
p	0,00001	0,0	0,00323	0,6	0,00125	0,3	0,00387	0,7	0,01031	2,1	0,00194	0,4	0,07988	1,2
z	-0,00065	0,0	-0,00045	0,0	0,00021	0,0	0,0016	0,3	0,00221	0,4	0,00032	0,1	0,01594	0,2
mp	0,10544	25,7	0,10254	19,1	0,10836	24,2	0,11087	20,9	0,10123	20,5	0,15735	29,0	1,03678	15,7
mz	0,00418	1,0	0,00206	0,4	0,01133	2,5	0,00961	1,8	0,01606	3,3	0,01314	2,4	0,16996	2,6
pz	0,00311	0,8	0,00204	0,4	-0,00001	0,0	0,00202	0,4	-0,00023	0,0	0,00016	0,0	0,00660	0,1
mpz	0,05554	13,5	0,06687	12,5	0,0609	13,6	0,05279	9,9	0,10248	20,8	0,05754	10,6	0,49725	7,5
Top		100		100		100		100		100		100		100

m: madde, p: puanlayıcı, z: zaman

Bir madde bir uygulamada daha yüksek puan alırken bir diğerinde daha düşük puan alabilmektedir. Farklı uygulamalardaki farklı madde sıralamaları, maddeler ile uygulamalar arasındaki etkileşimden kaynaklanmaktadır. $m \times z$ ortak etkisi varyans oranının oldukça küçük olması madde x zaman ortak etkisinden kaynaklanan farklılığın büyük olmadığını ve belli maddelerin bağıl durumlarının bir uygulamadan diğerine çok değişmediğini göstermektedir. Diğer bir deyişle farklı uygulamalar için maddelerin bağıl durumları hakkında farklılık olmadığı söylenebilir.

Puanlayıcı x zaman (pz) ortak etkisi için kestirilen varyans bileşeni (0,00311) toplam varyansın % 0,8'ini açıklamaktadır. $p \times z$ ortak etkisi varyansı puanlayıcıların maddeleri bir uygulamadan diğerine puanlamalarındaki kararlılığı göstermektedir. Diğer bir deyişle, bazı puanlayıcıların bazı uygulamalarda verdikleri puanlar arasındaki tutarsızlıkları ifade etmektedir. Bir puanlayıcı bir uygulamada daha cömert davranırken bir diğerinde daha katı davranabilmektedir. Puanlayıcıların farklı uygulamalardaki farklı madde sıralamaları puanlayıcılar ile uygulamalar arasındaki etkileşimlerden kaynaklanmaktadır (Güler, Uyanık ve Teker, 2012, s. 67). $p \times z$ ortak etkisi varyans oranının oldukça küçük olması puanlayıcı x zaman ortak etkisinden kaynaklanan farklılığın büyük olmadığını ve puanlayıcıların maddeleri farklı uygulamalarda kararlı bir şekilde puanladıklarını göstermektedir. Diğer bir deyişle, bazı puanlayıcıların verdikleri puanlar uygulamadan uygulamaya değişmemektedir. Bununla birlikte $p \times z$ ortak etkisi için kestirilen varyans oranının diğer alt kategorilere kıyasla iletişim alt kategorisinde daha yüksek olduğu görülmektedir. O halde, bazı puanlayıcıların verdikleri puanların bir uygulamadan diğerine değişiminin en çok iletişim alt kategorisinde gerçekleştiği söylenebilir.

Madde x puanlayıcı x zaman (mpz) (artık) ortak etkisi için kestirilen varyans bileşeni (0,05554) toplam varyansın % 13,5'ini açıklamaktadır. Artık varyans oranı iletişim alt kategorisi için en yüksek orana sahip üçüncü varyans bileşenidir. $m \times p \times z$ ortak etkisi varyansı tesadüfi hatalardan ve madde, puanlayıcı ve zaman etkileşiminden oluşmaktadır (Güler, Uyanık ve Teker, 2012, s. 67). Artık varyans oranının büyük olması madde puanlayıcı ve zaman ortak etkisinin ve tesadüfi hataların büyük olabileceğini işaret ettiğinden bu oranın küçük olması istenilen bir durumdur.

Strateji planlama: Çok değişkenli G çalışması sonucunda strateji planlama alt kategorisi için kestirilen varyanslar ve toplam varyansı açıklama yüzdeleri incelendiğinde ölçme objesi olarak ele alınan madde ana etkisi (m) için kestirilen varyans bileşeninin (0,35911) toplam

varyansın % 67'sini açıkladığı ve en yüksek orana sahip varyans bileşeni olduğu görülmektedir. Bu durum strateji planlama yeterliliği gereksinimleri açısından maddelerin birbirinden farklılaştığını göstermektedir. Puanlayıcı (p) ana etkisi için kestirilen varyans bileşeni (0,00323) toplam varyansın % 0,6'sını, zaman (z) ana etkisi için kestirilen varyans bileşeni (-0,00045) toplam varyansın % 0'ını açıklamaktadır. Puanlayıcı değişkenlik kaynağı için kestirilen varyans bileşeni puanlayıcılardan kaynaklanan farklılıkları, zaman değişkenlik kaynağı için kestirilen varyans bileşeni ise ölçülen özellik açısından uygulamalardan kaynaklanan farklılıkları göstermektedir. Bu varyans bileşenlerinin toplam varyansı açıklama oranının sıfıra çok yakın bir değer alması, puanlayıcıların bütün maddeler için verdikleri puanlarda katılık/cömertliklerinin birbirlerine çok yakın olduğuna ve bütün maddeler için uygulamalarda verilen puanlar arasında farklılığın olmadığına işaret etmektedir. Ayrıca zaman değişkenlik kaynağı için kestirilen varyans bileşeninin negatif değer aldığı görülmektedir.

Madde x puanlayıcı (mp) ortak etkisi için kestirilen varyans bileşeninin (0,10254) toplam varyansın % 19,1'ini açıkladığı ve toplam varyans içerisinde en yüksek ikinci orana sahip varyans bileşeni olduğu görülmektedir. Bu varyans bileşeni belli bir maddenin bir puanlayıcıdan diğerine bağlı durumunun farklı olup olmadığını göstermektedir. m x p ortak etkisi varyans oranının yüksek olması madde x puanlayıcı ortak etkisinden kaynaklanan farklılığın büyük olduğunu ve strateji planlama yeterliliği açısından belli maddelerin bağlı durumlarının bir puanlayıcıdan diğerine çok farklılaştığını göstermektedir. Diğer bir deyişle puanlayıcıların maddelerin bağlı durumları hakkında uzlaşmazlıkları olduğu söylenebilir. Madde x zaman (mz) ortak etkisi için kestirilen varyans bileşeni (0,00206) ile puanlayıcı x zaman (pz) ortak etkisi için kestirilen varyans bileşeninin (0,00204) her ikisi de toplam varyansın % 0,4'ünü açıklamaktadır. m x z ortak etkisi varyansı belli bir maddenin bir uygulamadan diğerine bağlı durumunun farklı olup olmadığını, p x z ortak etkisi varyansı ise puanlayıcıların maddeleri bir uygulamadan diğerine puanlamalarındaki kararlılığı göstermektedir. m x z ve p x z ortak etkisi varyans oranlarının oldukça küçük olması madde x zaman ve puanlayıcı x zaman ortak etkilerinden kaynaklanan farklılıkların büyük olmadığını, belli maddelerin bağlı durumlarının bir uygulamadan diğerine çok değişmediğini ve puanlayıcıların maddeleri farklı uygulamalarda kararlı bir şekilde puanladıklarını göstermektedir. Diğer bir deyişle farklı uygulamalar için maddelerin bağlı durumları hakkında farklılık olmadığı ve bazı puanlayıcıların verdikleri puanların uygulamadan uygulamaya değişmediği söylenebilir.

Madde x puanlayıcı x zaman (mpz) (artık) ortak etkisi için kestirilen varyans bileşeninin (0,06687) toplam varyansın % 12,5'ini açıkladığı ve toplam varyans içerisinde en yüksek üçüncü orana sahip varyans bileşeni olduğu görülmektedir. Artık varyans oranının büyük olması madde puanlayıcı ve zaman ortak etkisinin ve bu çalışmada ölçülemeyen tesadüfi hataların büyük olabileceğini işaret ettiğinden bu oranın küçük olması istenilen bir durumdur. Burada artık varyans oranı diğer varyans bileşenlerinin oranlarına göre nispeten daha büyük olsa da toplam varyans içindeki oranının kabul edilebilir bir düzeyde olduğu düşünülmektedir. Son olarak strateji planlama alt kategorisi için kestirilen varyans bileşenlerin ve toplam varyansı açıklama oranlarının iletişim alt kategorisine ilişkin bulgularla paralel olduğu görülmektedir.

Matematikleştirme: Çok değişkenli G çalışması sonucunda matematikleştirme alt kategorisi için kestirilen varyanslar ve toplam varyansı açıklama yüzdeleri incelendiğinde ölçme objesi olarak ele alınan madde ana etkisi (m) için kestirilen varyans bileşeninin (0,26611) toplam varyansın % 59,4'ünü açıkladığı ve en yüksek orana sahip varyans bileşeni olduğu görülmektedir. Bu durum matematikleştirme yeterliliği gereksinimleri açısından maddelerin birbirinden farklılaştığını göstermektedir. Puanlayıcı (p) ana etkisi için kestirilen varyans bileşeni (0,00125) toplam varyansın % 0,3'ünü, zaman (z) ana etkisi için kestirilen varyans bileşeni (0,00021) toplam varyansın % 0'ını açıklamaktadır. Puanlayıcı değişkenlik kaynağı için kestirilen varyans bileşeni puanlayıcılardan kaynaklanan farklılıkları, zaman değişkenlik kaynağı için kestirilen varyans bileşeni ise ölçülen özellik açısından uygulamalardan kaynaklanan farklılıkları göstermektedir. Bu varyans bileşenlerinin toplam varyansı açıklama oranının sıfır olması ya da sıfıra çok yakın bir değer alması, puanlayıcıların bütün maddeler için verdikleri puanlarda katılık/cömertliklerinin birbirlerine çok yakın olduğuna ve bütün maddeler için uygulamalarda verilen puanlar arasında farklılığın olmadığına işaret etmektedir. O halde, matematikleştirme alt kategorisi açısından puanlayıcıdan ya da uygulamalardan kaynaklanan farklılıklar yok denecek kadar azdır.

Madde x puanlayıcı (mp) ortak etkisi için kestirilen varyans bileşeninin (0,10836) toplam varyansın % 24,2'sini açıkladığı ve toplam varyans içerisinde en yüksek ikinci orana sahip varyans bileşeni olduğu görülmektedir. m x p ortak etkisi varyans oranının yüksek olması madde x puanlayıcı ortak etkisinden kaynaklanan farklılığın büyük olduğunu ve matematikleştirme yeterliliği açısından da yine belli maddelerin bağıl durumlarının bir puanlayıcıdan diğerine çok farklılaştığını göstermektedir. Diğer bir deyişle puanlayıcıların maddelerin bağıl durumları hakkında uzlaşmazlıkları olduğu söylenebilir. Madde x zaman

(mz) ortak etkisi için kestirilen varyans bileşeni (0,01133) toplam varyansın %2,5'ini puanlayıcı x zaman (pz) ortak etkisi için kestirilen varyans bileşeni (-0,00001) ise toplam varyansın % 0'ını açıklamaktadır. m x z ve p x z ortak etkisi varyans oranlarının oldukça küçük olması madde x zaman ve puanlayıcı x zaman ortak etkilerinden kaynaklanan farklılıkların büyük olmadığını, belli maddelerin bağıl durumlarının bir uygulamadan diğerine çok değişmediğini ve puanlayıcıların maddeleri farklı uygulamalarda kararlı bir şekilde puanladıklarını göstermektedir. Diğer bir deyişle farklı uygulamalar için maddelerin bağıl durumları hakkında farklılık olmadığı ve bazı puanlayıcıların verdikleri puanların uygulamadan uygulamaya değişmediği söylenebilir. Bununla birlikte m x z ortak etkisi varyans oranının iletişim ve strateji planlama alt kategorilerine göre nispeten daha yüksek olduğu dolayısıyla bu alt kategori için belli bir maddenin bir uygulamadan diğerine bağıl durumundaki farklılaşmanın daha yüksek olduğu söylenebilir.

Madde x puanlayıcı x zaman (mpz) (artık) ortak etkisi için kestirilen varyans bileşeninin (0,0609) toplam varyansın %13,6'sını açıkladığı ve toplam varyans içerisinde en yüksek üçüncü orana sahip varyans bileşeni olduğu görülmektedir. Artık varyans etkisi oranı madde puanlayıcı ve zaman ortak etkisini ve bu çalışmada ölçülemeyen tesadüfi hataların olabileceğini işaret etmektedir. Burada artık varyans oranı diğer varyans bileşenlerinin oranlarına göre nispeten daha büyük olsa da toplam varyans içindeki oranının kabul edilebilir bir düzeyde olduğu düşünülmektedir. Son olarak matematikleştirme alt kategorisi için kestirilen varyans bileşenlerin ve toplam varyansı açıklama oranlarının iletişim ve strateji planlama alt kategorilerine ilişkin bulgularla paralel olduğu görülmektedir.

Temsil: Çok değişkenli G çalışması sonucunda temsil alt kategorisi için kestirilen varyanslar ve toplam varyansı açıklama yüzdeleri incelendiğinde, ölçme objesi olarak ele alınan madde ana etkisi (m) için kestirilen varyans bileşeninin (0,35088) toplam varyansın % 66'sını açıkladığı ve diğer alt kategorilerde olduğu gibi en yüksek orana sahip varyans bileşeni olduğu görülmektedir. Bu durum temsil yeterliliği gereksinimleri açısından maddelerin birbirinden farklılaştığını göstermektedir. Puanlayıcı (p) ana etkisi için kestirilen varyans bileşeni (0,00387) toplam varyansın % 0,7'sini, zaman (z) ana etkisi için kestirilen varyans bileşeni (0,0016) ise toplam varyansın % 0,3'ünü açıklamaktadır. Puanlayıcı değişkenlik kaynağı için kestirilen varyans bileşeni puanlayıcılardan kaynaklanan farklılıkları, zaman değişkenlik kaynağı için kestirilen varyans bileşeni ise ölçülen özellik açısından uygulamalardan kaynaklanan farklılıkları göstermektedir. Bu varyans bileşenlerinin toplam varyansı açıklama oranlarının sıfıra çok yakın bir değer alması, puanlayıcıların bütün

maddeler için verdikleri puanlarda katılık/cömertliklerinin birbirlerine çok yakın olduğuna ve bütün maddeler için uygulamalarda verilen puanlar arasında farklılığın olmadığına işaret etmektedir. O halde, temsil alt kategorisi açısından diğer alt kategorilerde olduğu gibi puanlayıcıdan ya da uygulamalardan kaynaklanan farklılıklar yok denecek kadar azdır.

Madde x puanlayıcı (mp) ortak etkisi için kestirilen varyans bileşeninin (0,11087) toplam varyansın % 20,9'unu açıkladığı ve toplam varyans içerisinde en yüksek ikinci orana sahip varyans bileşeni olduğu görülmektedir. m x p ortak etkisi varyans oranının yüksek olması madde x puanlayıcı ortak etkisinden kaynaklanan farklılığın büyük olduğunu ve temsil yeterliliği açısından da yine belli maddelerin bağlı durumlarının bir puanlayıcıdan diğerine çok farklılaştığını göstermektedir. Diğer bir deyişle puanlayıcıların maddelerin bağlı durumları hakkında uzlaşmazlıkları olduğu söylenebilir. Madde x zaman (mz) ortak etkisi için kestirilen varyans bileşeni (0,00961) toplam varyansın %1,8'ini, puanlayıcı x zaman (pz) ortak etkisi için kestirilen varyans bileşeni (0,00202) ise toplam varyansın % 0,4'ünü açıklamaktadır. m x z ve p x z ortak etkisi varyans oranlarının oldukça küçük olması madde x zaman ve puanlayıcı x zaman ortak etkilerinden kaynaklanan farklılıkların büyük olmadığını, belli maddelerin bağlı durumlarının bir uygulamadan diğerine çok değişmediğini ve puanlayıcıların maddeleri farklı uygulamalarda kararlı bir şekilde puanladıklarını göstermektedir. Diğer bir deyişle farklı uygulamalar için maddelerin bağlı durumları hakkında farklılık olmadığı ve bazı puanlayıcıların verdikleri puanların uygulamadan uygulamaya değişmediği söylenebilir. Bununla birlikte temsil alt kategorisinde m x z ortak etkisi varyans oranının iletişim ve strateji planlama alt kategorilerine göre nispeten daha yüksek, matematikleştirme alt kategorisine göre nispeten daha düşük olduğu görülmektedir. Dolayısıyla temsil alt kategorisi için belli bir maddenin bir uygulamadan diğerine bağlı durumundaki farklılaşmanın iletişim ve strateji planlama alt kategorilerine göre nispeten daha fazla, matematikleştirme alt kategorisine göre ise nispeten daha az olduğu söylenebilir.

Madde x puanlayıcı x zaman (mpz) (artık) ortak etkisi için kestirilen varyans bileşeninin (0,05279) toplam varyansın %9,9'unu açıkladığı ve toplam varyans içerisinde en yüksek üçüncü orana sahip varyans bileşeni olduğu görülmektedir. Artık varyans etkisi oranı madde puanlayıcı ve zaman ortak etkisini ve bu çalışmada ölçülemeyen tesadüfi hataların olabileceğini işaret etmektedir. Burada artık varyans oranı diğer varyans bileşenlerinin oranlarına göre nispeten daha büyük olsa da toplam varyans içindeki oranının kabul edilebilir bir düzeyde olduğu düşünülmektedir. Bununla birlikte temsil kategorisine ilişkin

m x p x z (artık) ortak etkisi varyans oranının, diğer alt kategorilere kıyasla en düşük orana sahip olduğu göze çarpmaktadır. O halde madde puanlayıcı ve zaman ortak etkisinin ve bu çalışmada ölçülemeyen tesadüfi hataların en düşük olduğu alt kategorinin temsil kategorisi olduğu söylenebilir. Buna ek olarak temsil alt kategorisi için kestirilen varyans bileşenlerin ve toplam varyansı açıklama oranları, iletişim, strateji planlama ve matematikleştirme alt kategorilerine ilişkin bulgularla benzerlik göstermektedir.

Sembollerin, İşlemlerin Ve Formal Dilin Kullanımı: Çok değişkenli G çalışması sonucunda bu alt kategori için kestirilen varyanslar ve toplam varyansı açıklama yüzdeleri incelendiğinde ölçme objesi olarak ele alınan madde ana etkisi (m) için kestirilen varyans bileşeninin (0,26085) toplam varyansın %52,9'unu açıkladığı görülmektedir. Bu varyans bileşeni oranı diğer alt kategorilerde olduğu gibi bu alt kategori için de en yüksek orana sahip varyans bileşenidir. Ancak tüm alt kategoriler arasında madde ana etkisi için kestirilen varyans oranının en düşük olduğu kategorinin bu alt kategori olduğu göze çarpmaktadır. Bu durum diğer yeterliliklere göre sembollerin, işlemlerin ve formal dilin kullanımı yeterliliği gereksinimleri açısından maddelerin birbirinden daha az farklılaştığını göstermektedir. Puanlayıcı (p) ana etkisi için kestirilen varyans bileşeni (0,01031) toplam varyansın %2,1'ini, zaman (z) ana etkisi için kestirilen varyans bileşeni (0,00221) toplam varyansın %0,4'ünü açıklamaktadır. Puanlayıcı değişkenlik kaynağı için kestirilen varyans bileşeni puanlayıcılardan kaynaklanan farklılıkları, zaman değişkenlik kaynağı için kestirilen varyans bileşeni ise ölçülen özellik açısından uygulamalardan kaynaklanan farklılıkları göstermektedir. Bu varyans bileşenlerinin toplam varyansı açıklama oranının sıfıra çok yakın bir değer alması, puanlayıcıların bütün maddeler için verdikleri puanlarda katılık/cömertliklerinin birbirlerine çok yakın olduğuna ve bütün maddeler için uygulamalarda verilen puanlar arasında farklılığın olmadığına işaret etmektedir. Bununla birlikte tüm alt kategoriler arasında puanlayıcı ve zaman ana etkileri için kestirilen varyans oranlarının nispeten daha fazla olduğu kategorinin bu alt kategori olduğu göze çarpmaktadır. Bu durum diğer yeterliliklere göre sembollerin, işlemlerin ve formal dilin kullanımı yeterliliği gereksinimleri açısından puanlayıcılardan ve uygulamalardan kaynaklanan farklılıkların nispeten daha fazla olduğunu göstermektedir.

Madde x puanlayıcı (mp) ortak etkisi için kestirilen varyans bileşeninin (0,10123) toplam varyansın %20,5'ini açıkladığı ve toplam varyans içerisinde en yüksek üçüncü orana sahip varyans bileşeni olduğu görülmektedir. Bu varyans bileşeni belli bir maddenin bir puanlayıcıdan diğerine bağlı durumunun farklı olup olmadığını göstermektedir. m x p ortak

etkisi varyans oranının yüksek olması madde x puanlayıcı ortak etkisinden kaynaklanan farklılığın büyük olduğunu ve sembollerin, işlemlerin ve formal dilin kullanımı yeterliliği gereksinimleri açısından belli maddelerin bağıl durumlarının bir puanlayıcıdan diğerine çok farklılaştığını göstermektedir. Diğer bir deyişle puanlayıcıların maddelerin bağıl durumları hakkında uzlaşmazlıkları olduğu söylenebilir. Madde x zaman (mz) ortak etkisi için kestirilen varyans bileşeni (0,01606) %3,3'ünü puanlayıcı x zaman (pz) ortak etkisi için kestirilen varyans bileşeninin (-0,00023) toplam varyansın % 0'ını açıklamaktadır. m x z ortak etkisi varyansı belli bir maddenin bir uygulamadan diğerine bağıl durumunun farklı olup olmadığını, p x z ortak etkisi varyansı ise puanlayıcıların maddeleri bir uygulamadan diğerine puanlamalarındaki kararlılığı göstermektedir. Negatif değer alan varyans sıfır kabul edildiğinden m x z ve p x z ortak etkisi varyans oranlarının oldukça küçük olması, madde x zaman ve puanlayıcı x zaman ortak etkilerinden kaynaklanan farklılıkların büyük olmadığını, belli maddelerin bağıl durumlarının bir uygulamadan diğerine çok değişmediğini ve puanlayıcıların maddeleri farklı uygulamalarda kararlı bir şekilde puanladıklarını göstermektedir. Diğer bir deyişle farklı uygulamalar için maddelerin bağıl durumları hakkında farklılık olmadığı ve bazı puanlayıcıların verdikleri puanların uygulamadan uygulamaya değişmediği söylenebilir.

Madde x puanlayıcı x zaman (mpz) (artık) ortak etkisi için kestirilen varyans bileşeninin (0,10248) toplam varyansın % 20,8'ini açıkladığı ve toplam varyans içerisinde en yüksek ikinci orana sahip varyans bileşeni olduğu görülmektedir. Artık varyans oranının büyük olması madde puanlayıcı ve zaman ortak etkisinin ve bu çalışmada ölçülemeyen tesadüfi hataların büyük olabileceğini işaret ettiğinden bu oranın büyük olması istenilen bir durum değildir. Burada artık varyans oranı diğer varyans bileşenlerinin oranlarına göre nispeten daha büyüktür. Ayrıca artık varyans oranının en yüksek olduğu kategori bu kategoridir. Son olarak sembollerin, işlemlerin ve formal dilin kullanımı alt kategorisi için kestirilen varyans bileşenlerin ve toplam varyansı açıklama oranlarının diğer alt kategorilere ilişkin bulgulardan farklılık göstermektedir. Öncelikle bu alt kategori artık varyans oranının en yüksek ikinci orana sahip olduğu tek alt kategoridir. Bu oran diğer alt kategorilerdeki artık varyans oranlarına göre de nispeten daha yüksektir. Ayrıca, puanlayıcı ve zaman ana etkileri ile madde x zaman ortak etkisi için kestirilen varyans oranlarının da diğer alt kategorilerdeki ilişkili oranlara göre nispeten daha yüksek olduğu göze çarpmaktadır.

Muhakeme ve argüman: Çok değişkenli G çalışması sonucunda muhakeme ve argüman alt kategorisi için kestirilen varyanslar ve toplam varyansı açıklama yüzdeleri incelendiğinde

ölçme objesi olarak ele alınan madde ana etkisi (m) için kestirilen varyans bileşeninin (0,31209) toplam varyansın % 57,5'ini açıkladığı ve en yüksek orana sahip varyans bileşeni olduğu görülmektedir. Bu durum muhakeme ve argüman yeterliliği gereksinimleri açısından maddelerin birbirinden farklılaştığını göstermektedir. Puanlayıcı (p) ana etkisi için kestirilen varyans bileşeni (0,00194) toplam varyansın % 0,4'ünü, zaman (z) ana etkisi için kestirilen varyans bileşeni (0,00032) toplam varyansın % 0,1'ini açıklamaktadır. Puanlayıcı değişkenlik kaynağı için kestirilen varyans bileşeni puanlayıcılardan kaynaklanan farklılıkları, zaman değişkenlik kaynağı için kestirilen varyans bileşeni ise ölçülen özellik açısından uygulamalardan kaynaklanan farklılıkları göstermektedir. Bu varyans bileşenlerinin toplam varyansı açıklama oranının sıfıra çok yakın değerler alması, puanlayıcıların bütün maddeler için verdikleri puanlarda katılık/cömertliklerinin birbirlerine çok yakın olduğuna ve bütün maddeler için uygulamalarda verilen puanlar arasında farklılığın olmadığına işaret etmektedir.

Madde x puanlayıcı (mp) ortak etkisi için kestirilen varyans bileşeninin (0,15735) toplam varyansın %29'unu açıkladığı ve toplam varyans içerisinde en yüksek ikinci orana sahip varyans bileşeni olduğu görülmektedir. Bu varyans bileşeni belli bir maddenin bir puanlayıcıdan diğerine bağıl durumunun farklı olup olmadığını göstermektedir. m x p ortak etkisi varyans oranının yüksek olması madde x puanlayıcı ortak etkisinden kaynaklanan farklılığın büyük olduğunu ve muhakeme ve argüman yeterliliği açısından belli maddelerin bağıl durumlarının bir puanlayıcıdan diğerine çok farklılaştığını göstermektedir. Diğer bir deyişle puanlayıcıların maddelerin bağıl durumları hakkında uzlaşmazlıkları olduğu söylenebilir. Bununla birlikte m x p ortak etkisi varyans oranının diğer alt kategorilerdeki m x p ortak etkisi varyans oranlarına göre nispeten daha yüksek olduğu göze çarpmaktadır. Bu durum belli maddelerin bağıl durumlarının bir puanlayıcıdan diğerine göre en çok farklılaştığı alt kategorinin muhakeme ve argüman olduğunu ortaya koymaktadır.

Madde x zaman (mz) ortak etkisi için kestirilen varyans bileşeni (0,01314) toplam varyansın %2,4'ünü, puanlayıcı x zaman (pz) ortak etkisi için kestirilen varyans bileşeni (0,00016) toplam varyansın %0'ını açıklamaktadır. m x z ortak etkisi varyansı belli bir maddenin bir uygulamadan diğerine bağıl durumunun farklı olup olmadığını, p x z ortak etkisi varyansı ise puanlayıcıların maddeleri bir uygulamadan diğerine puanlamalarındaki kararlılığı göstermektedir. m x z ve p x z ortak etkisi varyans oranlarının oldukça küçük olması madde x zaman ve puanlayıcı x zaman ortak etkilerinden kaynaklanan farklılıkların büyük olmadığını, belli maddelerin bağıl durumlarının bir uygulamadan diğerine çok

değişmediğini ve puanlayıcıların maddeleri farklı uygulamalarda kararlı bir şekilde puanladıklarını göstermektedir. Diğer bir deyişle farklı uygulamalar için maddelerin bağlı durumları hakkında farklılık olmadığı ve bazı puanlayıcıların verdikleri puanların uygulamadan uygulamaya değişmediği söylenebilir.

Madde x puanlayıcı x zaman (mpz) (artık) ortak etkisi için kestirilen varyans bileşeninin (0,05754) toplam varyansın %10,6'sını açıkladığı ve toplam varyans içerisinde en yüksek üçüncü orana sahip varyans bileşeni olduğu görülmektedir. Artık varyans oranının büyük olması madde puanlayıcı ve zaman ortak etkisinin ve bu çalışmada ölçülemeyen tesadüfi hataların büyük olabileceğini işaret ettiğinden bu oranın küçük olması istenilen bir durumdur. Burada artık varyans oranı diğer varyans bileşenlerinin oranlarına göre nispeten daha büyük olsa da toplam varyans içindeki oranının kabul edilebilir bir düzeyde olduğu düşünülmektedir. Ayrıca, muhakeme ve argüman alt kategorisi için kestirilen varyans bileşenlerin ve toplam varyansı açıklama oranlarının sembollerin, işlemlerin ve formal dilin kullanımı alt kategorisi dışındaki diğer alt kategorilerdeki ilişkili bulgularla paralel olduğu görülmektedir.

Tablo 15'te verilen tek değişkenli G çalışması sonucunda toplam puanlar için kestirilen varyanslar ve toplam varyansı açıklama yüzdeleri incelendiğinde, madde ana etkisi için kestirilen varyans bileşeninin (4,78919) toplam varyansın % 72,6'sını açıkladığı ve en yüksek orana sahip varyans bileşeni olduğu görülmektedir. Bu durum tek değişkenli ve çok değişkenli modellerle maddeler için kestirilen varyans bileşeninin toplam varyans içindeki oranlarının diğer bileşenlerin toplam varyans içindeki oranlarından çok daha yüksek olduğunu göstermiştir. Bununla birlikte tek değişkenli modellerle kestirilen madde ana etkisi varyans oranının çok değişkenli modellerle kestirilen madde ana etkisi varyans oranlarından daha büyük çıkma eğiliminde olduğu görülmüştür. Puanlayıcı (p) ana etkisi için kestirilen varyans bileşeni (0,07988) toplam varyansın %1,2'sini açıklamaktadır. Bu varyans bileşeninin toplam varyansı açıklama oranının sıfıra çok yakın bir değer alması, puanlayıcıların bütün maddeler için verdikleri toplam yeterlilik puanlarında katılık/cömertliklerinin birbirlerine çok yakın olduğuna işaret etmektedir. Tek değişkenli modellerle kestirilen puanlayıcı ana etkisi varyans bileşeninin toplam varyans içindeki oranının çok değişkenli modellerle sembollerin, işlemlerin ve formal dilin kullanımı alt kategorisi dışındaki diğer alt kategoriler için kestirilenlerden daha büyük olduğu görülmektedir. Diğer bir deyişle, bu durum tek değişkenli ve çok değişkenli modellerle kestirilen puanlayıcı ana etkisi varyans oranlarının tutarlılık gösterdiği, puanlayıcı ana etkisi varyans oranlarının

modelden bu ölçme için etkilenmediği şeklinde yorumlanabilir. Zaman (z) ana etkisi için kestirilen varyans bileşeni (0,01594) toplam varyansın % 0,2'ini açıklamaktadır. Bu varyans bileşeninin toplam varyansı açıklama oranının sıfıra çok yakın bir değer alması, bütün maddeler için uygulamalarda verilen puanlar arasında farklılığın olmadığına işaret etmektedir. Tek değişkenli modelle kestirilen zaman ana etkisi varyans bileşeninin toplam varyans içindeki oranının çok değişkenli modelle diğer alt kategoriler için kestirilenlerle paralellik gösterdiği görülmektedir. Diğer bir deyişle, bu durum tek değişkenli ve çok değişkenli modelle kestirilen zaman ana etkisi varyans oranlarının tutarlılık gösterdiği, zaman ana etkisi varyans oranlarının modelden bu ölçme için etkilenmediği şeklinde yorumlanabilir.

Madde x puanlayıcı (mp) ortak etkisi için kestirilen varyans bileşeninin (1,03678) toplam varyansın %15,7'sini açıkladığı ve toplam varyans içerisinde en yüksek ikinci orana sahip varyans bileşeni olduğu görülmektedir. Tek değişkenli modelle kestirilen madde x puanlayıcı ortak etkisi varyans oranının çok değişkenli modelle alt kategorilere göre kestirilen varyans oranından daha küçük çıkma eğiliminde olduğu görülmüştür. Buna dayanarak, bu ölçme için tek değişkenli modelle yapılan analizde çok değişkenli modele göre madde puanlayıcı etkileşiminden kaynaklanan farklılıkların daha düşük olduğu ya da toplam puanlar bazında puanlayıcıların maddelerin bağlı durumları hakkında uzlaşmazlıklarının daha az olduğu söylenebilir. Madde x zaman (mz) ortak etkisi için kestirilen varyans bileşeni (0,16996) toplam varyansın %2,6'ünü, puanlayıcı x zaman (pz) ortak etkisi için kestirilen varyans bileşeni (0,00660) toplam varyansın %0,1'ini açıklamaktadır. m x z ortak etkisi varyansı belli bir maddenin bir uygulamadan diğerine bağlı durumunun farklı olup olmadığını, p x z ortak etkisi varyansı ise puanlayıcıların maddeleri bir uygulamadan diğerine puanlamalarındaki kararlılığı göstermektedir. m x z ve p x z ortak etkisi varyans oranlarının oldukça küçük olması madde x zaman ve puanlayıcı x zaman ortak etkilerinden kaynaklanan farklılıkların büyük olmadığını, belli maddelerin bağlı durumlarının bir uygulamadan diğerine çok değişmediğini ve puanlayıcıların maddeleri farklı uygulamalarda kararlı bir şekilde puanladıklarını göstermektedir. Tek değişkenli modelle kestirilen madde x zaman ve puanlayıcı x zaman ortak etkisi varyans bileşenlerinin toplam varyans içindeki oranlarının çok değişkenli modelle diğer alt kategoriler için kestirilenlerle paralellik gösterdiği görülmektedir. Diğer bir deyişle, bu durum tek değişkenli ve çok değişkenli modelle kestirilen madde x zaman ve puanlayıcı x zaman ortak etkisi

varyans oranlarının tutarlılık gösterdiği, madde x zaman ve puanlayıcı x zaman ortak etkisi varyans oranlarının modelden bu ölçme için etkilenmediği şeklinde yorumlanabilir.

Madde x puanlayıcı x zaman (mpz) (artık) ortak etkisi için kestirilen varyans bileşeninin (0,49725) toplam varyansın %7,5'sını açıkladığı ve toplam varyans içerisinde en yüksek üçüncü orana sahip varyans bileşeni olduğu görülmektedir. Artık varyans oranının küçük olması madde puanlayıcı ve zaman ortak etkisinin ve bu çalışmada ölçülemeyen tesadüfi hataların az olabileceğini işaret ettiğinden bu oranın küçük olması istenilen bir durumdur. Burada artık varyans oranı diğer varyans bileşenlerinin oranlarına göre nispeten daha büyük olsa da toplam varyans içindeki oranının kabul edilebilir bir düzeyde olduğu düşünülmektedir. Tek değişkenli modelle kestirilen madde x puanlayıcı x zaman (artık) ortak etkisi varyans oranının çok değişkenli modelle alt kategorilere göre kestirilen varyans oranından daha küçük çıkma eğiliminde olduğu görülmüştür. Buna dayanarak, bu ölçme için tek değişkenli modelle yapılan analizde çok değişkenli modele göre madde x puanlayıcı x zaman etkileşiminden kaynaklanan farklılıkların ve çalışmada ölçülemeyen tesadüfi hataların daha düşük olduğu söylenebilir.

Genel olarak çok değişkenli ve tek değişkenli G çalışmaları ile alt kategorilere ve toplam puanlara ilişkin kestirilen varyanslar ve toplam varyansları açıklama oranları incelendiğinde her iki modelle kestirilen varyans oranlarının sembollerin, işlemlerin ve formal dilin kullanımı alt kategorisi için farklılıklar gösterdiği ortaya çıkmıştır. Diğer bir deyişle puanlayıcı ve zaman ana etkisinden gelen varyans oranları diğer kategorilerde daha düşükken, bu kategoride nispeten daha yüksek bulunmuştur. Tek değişkenli modelle yapılan G analizi ile çok değişkenli modelle yapılan G analizi bu ölçme için varyansları birbirinden farklı kestirmektedir.

3.2.Araştırma Sorusuna İlişkin Bulgular

“Orijinal puanlayıcı ve zaman sayıları ile puanlayıcı ve zaman sayılarının bir arttırılıp bir azaltılması senaryolarına göre çok değişkenli Karar (K) çalışmalarıyla kestirilen G ve Phi katsayıları alt kategorilere göre nasıl değişmektedir?”

Yayınlanmış PISA matematik okuryazarlığı maddelerinin yeterlilik gereksinim düzeylerinin belirlenmesine ilişkin olarak G kuramının çok değişkenli $m^* \times p^* \times z^*$ deseni ile puanlayıcı ve zaman sayılarının bir arttırılıp bir azaltılmasıyla oluşturulan senaryolara göre K çalışmaları yapılmıştır. 156 maddenin puanlandığı ölçme durumunda beş puanlayıcı iki

farklı zamanda PISA matematik okuryazarlığı maddelerini birbirinden bağımsız olarak Yeterlilik Şeması ile puanlamışlardır. K çalışmaları sonucunda alt kategoriler için kestirilen G ve Phi katsayıları Tablo 16, Şekil 5 ve Şekil 6’da verilmiştir.

İletişim alt kategorisinde orijinal puanlayıcı (5) ve zaman (2) sayıları için çok değişkenli model K çalışması ile kestirilen G ve Phi katsayıları sırasıyla 0,8941 ve 0,8931’dir. Puanlayıcı sayısı ($n_p=5$) sabit tutularak zaman sayısının bir azaltılması ($n_z=1$) durumunda kestirilen G ve Phi katsayıları sırasıyla 0,8696 ve 0,8677 olarak hesaplanmıştır. Puanlayıcı sayısı sabitken zaman sayısının bir azaltılması durumunda G ve Phi katsayılarının sırasıyla 0,0245 ve 0,0254 azaldığı görülmüştür. Puanlayıcı sayısı ($n_p=5$) sabit tutularak zaman sayısının bir arttırılması ($n_z=3$) durumunda ise kestirilen G ve Phi katsayıları ise sırasıyla 0,9026 ve 0,9019 olarak bulunmuş, G ve Phi katsayılarının sırasıyla 0,0085 ve 0,0088 arttığı görülmüştür.

Zaman sayısı ($n_z=2$) sabit tutularak puanlayıcı sayısının bir azaltılması ($n_p=4$) durumunda kestirilen G ve Phi katsayıları sırasıyla 0,8727 ve 0,8715, puanlayıcı sayısının bir arttırılması ($n_z=6$) durumunda kestirilen G ve Phi katsayıları ise sırasıyla 0,9090 ve 0,9081 olarak hesaplanmıştır. Zaman sayıları sabit tutularak; puanlayıcı sayısının bir azaltılması durumunda G ve Phi katsayılarının sırasıyla 0,0214 ve 0,0216 azaldığı, puanlayıcı sayısının bir arttırılması durumunda ise G ve Phi katsayılarının 0,0149 ve 0,0150 arttığı görülmüştür. Tablo 15, Şekil 4 ve Şekil 5’te görüldüğü gibi bu alt kategori için zaman ve puanlayıcı varyanslarının toplam varyans içindeki oranlarının (%0; %0) sıfır olmasından dolayı zaman/puanlayıcı sayısının arttırılmasının ya da azaltılmasının G ve Phi katsayılarını ciddi bir şekilde etkilemediği söylenebilir.

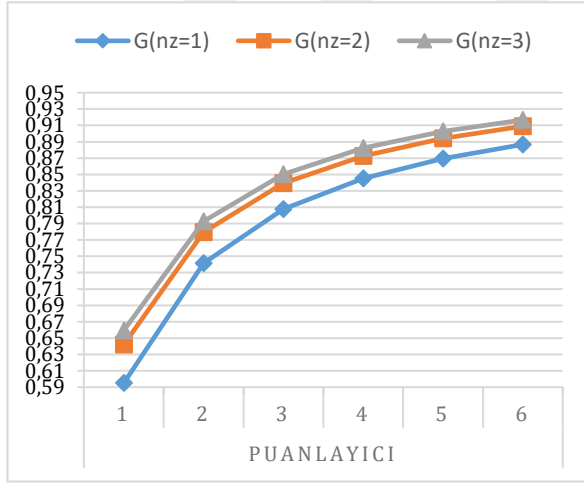
Puanlayıcı ve zaman sayısının birlikte azaltıldığı ($n_p=4$ ve $n_z=1$) durumda kestirilen G ve Phi katsayıları sırasıyla 0,8453 ve 0,8430 iken, puanlayıcı ve zaman sayısının birlikte arttırıldığı ($n_p=6$ ve $n_z=3$) durumda kestirilen G ve Phi katsayıları sırasıyla 0,9167 ve 0,9161 olarak hesaplanmıştır. Puanlayıcı ve zaman sayısı birlikte azaltıldığında ($n_p=4$ ve $n_z=1$) kestirilen G ve Phi katsayılarının sırasıyla 0,0488 ve 0,0501 azaldığı, puanlayıcı ve zaman sayısı birlikte arttırıldığında ($n_p=6$ ve $n_z=3$) kestirilen G ve Phi katsayılarının ise sırasıyla 0,0226 ve 0,0230 arttığı görülmüştür. Zaman ya da puanlayıcı sayılarından biri ya da her ikisi birlikte bir azaltıldığında G ve Phi katsayılarında meydana gelen değişim zaman ya da puanlayıcı sayılarından biri ya da her ikisi birlikte bir arttırıldığında meydana gelen değişimden daha fazladır.

Tablo 16

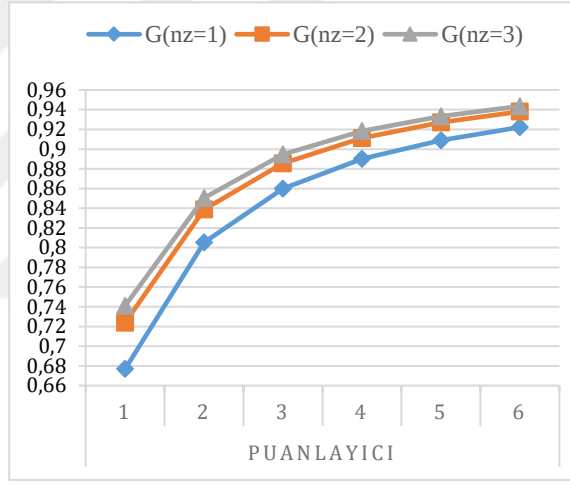
Yeterlilik Şeması ile Elde Edilen Ölçümlere İlişkin Çok Değişkenli K Çalışması Sonucunda Farklı Senaryolar için Kestirilen G ve Phi Katsayıları

np	nz	İletişim		Strateji Planlama		Matematikleştirme		Temsil		Sembolleri, İşlemlerin ve Formal Dilin Kullanımı		Muhakeme ve Argüman	
		G	ϕ	G	ϕ	G	ϕ	G	ϕ	G	Φ	G	Φ
1	1	0,5950	0,5905	0,6768	0,6702	0,5957	0,5938	0,6694	0,6600	0,5427	0,5290	0,5778	0,5752
1	2	0,6420	0,6394	0,7238	0,7177	0,6481	0,6460	0,7118	0,7037	0,6191	0,6028	0,6183	0,6156
1	3	0,6594	0,6575	0,7410	0,7351	0,6677	0,6655	0,7271	0,7196	0,6495	0,6322	0,6330	0,6304
2	1	0,7413	0,7378	0,8054	0,8007	0,7350	0,7333	0,7933	0,7852	0,6887	0,6756	0,7213	0,7190
2	2	0,7794	0,7774	0,8388	0,8347	0,7800	0,7783	0,8269	0,8207	0,7558	0,7424	0,7580	0,7559
2	3	0,7929	0,7916	0,8505	0,8466	0,7962	0,7945	0,8388	0,8332	0,7812	0,7677	0,7711	0,7690
3	1	0,8075	0,8047	0,8599	0,8563	0,7971	0,7956	0,8454	0,8382	0,7565	0,7443	0,7864	0,7844
3	2	0,8392	0,8377	0,8857	0,8826	0,8367	0,8353	0,8741	0,8688	0,8159	0,8045	0,8198	0,8180
3	3	0,8503	0,8493	0,8946	0,8917	0,8508	0,8495	0,8840	0,8795	0,8379	0,8268	0,8316	0,8298
4	1	0,8453	0,8430	0,8899	0,8871	0,8322	0,8309	0,8741	0,8675	0,7957	0,7842	0,8236	0,8217
4	2	0,8727	0,8715	0,9111	0,9087	0,8683	0,8671	0,8997	0,8950	0,8497	0,8396	0,8546	0,8530
4	3	0,8823	0,8814	0,9184	0,9161	0,8810	0,8799	0,9085	0,9046	0,8694	0,8599	0,8655	0,8640
5	1	0,8696	0,8677	0,9090	0,9066	0,8549	0,8536	0,8923	0,8860	0,8212	0,8103	0,8476	0,8459
5	2	0,8941	0,8931	0,9271	0,9251	0,8884	0,8874	0,9158	0,9116	0,8713	0,8622	0,8770	0,8756
5	3	0,9026	0,9019	0,9333	0,9314	0,9002	0,8992	0,9239	0,9204	0,8894	0,8810	0,8872	0,8859
6	1	0,8867	0,8850	0,9222	0,9201	0,8707	0,8695	0,9049	0,8989	0,8391	0,8287	0,8644	0,8628
6	2	0,9090	0,9081	0,9381	0,9364	0,9024	0,9014	0,9269	0,9229	0,8864	0,8780	0,8925	0,8912
6	3	0,9167	0,9161	0,9435	0,9419	0,9134	0,9126	0,9344	0,9312	0,9033	0,8957	0,9023	0,9011

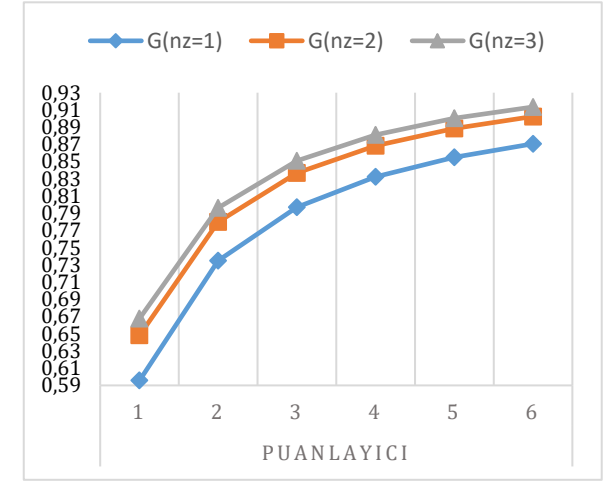
Not: Koyu yazılmış değerler çalışmada yer verilen zaman ve puanlayıcı sayılarına göre kestirilmiştir.



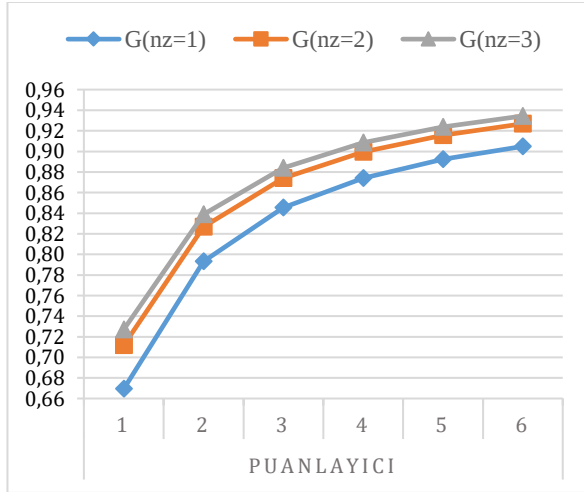
Şekil 5.a. İletişim Kategorisi İçin G Katsayıları Grafiği



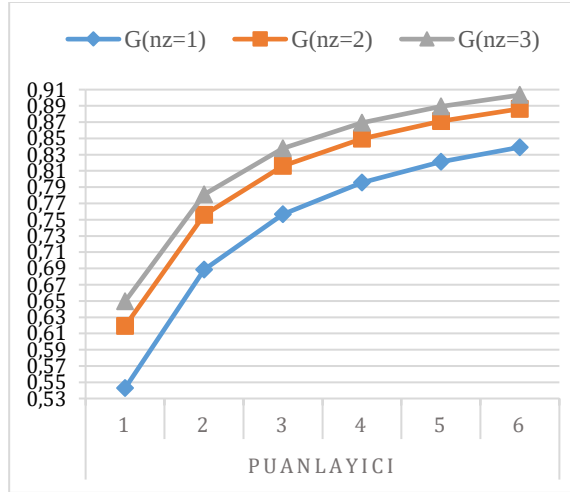
Şekil 5.b. Strateji Planlama Kategorisi İçin G Katsayıları Grafiği



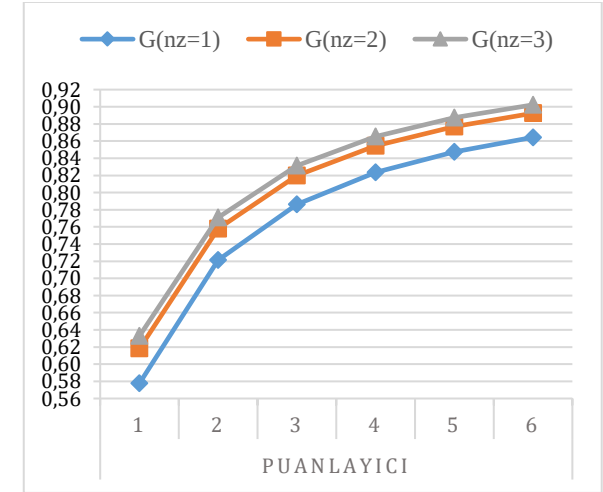
Şekil 5.c. Matematikleştirme Kategorisi İçin G Katsayıları Grafiği



Şekil 5.d. Temsil Kategorisi İçin G Katsayıları Grafiği

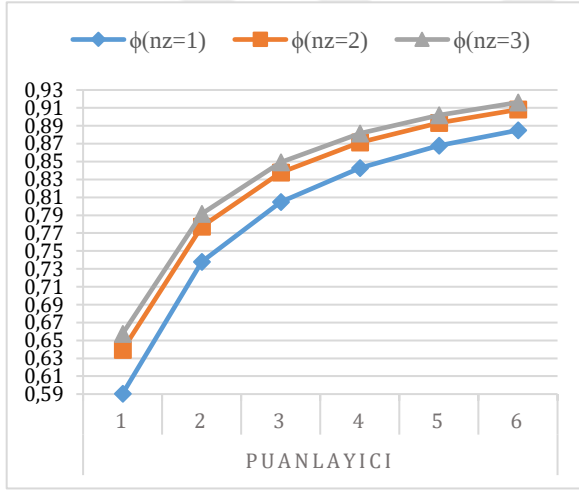


Şekil 5.e. Sembollerin vs. Kullanımı Kategorisi İçin G Katsayıları Grafiği

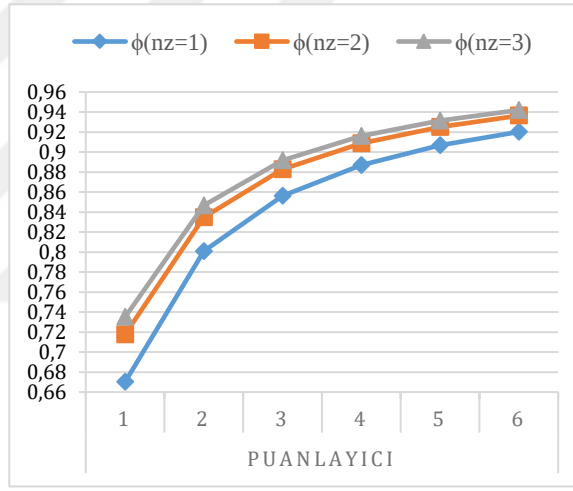


Şekil 5.f. Muhakeme ve Argüman Kategorisi İçin G Katsayıları Grafiği

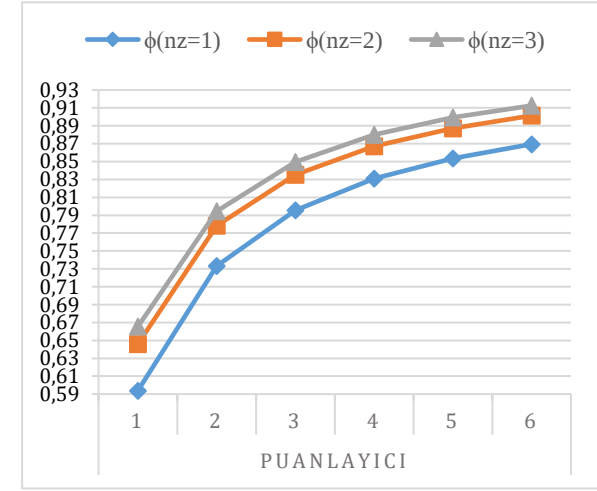
Şekil 5. Çok değişkenli K çalışmaları sonucunda alt kategoriler için kestirilen G katsayıları



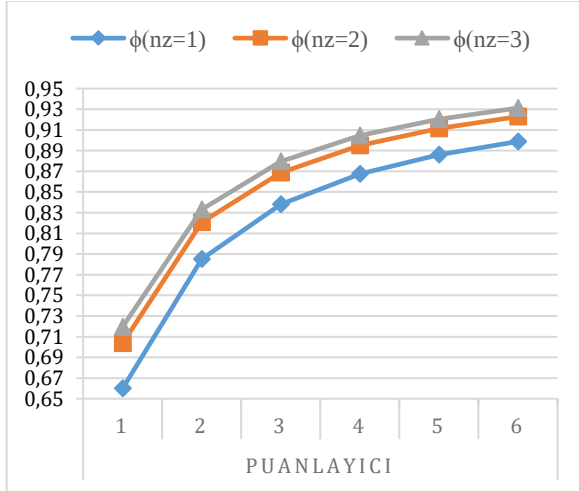
Şekil 4.a. İletişim Kategorisi İçin Phi Katsayıları Grafiği



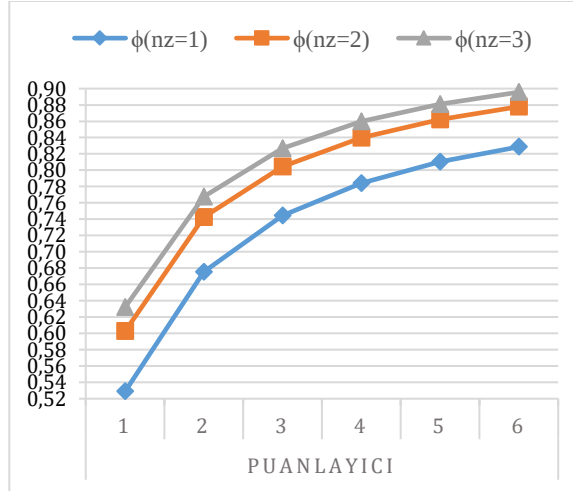
Şekil 4.b. Strateji Planlama Kategorisi İçin Phi Katsayıları Grafiği



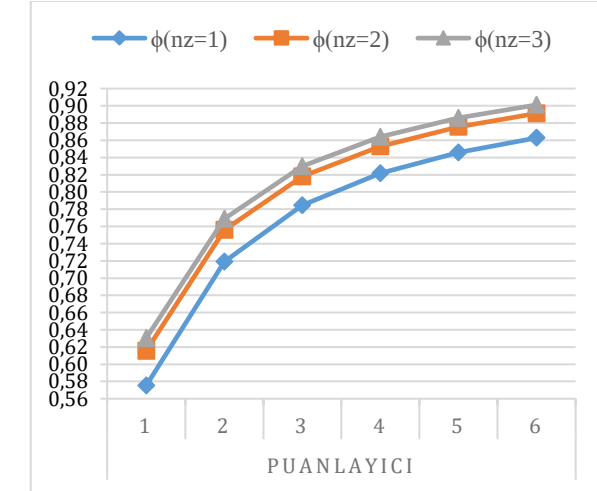
Şekil 4.c. Matematikleştirme Kategorisi İçin Phi Katsayıları Grafiği



Şekil 4.d. Temsil Kategorisi İçin Phi Katsayıları Grafiği



Şekil 4.e. Sembollerin vs. Kullanımı Kategorisi İçin Phi Katsayıları Grafiği



Şekil 4.f. Muhakeme ve Argüman Kategorisi İçin Phi Katsayıları Grafiği

Şekil 6. Çok değişkenli K çalışmaları sonucunda alt kategoriler için kestirilen Phi katsayıları

Strateji planlama alt kategorisi için orijinal puanlayıcı (5) ve zaman (2) sayıları için çok değişkenli model K çalışması ile kestirilen G ve Phi katsayıları 0,9271 ve 0,9251 olarak hesaplanmıştır. Bu alt kategori için puanlayıcı sayısı ($n_p=5$) sabit tutularak zaman sayısının bir azaltılması ($n_z=1$) durumunda kestirilen G ve Phi katsayılarının sırasıyla 0,9090 ve 0,9066 olduğu, G katsayısının 0,0181 ve Phi katsayısının 0,0185 kadar azaldığı görülmüştür. Puanlayıcı sayılarının sabit tutularak zaman sayısının bir arttırılması ($n_z=3$) durumunda ise kestirilen G ve Phi katsayıları sırasıyla 0,9333 ve 0,9314 olarak hesaplanmış, G katsayısının 0,0062 ve Phi katsayısının 0,0063 arttığı görülmüştür.

Zaman sayısı ($n_z=2$) sabit tutularak puanlayıcı sayısının bir azaltılması ($n_p=4$) durumunda kestirilen G ve Phi katsayılarının sırasıyla 0,9111 ve 0,9087 olduğu, G katsayısının 0,016 ve Phi katsayısının 0,0164 azaldığı görülmüştür. Zaman sayısı ($n_z=2$) sabit tutularak puanlayıcı sayısının bir arttırılması ($n_p=6$) durumundaysa G ve Phi katsayıları sırasıyla 0,9381 ve 0,9364 olarak kestirilmiş, G ve Phi katsayılarında sırasıyla 0,011 ve 0,0113 artış görülmüştür.

Puanlayıcı ve zaman sayısının birlikte azaltılması durumunda ($n_p=4$ ve $n_z=1$) G ve Phi katsayıları 0,8899 ve 0,8871 olarak hesaplanmıştır. Orijinal puanlayıcı ve zaman sayılarına göre kestirilen G ve Phi katsayıları ile karşılaştırıldığında puanlayıcı ve zaman sayısının birlikte azaltılması durumunda ($n_p=4$ ve $n_z=1$) kestirilen G ve Phi katsayılarının 0,0372 ve 0,038 azaldığı görülmüştür. Puanlayıcı ve zaman sayısının birlikte arttırılması ($n_p=6$ ve $n_z=3$) durumundaysa, G ve Phi katsayıları sırasıyla 0,9435 ve 0,9419 olarak kestirilmiş, G ve Phi katsayılarının 0,0164 ve 0,0168 kadar arttığı bulunmuştur.

İletişim kategorisinde olduğu gibi strateji planlama kategorisinde de zaman ve puanlayıcı varyansının toplam varyans içindeki oranlarının (%0; %0,6) sıfır/ sıfıra yakın olmasından dolayı zaman/puanlayıcı sayısının arttırılmasının ya da azaltılmasının G ve Phi katsayılarını ciddi bir şekilde etkilemediği söylenebilir. Puanlayıcı sayısı sabit tutulurken zaman sayısının ya da puanlayıcı-zaman sayılarının birlikte bir azaltılması durumlarında G ve Phi katsayılarında meydana gelen değişim bir arttırıldığı durumlarda meydana gelen değişimden daha fazladır. Ancak zaman sayısı sabit tutularak puanlayıcı sayısının bir arttırılıp azaltılması durumlarında G ve Phi katsayılarında meydana gelen değişimlerin birbirine daha yakın olduğu görülmüştür.

Matematikleştirme alt kategorisinde orijinal puanlayıcı (5) ve zaman (2) sayıları için çok değişkenli model K çalışması ile kestirilen G ve Phi katsayıları 0,8884 ve 0,8874 olarak

hesaplanmıştır. Bu alt kategori için puanlayıcı sayısı ($n_p=5$) sabit tutularak zaman sayısının bir azaltılması ($n_z=1$) durumunda G ve Phi katsayıları sırasıyla 0,8549 ve 0,8536 olarak kestirilmiş, G ve Phi katsayılarının sırasıyla 0,0335 ve 0,0338 kadar azaldığı görülmüştür. Puanlayıcı sayılarının sabit tutularak zaman sayısının bir arttırılması ($n_z=3$) durumunda ise kestirilen G ve Phi katsayıları sırasıyla 0,9002 ve 0,8992 olurken, G ve Phi katsayılarının her ikisi de 0,0118 artmıştır. Puanlayıcı sayısı sabit tutulurken zaman sayısının arttırıp azaltılması durumunda matematikleştirme ve iletişim alt kategorisine ilişkin sonuçlar benzerlik göstermektedir.

Zaman sayısı ($n_z=2$) sabit tutularak, puanlayıcı sayısının bir azaltılması ($n_p=4$) durumunda G ve Phi katsayıları 0,8683 ve 0,8671 olarak kestirilmiş, G ve Phi katsayılarının sırasıyla 0,0201 ve 0,0203 azaldığı görülmüştür. Zaman sayısı ($n_z=2$) sabit tutularak, puanlayıcı sayısının bir arttırılması ($n_p=6$) durumunda ise kestirilen G ve Phi katsayıları 0,9024 ve 0,9014 olarak hesaplanmış, G ve Phi katsayılarının her ikisi de 0,014 artmıştır.

Puanlayıcı ve zaman sayısının birlikte azaltılması ($n_p=4$ ve $n_z=1$) durumunda G ve Phi katsayıları 0,8322 ve 0,8309 olarak hesaplanmış, G ve Phi katsayılarının sırasıyla 0,0562 ve 0,0564 azaldığı görülmüştür. Puanlayıcı ve zaman sayısının birlikte arttırılması ($n_p=6$ ve $n_z=3$) durumunda G ve Phi katsayıları 0,9134 ve 0,9126 olarak kestirilmiş, G ve Phi katsayılarında sırasıyla 0,025 ve 0,0252 artış gözlenmiştir.

Diğer alt kategorilerde olduğu gibi bu alt kategori için de zaman ve puanlayıcı varyanslarının toplam varyans içindeki oranlarının (%0;%0,3) sıfır/ sıfıra yakın olmasından dolayı zaman/puanlayıcı sayısının arttırılmasının ya da azaltılmasının G ve Phi katsayılarını ciddi bir şekilde etkilemediği söylenebilir. Zaman ya da puanlayıcı sayılarından biri ya da her ikisi birlikte bir azaltıldığında G ve Phi katsayılarında meydana gelen değişim, bunlardan biri ya da her ikisi birlikte bir arttırıldığında meydana gelen değişimden daha fazladır.

Temsil alt kategorisinde orijinal puanlayıcı (5) ve zaman (2) sayıları için çok değişkenli model K çalışması ile kestirilen G ve Phi katsayıları sırasıyla 0,9158 ve 0,9116 olarak hesaplanmıştır. Bu alt kategori için puanlayıcı sayısı ($n_p=5$) sabit tutularak zaman sayısının bir azaltılması ($n_z=1$) durumunda kestirilen G ve Phi katsayılarının sırasıyla 0,8923 ve 0,8860 olduğu, G ve Phi katsayılarının sırasıyla 0,0235 ve 0,0256 azaldığı görülmüştür. Puanlayıcı sayısı ($n_p=5$) sabitken zaman sayısının bir arttırılması ($n_z=3$) durumunda ise G ve Phi katsayıları 0,9239 ve 0,9204 olarak hesaplanmış, G ve Phi katsayılarının sırasıyla 0,0081 ve 0,0088 arttığı gözlenmiştir. Puanlayıcı sayısı sabitken zaman sayısının bir azaltılıp

arttırılması durumlarında zaman sayısının bir azaltılmasının bir arttırılmasına göre güvenilirlik parametreleri üzerinde daha etkili olduğu ortaya çıkmıştır.

Zaman sayısı ($n_z=2$) sabit tutularak puanlayıcı sayısı bir azaltıldığında ($n_p=4$) G ve Phi katsayılarının sırasıyla 0,8997 ve 0,8950 olarak kestirildiği, G ve Phi katsayılarının sırasıyla 0,0161 ve 0,0166 azaldığı görülmüştür. Zaman sayısı ($n_z=2$) sabitken puanlayıcı sayısının bir arttırılması ($n_z=6$) durumundaysa G ve Phi katsayılarının sırasıyla 0,9269 ve 0,9229'a yükseldiği, G ve Phi katsayılarında sırasıyla 0,011 ve 0,0113 artış olduğu görülmüştür.

Puanlayıcı ve zaman sayısının birlikte azaltılması ($n_p=4$ ve $n_z=1$) durumunda kestirilen G ve Phi katsayıları sırasıyla 0,8741 ve 0,8675'e düşerken G ve Phi katsayılarının sırasıyla 0,0417 ve 0,0441 azaldığı görülmüştür. Puanlayıcı ve zaman sayısının birlikte arttırılması ($n_p=6$ ve $n_z=3$) durumundaysa kestirilen G ve Phi katsayıları 0,9344 ve 0,9312'a yükselmiş, G ve Phi katsayılarının sırasıyla 0,0186 ve 0,0196 arttığı görülmüştür.

Orijinal puanlayıcı ve zaman sayılarına göre kestirilen G ve Phi katsayıları ile karşılaştırıldığında, puanlayıcı ve zaman sayılarının bir arttırıp bir azaldığı senaryolara göre kestirilen G ve Phi katsayılarında büyük bir değişim olmadığı; puanlayıcı ve zaman sayısı birlikte azaltıldığında kestirilen G ve Phi katsayılarındaki azalma miktarlarının (0,0417;0,0441) diğerlerine göre nispeten daha fazla olduğu; artış miktarlarının ise birbirine çok yakın ve düşüş miktarlarından daha az olduğu görülmüştür. Bununla birlikte zaman sayısı sabit tutularak puanlayıcı sayısının bir arttırılıp azaltılması durumlarında G ve Phi katsayılarında meydana gelen değişimlerin birbirine daha yakın olduğu göze çarpmıştır. Ayrıca diğer alt kategorilerde olduğu gibi bu alt kategori için de zaman ve puanlayıcı varyanslarının toplam varyans içindeki oranlarının (%0,3;%0,7) sifra çok yakın olmasından dolayı zaman/puanlayıcı sayısının arttırılmasının ya da azaltılmasının G ve Phi katsayılarını ciddi bir şekilde etkilemediği söylenebilir.

Sembollerin, işlemlerin ve formal dilin kullanımı alt kategorisinde orijinal puanlayıcı (5) ve zaman (2) sayıları için çok değişkenli model K çalışması ile kestirilen G katsayısı 0,8713 ve Phi katsayısı 0,8622 olarak hesaplanmıştır. Bu alt kategori için puanlayıcı sayısı ($n_p=5$) sabit tutularak zaman sayısının bir azaltılması ($n_z=1$) durumunda kestirilen G katsayısı 0,8212 ve Phi katsayısı 0,8103 olup, G ve Phi katsayılarının sırasıyla 0,0501 ve 0,0519 azaldığı görülmüştür. Puanlayıcı sayısı ($n_p=5$) sabitken zaman sayısı bir arttırıldığında ($n_z=3$) G katsayısı 0,8894 ve Phi katsayısı 0,8810 olarak kestirilmiş, G ve Phi katsayılarının sırasıyla 0,0181 ve 0,0188 arttığı gözlenmiştir. Puanlayıcı sayısı sabitken zaman sayısının bir azaltılıp

arttırılması durumlarında zaman sayısının bir azaltılmasının bir arttırılmasına göre güvenilirlik parametreleri üzerinde daha etkili olduğu ortaya çıkmıştır.

Zaman sayısı ($n_z=2$) sabit tutularak, puanlayıcı sayısının bir azaltılması ($n_p=4$) durumundaysa G ve Phi katsayıları sırasıyla 0,8497 ve 0,8396 olarak kestirilmiş, G ve Phi katsayılarının sırasıyla 0,0216 ve 0,0226 azaldığı ortaya çıkmıştır. Zaman sayısı ($n_z=2$) sabit tutularak puanlayıcı sayısının bir arttırıldığında ($n_z=6$) kestirilen G ve Phi katsayılarının sırasıyla 0,8864 ve 0,8780'e yükseldiği, G ve Phi katsayılarının sırasıyla 0,0151 ve 0,0158 arttığı görülmüştür.

Puanlayıcı ve zaman sayılarının birlikte azaltılması ($n_p=4$ ve $n_z=1$) durumunda G katsayısı 0,7957 ve Phi katsayısı 0,7842 olarak hesaplanırken, bu katsayıların sırasıyla 0,0756 ve 0,078 azaldığı bulunmuştur. Puanlayıcı ve zaman sayıları birlikte arttırıldığında ($n_p=6$ ve $n_z=3$) G ve Phi katsayılarının 0,9033 ve 0,8957'ye yükseldiği, G ve Phi katsayılarında 0,032 ve 0,0335 artış olduğu görülmüştür.

Diğer kategorilerde olduğu gibi bu alt kategoride de zaman ve puanlayıcı varyanslarının toplam varyans içindeki oranlarının (%0,4;%2,1) sifıra yakın olmasından dolayı zaman/puanlayıcı sayısının arttırılması ya da azaltılması G ve Phi katsayılarını ciddi bir şekilde etkilememiştir. Puanlayıcı ve zaman sayısı birlikte azaltıldığında kestirilen G ve Phi katsayılarındaki azalma miktarlarının (0,0756; 0,078) puanlayıcı ve zaman sayılarının bir arttırıp bir azaldığı senaryolara göre kestirilen G ve Phi katsayılarına göre nispeten daha fazla olduğu; artış miktarlarının ise birbirine çok yakın ve düşüş miktarlarından daha az olduğu görülmüştür. Ayrıca puanlayıcı ve zaman sayısı birlikte azaltıldığında kestirilen G ve Phi katsayılarının 0,80'in altına düştüğü, diğer bir deyişle kabul edilebilir düzeyin altına olduğu ortaya çıkmıştır.

Muhakeme ve argüman alt kategorisi için orijinal puanlayıcı (5) ve zaman (2) sayıları için çok değişkenli model K çalışması ile kestirilen G katsayısı 0,8770 ve Phi katsayısı 0,8756 olarak hesaplanmıştır. Puanlayıcı sayısı ($n_p=5$) sabit tutularak zaman sayısının bir azaltılması ($n_z=1$) durumunda G katsayısı 0,8476 ve Phi katsayısı 0,8459 olarak kestirilmiş olup, G ve Phi katsayılarında 0,0294 ve 0,0297 azalma olduğu görülmüştür. Puanlayıcı sayısı ($n_p=5$) sabit tutularak zaman sayısının bir arttırıldığında ($n_z=3$) kestirilen G ve Phi katsayıları 0,8872 0,8859'a yükselmiş, G ve Phi katsayılarındaki artış miktarları 0,0102 ve 0,0104 olarak hesaplanmıştır.

Zaman sayısı ($n_z=2$) sabit tutularak, puanlayıcı sayısı bir azaltıldığında ($n_p=4$) G ve Phi katsayıları 0,8546 ve 0,8530 olarak kestirilmiş, G ve Phi katsayıların sırasıyla 0,0224 ve 0,0226 azaldığı görülmüştür. Zaman sayısı ($n_z=2$) sabit tutularak, puanlayıcı sayısının bir artırılması ($n_z=6$) durumunda kestirilen G ve Phi katsayıları ise 0,8925 ve 0,8912'a yükselmiş, G ve Phi katsayılarında 0,0155 ve 0,0156 artış gözlenmiştir.

Puanlayıcı ve zaman sayılarının birlikte azaltılması durumunda ($n_p=4$ ve $n_z=1$) G katsayısı 0,8236 ve Phi katsayısı 0,8217 olarak kestirilirken G ve Phi katsayılarının sırasıyla 0,0534 ve 0,0539 azaldığı görülmüştür. Puanlayıcı ve zaman sayıları birlikte arttırıldığında ($n_p=6$ ve $n_z=3$) kestirilen G ve Phi katsayıları sırasıyla 0,9023 ve 0,9011'a yükselmiş, G ve Phi katsayılarında sırasıyla 0,0253 ve 0,0255 artış gözlenmiştir.

Diğer kategorilerde olduğu gibi bu kategoride de zaman ve puanlayıcı varyanslarının toplam varyans içindeki oranlarının (%0,1;%0,4) sıfıra yakın olmasından dolayı zaman/puanlayıcı sayısının arttırılması ya da azaltılması G ve Phi katsayılarını ciddi bir şekilde etkilememiştir. Puanlayıcı ve zaman sayısı birlikte azaltıldığında kestirilen G ve Phi katsayılarındaki azalma miktarlarının (0,0534 ve 0,0539) puanlayıcı ve zaman sayılarının bir arttırıp bir azaldığı senaryolara göre kestirilen G ve Phi katsayılarına göre nispeten daha fazla olduğu; artış miktarlarının ise birbirine çok yakın ve düşüş miktarlarından daha az olduğu görülmüştür. Ayrıca puanlayıcı ve zaman sayısı birlikte arttırıldığında kestirilen G ve Phi katsayılarının 0,90'ın üstüne çıktığı gözlenmiştir.

Genel olarak tüm alt kategoriler için G ve Phi katsayılarının birbirine yakın değerler aldığı ve paralellik gösterdiği, G katsayılarının Phi katsayılarından daha yüksek çıkma eğiliminde olduğu görülmektedir. Farklı puanlayıcı zaman senaryolarına göre kestirilen G katsayılarındaki artış ya da azalışa bağlı olarak Phi katsayılarının da arttığı ya da azaldığı görülmektedir. Zaman sayılarının azaltılması ile G ve Phi katsayılarında meydana gelen azalma puanlayıcı sayısının azaltılması ile meydana gelen azalmadan daha fazla olmuştur. Diğer yandan G ve Phi katsayılarında puanlayıcı sayısının arttırılması ile meydana gelen artış zaman sayısının arttırılması ile meydana gelen artıştan daha yüksektir. Zaman ve ekonomi dikkate alındığında, G ve Phi katsayılarının arttırılması için puanlayıcı sayısının arttırılmasının daha doğru olacağı söylenebilir.

3.3.Araştırma Sorusuna İlişkin Bulgular

“Orijinal puanlayıcı ve zaman sayıları ile puanlayıcı ve zaman sayılarının bir arttırılıp bir azaltılması senaryolarına göre çok değişkenli ve tek değişkenli Karar (K) çalışmalarıyla toplam puanlara ilişkin kestirilen G ve Phi katsayıları modellere göre nasıl değişmektedir?”

Toplam puanlar için puanlayıcı ve zaman sayılarının bir arttırılıp bir azaltıldığı senaryolara göre tek değişkenli ve çok değişkenli Karar çalışmaları ile kestirilen G ve Phi katsayıları Tablo 17’de ve bu katsayılarla ilişkin grafikler Şekil 6’de sunulmuştur.

Tablo 17

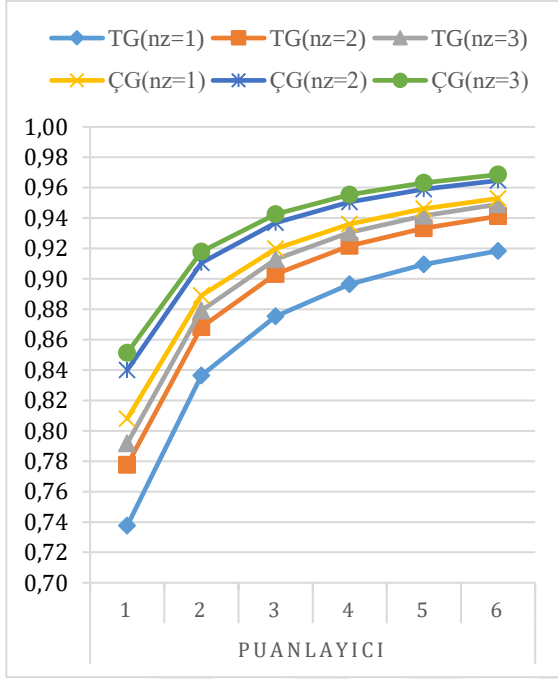
Toplam Puanlara için Zaman ve Puanlayıcı Sayıları Senaryolarına Göre Çok Değişkenli ve Tek Değişkenli K Çalışması ile Kestirilen G ve Phi Katsayıları

Model	Puanlayıcı	Nz=1		Nz=2		Nz=3	
	Sayısı	G	ϕ	G	ϕ	G	Φ
Çok Değişkenli	4	0,9361	0,9340	0,9506	0,9490	0,9555	0,9541
	5	0,9461	0,9443	0,9590	0,9576	0,9633	0,9622
	6	0,9529	0,9512	0,9647	0,9634	0,9686	0,9676
Tek Değişkenli	4	0,8964	0,8902	0,9218	0,9167	0,9306	0,9259
	5	0,9095	0,9038	0,9333	0,9289	0,9416	0,9376
	6	0,9184	0,9131	0,9412	0,9372	0,9491	0,9455

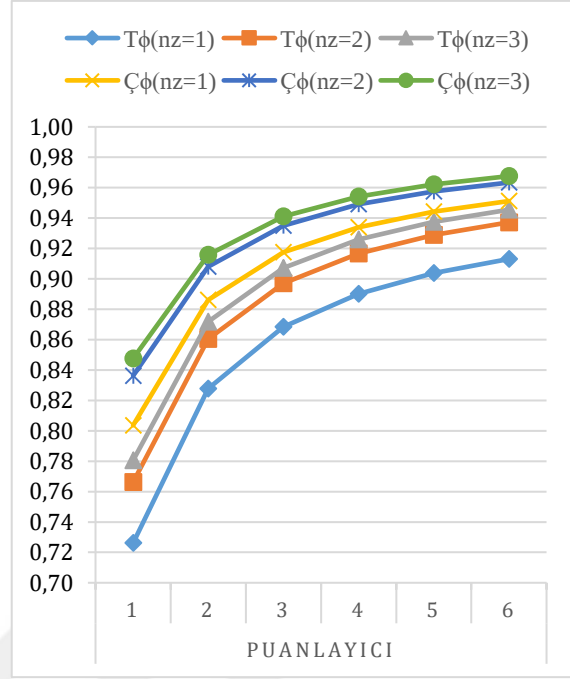
Not: Koyu yazılmış değerler çalışmada yer verilen zaman ve puanlayıcı sayılarına göre kestirilmiştir.

Tablo 16 ve Şekil 7’de görüldüğü gibi beş puanlayıcının iki farklı zamanda yaptığı puanlamalara ilişkin olarak çok değişkenli model K çalışması ile kestirilen G ve Phi katsayıları 0,9590;0,9576 ve tek değişkenli model K çalışması ile kestirilen G ve Phi katsayıları 0,9333;0,9289 olarak hesaplanmıştır. Çok değişkenli modelle hesaplanan G ve Phi katsayılarının tek değişkenli modelle hesaplanan G ve Phi katsayılarından 0,0257;0,0287 daha fazla olduğu göze çarpmaktadır.

Puanlayıcı sayısı ($n_p=5$) sabit tutularak zaman sayısının bir azaltılması ($n_z=1$) durumunda çok değişkenli modelle kestirilen G ve Phi katsayıları 0,9461 ve 0,9443 iken tek değişkenli ile modelle kestirilen G ve Phi katsayılarının 0,9095 ve 0,9038 olduğu görülmektedir.



Şekil 7.a. Toplam Puanlar İçin Çok Değişkenli ve Tek Değişkenli Model G Katsayıları Grafiği



Şekil 7.b. Toplam Puanlar İçin Çok Değişkenli ve Tek Değişkenli Model Phi Katsayıları Grafiği

Şekil 7. Toplam puanlara için zaman ve puanlayıcı sayıları senaryolarına göre çok değişkenli ve tek değişkenli K çalışması ile kestirilen G ve Phi katsayıları

Puanlayıcı sayısı ($n_p=5$) sabit tutularak zaman sayısının ($n_z=1$) azaltıldığı durumda çok değişkenli modelle hesaplanan G ve Phi katsayılarının tek değişkenli modelle hesaplanan G ve Phi katsayılarından 0,0366; 0,0405 daha fazla olduğu sonucuna ulaşılmıştır. Puanlayıcı sayısı ($n_p=5$) sabitken zaman sayısının bir arttırılması ($n_z=3$) durumunda çok değişkenli modelle kestirilen G ve Phi katsayıları 0,9633; 0,9622 iken tek değişkenli modelle kestirilenler 0,9416; 0,9376 olarak hesaplanmıştır. Puanlayıcı sayısı ($n_p=5$) sabitken zaman sayısının bir arttırılması ($n_z=3$) durumunda çok değişkenli modelle kestirilen G ve Phi katsayılarının tek değişkenli modelle kestirilenlerden 0,0217; 0,0246 daha fazla görülmüştür.

Zaman sayısı ($n_z=2$) sabit tutulup puanlayıcı sayısının bir azaltılması ($n_p=4$) durumunda çok değişkenli modelle G ve Phi katsayıları 0,9506 ve 0,9490 olarak kestirilirken, tek değişkenli ile modelle 0,9218 ve 0,9167 olarak kestirilmiştir. Çok değişkenli modelle kestirilen G ve Phi katsayılarının tek değişkenli modelle kestirilenlerden 0,0288 ve 0,0323 daha fazla olduğu görülmüştür. Zaman sayısı ($n_z=2$) sabitken puanlayıcı sayısının bir arttırılması ($n_z=6$) durumunda G ve Phi katsayıları çok değişkenli modelle 0,9647 ve 0,9634 olarak, tek değişkenli modelle 0,9412 ve 0,9372 olarak kestirilmiştir. Bu senaryo için çok değişkenli

modelle kestirilen G ve Phi katsayılarının tek değişkenli modelle kestirilenlerden 0,0235 ve 0,0262 kadar daha yüksek hesaplandığı görülmüştür.

Puanlayıcı ve zaman sayısının birlikte azaltılması ($n_p=4$ ve $n_z=1$) durumunda kestirilen G ve Phi katsayıları çok değişkenli modelle 0,9361 ve 0,9340 olarak hesaplanırken, tek değişkenli modelle 0,8964 ve 0,8902 olarak hesaplanmıştır. Puanlayıcı ve zaman sayısının birlikte azaltıldığı ($n_p=4$ ve $n_z=1$) senaryo için çok değişkenli modelle kestirilen G ve Phi katsayılarının tek değişkenli modelle kestirilenlerden 0,0195 ve 0,0221 daha fazla olduğu görülmektedir. Puanlayıcı ve zaman sayısının birlikte arttırılması ($n_p=6$ ve $n_z=3$) durumunda G ve Phi katsayıları çok değişkenli modelle 0,9686; 0,9676 ve tek değişkenli modelle 0,9491;0,9455 olarak kestirilmiştir. Puanlayıcı ve zaman sayısının birlikte arttırılması ($n_p=6$ ve $n_z=3$) durumunda çok değişkenli modelle kestirilen G ve Phi katsayılarının tek değişkenli modelle kestirilenlerden 0,0397 ve 0,0438 daha fazla olduğu görülmüştür.

Dördüncü Araştırma Sorusuna İlişkin Bulgular

“Yeterlilik Şeması ile elde edilen her bir kategoriye ait puanlar için çok değişkenli model K çalışmalarıyla kabul edilebilir düzeyde bir G katsayısı elde etmek için gerekli minimum puanlayıcı ve zaman sayısı nedir?”

Araştırmanın dördüncü sorusuna yönelik olarak Yeterlilik Şeması ile elde edilen her bir kategoriye ait puanlar için zaman sayılarının bir ile üç, puanlayıcı sayılarının bir ile altı arasında değiştiği senaryolar için çok değişkenli model K çalışmaları yapılmıştır. K çalışmaları sonucunda farklı senaryolara göre kestirilen G ve Phi katsayıları incelenerek kabul edilebilir düzeyde bir G katsayısı elde etmek için gerekli minimum puanlayıcı ve zaman sayıları önerilmiştir.

Yeterlilik Şeması’nın her bir alt kategorisi matematik okuryazarlığının ölçülmesinde göz önünde tutulan temel matematiksel yeterliliklere karşılık gelmektedir. Maddelere ilişkin elde edilen toplam puanlar bir maddenin matematik okuryazarlığı gereksinimine ilişkin bilgi verirken, alt kategorilere ilişkin puanlar her bir yeterliliğe ilişkin gereksinimler bazında bilgi vermektedir. Örneğin bir maddenin temsil yeterliliği gereksinimi düşük olabilirken (0) matematikleştirme yeterliliği gereksinimi (2) ya da matematik okuryazarlığı gereksinimi (7) daha yüksek olabilmektedir. Araştırmacıların ilgilendikleri alt kategoriler bazında güvenilir ölçümler elde etmelerine ilişkin önerilerde bulunulabilmesi için her bir alt kategoriye ilişkin

gerekli minimum puanlayıcı ve zaman sayısı aşağıdaki Tablo 16, Şekil 5 ve Şekil 6’da verilen katsayılar üzerinden yorumlanmıştır.

Ölçmelerin kabul edilebilir güvenilirlik düzeyinde olması için G ve Phi katsayılarının en az 0,80 (Cardinet vd., 2010), yüksek güvenilirliğe işaret etmesi içinse en az 0,90 ve üzeri (Güler, 2008; Nalbantoğlu, 2009) olması gerektiği belirtilmektedir. Puanlayıcı sayısının bir ile altı arasında zaman sayısının bir ile üç arasında değiştiği senaryolar için kestirilen G ve Phi katsayıları bu ölçütlere göre incelendiğinde, araştırmada tasarlanan senaryo için ölçümlerin güvenilirliğinin strateji planlama ve temsil alt kategorilerinde yüksek diğer alt kategorilerde ise kabul edilebilir düzeyin üstünde olduğu görülmüştür.

İletişim alt kategorisine yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının üç ve zaman sayısının bir olmasının yeterli olduğu görülmektedir. Puanlayıcı sayısının üç ya da üçten fazla ve zaman sayısının bir ya da birden fazla olduğu tüm senaryolarda güvenirliliğin 0,80’in üstünde olduğu görülmektedir.

Strateji planlama alt kategorisine yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının iki ve zaman sayısının bir olmasının yeterli olduğu görülmektedir. Puanlayıcı sayısının iki ya da üstünde ve zaman sayısının bir ya da daha fazla olduğu tüm senaryolarda güvenirliliğin 0,80’in üstünde olduğu görülmektedir.

Matematikleştirme alt kategorisine yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının iki ve zaman sayısının üç ya da minimum puanlayıcı sayısının üç ve zaman sayısının bir olmasının yeterli olduğu görülmektedir. Bu noktada araştırmacılar puanlayıcı sayısı ve zaman sayısına yönelik olarak seçim yapmada kendileri için en uygun olan ekonomiklik ölçütünü (puanlayıcı ya da zaman) göz önünde tutarak uygun senaryoya karar verebilirler. G ve Phi katsayılarını arttırmada puanlayıcı sayısının üç ya da üçten fazla ve zaman sayısının bir ya da birden fazla olduğu tüm senaryolarda güvenirliliğin 0,80’in üstünde olduğu görülmektedir.

Temsil alt kategorisine yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının iki ve zaman sayısının iki ya da minimum puanlayıcı sayısının üç ve zaman sayısının bir olmasının yeterli olduğu görülmektedir. Bu noktada araştırmacılar puanlayıcı sayısı ve zaman sayısına yönelik olarak seçim yapmada kendileri için en uygun olan ekonomiklik ölçütünü (puanlayıcı ya da zaman) göz önünde tutarak uygun senaryoya karar verebilirler. G ve Phi katsayılarını arttırmada puanlayıcı

sayısının üç ya da üçten fazla ve zaman sayısının bir ya da birden fazla olduğu tüm senaryolarda güvenilirliğin 0,80'in üstünde olduğu görülmektedir.

Sembollerin, işlemlerin ve formal dilin kullanımı alt kategorisine yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının üç ve zaman sayısının iki olmasının yeterli olduğu görülmektedir. Zaman sayısının bir olarak seçilmesi söz konusu olduğunda ise dört puanlayıcı ile elde edilen ölçümlerin G katsayısı kabul edilebilir düzeyde iken Phi katsayı 0,80'in altında kaldığı görülmektedir. Bu noktada araştırmacılar puanlayıcı sayısı ve zaman sayısına yönelik olarak seçim yapmada kendileri için en uygun olan ekonomiklik ölçütünü (puanlayıcı ya da zaman) göz önünde tutarak uygun senaryoya karar verebilirler. G ve Phi katsayılarını arttırmada puanlayıcı sayısının beş ya da beşten fazla ve zaman sayısının bir ya da birden fazla olduğu tüm senaryolarda güvenilirliğin 0,80'in üstünde olduğu görülmektedir.

Muhakeme ve argüman alt kategorisine yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının üç ve zaman sayısının iki ya da minimum puanlayıcı sayısının dört ve zaman sayısının bir olmasının yeterli olduğu görülmektedir. Bu noktada araştırmacılar puanlayıcı sayısı ve zaman sayısına yönelik olarak seçim yapmada kendileri için en uygun olan ekonomiklik ölçütünü (puanlayıcı ya da zaman) göz önünde tutarak uygun senaryoya karar verebilirler. G ve Phi katsayılarını arttırmada puanlayıcı sayısının üç ya da üçten fazla ve zaman sayısının bir ya da birden fazla olduğu tüm senaryolarda güvenilirliğin 0,80'in üstünde olduğu görülmektedir.

Genel olarak tüm alt kategoriler için tek bir puanlayıcının farklı zamanlarda yaptığı puanlamaların kabul edilebilir güvenilirlik düzeyinin altında sonuçlar verdiği, tüm kategoriler için minimum puanlayıcı sayısı üç ve zaman sayısı iki olduğunda kabul edilebilir düzeyde güvenilir ölçümler yapıldığı söylenebilir.

Beşinci Araştırma Sorusuna İlişkin Bulgular

“Yeterlilik Şeması ile elde edilen toplam puanlar için çok değişkenli ve tek değişkenli model K çalışmalarıyla kabul edilebilir düzeyde bir G katsayısı elde etmek için gerekli minimum puanlayıcı ve zaman sayısı modellere göre nasıl değişmektedir?”

Araştırmanın beşinci sorusuna yönelik olarak Yeterlilik Şeması ile elde edilen toplam puanlar için zaman sayılarının bir ile üç, puanlayıcı sayılarının bir ile altı arasında değiştiği

senaryolar için çok değişkenli ve tek değişkenli model K çalışmaları yapılmıştır. K çalışmaları sonucunda farklı senaryolara ve modellere göre kestirilen G katsayıları incelenerek kabul edilebilir düzeyde bir G katsayısı elde etmek için gerekli minimum puanlayıcı ve zaman sayıları önerilmiştir.

PISA matematik okuryazarlığı maddelerinin matematik okuryazarlığı gereksinimlerine ilişkin ölçümler için puanlayıcı sayısının bir ile altı zaman sayısının bir ile üç arasında değiştiği senaryolara göre tek değişkenli ve çok değişkenli K çalışmaları ile kestirilen G ve Phi katsayıları Tablo 18 ve bu katsayıların değişimi Şekil 7’de sunulmuştur.

Tablo 18

Toplam Puanlar için Çok Değişkenli ve Tek Değişkenli K Çalışmalarıyla Kestirilen G ve Phi Katsayıları

np	nz	Tek Değişkenli Model		Çok Değişkenli Model	
		G	Φ	G	Φ
1	1	0,7376	0,7261	0,8080	0,8036
1	2	0,7775	0,7662	0,8400	0,8362
1	3	0,7918	0,7805	0,8513	0,8476
2	1	0,8364	0,8278	0,8891	0,8861
2	2	0,8681	0,8604	0,9106	0,9081
2	3	0,8792	0,8718	0,9180	0,9157
3	1	0,8755	0,8684	0,9199	0,9175
3	2	0,9032	0,8971	0,9369	0,9350
3	3	0,9128	0,9072	0,9426	0,9410
4	1	0,8964	0,8902	0,9361	0,9340
4	2	0,9218	0,9167	0,9506	0,9490
4	3	0,9306	0,9259	0,9555	0,9541
5	1	0,9095	0,9038	0,9461	0,9443
5	2	0,9333	0,9289	0,9590	0,9576
5	3	0,9416	0,9376	0,9633	0,9622
6	1	0,9184	0,9131	0,9529	0,9512
6	2	0,9412	0,9372	0,9647	0,9634
6	3	0,9491	0,9455	0,9686	0,9676

Tek deęişkenli K alışması sonuçlarına göre, matematik okuryazarlığı gereksinim düzeylerine dięer bir deyişle toplam puanlara yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının iki ve zaman sayısının bir olmasının yeterli olduęu görölmektedir. Puanlayıcı sayısının iki ve zaman sayısının birden fazla olduęu tüm senaryolarda G ve Phi katsayılarının 0,80'in üstünde olduęu görölmektedir. Bununla birlikte güvenilirlik düzeyinin yüksek olması istendiğinde minimum puanlayıcı sayısının üç ve zaman sayısının iki olması gerektięi görölmektedir.

Çok deęişkenli K alışması sonuçlarına göre toplam puanlara yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için gerekli minimum puanlayıcı ve zaman sayılarının tek deęişkenli K alışması sonuçlarına göre gerekli olanlardan daha az olduęu ortaya çıkmıştır. Çok deęişkenli K alışmasına göre bir puanlayıcının bir zamanda puanlama yapmasıyla dahi kabul edilebilir güvenilirlik düzeyine ulaşılabilir. Bununla birlikte güvenilirlik düzeyinin yüksek olması istendiğinde minimum puanlayıcı sayısının iki ve zaman sayısının iki olmasının yeterli olduęu görölmektedir.

Altıncı Araştırma Sorusuna İlişkin Bulgular

“Yeterlilik Şeması ile elde edilen toplam puanlar ve her bir kategoriye ait puanlar için KTK ve G Kuramı çerçevesinde hesaplanan güvenilirlik katsayıları iç tutarlılık ölçütlerine göre tutarlı mıdır?”

Araştırmanın altıncı sorusuna yönelik olarak Yeterlilik Şeması ile elde edilen toplam puanlar ve her bir kategoriye ait puanlar için KTK ve G Kuramı çerçevesinde çok ve tek deęişkenli model G alışması ile kestirilen güvenilirlik katsayıları karşılaştırılmıştır. Bu güvenilirlik katsayıları Tablo 19’da verilmiştir.

Tablo 19 incelendiğinde Yeterlilik Şeması ile elde edilen puanlamalar için hesaplanan güvenilirlik katsayılarının deęişkenlik gösterdięi görölmektedir. İç ölçütler bağlamında Spearman rho ve Kendall W katsayıları için 0,70, G ve Phi katsayıları için 0,80 genel kabullerine uyularak güvenilirlik katsayılarının tutarlılıkları yorumlanmıştır. Tüm alt kategoriler için Kendall W, G ve Phi katsayılarının beklenen deęerlerden yüksek olduęu görölmektedir. Korelasyon katsayılarına bakıldığında sembollerin, işlemlerin ve formal dilin kullanımı kategorisi bağlamında P1 için hesaplanan puanlayıcı içi güvenilirlik katsayısı dışında dięerlerinin tamamının beklenen deęerde ya da beklenen deęerin üstünde olduęu söylenebilir.

Tablo 19

Yeterlilik Şeması ile Elde Edilen Toplam Puanlar ve Her Bir Kategoriyeye Ait Puanlar İçin KTK ve G Kuramı Çerçevesinde Hesaplanan Güvenirlilik Katsayıları

Kategori	KTK Puanlayıcı içi güvenirlilik (r_s)	Puanlayıcılar arası güvenirlilik (Kendall W)	G Kuramı Genellenebilirlik Katsayısı (G)	Güvenirlilik katsayısı (Phi)
İletişim	0,845	0,711	0,8941	0,8931
	0,936			
	0,889			
	0,829			
	0,725			
Strateji Planlama	0,934	0,784	0,9271	0,9251
	0,907			
	0,919			
	0,852			
	0,699			
Matematikleştirme	0,930	0,721	0,8884	0,8874
	0,842			
	0,911			
	0,800			
	0,705			
Temsil	0,951	0,781	0,9158	0,9116
	0,954			
	0,904			
	0,936			
	0,756			
Sembollerin, İşlemlerin ve Formal Dilin Kullanımı	0,421	0,704	0,8713	0,8622
	0,892			
	0,928			
	0,877			
	0,698			
Muhakeme Argüman	0,919	0,709	0,8770	0,8756
	0,861			
	0,909			
	0,879			
	0,752			
Toplam Puan	0,894	0,833	0,9333	0,9289
	0,939			
	0,946			
	0,902			
	0,799			

Bu durum KTK ve G Kuramı çerçevesinde yapılan yorumların farklılaşmasına neden olmaktadır. G kuramı ile kestirilen güvenilirlik parametrelerine dayanarak Yeterlilik Şemasının alt kategorileri için elde edilen puanların kabul edilebilir güvenilirlik düzeyinin üstünde olduğu söylenebilir. Ancak KTK ile kestirilen güvenilirlik katsayılarına bakıldığında puanlayıcılar arası güvenilirliğin güçlü, puanlayıcı içi güvenilirliğin bir durum dışında yüksek olduğu şeklinde bir sonuca ulaşılmaktadır. Bu durumun nedeninin G Kuramı'nın tüm hata kaynaklarını birlikte ele alarak değerlendirilmesinden kaynaklandığı düşünülmüştür. G kuramı puanlayıcı içi ve puanlayıcılar arası uyumu birlikte değerlendirerek tek bir güvenilirlik katsayısı sunmaktadır.



BÖLÜM V

SONUÇ, TARTIŞMA VE ÖNERİLER

Bu bölümde araştırmanın bulgularından ulaşılan sonuçlar her bir araştırma sorusu için ayrı ayrı özetlenmiş ve bu sonuçlar ilgili araştırmaların sonuçları doğrultusunda tartışılmıştır. Araştırma sonuçlarına göre Yeterlilik Şeması puanlama anahtarına, kullanıcılarına ve benzer çalışmalara yönelik önerilerde bulunulmuştur.

Araştırmada PISA matematik okuryazarlığı alanından yayınlanmış problemlerin tümü olan 156 maddenin matematik okuryazarlığı gereksinim düzeyleri beş puanlayıcı tarafından iki farklı zamanda puanlanmış ve bu puanlamaların geçerliği ölçüt ve yapı geçerliği ile güvenilirliği ise KTK ve G kuramı çerçevesinde incelenmiştir. Her bir bulguya ilişkin ulaşılan sonuçlar aşağıda verilmiştir.

Birinci Araştırma Sorusuna İlişkin Sonuç ve Tartışma

Bu bölümde birinci araştırma sorusuna ilişkin sonuçlar literatürle desteklenerek tartışılmıştır. Yeterlilik Şeması ile PISA matematik okuryazarlığı maddelerinin gereksinim düzeylerine ilişkin farklı puanlayıcıların farklı zamanlarda vermiş oldukları puanların ortalamasının ölçüt ve yapı geçerliğini sağladığı görülmüştür. Bu sonuç Turner ve Adams (2012) ile Turner vd. (2013)'ün Yeterlilik Şeması ile yapılan ölçümlerin geçerliğine ilişkin sunduğu sonuçları desteklemektedir.

1.1.Araştırma Sorusuna İlişkin Sonuç ve Tartışma

“Yeterlilik Şeması ile elde edilen ölçümlerin ölçüt geçerliliği ne düzeydedir?”

Bu araştırma sorusu kapsamında Yeterlilik Şeması ile elde edilen puanlamaların ölçüt geçerliliği incelenmiştir. Ölçüt olarak madde güçlük puanları ele alınmış olup yalnızca esas uygulamalarda kullanılan ve yayınlanmış 79 madde için madde güçlük puanları bilindiğinden ölçüt geçerliliği bu 79 maddeye ilişkin verilen puanlar üzerinden hesaplanmıştır. Esas uygulamalarda yer verilen ve yayınlanmış 79 PISA matematik okuryazarlığı maddesinin matematik okuryazarlığı gereksinim düzeyleri için beş puanlayıcıdan iki farklı zamanda elde edilen madde toplam puanları ortalaması ile OECD tarafından yayınlanan bu maddelerin deneysel güçlük puanları arasındaki korelasyon katsayısı 0,87 ($p=0,01$) olarak bulunmuştur. 0,70 ve üstü korelasyon katsayılarının pozitif yönde ve kuvvetli ilişkiyi işaret ettiği genel kabulüne göre esas uygulamalarda kullanılan ve deneysel güçlük puanları yayınlanmış PISA matematik okuryazarlığı maddelerinin yeterlilik düzeyleri ile deneysel güçlük puanları arasında pozitif yönlü ve kuvvetli bir ilişki olduğu sonucuna varılmıştır. Ayrıca yapılan basit doğrusal regresyon analizine göre maddelerin Yeterlilik Şeması ile belirlenen yeterlilik gereksinimi toplam puanlarının maddelerin deneysel güçlük puanlarını anlamlı olarak yordadığı ve güçlük puanlarındaki varyansın %74,7'sinin yeterlilik gereksinimi toplam puanı ile açıklanabildiği görülmüştür. Bu sonuç Turner ve Adams (2012) ile Turner vd. (2013)'in sonucunu desteklemektedir.

Turner vd. (2013) PISA 2003 ve PISA 2006 maddelerinin yeterlilik puanları ile güçlük puanları arasındaki ilişkiyi çoklu regresyon analizi ile incelemişlerdir. Madde güçlüğündeki değişkenliğin, PISA 2003 maddeleri üzerinden yapılan çoklu regresyon analizi sonucunda %70,5'inin (Düzeltilmiş R^2), PISA 2006 maddeleri üzerinden yapılan analiz sonucunda %71,4'ünün (Düzeltilmiş R^2) açıklanabildiği görülmüştür. Her iki durum içinde Durbin-Watson istatistiği sonucunda artıkların ilişkisiz olduğu belirtilmiştir. Turner ve Adams (2012)'nin çalışmasında da çoklu regresyon çalışmaları ile madde güçlüğü'nün %74 ile önemli bir ölçüde yordanabildiği görülmüştür. Bu sonuçlar PISA maddelerinin yeterlilik gereksinimlerinin madde güçlüğü'nün önemli bir yordayıcısı olduğunu desteklemektedir.

1.2.Araştırma Sorusuna İlişkin Sonuç ve Tartışma

“Yeterlilik Şeması puanlama anahtarının yapı geçerliği ne düzeydedir?”

Yeterlilik Şeması'nın yapı geçerliğinin incelenmesine yönelik olarak her bir puanlayıcının her bir maddeye her bir kategori için iki farklı zamanda verdiği puanların ortalamaları göstergeler olarak ele alınmıştır. Böylelikle 156 maddeyi beş puanlayıcının altı kategori bazında puanlamasıyla toplamda 30 gösterge olduğu kabul edilerek öncelikle açıklayıcı faktör analizi yapılmıştır. Açıklayıcı faktör analizinde iki göstergenin (P3_MU, P5_SP) birden fazla faktörde yüklendiği görüldüğünden analiz dışı bırakılmasıyla 28 gösterge için temel bileşenler analizi ve varimax döndürme yöntemi seçilmiştir. Puanlama anahtarının altı faktörlü bir yapı sergilediği, faktörlerde yer alan göstergelerin Yeterlilik Şeması alt kategorilerine paralel olduğu ve bu altı faktörün puanlamalardaki varyansın toplamda %77,20'sinin açıklanabildiği ortaya çıkmıştır. Bu sonuç açıklanan varyans oranları açısından Turner vd. (2013)'ün sonuçlarıyla benzerlik göstermektedir. Turner vd. (2013), 48 madde üzerinden temel bileşenler analizi ile döndürme yapılmaksızın yaptıkları açıklayıcı faktör analizi sonucunda özdeğeri 1'den büyük iki faktör olduğu ve bu iki faktörün puanlamalardaki varyansın toplamda %64'ünü açıkladığı belirtilmiştir. Puanlama anahtarının faktör yapısı açısından bu iki çalışmanın bulgularının birbirini desteklemediği görülmektedir. Bu durumun, puanlama anahtarının ilk versiyonu ile bu çalışmada kullanılan üçüncü versiyonu arasındaki farklılıklardan, puanlayıcı grubundan ya da maddde örnekleminde kaynaklanabileceği düşünülmüştür.

Açıklayıcı faktör analizi sonuçları ile karşılaştırmak için doğrulayıcı faktör analizi yapılmıştır. Doğrulayıcı faktör analizi sonucunda verinin modele iyi uyum sağladığı ve hiçbir göstergenin çıkarılmasına gerek olmadığı görülmüştür. Yapı geçerliği çalışmaları sonucunda Yeterlilik Şeması ile elde edilen puanlar için verinin modele iyi uyum sağladığı ve modelin doğrulandığı görülmüştür.

İkinci Araştırma Sorusuna İlişkin Sonuç ve Tartışma

Bu bölümde ikinci araştırma sorusuna yönelik sonuçlar literatürle desteklenerek tartışılmıştır. Araştırmanın ikinci sorusuna yönelik olarak KTK çerçevesinde Yeterlilik Şeması ile toplanan verilerin puanlayıcı içi güvenilirlik katsayıları, puanlayıcılar arası güvenilirlik katsayılarına göre daha yüksektir. Bununla birlikte puanlayıcı içi güvenilirlik katsayıları toplam puanlar ve kategoriler bazında puanlayıcıların çoğunun yaptığı

puanlamalar için yüksek güvenilirliğe işaret etmektedir. Puanlayıcılar arası güvenilirlik değerlerinin ise tüm kategoriler ve toplam puanlar bazında birbirine daha yakın olduğu ve puanlayıcılar arasında güçlü bir uyum olduğu görülmüştür. Toplam puanlara ilişkin güvenilirlik katsayılarının her iki durumda da kategorilere ilişkin güvenilirlik katsayılarından daha yüksek olduğu, kategoriler bazında bakıldığında en düşük güvenilirlik katsayılarının sembollerin, işlemlerin ve formal dilin kullanımı kategorisine, en yüksek güvenilirlik katsayılarının ise temsil kategorisine ait olduğu görülmüştür. Ayrıca beşinci puanlayıcının puanlamalar arasındaki tutarlılığının diğer puanlayıcılara göre her bir kategori ve toplam puanlar için daha düşük olduğu ortaya çıkmıştır.

2.1.Araştırma Sorusuna İlişkin Sonuç ve Tartışma

“Yeterlilik Şeması ile elde edilen toplam puanlar ve her bir kategoriye ait puanlar için puanlayıcı içi tutarlılık derecesi nedir?”

KTİK çerçevesinde puanlayıcı içi güvenilirliğin araştırıldığı bu soruda puanlayıcı içi güvenilirliği belirlemeye yönelik olarak Spearman rho korelasyon katsayısı kullanılmıştır. Analiz sonuçları toplam puanlar için kestirilen güvenilirlik katsayılarının kategoriler için kestirilenlerden daha yüksek olduğunu göstermiştir. Bu sonuç Yelboğa (2007) ile Yelboğa ve Tavşancıl (2010)’un sonuçları ile benzerlik göstermektedir. Bu durum matematik okuryazarlığı yeterlilik gereksinimleri için puanlayıcıların farklı zamanlarda verdikleri puanlardaki tutarlılığın alt yeterlilik gereksinimlerine verdikleri puanlar arasındaki tutarlılıktan daha yüksek olduğunu ortaya çıkarmıştır. Bununla birlikte tüm alt yeterlilikler bazında puanlamalar arasında genellikle kararlılık anlamında tutarlılık olduğu görülmüştür. Bu sonuç Turner vd. (2013)’ün sonuçları ile benzerlik göstermektedir. Burada sadece bir puanlayıcının (P1_SM) sembollerin, işlemlerin ve formal dilin kullanımı yeterliliği için iki farklı zamanda puanlamalarındaki tutarlılığın zayıf olduğu görülmüştür. Bununla birlikte toplam puanlar ve alt kategoriler bazında puanlayıcılardan P5’in puanlamalarındaki tutarlılık diğer puanlayıcılara göre daha düşüktür.

Her ne kadar puanlamalar arasında tutarlılık olduğu görülse de puanlayıcıların farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı farklılık olan ve olmayan durumların olduğu görülmektedir. Bu sonuç korelasyon katsayılarının hesaplandığı çalışmalarda ortalamalar arasındaki farkların test edildiği çalışmaların sonuçları ile benzerlik göstermektedir (Goodwin, 2001; Güler, 2008; Güler ve Teker, 2015).

Kategoriler bazında bakıldığında iletişim kategorisinde tüm puanlayıcıların farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı bir farklılık varken strateji planlama kategorisi için yalnızca P5'in matematikleştirme kategorisinde yalnızca P3'ün farklı zamanlarda verdiği puanların ortalamaları arasında anlamlı farklılık olduğu görülmektedir. Temsil kategorisinde yalnızca P1 ve P5'in farklı zamanlarda verdikleri puanların ortalamaları arasında farklılık arasında anlamlı farklılık görülürken sembollerin, işlemlerin ve formal dilin kullanımı kategorisinde yalnızca P1 ve P5'in farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı bir farklılık olmadığı görülmüştür. Muhakeme kategorisi için tüm puanlayıcıların, toplam puanlar bazında ise yalnızca P4 ve P5'in farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı bir farklılık olmadığı sonucuna varılmıştır.

Genel olarak bakıldığında, sembollerin, işlemlerin ve formal dilin kullanımı kategorisi ile toplam puanlar için puanlayıcıların çoğunun farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı bir farklılık olduğu, diğer kategoriler içinse puanlayıcıların çoğunun farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı bir farklılık olmadığı söylenebilir. Turner vd., (2013) puanlayıcı içi güvenilirlik bağlamında hesaplanan korelasyonların 0,37 ile 0,91 arasında değiştiği ve 0,05 düzeyinde anlamlı olduğunu ve Cronbach α değerlerin 0,62 ile 0,95 arasında değiştiğini belirtmiştir. Yaptıkları çalışmada güvenilirliğin en düşük olduğu yeterliliğin muhakeme ve argüman en yüksek olduğu yeterliliğin ise iletişim olduğu görülmüştür. Bu sonuçlara dayanarak muhakeme ve argüman yeterliliği dışındaki yeterlilikler için puanlayıcılar arası güvenilirliğin yüksek bulunduğu belirtilmiştir. Bu iki çalışmanın sonuçları arasında ortaya çıkan bu farkın şemanın farklı versiyonlarının kullanımından ya da farklı puanlayıcı gruplarının, farklı matematiksel yeterlilikler bazında görüş birliğine sahip olmamasından kaynaklanabileceği düşünülmektedir.

2.2.Araştırma Sorusuna İlişkin Sonuç ve Tartışma

“Yeterlilik Şeması ile elde edilen toplam puanlar ve her bir kategoriye ait puanlar için puanlayıcılar arası tutarlılık derecesi nedir?”

Bu araştırma sorusu için KTK çerçevesinde puanlayıcılar arası tutarlılık hem birinci hem ikinci hem de iki uygulamada verilen puanların ortalaması üzerinden incelenmiştir. Birinci uygulama için puanlayıcılar arası tutarlılığın ikinci uygulama için ve iki uygulamada verilen

puanların ortalaması için hesaplananlara göre daha düşük olduğu görülmüştür. Bu sonuç Yeterlilik Şeması puanlama anahtarını kullanma konusundaki deneyimin puanlama güvenilirliği arttırabileceği görüşünü desteklemektedir (Turner ve Adams, 2012; Turner vd.,2013; Turner vd., 2015). Buna ek olarak puanlayıcılar arası güvenirliliğin hesaplandığı çalışmaların sonuçları ile de uyumludur (Deliceoğlu, 2009; Yelboğa, 2007; Yelboğa ve Tavşancıl, 2010). Birinci uygulama için puanlayıcılar arası tutarlılık strateji planlama, temsil kategorileri ve toplam puanlar için yüksek iken, diğer kategoriler için orta düzeydedir. İkinci uygulama içinse tüm kategoriler ve toplam puanlar bazında puanlayıcılar arası tutarlılığın yüksek olduğu bulunmuştur. İki uygulamada verilen puanların ortalaması için yine ikinci puanlamada olduğu gibi tüm kategoriler ve toplam puanlar bazında puanlayıcılar arası tutarlılığın yüksek olduğu ortaya çıkmıştır. Bu sonuçlar Turner ve Adams (2012) ile Turner vd. (2013)'nin sonuçları ile benzerlik göstermektedir.

Üçüncü Araştırma Sorusuna İlişkin Sonuç ve Tartışma

Araştırmanın üçüncü sorusuna yönelik olarak PISA matematik okuryazarlığı maddelerinin gereksinim düzeylerine ilişkin Yeterlilik Şeması puanlama anahtarı ile beş puanlayıcıdan iki farklı zamanda elde edilen ölçümler için Genellenebilirlik Kuramı çerçevesinde iki farklı modele başvurulmuştur. Bu bağlamda $m^* \times p^* \times z^*$ desenli çok değişkenli model ve $m \times p \times z$ desenli tek değişkenli model G ve K çalışmaları ile kestirilen varyansların, toplam varyansı açıklama yüzdelерinin, G ve Phi katsayılarının değişkenlik gösterdiği ortaya çıkmıştır.

3.1.Araştırma Sorusuna İlişkin Sonuç ve Tartışma

“Çok değişkenli ve tek değişkenli G çalışmaları sonucunda her biri kategoriye ve toplam puanlara ilişkin kestirilen varyanslar ve toplam varyansı açıklama yüzdeleri nelerdir?”

Çok değişkenli ve tek değişkenli G çalışmaları ile alt kategoriler ve toplam puanlar bazında kestirilen varyansların ve toplam varyansı açıklama oranlarının birbirine paralel sonuçlar verdiği bulunmuştur. Tüm alt kategoriler ve toplam puanlar bazında toplam varyansı açıklama oranı en yüksek olan etkinin bu çalışmada ölçmenin objesi olan madde ana etkisi olduğu ve toplam varyansın en az yarısından fazlasını açıkladığı sonucuna ulaşılmıştır. Bu durum maddelerin yeterlilik gereksinimleri açısından birbirlerinden farklılaştıklarına işaret etmektedir. Bu sonuç deneysel güçlük düzeylerinin açıklandığı OECD (2009a) raporunda

verilen bilgilerle örtüşmektedir. Toplam varyans içinde en yüksek ikinci paya sahip olan etki ise yine tüm kategoriler ve toplam puanlar bazında madde x puanlayıcı ortak etkisidir. Bu durum bazı maddeleri bazı puanlayıcıların farklı puanladıklarına işaret etmektedir. Diğer bir deyişle bir puanlayıcıdan diğerine bir maddenin bağıl durumunun farklılaştığını ortaya çıkarmıştır. Puanlayıcı ve zaman ana etkilerinin toplam varyansı açıklama oranları önemsenmeyecek kadar küçük olmakla birlikte iletişim ile muhakeme ve argüman kategorileri dışındaki kategoriler ve toplam puanlar için puanlayıcı ana etkisinin zaman ana etkisine göre toplam varyansı açıklama oranının daha yüksek, muhakeme ve argüman kategorisi için daha düşük olduğu sonucuna ulaşılmıştır.

3.2.Araştırma Sorusuna İlişkin Sonuç ve Tartışma

“Orijinal puanlayıcı ve zaman sayıları ile puanlayıcı ve zaman sayılarının bir arttırılıp bir azaltılması senaryolarına göre çok değişkenli Karar (K) çalışmalarıyla kestirilen G ve Phi katsayıları alt kategorilere göre nasıl değişmektedir?”

Bu alt problem için çok değişkenli K çalışmaları ile orijinal puanlayıcı ve zaman sayılarının bir arttırılıp bir azaltıldığı senaryolara göre kestirilen G ve Phi katsayılarının kategoriler bazında farklılık gösterdiği sonucuna ulaşılmıştır. Orijinal puanlayıcı ve zaman sayılarına göre kestirilen G ve Phi katsayılarının temsil ve strateji planlama alt kategorileri için 0,90’ın üstünde olduğu ve yüksek güvenilirliğe işaret ettiği, diğer kategoriler için ise 0,80 ile 0,90 arasında olup kabul edilebilir güvenilirlik düzeyinin üstünde olduğu ortaya çıkmıştır. Bu sonuç Turner ve Adams (2012) ile Turner vd. (2013)’nin sonuçları ile benzerlik göstermektedir. Bu durum strateji planlama ve temsil kategorisine ilişkin puanlamaların diğer kategorilere ilişkin puanlamalardan daha güvenilir olduğunu göstermektedir. Zaman sayısı sabit tutularak puanlayıcı sayılarının bir azaltıldığı senaryolara göre kestirilen G ve Phi katsayılarındaki değişim, puanlayıcı sayısı sabit tutularak zaman sayısının bir azaltıldığı senaryolar için kestirilen G ve Phi katsayılarındaki değişimden daha azdır. Diğer yandan zaman sayısı sabit tutularak puanlayıcı sayılarının bir arttırıldığı senaryolara göre kestirilen G ve Phi katsayılarındaki değişim, puanlayıcı sayısı sabit tutularak zaman sayısının bir arttırıldığı senaryolar için kestirilen G ve Phi katsayılarındaki değişimden sembollerin, işlemlerin ve formal dilin kullanımı kategorisi dışındaki tüm kategorilerde daha fazladır. Puanlayıcı ve zaman sayılarının her ikisinin birer azaltılması senaryosu için kestirilen G ve Phi katsayılarının sembollerin, işlemlerin ve formal dilin kullanımı kategorisi dışındaki tüm

kategorilerde kabul edilebilir düzeyde kaldığı bulunmuştur. Puanlayıcı ve zaman sayılarının her ikisinin birer arttırılması senaryosu için kestirilen G ve Phi katsayılarının tüm kategorilerde 0,90'ın üzerine çıktığı ve yüksek güvenilirliğe işaret ettiği sonucuna ulaşılmıştır. Bu sonuç literatürdeki diğer çalışmaların sonuçları (Atılğan, 2004; Bağcı, 2015; Deliceoğlu, 2009; Güler, 2008; Yelboğa, 2007; Yelboğa ve Tavşancıl, 2010) ile tutarlıdır. Bununla birlikte tüm kategorilerde G katsayıların Phi katsayılarından daha yüksek olduğu, G ve Phi katsayılarının puanlayıcı ve zaman sayılarındaki artış ve düşüşe paralel olarak artma ve azalma eğiliminde olduğu görülmüştür.

3.3.Araştırma Sorusuna İlişkin Sonuç ve Tartışma

“Orijinal puanlayıcı ve zaman sayıları ile puanlayıcı ve zaman sayılarının bir arttırılıp bir azaltılması senaryolarına göre çok değişkenli ve tek değişkenli Karar (K) çalışmalarıyla toplam puanlara ilişkin kestirilen G ve Phi katsayıları modellere göre nasıl değişmektedir?”

Çok değişkenli ve tek değişkenli K çalışmaları ile orijinal puanlayıcı ve zaman sayılarının bir arttırılıp bir azaltıldığı senaryolara göre kestirilen G ve Phi katsayılarının toplam puanlar bazında farklılık gösterdiği sonucuna ulaşılmıştır. Tüm senaryolar için çok değişkenli modelle hesaplanan G ve Phi katsayılarının tek değişkenli modelle hesaplanan G ve Phi katsayılarından daha yüksek çıkma eğiliminde olduğu sonucuna ulaşılmıştır. Orijinal puanlayıcı ve zaman sayılarına göre toplam puanlar için her iki modelle kestirilen G ve Phi katsayılarının 0,90'ın üstünde olduğu ve yüksek güvenilirliğe işaret ettiği bulunmuştur. Ayrıca maddelerin matematik okuryazarlığı gereksinimlerine ilişkin puanlamaların alt kategorilere ilişkin puanlamalardan daha güvenilir olduğu söylenebilir. Her iki modelle kestirilen G ve Phi katsayılarını zaman sayısının bir azaltılmasının bir arttırılmasına göre daha çok etkilediği görülmüştür. Puanlayıcı ve zaman sayılarının her ikisinin birer azaltılması senaryosu için kestirilen G ve Phi katsayılarının dahi her iki modelde de kabul edilebilir güvenilirlik düzeyinin üstünde olduğu bulunmuştur. Puanlayıcı ve zaman sayılarının her ikisinin birer arttırılması senaryosu için kestirilen G ve Phi katsayılarının her iki modelde de 0,95'in üzerine çıktığı ve yüksek güvenilirliğe işaret ettiği sonucuna ulaşılmıştır. Bununla birlikte tüm kategorilerde G katsayıların Phi katsayılarından daha yüksek olduğu, G ve Phi katsayılarının puanlayıcı ve zaman sayılarındaki artış ve düşüşe paralel olarak artma ve azalma eğiliminde olduğu görülmüştür. Bu sonuç literatürdeki diğer çalışmaların sonuçları

(Atılğan, 2004; Bağcı, 2015; Deliceoğlu, 2009; Güler, 2008; Yelboğa, 2007; Yelboğa ve Tavşancıl, 2010) ile tutarlıdır.

Dördüncü Araştırma Sorusuna İlişkin Sonuç ve Tartışma

“Yeterlilik Şeması ile elde edilen her bir kategoriye ait puanlar için çok değişkenli model K çalışmalarıyla kabul edilebilir düzeyde bir G katsayısı elde etmek için gerekli minimum puanlayıcı ve zaman sayısı nedir?”

Bu alt problem için puanlayıcı sayısının bir ile altı, zaman sayısının bir ile üç arasında değiştiği senaryolara göre çok değişkenli K çalışmaları ile alt kategoriler için kestirilen G ve Phi katsayıları iç ölçütler bağlamında değerlendirilerek minimum puanlayıcı ve zaman sayıları belirlenmiştir. Bu bağlamda iletişim alt kategorisine yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının üç ve zaman sayısının bir olmasının yeterli olduğu görülmektedir. Strateji planlama alt kategorisine yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının iki ve zaman sayısının bir olmasının yeterli olduğu görülmektedir. Matematikleştirme alt kategorisine yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının iki ve zaman sayısının üç ya da minimum puanlayıcı sayısının üç ve zaman sayısının bir olmasının yeterli olduğu görülmektedir. Temsil alt kategorisine yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının iki ve zaman sayısının iki ya da minimum puanlayıcı sayısının üç ve zaman sayısının bir olmasının yeterli olduğu görülmektedir. Sembollerin, işlemlerin ve formal dilin kullanımı alt kategorisine yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının üç ve zaman sayısının iki olmasının yeterli olduğu görülmektedir. Muhakeme ve argüman alt kategorisine yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının üç ve zaman sayısının iki ya da minimum puanlayıcı sayısının dört ve zaman sayısının bir olmasının yeterli olduğu görülmektedir. Genel olarak tüm alt kategoriler için tek bir puanlayıcının farklı zamanlarda yaptığı puanlamaların kabul edilebilir güvenilirlik düzeyinin altında sonuçlar verdiği, tüm kategoriler için minimum puanlayıcı sayısı üç ve zaman sayısı iki olduğunda kabul edilebilir düzeyde güvenilir ölçümler yapıldığı sonucuna ulaşılmıştır.

Beşinci Araştırma Sorusuna İlişkin Sonuç ve Tartışma

“Yeterlilik Şeması ile elde edilen toplam puanlar için çok değişkenli ve tek değişkenli model K çalışmalarıyla kabul edilebilir düzeyde bir G katsayısı elde etmek için gerekli minimum puanlayıcı ve zaman sayısı modellere göre nasıl değişmektedir?”

Bu alt problem için puanlayıcı sayısının bir ile altı, zaman sayısının bir ile üç arasında değiştiği senaryolara göre çok değişkenli ve tek değişkenli K çalışmaları ile toplam puanlar için kestirilen G ve Phi katsayıları iç ölçütler bağlamında değerlendirilerek minimum puanlayıcı ve zaman sayıları belirlenmiştir. Bu bağlamda tek değişkenli K çalışması sonuçlarına göre, maddelerin matematik okuryazarlığı gereksinim düzeylerine yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için minimum puanlayıcı sayısının iki ve zaman sayısının bir olmasının yeterli olduğu görülmektedir. Çok değişkenli K çalışması sonuçlarına göre toplam puanlara yönelik olarak kabul edilebilir düzeyde güvenilirlik katsayısı elde etmek için gerekli minimum puanlayıcı ve zaman sayılarının tek değişkenli K çalışması sonuçlarına göre gerekli olanlardan daha az olduğu ortaya çıkmıştır. Bu sonuç çok değişkenli G çalışmalarının yapıldığı çalışmaların sonuçları ile benzerlik göstermektedir (Atılgan, 2004; Deliceoğlu, 2009; Yelboğa, 2007; Yelboğa ve Tavşancıl, 2010) Çok değişkenli K çalışmasına göre bir puanlayıcının bir zamanda puanlama yapmasıyla dahi kabul edilebilir güvenilirlik düzeyine ulaşılabilir.

Altıncı Araştırma Sorusuna İlişkin Sonuç ve Tartışma

“Yeterlilik Şeması ile elde edilen toplam puanlar ve her bir kategoriye ait puanlar için KTK ve G Kuramı çerçevesinde hesaplanan güvenilirlik katsayıları iç tutarlılık ölçütlerine göre tutarlı mıdır?”

Yeterlilik Şeması ile elde edilen puanlamalar için KTK ve G Kuramı bağlamında hesaplanan güvenilirlik katsayılarının değişkenlik gösterdiği görülmektedir. Bu güvenilirlik katsayıları iç ölçütlere göre incelendiğinde ölçümlerin puanlayıcı içi güvenirliliğinin bir durum (sembollerin, işlemlerin ve formal dilin kullanımı kategorisi bağlamında birinci puanlayıcı için hesaplanan puanlayıcı içi güvenirlilik katsayısı) hariç genellikle yüksek olduğu sonucuna varılmıştır. Ayrıca puanlayıcıların farklı zamanlardaki puanlamaları arasında çoğunlukla anlamlı bir farklılık olmadığı görülmektedir. Yalnızca iletişim kategorisinde puanlayıcıların tamamının iki farklı zamanlarda verdikleri puanların ortalamaları arasında anlamlı farklılık bulunmuştur. Sonuç olarak puanlayıcılar arası güvenirliliğin, genellenebilirlik katsayısının ve

güvenirlilik katsayısının birbirine paralel güvenirlilik düzeylerine işaret ettiği ortaya çıkmıştır. Ancak puanlayıcı içi güvenirliliğin puanlayıcılar ve kategoriler bazında kendi içinde farklı güvenirlilik düzeylerine karşılık geldiği bulunmuştur. Bu durumun nedeninin G Kuramı'nın tüm hata kaynaklarını birlikte ele alarak değerlendirilmesinden kaynaklandığı düşünülmüştür. G kuramı puanlayıcı içi ve puanlayıcılar arası uyumu birlikte değerlendirerek tek bir güvenirlilik katsayısı sunmaktadır. Yeterlilik Şeması'na ilişkin KTK ve G kuramı analizlerine göre elde edilen bu bulgular, Goodwin (2001), Yelboğa (2007) ve Delicoğlu (2009)'un çalışmalarındaki KTK ve G Kuramı güvenirlilik katsayılarının iç ölçütlere göre beklenen değerlerinden yüksek sonuçlar ürettiği şeklindeki bulgular ile tutarlılık göstermektedir.

Özet olarak, incelenen araştırma problemlerine göre bu ölçme için Yeterlilik Şeması puanlama anahtarı ile elde edilen ölçümlerin ölçt ve yapı geçerliğini sağladığı ve güvenirlilik düzeylerinin kabul edilebilir düzeyin üstünde olduğu bulunmuştur. KTK ve G kuramı çerçevesinde güvenirliliğe ilişkin elde edilen sonuçların birbirine paralellik gösterdiği sonucuna ulaşılmıştır. G Kuramı çerçevesinde puanlayıcı ve zaman sayılarının bir artırılıp azaltıldığı senaryolara göre tek değişkenli modelle kestirilen güvenirlilik parametrelerinin çok değişkenli modelle kestirilenlere göre daha düşük çıkma eğiliminde olduğu görülmüştür. Toplam puanlar ve alt kategoriler bazında G Kuramı çerçevesinde kabul edilebilir düzeyde güvenirlilik katsayılarının elde edilebilmesi için en az üç puanlayıcının iki farklı zamanda puanlama yapmasının yeterli olduğu sonucuna ulaşılmıştır. KTK çerçevesinde puanlayıcı içi güvenirliliğine ilişkin elde edilen sonuçların puanlayıcıdan puanlayıcıya ve kategoriden kategoriye farklılık gösterdiği ortaya çıkmıştır. Ayrıca puanlayıcı içi güvenirlilik bağlamında yüksek güvenirliliğe ve farklı zamanlardaki puanlamalar arasında anlamlı bir ilişki olduğuna işaret eden korelasyon katsayıları elde edilmiştir. Yüksek güvenirliliğe işaret eden korelasyon katsayılarına karşın ilişkili örneklem t testi ile bazı puanlayıcıların bazı kategoriler için farklı zamanlardaki puanlamalarının ortalamaları arasında anlamlı farklılıkların olduğu ortaya çıkmıştır.

Öneriler

Bu araştırmada incelenen alt problemlerin sonuçlarına göre KTK ve G Kuramı çerçevesinde kestirilen güvenirlilik parametreleri benzerlik göstermektedir. Öneriler araştırma sonuçlarına

ve ileride yapılacak arařtırmalara yönelik olmak üzere iki ayrı bařlık altında ařağıda sunulmuřtur.

Arařtırma Sonularına Yönelik Öneriler

1. PISA matematik okuryazarlığı maddelerinin yeterlilik düzeylerinin belirlenmesinde geçerli ve güvenilir ölçmeler elde etmek için Yeterlilik Şeması puanlama anahtarı kullanılabilir. Yeterlilik Şeması ile puanlama yapılması durumunda matematik okuryazarlığının birden fazla alt yeterliliğı ile ilgileniliyorsa en az üç puanlayıcının iki farklı zamanda puanlama yapması tercih edilmelidir.
2. Puanlayıcı içi güvenilirliğın incelenmesinde yalnızca korelasyon katsayılarının ve anlamlılık düzeylerinin incelenmesi yeterli olmamaktadır. Farklı zamanlardaki puanlamaların ortalamaları arasındaki farklılıkların test edilmesi önerilmektedir.
3. G Kuramı çerçevesinde yapılan arařtırmalarda genellikle ölçme objesi olarak bireylerin seçildiğı ve bireylerin davranıřlarındaki ve özelliklerindeki gözlemlerin incelendiğı göze çarpmaktadır. Bu çalışma da ölçme objesi olarak maddeler ele alınmıřtır. G kuramı çerçevesinde birey dıřındaki değıřkenlik kaynaklarının da ölçme objesi olarak ele alınabileceğı önerilmektedir.
4. G Kuramı çerçevesinde yapılan analizler KTK çerçevesinde yapılanlara göre daha fazla bilgi vermekle birlikte KTK ile yapılan analizlerde G Kuramı ile ortaya konulamayan bilgilere (hangi puanlayıcıların puanlamaları arasında tutarsızlık olduğı gibi) ulařılabildiğı görölmektedir. Bu bağlamda özellikle puanlayıcı içi güvenilirliğın incelenmesi söz konusu olduğunda ve kullanılan ölçme aracı tek boyutlu olmadığında her iki kuramın birlikte kullanılması önerilmektedir.

İleride Yapılacak Arařtırmalara Yönelik Öneriler

1. Bu çalışma bağlamında puanlayıcılar evreninden seçilen beř puanlayıcı matematik eğitimi alanında en az doktora yeterliliğini almıř uzmanlardan seçilmiřtir. Yeterlilik Şeması puanlama anahtarı kullanımının belli bir düzeyde uzmanlık gerektirip gerektirmediğini test etmek için aynı çalışma farklı puanlayıcılar evreninden seçilen puanlayıcılar ile tekrarlanabilir.
2. Bu çalışmada yayınlanmış PISA maddelerinin matematik okuryazarlığı gereksinimlerinin ölçülmesi ile ilgilenilmiřtir. Yeterlilik Şeması puanlama anahtarı

kullanılarak matematik okuryazarlığının ölçülmesinin hedeflendiği maddeler için benzer bir çalışma yapılabilir.

3. Yeterlilik Şeması puanlama anahtarı ile ölçülen matematik maddelerinin yeterlilik gereksinimlerinin maddelerin deneysel güçlük puanlarını yordadığı görülmüştür. Yeterlilik Şeması kuramsal dayanağı olan bir çerçeve olarak ele alınarak matematik okuryazarlığı ve içerdiği alt yeterlilikler bağlamında farklı güçlüklerde madde yazımında kullanılabilir.
4. Yeterlilik Şeması puanlama anahtarının güvenirliği Madde Tepki Kuramı çerçevesinde incelenebilir.



KAYNAKLAR

- Anıl, D., & Büyükkıdık, S. (2012). Genellenebilirlik kuramında dört facetli karışık desen kullanımı için örnek bir uygulama. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 3(2) 291-296.
- Atılğan, H. (2004). *Genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch modelinin karşılaştırılmasına ilişkin bir araştırma*. Doktora Tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.
- Atılğan H.(2005). Genellenebilirlik kuramı ve puanlayıcılar arası güvenirlik için örnek bir uygulama. *Eğitim Bilimleri ve Uygulama*, 4(7), 95-108.
- Bağcı, V. (2015). *Matematiksel muhakeme becerisinin değerlendirilmesinde klasik test kuramı ile genellenebilirlik kuramındaki farklı desenlerin karşılaştırılması*. Yüksek Lisans Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.
- Berberoğlu, G. (2007). Türk bakış açısından PISA araştırma sonuçları. *Konrad Adenauer Stiftung*. <http://konrad.org.tr/Egitimturk/07girayberberoglu.pdf> sayfasından erişilmiştir.
- Brennan, R. L. (2001). *Generalizability theory*. USA: Springer-Verlag New York Inc.
- Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: a practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, 52,13–35.
- Büyükkıdık, S., & Anıl, D. (2015). Performansa dayalı durum belirlemede güvenirliğin genellenebilirlik kuramında farklı desenlerle incelenmesi. *Eğitim ve Bilim*, 40(177).
- Büyüköztürk, Ş. (2002). *Sosyal bilimler için veri analizi el kitabı* (16. Baskı.) Pegem akademi.

- Cardinet, J., Tourneur, W., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, 13(2), 119-135.
- Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. London: Routledge.
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage.
- Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2014). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları* (3.Baskı). Pegem Akademi.
- Daniel, R. C., & Embretson, S. E. (2010). Designing cognitive complexity in mathematical problem-solving items. *Applied Psychological Measurement*, 34(5), 348-364.
- Deliceoğlu, G. (2009). *Futbol yetilerine ilişkin dereceleme ölçeğinin genellenebilirlik ve klasik test kuramına dayalı güvenirliklerinin karşılaştırılması*. Doktora Tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.
- Deliceoğlu G. & Demirtaşlı, N. Ç. (2012). Futbol yetilerine ilişkin dereceleme ölçeğinin güvenirliğinin genellenebilirlik kuramına ve klasik test kuramına dayalı olarak karşılaştırılması. *Spor Bilimleri Dergisi*, 23(1), 1-12.
- Eser, D. Ç. & Gelbal, S. (2013). Genellenebilirlik kuramı ve lojistik regresyona dayalı hesaplanan puanlayıcılar arası tutarlığın karşılaştırılması. *Kastamonu Eğitim Dergisi*, 21(2), 421-438.
- Field, A. (2009). *Discovering statistics using SPSS* (3. baskı). London: Sage.
- Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercises Science*, 5(1), 13-14.
- Güler, N. (2008). *Klasik test kuramı genellenebilirlik kuramı ve rasch modeli üzerine bir araştırma*. Doktora Tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.
- Güler, N. & Gelbal, S. (2010). Açık uçlu matematik sorularının güvenirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri*, 10(2), 989-1019.
- Güler, N. (2011). Rasgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramına göre güvenirliğin karşılaştırılması. *Eğitim ve Bilim Dergisi*, 36(162), 225-234.

- Güler, N., Kaya Uyanık, G., & Taşdelen Teker, G. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi.
- Güler, N., & Teker, G. T. (2015). Açık uçlu maddelerde farklı yaklaşımlarla elde edilen puanlayıcılar arası güvenirliğin değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 6(1), 12-24.
- Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, 72(4), 665-686.
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, 6(1), 53-60.
- Howell, D. C. (2013). *Statistical methods for psychology*. USA. Wadsworth, Cengage Learning.
- Höfer, T., & Beckmann, A. (2009). Supporting mathematical literacy: Examples from a cross-curricular project. *ZDM*, 41(1-2), 223-230.
- Jablonka, E. (2003). Mathematical literacy. A. J. Bishop, M.A. Clements, C. Keitel, J. Kilpatrick & F. K. S. Leung (Ed.), *Second international handbook of mathematics education* içinde (s. 75-102). Dordrecht: Kluwer.
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144.
- Kaiser, G., & Willander, T. (2005). Development of mathematical literacy: Results of an empirical study. *Teaching Mathematics and its Applications*, 24(2-3), 48-60.
- Milli Eğitim Bakanlığı. (2005). *PISA 2003 Uluslararası öğrenci değerlendirme projesi, Ulusal nihai rapor*. Ankara Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü.
- Milli Eğitim Bakanlığı. (2012). *PISA Türkiye*. Ankara Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü. http://www.selahattinakbileklisesi.k12.tr/PISA_kitabi.pdf adresinden 21. 04. 2015 tarihinde erişilmiştir.
- Milli Eğitim Bakanlığı. (2013). *PISA 2012 ulusal ön raporu*. Ankara Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü.

- Nalbantoğlu, F. (2009). *Performans ölçümlerinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması*. Yüksek Lisans Tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.
- Niss, M., & Højgaard, T. (Eds.). (2011). Competencies and mathematical learning. Ideas and inspiration for the development of mathematics teaching and learning in Denmark (Tekster fra IMFUFA, no 485). Roskilde: Roskilde University, IMFUFA.
- Niss, M., & Jablonka, E. (2014). Mathematical literacy. S. Lerman (Ed.), *Encyclopedia of mathematics education* içinde (s. 391-396). Netherlands: Springer.
- Organisation for Economic Co-operation and Development. (2004a). *Learning for Tomorrow's World: First Results from PISA 2003*. Paris: OECD.
- Organisation for Economic Co-operation and Development. (2004b). The PISA 2003 assessment framework: mathematics, reading, science and problem solving knowledge and skills. Paris: OECD.
- Organisation for Economic Co-operation and Development. (2007). *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis*. Paris: OECD.
- Organisation for Economic Co-operation and Development (OECD). (2009a). *PISA: Take the test*. Paris: OECD. <http://www.oecd.org/pisa/pisaproducts/Take%20the%20test%20e%20book.pdf>. sayfasından erişilmiştir.
- Organisation for Economic Co-operation and Development. (2010a). PISA 2009 Results: What students know and can do – Student performance in reading, mathematics and science (Volume I). Paris: OECD.
- Organisation for Economic Co-operation and Development. (2010b). *Learning mathematics for life: A view perspective from PISA*. Paris: OECD.
- Organisation for Economic Co-operation and Development. (2013a). PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy. Paris: OECD.
- Organisation for Economic Co-operation and Development. (2013b). *PISA 2012 released mathematics items*. <http://www.oecd.org/pisa/pisaproducts/pisa2012-2006-rel-items-mathsENG.pdf>. sayfasından erişilmiştir.

- Organisation for Economic Co-operation and Development (2014). PISA 2012 Results: What Students Know and Can Do (Cilt I, Düzenlenmiş Baskı, Şubat 2014): Student Performance in Mathematics, Reading and Science. Paris: OECD.
- Özgen, K., & Bindak, R. (2008). Matematik okuryazarlığı öz-yeterlik ölçeğinin geliştirilmesi. *Kastamonu Eğitim Dergisi*, 16(2), 517-528.
- Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation*, 13(4), 1-11.
- Popham, W. J. (1997). What's wrong – and what's right – with rubrics. *Educational Leadership*, 55(2), 72–75.
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18-39.
- Schleicher, A. (2010). *PISA 2009 Evaluating Systems to Improve Education* (PowerPoint slaytı). <http://www.slideshare.net/OECDDEDU/pisa-2012-evaluating-school-systems-to-improve-education> sayfasından erişilmiştir.
- Shavelson, R. J., & Webb, M. N. (1991). *Generalizability theory: A Prime*. SAGE Publication, Inc., California.
- Smith, V. E. & Kulikowich, J. M. (2004). An application of generalizability theory and many facet rasch measurement using a complex problem solving skills assesment. *Educational and Psychological Measurement*, 64(4), 617-639.
- Tan, Ş. (2016). *SPSS ve Excel uygulamalı temel istatistik-1*. Ankara: Pegem Akademi.
- Tan, Ş. (2013). *Öğretimde ölçme ve değerlendirme KPSS el kitabı* (8. Baskı). Ankara: Pegem Akademi.
- Taylor, M. A. & Pastor, D. A. (2013). An application of generalizability theory to evaluate the technical quality of an alternate assessment. *Applied Measurement in Education*, 26(4), 279-297.
- Turgut M. F., & Baykul, Y. (2013). *Eğitimde ölçme ve değerlendirme* (5. Baskı). Ankara: Pegem Akademi.
- Turner, R. (2012). *Some drivers of test item difficulty in mathematics*. Annual Meeting of the American Educational Research Association (AERA)' da sunulmuş bildiri, Vancouver. http://works.bepress.com/ross_turner/17. sayfasından erişilmiştir.

- Turner, R., & Adams, R. J. (2012). *Some drivers of test item difficulty in mathematics: An analysis of the competency rubric*. Annual Meeting of the American Educational Research Association (AERA)' da sunulmuş bildiri, Vancouver. <http://research.acer.edu.au/pisa/7/>. sayfasından erişilmiştir.
- Turner, R., Dossey, J., Blum, W., & Niss, M. (2013). Using mathematical competencies to predict item difficulty in PISA: a MEG study. Prenzel, M. (Ed.), *Research on PISA* içinde (s. 23-37). Springer Netherlands.
- Turner, R., Blum, W., & Niss, M. (2015). Using competencies to explain mathematical item demand: A work in progress. K. Stacey, & R. Turner (Ed.), *Assessing mathematical literacy: The PISA experience* içinde (s. 85-115). New York, NY: Springer.
- Ünlü, A., Kasper, D., Trendtel, M., & Schurig, M. (2014). The OECD's programme for international student assessment (PISA) study: A review of its basic psychometric concepts. M. Spiliopoulou, L. Schmidt-Thieme, & R. Janning (Ed.), *Data analysis, machine learning and knowledge discovery* içinde (s. 417-425). Springer International Publishing.
- Webb, N. M. & Shavelson, R. J. (2005). Generalizability theory: Overview. B. S.Everitt & D. C.Howell (Ed.), *Encyclopedia of statistics in behavioral science* içinde (c. 2, ss. 717-719). Chichester, UK: John Wiley & Sons.
- Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and Generalizability theory. *Handbook of statistics* içinde (s.81-124).
- Wilhelm, A. G., & Kim, S. (2015). Generalizing from observations of mathematics teachers' instructional practice using the instructional quality assessment. *Journal for Research in Mathematics Education*, 46(3), 270-279.
- Yelboğa, A. (2007). *Klasik test kuramı ve genellenebilirlik kuramına göre güvenilirliğin bir iş performansı ölçeği üzerinde incelenmesi*. Doktora Tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.
- Yelboğa, A., & Tavşancıl, E. (2010). Klasik test ve genellenebilirlik kuramına göre güvenilirliğin bir iş performansı ölçeği üzerinde incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri*, 10 (3), 1825-1854.

Yılmaz, F. N. & Gelbal, S. (2011). İletişim becerileri istasyonu örneğinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 41, 509-518.



EKLER



EK 1. Yeterlilik Şeması

Yeterlilik Tanımları ve Düzey Açıklamaları (2013)

İletişim

İletişim yeterliliği hem “algısal” hem “ yapısal” bileşenlere sahiptir. Algısal bileşen, kullanılan matematiksel dili, hangi bilgilerin ilişkili olduğunu ve gerekli yanıtın doğasının ne olduğunu kapsayan görevin matematiksel amaçlarıyla ilişkili olarak belirtilenlerin ve gösterilenlerin anlaşılmasını içerir. Yapısal bileşen çözüm aşamalarını, yapılan muhakemenin açıklanmasını ve verilen yanıtın doğrulanmasını içerebilen yanıtın sunumundan oluşmaktadır.

Yazılı ve bilgisayar tabanlı maddelerde algısal iletişim, metnin ve hareketli ve hareketsiz şekillerin anlaşılması ile ilgilidir. Metin sözel olarak verilen matematiksel ifadeler içerir ve aynı zamanda matematiksel temsillerin (örneğin; başlıklar, etiketler ve grafik ve diyagramlardaki açıklayıcı bilgiler) içinde de bulunabilir.

İletişim probleme nasıl yaklaşılacağını ya da problemin nasıl çözüleceğini, verilen belirli bilgilerden nasıl faydalanılacağını, elde edilen yanıtın nasıl muhakeme yapılacağını ya da savunulacağını bilmeyi içermez; daha çok ilgili bilgilerin anlaşılması ya da sunumu ile ilgilidir. Ayrıca iletişim temsillerden matematiksel bilgilerin çıkarılması ya da işlenmesi ile ilgili değildir.

Bilgisayar-tabanlı maddelerde, navigasyon ve bilgisayar ortamı ile ilgili diğer konular hakkındaki yönergeler genel görev gereksinimlerini arttırabilir ancak bu iletişim yeterliliğinin bir parçası değildir.

Bu yeterliliğin algısal boyutuna duyulan gereksinim, görevin anlaşılmasında yorumlanacak olan materyalin karmaşıklığına; çoklu bilgi kaynaklarını ilişkilendirme ya da bilgi parçaları arasında ileriye ve geriye gitme (döngü) gerekliliğine bağlı olarak artar. Yapısal boyut detaylandırılmış yazılı çözüm ya da açıklama sağlanması gerekliliği ile artar.

Tanım: Açıklamaları, soruları, yönergeleri, görevleri, şekilleri ve nesneleri okuma ve yorumlama; sunulan durumu hayalinde canlandırma ve anlama ve bahsedilen matematiksel terimleri içeren bilgileri anlamlandırma; bir kişinin matematiksel çalışmasını ya da muhakemesini sunma ve açıklama.

0: Tüm bilgilerin görevle doğrudan ilişkili olduğu ve bilgilerin sırası ile görevde istenilenleri anlamak için gerekli düşünme adımlarının eşleştiği, bağlama doğrudan erişim veren iyi bilinen kavramlarla ilişkili bir cümleyi ya da söz öbeğini anlama. Yapısal iletişim sadece tek bir kelimenin ya da sayısal sonucun sunumunu kapsar.

1: Sunulan materyal kısa cümlelerden ya da söz öbeklerinden daha karmaşık ve geniş olduğunda ya da bazı dışsal bilgiler sunulabildiğinde, metindeki ve ilişkili diğer temsillerdeki ilgili bilgi parçalarını tanıma ve aralarında bağlantı kurma. Gerekli olan yapısal iletişim basittir. Örneğin kısa bir açıklama/hesap yapmayı ya da bir aralığı/bir dizi değerleri açıklamayı kapsayabilir.

2: Görevi anlamak için sunulan materyalde tekrarlı döngüler gerekli olduğunda bağlantı kurulacak öğeleri tanıma ve seçme; ya da bağlamın/görevin ya da bunlar arasındaki bağlantıların birden çok ögesini anlama. Herhangi bir yapısal iletişim kısa ve öz bir betimleme/ açıklama sağlamayı ya da bir dizi hesaplama adımlarının sunumunu içerir.

3: Mantıksal olarak karmaşık ilişkiler (şartlı ya da iç içe geçmiş açıklamalar gibi) içeren birden çok bağlam ya da görev öğelerini ve bunlar arasındaki bağlantıları tanıma, seçme ve

anlama. Herhangi bir yapısal iletişim, problemin ya da çözümünün birden çok ögesi arasında bağlantının kurulduğu argümantasyonun sunumunu içerecektir.

Strateji Planlama

Bu yeterliliğin odağı matematiksel problem çözmenin stratejik boyutu üzerinedir: bir çözüm stratejisini seçme/ yapılandırma/ etkinleştirme ve ilişkili süreçlerin uygulanmasını izleme ve denetleme. “Strateji”, problemi çözmek için gerekli genel planı şekillendiren bir aşamalar dizisi anlamında kullanılmıştır. Her aşama bir alt amaç ve ilişkili adımlardan oluşmaktadır. Örneğin, veri toplamak, verileri dönüştürmek ve bunları farklı bir yolla sunmak için geliştirilen plan normal olarak üç ayrı aşamadan oluşacaktır. Çözüm sürecini fiilen gerçekleştirmek için gerekli bilgi, teknik prosedürler, matematikleştirme ve muhakeme ait oldukları diğer yeterlilikler altında ele alınır.

Bu yeterliliğe duyulan gereksinim, uygun stratejinin tanınmasında yer alan buluş ve yaratıcılık düzeyiyle, çözüm sürecinin artan karmaşıklığıyla (örneğin bir stratejide gerekli olan basamakların sayısı, aralığı ve karmaşıklığı) ve buna bağlı olarak bir çözüme ulaşmak için stratejinin uygulanmasının üst düzey üst bilişsel kontrol gerekliliği ile birlikte artar.

Tanım: Bir problemi çözmek için matematiksel bir strateji seçmenin ya da planlamanın yanı sıra stratejinin uygulanmasının izlenmesi ve denetlenmesi

- 0: Gerekli çözüm süreci açıkça belirtildiğinde ya da aşikâr olduğunda doğrudan adım atma
- 1: Verilen bilgileri birleştirmek ya da kullanmak için basit bir strateji (genellikle tek aşamalı) bulma
- 2: Çok aşamalı basit bir strateji oluşturma, örneğin bir dizi doğrusal aşamayı ya da hedeflenmiş ve denetlenmiş uygulama gerektiren belirlenmiş bir stratejiyi tekrarlı olarak kullanmayı içeren stratejiler
- 3: Karmaşık çok aşamalı bir strateji oluşturma, örneğin birden çok alt amacı bir araya getirmeyi gerektiren ya da strateji kullanımının çözüm sürecini önemli düzeyde izleme ve denetleme gerektirmesi; ya da stratejilerin karşılaştırılması ya da değerlendirilmesi.

Matematikleştirme

Bu yeterliliğin odak noktası, modelleme döngüsünün matematik dışı bir bağlam ile bazı matematiksel alanlar arasında bağlantı kuran yönü üzerinedir. Buna göre, matematikleştirme yeterliliğinin iki bileşeni vardır. Matematik dışı bir durum matematiksel işlemlerin uygulanabileceği bir forma dönüştürme gerektirebilir. Bu, varsayımların sadeleştirilmesini, bağlamda sunulan değişkenlerin ve bunlar arasındaki ilişkilerin belirlenmesini ve bu değişkenlerin matematiksel bir formda ifade edilmesini içerir. Bu dönüştürme yer yer matematikleştirme olarak adlandırılmaktadır. Diğer taraftan, bir matematiksel varlık ya da sonuç, matematik dışı bir durum ya da bağlamla ilişkili olarak yorumlanmayı gerektirebilir. Bu, bağlamın belirli parçaları ile ilişkili olarak matematiksel sonuçların dönüştürülmesini ve bulunan sonucun bağlama uygunluğunun doğrulanmasını içerir. Bu süreç yer yer tersine-matematikleştirme olarak adlandırılmaktadır. Matematiksel alan içinde bundan sonra gelen konuların ve sorunların matematik-içi ele alınış biçimi diğer yeterlilikler altında ele alınmaktadır. Bu nedenle, matematikleştirme yeterliliği matematik dışı bağlamların matematiksel varlıklar yoluyla temsil edilmesi ile ilgiliyken, matematiksel varlıkların temsili ile temsil yeterliliği altında ilgilenilmektedir.

Bu yeterliliğin etkinleştirilmesine duyulan gereksinimi, bağlam öğeleri ve problemin matematiksel yapısı arasında dönüştürme yapmak için gerekli yaratıcılık, iç görü ve bilgi düzeyi ile birlikte artar.

Tanım: Matematik dışı bir durumu matematiksel bir modele dönüştürme, problem durumu ile ilişkili bir model kullanımının sonuçlarını yorumlama, ya da problem durumuna ilişkin modelin uygunluğunu doğrulama

0: Durum ya tamamen matematik içidir ya da matematik dışı durum ve model arasındaki ilişki problemin çözümü ile ilgili değildir.

1: Gerekli varsayımlar, değişkenler, ilişkiler ve sınırlılıklar verildiğinde bir model oluşturma; ya da verilen bir modelden ya da matematiksel sonuçlardan durum hakkında doğrudan sonuç çıkarma

2: Gerekli varsayımlar, değişkenler, ilişkiler ve sınırlılıkların hali hazırda tanımlanabildiği bir model oluşturma; ya da değişen koşulları karşılamak için verilen bir model üzerinde değişiklik yapma; ya da problem durumunun gözden geçirilmesinin zorunlu olduğu bir modeli ya da matematiksel sonuçları yorumlama

3: Varsayımların, değişkenlerin, ilişkilerin ve sınırlılıkların tanımlanması gereken bir durumda bir model oluşturma; ya da problem durumuna ilişkin modelleri doğrulama ya da değerlendirme; ya da farklı modeller arasında bağlantı kurma ya da karşılaştırma yapma

Temsil

Bu yeterliliğin odak noktası, matematiksel varlıkların temsillerinin çözümlenmesi, oluşturulması ve manipüle edilmesi ya da matematiksel bir çözüm gerçekleştirmek için farklı temsiller arasında bağ kurma üzerinedir. “Bir matematiksel varlığın temsili” ifadesinden bir matematiksel kavram, nesne, ilişki, süreç ya da eylemin somut bir ifadesini (eşleştirmesini) anlıyoruz. Bu fiziksel, sözel, sembolik, grafiksel, tablo şeklinde, diyagram şeklinde ya da şekilsel olabilir. Matematiksel görevler sıklıkla metin biçiminde, bazen de sadece bağlamı oturtmaya yardımcı grafik ile sunulurlar. Sözlü/metin yönergelerini ve bilgilerini, fotoğrafları, grafikleri anlama genellikle temsil yeterliliğine ait değildir- bu iletişim yeterliliğinin bir parçasıdır. Benzer şekilde, yalnızca sembolik temsillerle çalışma sembollerin, işlemlerin ve formal dilin kullanımı yeterliliğinde yer almaktadır. Diğer yandan, farklı temsiller arasındaki dönüşüm her zaman temsil yeterliliğinin bir parçasıdır. Örneğin, ilgili metin öğelerinden çıkarsanan matematiksel bilgiyi sözel olmayan temsillere dönüştürme eylemi temsilin uygulanmaya başladığı yerdir. Temsil yeterliliği matematiksel varlıkların diğer varlıklar (matematiksel ya da matematik-dışı) yoluyla temsil edilmesi ile ilgiliyken, matematik-dışı bağlamların matematiksel varlıklar yoluyla temsil edilmesi matematikleştirme yeterliliği altında ele alınmaktadır.

Bu yeterliliğe duyulan gereksinim, seçilip çıkarılacak bilgi miktarıyla, birden çok temsildeki bilginin bir araya getirilmesi gerekliliği ile ve verilen temsilleri kullanmaktan çok temsiller oluşturma gerekliliği ile birlikte artar. Gereksinim ayrıca, minimal çözümleme gerektiren basit ve standart temsillerden (çubuk grafiği ya da Kartezyen grafiği gibi), birden çok bileşeni kapsayan ve muhtemelen belli bir amaç için oluşturulmuş önemli düzeyde çözümleme gerektiren karmaşık ve daha az standart temsillere (nüfus piramidi ya da bir binanın yandan görünümü gibi) doğru temsilin ya da çözümlemesinin karmaşıklığının artması ile birlikte artar.

Tanım: Bir çözüm arayışında verilen temsilleri çözümleme, birbirine dönüştürme ve amaçlı bir şekilde kullanma; bir durumu yansıtmak ya da bir kişinin çalışmasını sunmak için temsiller seçme ya da oluşturma

0: Ya hiç temsil içermez; ya da basit bir temsilden -örneğin bir koordinat sistemi, tablo ya da çubuk grafiği- soyutlanmış değerleri okuma; ya da bu değerleri işaretleme; ya da metinden doğrudan soyutlanmış sayısal değerleri okuma

1: İlişkileri ya da eğilimleri yorumlamak için verilen basit ve standart bir temsil kullanma, örneğin değerleri karşılaştırmak için bir tablodan veri çıkarma ya da bir grafikte gösterilen zamana bağlı değişimi yorumlama; ya da karmaşık bir temsildeki soyutlanmış değerleri işaretleme ya da okuma; ya da basit bir temsil yapılandırma

2: Karmaşık bir temsili anlama ya da kullanma ya da gerekli bazı yapıların sağlandığı karmaşık bir temsil yapılandırma; ya da bir temsili yeniden düzenlemeyi de içeren bir matematiksel varlığın farklı basit temsillerini birbirine dönüştürme ve kullanma

3: Matematiksel varlıkların birden çok karmaşık temsillerini anlama, kullanma, ilişkilendirme ya da birbirine dönüştürme; ya da temsilleri karşılaştırma ya da değerlendirme; ya da karmaşık bir matematiksel varlığı yansıtan bir temsil oluşturma

Sembollerin, İşlemlerin Ve Formal Dilin Kullanımı

Bu yeterlilik matematiksel tanımlar, sonuçlar (gerçekler), kurallar, algoritmalar ve prosedürler gibi matematiksel alan bilgisini etkinleştirme ve kullanmayla, sembolik ifadeleri hatırlama ve kullanmayla, formülleri, fonksiyonel ilişkileri ya da diğer cebirsel ifadeleri anlama ve manipüle etmeyle ve işlemlerin formal kurallarını (örn; aritmetik hesaplamalar ya da denklemlerin çözümü) kullanmayla ilgili yeterlilikleri yansıtmaktadır. Bu yeterlilik aynı zamanda ölçme birimleri ve “hız” ve “yoğunluk” gibi türetilmiş nicelikler ile çalışmayı da kapsar. Matematik-dışı durumların sembolik formülasyonlarını geliştirme matematikleştirmenin bir parçasıdır. Örneğin, matematik-dışı bir durumun anahtar bileşenlerini yansıtmak için bir denklem kurma matematikleştirmeye aitken, bu denklemi çözme sembollerin, işlemlerin ve formal dilin kullanımı yeterliliğinin bir parçasıdır. Sembolik ifadeler matematiksel temsiller olmasına rağmen, sembolik ifadelerin manipüle edilmesi sembollerin, işlemlerin ve formal dilin kullanımı yeterliliğine aittir. Bununla birlikte, sembolik ve diğer temsillerin birbirine dönüştürülmesi temsil yeterliliğine aittir. Burada “değişken” terimi belirsiz bir sayıyı ya da değişen bir niceliği, örneğin $C = 2\pi r$ formülündeki C ve r , temsil eden bir sembolü adlandırmak için kullanılmıştır.

Bu yeterliliğe duyulan gereksinim gerekli matematiksel içerik ve prosedürel bilginin karmaşıklığı ve kapsamlılığı ile birlikte artar.

Tanım: Matematiksel prosedürleri ve dili (sembolik ifadeler, prosedürler ve cebirsel işlemler içeren) anlama ve uygulama, bunlara hükmeden matematiksel usulleri ve kuralları kullanma; tanımların, sonuçların, kuralların ve formal sistemlerin bilgisini etkinleştirme ve kullanma

0: Başlangıç düzeyindeki matematiksel gerçekleri ve tanımları belirtme ve kullanma; ya da sadece kolayca işlem yapılabilen sayıları içeren kısa aritmetiksel hesaplamalar yapma. Örneğin kenar uzunlukları verilen bir dikdörtgenin alanını bulma ya da bir dikdörtgenin alan formülünü yazma

1: Değişkenler içeren basit bir matematiksel ilişkinin doğrudan amaçlı bir şekilde kullanımı (örneğin, bir doğrusal ilişkide yerine koyma); kesirleri ve ondalıkları içeren aritmetiksel hesaplamaları kullanma; sıfıncı düzeyden tekrarlı ya da sürdürülebilir hesaplamalar

kullanma; bir matematiksel tanımın, gerçeğin ya da usulün amaçlı bir şekilde kullanımı, örneğin bir üçgenin iç açılarının toplamı bilgisini bilinmeyen bir açıyı bulmak için kullanma

2: Birden çok bileşeni olan ve değişken içeren ifadeleri kullanma ve manipüle etme (örneğin bir formülü cebirsel olarak yeniden düzenleyerek); birden çok kuralı, tanımı, sonucu, usulü, prosedürü ya da formülü birlikte kullanma; birinci düzeyden tekrarlı ya da sürdürülebilir hesaplamalar kullanma

3: Çeşitli kuralları, gerçekleri, tanımları ve teknikleri bir araya getiren çok adımlı formal matematiksel prosedürleri uygulama; değişken içeren karmaşık matematiksel ilişkilerle esnek bir şekilde çalışma, örneğin belirli bir amaç için cebirsel ifadenin hangi formunun daha iyi olacağına karar vermek için iç gövüyü kullanma.

Muhakeme ve argüman

Bu yeterlilik, sağlam sonuçlar elde etmek için gerekli olan matematiksel bilginin iç zihinsel işlenmesine dayalı geçerli çıkarımlar yapmayla ve bir sonucu doğrulamak ya da daha kesin bir şekilde kanıtlamak için bu çıkarımları bir araya getirmeye ilişkilidir. Görevleri gerçekleştirirde yer alan zihinsel işlemenin ve yansımanın diğer formları diğer yeterliliklerin her birine zemin teşkil etmektedir. Örneğin bir problemi çözmek için bir yaklaşım seçmede ya da oluşturmada gerekli düşünme, strateji oluşturma yeterliliği altında ele alınmakta ve bağlamsal öğelerin matematiksel bir forma dönüştürülmesine ilişkin düşünme matematikleştirme yeterliliğinde hesaba katılmaktadır.

Çıkarım yapmada uygulamaya dökülmesi gereken öğelerin doğası, sayısı ve karmaşıklığı ve gerekli çıkarım zincirinin uzunluğu ve karmaşıklığı bu yeterliliğe duyulan gereksinim artışında önemli katkı sağlayıcılar olacaktır.

Tanım: Argümanları ve sonuçları biçimlendirmek, irdelemek ya da doğrulamak için problemin öğelerini araştıran ve birbirine bağlayan mantıksal olarak kökleşmiş düşünme süreçlerini kullanarak çıkarım yapma

0: Verilen bilgilerden ve yönergelerden doğrudan çıkarımlar yapma

1: Basit matematiksel varlıkları içeren bir problemin bir boyutunda muhakeme adımlarını kullanarak çıkarımlar yapma

2: Problemin farklı boyutlarındaki bilgi parçalarını birleştirerek ya da problemdeki karmaşık varlıklarla ilişkili olarak çıkarımlar yapma; çok adımlı bir argümanı izlemek ya da oluşturmak için bir çıkarım zinciri meydana getirme

3: İlişkilendirilmiş çıkarım zincirleri kullanma ya da oluşturma; ya da karmaşık çıkarımları kontrol etme ya da doğrulama; ya da sürdürülebilir ve güdümlü bir yolla, karmaşık bilginin birden çok ögesini kullanarak ve birleştirerek kararları ve çıkarımları sentezleme ve değerlendirme.

EK 2: Kolmogorov-Smirnov Normal Dağılım Testi Sonuçları

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
P1_ILT1	,295	156	,000	,829	156	,000
P1_ILT2	,329	156	,000	,805	156	,000
P2_ILT1	,355	156	,000	,774	156	,000
P2_ILT2	,366	156	,000	,764	156	,000
P3_ILT1	,325	156	,000	,805	156	,000
P3_ILT2	,332	156	,000	,806	156	,000
P4_ILT1	,322	156	,000	,810	156	,000
P4_ILT2	,358	156	,000	,763	156	,000
P5_ILT1	,376	156	,000	,740	156	,000
P5_ILT2	,396	156	,000	,707	156	,000
P1_SP1	,290	156	,000	,839	156	,000
P1_SP2	,290	156	,000	,825	156	,000
P2_SP1	,285	156	,000	,837	156	,000
P2_SP2	,288	156	,000	,822	156	,000
P3_SP1	,230	156	,000	,851	156	,000
P3_SP2	,232	156	,000	,846	156	,000
P4_SP1	,269	156	,000	,850	156	,000
P4_SP2	,268	156	,000	,841	156	,000
P5_SP1	,296	156	,000	,814	156	,000
P5_SP2	,292	156	,000	,812	156	,000
P1_MT1	,353	156	,000	,786	156	,000
P1_MT2	,348	156	,000	,792	156	,000
P2_MT1	,385	156	,000	,738	156	,000
P2_MT2	,365	156	,000	,769	156	,000
P3_MT1	,316	156	,000	,819	156	,000
P3_MT2	,307	156	,000	,829	156	,000
P4_MT1	,322	156	,000	,811	156	,000
P4_MT2	,318	156	,000	,813	156	,000
P5_MT1	,338	156	,000	,801	156	,000
P5_MT2	,351	156	,000	,784	156	,000
P1_TM1	,319	156	,000	,776	156	,000
P1_TM2	,296	156	,000	,796	156	,000
P2_TM1	,301	156	,000	,785	156	,000
P2_TM2	,295	156	,000	,792	156	,000
P3_TM1	,307	156	,000	,763	156	,000
P3_TM2	,287	156	,000	,783	156	,000
P4_TM1	,270	156	,000	,803	156	,000
P4_TM2	,263	156	,000	,808	156	,000

P5_TM1	,280	156	,000	,772	156	,000
P5_TM2	,231	156	,000	,814	156	,000
P1_SM1	,289	156	,000	,822	156	,000
P1_SM2	,277	156	,000	,826	156	,000
P2_SM1	,328	156	,000	,761	156	,000
P2_SM2	,306	156	,000	,786	156	,000
P3_SM1	,280	156	,000	,795	156	,000
P3_SM2	,268	156	,000	,818	156	,000
P4_SM1	,264	156	,000	,825	156	,000
P4_SM2	,262	156	,000	,831	156	,000
P5_SM1	,310	156	,000	,802	156	,000
P5_SM2	,281	156	,000	,823	156	,000
P1_MU1	,271	156	,000	,816	156	,000
P1_MU2	,270	156	,000	,816	156	,000
P2_MU1	,291	156	,000	,811	156	,000
P2_MU2	,314	156	,000	,800	156	,000
P3_MU1	,245	156	,000	,821	156	,000
P3_MU2	,277	156	,000	,817	156	,000
P4_MU1	,253	156	,000	,840	156	,000
P4_MU2	,265	156	,000	,836	156	,000
P5_MU1	,270	156	,000	,840	156	,000
P5_MU2	,283	156	,000	,821	156	,000
P1_TOP1	,147	156	,000	,956	156	,000
P1_TOP2	,150	156	,000	,959	156	,000
P2_TOP1	,106	156	,000	,977	156	,011
P2_TOP2	,119	156	,000	,969	156	,001
P3_TOP1	,139	156	,000	,956	156	,000
P3_TOP2	,101	156	,001	,963	156	,000
P4_TOP1	,138	156	,000	,970	156	,002
P4_TOP2	,100	156	,001	,978	156	,013
P5_TOP1	,124	156	,000	,975	156	,006
P5_TOP2	,125	156	,000	,974	156	,005
TOP_ORT	,069	156	,068	,984	156	,076
ILT_ORT	,198	156	,000	,939	156	,000
SP_ORT	,114	156	,000	,971	156	,002
MT_ORT	,175	156	,000	,948	156	,000
TM_ORT	,130	156	,000	,934	156	,000
SM_ORT	,108	156	,000	,970	156	,002
MU_ORT	,117	156	,000	,968	156	,001

a, Lilliefors Significance Correction

EK 3: Betimsel İstatistikler

Puan Kategorisi ve Puanlayıcı	N	Mean	Median	Mode	Std. Deviation	Skewness	Std. Error of Skew-ness	Skew-ness /SEoS	Kurtosis	Std. Error of Kurtosis	Kurtosis /SEoK
P1_ILT1	156	1,12	1	1	0,718	0,249	0,194	1,28	-0,106	0,386	-0,27
P1_ILT2	156	1,2	1	1	0,666	0,281	0,194	1,45	0,216	0,386	0,56
P2_ILT1	156	1,19	1	1	0,62	0,35	0,194	1,80	0,522	0,386	1,35
P2_ILT2	156	1,22	1	1	0,618	0,483	0,194	2,49	0,676	0,386	1,75
P3_ILT1	156	1,1	1	1	0,679	0,38	0,194	1,96	0,403	0,386	1,04
P3_ILT2	156	1,19	1	1	0,68	0,376	0,194	1,94	0,351	0,386	0,91
P4_ILT1	156	1,31	1	1	0,678	0,287	0,194	1,48	0,093	0,386	0,24
P4_ILT2	156	1,22	1	1	0,596	0,25	0,194	1,29	0,261	0,386	0,68
P5_ILT1	156	1,24	1	1	0,572	0,372	0,194	1,92	0,382	0,386	0,99
P5_ILT2	156	1,17	1	1	0,535	0,392	0,194	2,02	0,757	0,386	1,96
P1_SP1	156	1,36	1	1	0,736	0,204	0,194	1,05	-0,159	0,386	-0,41
P1_SP2	156	1,39	1	1	0,696	0,107	0,194	0,55	-0,155	0,386	-0,40
P2_SP1	156	1,23	1	1	0,726	0,129	0,194	0,66	-0,246	0,386	-0,64
P2_SP2	156	1,24	1	1	0,692	0,002	0,194	0,01	-0,329	0,386	-0,85
P3_SP1	156	1,19	1	1	0,796	0,029	0,194	0,15	-0,717	0,386	-1,86
P3_SP2	156	1,18	1	1	0,783	-0,001	0,194	-0,01	-0,746	0,386	-1,93
P4_SP1	156	1,26	1	1	0,763	0,133	0,194	0,69	-0,34	0,386	-0,88
P4_SP2	156	1,31	1	1	0,734	0,023	0,194	0,12	-0,333	0,386	-0,86
P5_SP1	156	1,37	1	1	0,673	0,053	0,194	0,27	-0,167	0,386	-0,43
P5_SP2	156	1,24	1	1	0,673	-0,065	0,194	-0,34	-0,39	0,386	-1,01
P1_MT1	156	1,19	1	1	0,67	0,547	0,194	2,82	0,72	0,386	1,87
P1_MT2	156	1,21	1	1	0,672	0,497	0,194	2,56	0,568	0,386	1,47
P2_MT1	156	1,19	1	1	0,599	0,638	0,194	3,29	1,197	0,386	3,10
P2_MT2	156	1,22	1	1	0,635	0,547	0,194	2,82	0,77	0,386	1,99
P3_MT1	156	1,13	1	1	0,707	0,368	0,194	1,90	0,222	0,386	0,58
P3_MT2	156	1,2	1	1	0,722	0,305	0,194	1,57	0,038	0,386	0,10
P4_MT1	156	1,27	1	1	0,675	0,252	0,194	1,30	0,09	0,386	0,23
P4_MT2	156	1,28	1	1	0,678	0,229	0,194	1,18	0,049	0,386	0,13
P5_MT1	156	1,26	1	1	0,673	0,403	0,194	2,08	0,319	0,386	0,83
P5_MT2	156	1,26	1	1	0,644	0,43	0,194	2,22	0,426	0,386	1,10
P1_TM1	156	0,84	1	1	0,627	0,29	0,194	1,49	0,169	0,386	0,44
P1_TM2	156	0,89	1	1	0,668	0,391	0,194	2,02	0,237	0,386	0,61
P2_TM1	156	0,84	1	1	0,657	0,457	0,194	2,36	0,434	0,386	1,12

P2_TM2	156	0,86	1	1	0,667	0,432	0,194	2,23	0,293	0,386	0,76
P3_TM1	156	0,65	0	0	0,776	1,027	0,194	5,29	0,499	0,386	1,29
P3_TM2	156	0,7	1	0	0,766	0,836	0,194	4,31	0,053	0,386	0,14
P4_TM1	156	0,78	1	0	0,8	0,658	0,194	3,39	-0,455	0,386	-1,18
P4_TM2	156	0,8	1	0	0,807	0,602	0,194	3,10	-0,567	0,386	-1,47
P5_TM1	156	0,67	1	0a	0,694	0,791	0,194	4,08	0,358	0,386	0,93
P5_TM2	156	0,85	1	1	0,752	0,342	0,194	1,76	-0,863	0,386	-2,24
P1_SM1	156	1,01	1	1	0,709	0,311	0,194	1,60	-0,063	0,386	-0,16
P1_SM2	156	1,13	1	1	0,707	0,035	0,194	0,18	-0,497	0,386	-1,29
P2_SM1	156	0,79	1	1	0,612	0,322	0,194	1,66	0,323	0,386	0,84
P2_SM2	156	0,85	1	1	0,644	0,3	0,194	1,55	-0,008	0,386	-0,02
P3_SM1	156	0,85	1	1	0,692	0,568	0,194	2,93	0,475	0,386	1,23
P3_SM2	156	0,9	1	1	0,725	0,457	0,194	2,36	-0,038	0,386	-0,10
P4_SM1	156	0,97	1	1	0,722	0,256	0,194	1,32	-0,436	0,386	-1,13
P4_SM2	156	1,04	1	1	0,73	0,132	0,194	0,68	-0,594	0,386	-1,54
P5_SM1	156	0,99	1	1	0,677	0,387	0,194	1,99	0,376	0,386	0,97
P5_SM2	156	1,03	1	1	0,704	0,18	0,194	0,93	-0,349	0,386	-0,90
P1_MU1	156	0,92	1	1	0,705	0,333	0,194	1,72	-0,218	0,386	-0,56
P1_MU2	156	0,97	1	1	0,695	0,16	0,194	0,82	-0,539	0,386	-1,40
P2_MU1	156	0,95	1	1	0,698	0,415	0,194	2,14	0,191	0,386	0,49
P2_MU2	156	1	1	1	0,672	0,388	0,194	2,00	0,426	0,386	1,10
P3_MU1	156	0,87	1	1	0,809	0,683	0,194	3,52	-0,016	0,386	-0,04
P3_MU2	156	0,92	1	1	0,77	0,659	0,194	3,40	0,307	0,386	0,80
P4_MU1	156	1,01	1	1	0,775	0,326	0,194	1,68	-0,443	0,386	-1,15
P4_MU2	156	1,05	1	1	0,743	0,204	0,194	1,05	-0,463	0,386	-1,20
P5_MU1	156	1,04	1	1	0,769	0,364	0,194	1,88	-0,241	0,386	-0,62
P5_MU2	156	1,01	1	1	0,7	0,211	0,194	1,09	-0,286	0,386	-0,74
P1_TOP1	156	6,44	6	6	2,463	0,497	0,194	2,56	-0,302	0,386	-0,78
P1_TOP2	156	6,79	6	6	2,478	0,493	0,194	2,54	-0,142	0,386	-0,37
P2_TOP1	156	6,18	6	6	2,485	0,265	0,194	1,37	-0,33	0,386	-0,85
P2_TOP2	156	6,38	6	6	2,474	0,407	0,194	2,10	0,438	0,386	1,13
P3_TOP1	156	5,79	5	4a	2,729	0,608	0,194	3,13	-0,077	0,386	-0,20
P3_TOP2	156	6,08	6	6	2,698	0,542	0,194	2,79	0,088	0,386	0,23
P4_TOP1	156	6,59	6	5	2,622	0,374	0,194	1,93	-0,138	0,386	-0,36
P4_TOP2	156	6,71	7	5	2,489	0,266	0,194	1,37	0,164	0,386	0,42
P5_TOP1	156	6,57	6	5	2,605	0,157	0,194	0,81	-0,353	0,386	-0,91
P5_TOP2	156	6,57	6	8	2,418	0,204	0,194	1,05	-0,024	0,386	-0,06
TOP_ORT	156	6,4103	6,2	7,2	2,26523	0,354	0,194	1,82	-0,147	0,386	-0,38
ILT_ORT	156	1,1955	1	1	0,52096	0,61	0,194	3,14	0,706	0,386	1,83
SP_ORT	156	1,2769	1,2	1	0,62236	0,043	0,194	0,22	-0,28	0,386	-0,73

MT_ORT	156	1,2199	1	1	0,5473	0,27	0,194	1,39	0,397	0,386	1,03
TM_ORT	156	0,7878	0,8	0	0,61898	0,491	0,194	2,53	-0,448	0,386	-1,16
SM_ORT	156	0,9564	1	1	0,54715	0,197	0,194	1,02	-0,004	0,386	-0,01
MU_ORT	156	0,9737	1	1	0,59656	0,295	0,194	1,52	-0,371	0,386	-0,96



EK 4: Uzman Görüş Formu

Değerli Uzman,

Aşağıda PISA matematik okuryazarlığı maddelerinin gerektirdiği matematik okuryazarlığı becerilerinin düzeylerini belirlemek amacıyla kullanılmak üzere Turner vd. (2013) tarafından geliştirilmiş olan ölçme aracının (analitik dereceli puanlama anahtarının) İngilizce ve Türkçe formları yer almaktadır. Bu ölçme aracının Türkçe 'ye uyarlanmasını da içeren yüksek lisans tezinde kullanmak amacıyla ölçme aracının Türkçe formunun dilimize ve matematik eğitimi alanına uygunluğunun belirlenmesinde görüşlerinize ihtiyaç duyulmaktadır. Ölçme aracının özgün formundaki her bir cümlemin Türkçe çevirileri karşılıklı olarak sunulmuştur. Yapılan çevirinin uygunluğuna ilişkin görüşlerinizi *Uygun*, *Düzeltilmeli* ve *Uygun Değil* seçeneklerinden birini işaretleyerek belirtmeniz ve varsa önerilerini belirtmeniz istenmektedir.

Bu çalışmadaki katkılarınız için teşekkür eder, saygılarımı sunarım.

Fatma KIZILTOPRAK
Gazi Üniversitesi
Eğitimde Ölçme ve Değerlendirme Anabilim
Dalı

Original Rubrik	Çeviri	Uygun	Düzeltilmeli	Uygun Değil	Öneri
Competency Definitions and Level Descriptions (May 2013)	Beceri Tanımları ve Düzey Açıklamaları (2013)				
Communication: The communication competency has both ‘receptive’ and ‘constructive’ components. The receptive component includes understanding what is being stated and shown related to the mathematical objectives of the task, including the mathematical language used, what information is relevant, and what is the nature of the response requested. The constructive component consists of presenting the response that may include solution steps, description of the reasoning used and justification of the answer provided.	İletişim: İletişim becerisi hem “algılayıcı” hem “yapıcı” bileşenlere sahiptir.				
	Algılayıcı bileşen, kullanılan matematiksel dili, ilgili bilgileri ve gerekli yanıtın doğasını kapsayan, görevin matematiksel amaçlarıyla ilişkili olarak belirtilenlerin ve gösterilenlerin anlaşılmasını içerir.				
	Yapıcı bileşen çözüm aşamalarını, yapılan muhakemenin açıklanmasını ve verilen yanıtın doğrulanmasını içerebilen yanıtın sunuluşunu kapsamaktadır.				
In written and computer-based items, receptive communication relates to understanding text and images, still and moving.	Yazılı ve bilgisayar tabanlı maddelerde algılayıcı iletişim, metnin ve hareketli ve hareketsiz şekillerin anlaşılması ile ilgilidir.				
Text includes verbally presented mathematical expressions and may also be found in mathematical representations (for example titles, labels and legends in graphs and diagrams).	Metin sözel olarak verilen matematiksel ifadeler içerir ve aynı zamanda matematiksel temsillerin (örneğin; başlıklar, etiketler ve grafik ve diyagramlardaki açıklayıcı bilgiler) içinde de bulunabilir.				

Communication does not include knowing how to approach or solve the problem, how to make use of particular information provided, or how to reason about or justify the answer obtained; rather it is the understanding or presenting of relevant information.

İletişim probleme nasıl yaklaşılacağını ya da problemin nasıl çözüleceğini, verilen belirli bilgilerin amaçlı bir şekilde nasıl kullanılacağını, elde edilen yanıtın nasıl muhakeme yapılacağını ya da savunulacağını bilmeyi içermez; daha çok ilgili bilgilerin sunumu ve anlaşılması ile ilgilidir.

It also does not apply to extracting or processing mathematical information from representations.

Ayrıca temsillerden matematiksel bilgilerin çıkarılması ve işlenmesi ile de ilgili değildir.

In computer-based items, the instructions about navigation and other issues related to the computer environment may add to the general task demand, but is not part of the communication competency.

Bilgisayar tabanlı maddelerde, navigasyon ve bilgisayar ortamı ile ilgili diğer konular hakkındaki yönergeler genel görev gereksinimlerini arttırabilir ancak bu iletişim becerisinin bir parçası değildir.

Demand for the receptive aspect of this competency increases according to the complexity of material to be interpreted in understanding the task; the need to link multiple information sources or to move backwards and forwards (to cycle) between information elements.

Bu becerinin algılayıcı boyutuna duyulan gereksinim, görevin anlaşılmasında yorumlanacak olan materyalin karmaşıklığına; çoklu bilgi kaynaklarını ilişkilendirme ya da bilgi parçaları arasında ileriye ve geriye gitme (devir yapma) gerekliliğine bağlı olarak artar.

The constructive aspect increases with the need to provide a detailed written solution or explanation.

Yapıcı boyut detaylı bir yazılı çözüm ya da açıklama sağlanması gerekliliği ile artar.