



T.C.

AKDENİZ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

EĞİTİM BİLİMLERİ ANA BİLİM DALI

YÜKSEK
LİSANS
TEZİ

ORTAÖĞRETİM İNGİLİZCE YETERLİĞİNİN
PUANLAYICILAR ARASI GÜVENİRLİĞİNİN
GENELLENEBİLİRLİK KURAMI VE RASCH
MODELİ İLE BELİRLENMESİ

AYKUT TÜZAN

EĞİTİMDE ÖLÇME VE DEĞERLENDİRME BİLİM DALI

Antalya, 2023

T.C.
AKDENİZ ÜNİVERSİTESİ EĞİTİM BİLİMLERİ ENSTİTÜSÜ
EĞİTİM BİLİMLERİ ANABİLİM DALI
EĞİTİMDE ÖLÇME VE DEĞERLENDİRME TEZLİ YÜKSEK LİSANS
PROGRAMI

**ORTAÖĞRETİM İNGİLİZCE YETERLİĞİNİN PUANLAYICILAR
ARASI GÜVENİRLİĞİNİN GENELLENEBİLİRLİK KURAMI VE
RASCH MODELİ İLE BELİRLENMESİ**

YÜKSEK LİSANS TEZİ

Aykut TÜZAN

Antalya, 2023

DOĞRULUK BEYANI

Yüksek lisans tezi olarak sunduğum bu çalışmayı, bilimsel ahlak ve geleneklere aykırı düşecek bir yol ve yardıma başvurmaksızın yazdığımı, yararlandığım eserlerin kaynakçalarda gösterilenlerden oluştuğunu ve bu eserleri her kullanışımında alıntı yaparak yararlandığımı belirtir, bunu onurumla doğrularım. Enstitü tarafından belli bir zamana bağlı olmaksızın, tezimle ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçlara katlanacağımı bildiririm.

Aykut TÜZAN



T.C.
AKDENİZ ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Aykut TÜZAN'ın bu çalışması **24/10/2023** tarihinde jürimiz tarafından **Eğitim Bilimleri** Ana Bilim Dalı **Eğitimde Ölçme ve Değerlendirme** Tezli Yüksek Lisans Programında **Yüksek Lisans Tezi** olarak **oy birliği** ile kabul edilmiştir.

İmza

Başkan : **Prof. Dr. Bayram BIÇAK**
Akdeniz Üniversitesi, Eğitim Fakültesi

Üye : **Doç. Dr. Sami PEKTAŞ**
Niğde Ömer Halisdemir Üniversitesi, Eğitim Fakültesi

Üye (Danışman) : **Doç. Dr. Hakan KOĞAR**
Akdeniz Üniversitesi, Eğitim Fakültesi

YÜKSEK LİSANS TEZİNİN ADI: Ortaöğretim İngilizce Yeterliğinin Puanlayıcılar Arası Güvenirliğinin Genellenebilirlik Kuramı ve Rasch Modeli İle Belirlenmesi

ONAY: Bu tez, Enstitü Yönetim Kurulunca belirlenen yukarıdaki jüri üyeleri tarafından uygun görülmüş ve Enstitü Yönetim Kurulunun tarihli ve sayılı kararıyla kabul edilmiştir.

Doç. Dr. Güçlü ŞEKERCİOĞLU
Enstitü Müdürü

TEŞEKKÜR

Çalışmam boyunca zaman fark etmeksizin istediğim her an ulaşabildiğim sorunlarıma çözüm bulmama özenle yardımcı olan ve her türlü desteği veren danışmanım sayın Doç. Dr. Hakan KOĞAR’a, veri analizi esnasında yaşadığım problemlere içtenlikle çözüm üreten sayın Prof. Dr. Neşe GÜLER hocama,

Akademik tecrübelerini paylaşan ve başarılarıyla ilham kaynağı olan dost bildiğim Dr. Gökhan GÜÇİN, Oğuz MERCAN ve Serkan YILMAZ’a, çalışmam boyunca her türlü kolaylığı sağlayan okul müdürüm Ahmet KOCABAŞ’a,

Verileri toplamamda bana yardımcı olan değerli meslektaşlarım Veli KAPLAN, Nevike DEMİRCAN, Nurgül ÜNAL ve Aysun KURT hocama,

Beni yetiştirirken hayatından maddi, manevi fedakarlıklar vermiş olan değerli annem Hatice TÜZAN, değerli babam Mustafa TÜZAN’a, yazılımsal problemlerimi çözen kardeşim Emre TÜZAN’a, hayatının belki de en zor döneminde bana sabırla destek olan ve motive eden biricik eşim Gizem CEYLAN TÜZAN’a ve tezimle beraber gelişip doğan, şans getirdiğine inandığım biricik kızım Nil Mina’ ya teşekkür ederim.

Aykut TÜZAN

ÖZET

ORTAÖĞRETİM İNGİLİZCE YETERLİĞİNİN PUANLAYICILAR ARASI GÜVENİRLİĞİNİN GENELLENEBİLİRLİK KURAMI VE RASCH MODELİ İLE BELİRLENMESİ

Tüzan, Aykut

Yüksek Lisans, Eğitimde Ölçme ve Değerlendirme Bölümü

Tez Danışmanı: Doç. Dr. Hakan KOĞAR

Haziran 2023, 98 sayfa

Bu araştırmada; 2022-2023 öğretim yılına ait ortaöğretim 12. sınıf öğrencilerinin İngilizce yeterliğine ait puanların farklı puanlayıcılar tarafından bütünsel puanlama anahtarı ve puanlama anahtarı olmadan puanlanması sonucunda “Genellenebilirlik Kuramı” ve “Çok Değişkenlik Kaynaklı Rasch Modeli” ile güvenirliklerinin ve tutarlılıkların belirlenmesi hedeflenmiştir. Genellenebilirlik Kuramında çapraz desen ele alınarak karar çalışmalarının geçerliği ve karşılaştırılması incelenmiştir. Rasch modeli üzerinden çok değişkenlik modeli ile veriler ele alınmıştır. Bu kuramlar kullanılırken okul, cinsiyet ve ölçekli-ölçeksiz değerlendirme değişkenleri dikkate alınmış olup analizler tamamen gerçek veriler üzerinden uygulanmıştır.

Araştırmaya konu olan İngilizce yeterliğinin değerlendirilmesi için 12. sınıf seviyesinde kompozisyon yazma becerisini puanlamak amacıyla 3 farklı uzman tarafından değerlendirmeleri yapılması planlanan 4 farklı özellikteki ortaöğretim okullarında 12. sınıftaki öğrenim gören 124 öğrenciye uygulanmıştır. Bu öğrencilerin verdiği cevaplar 1’i ölçekli diğer ikisi ölçeksiz olacak şekilde 3 farklı puanlayıcı tarafından puanlanması yapılmıştır. Araştırmanın sonraki aşamasında elde edilen puanların güvenirliği “Genellenebilirlik Kuramı” ve “Çok Değişkenlik Rasch Modeli Kuramı”na göre incelenmiştir. Genellenebilirlik Kuramında $b \times p$ (b: birey, p: puanlayıcı) çaprazlanmış desen kullanılmış olup güvenirlik katsayıları hesaplanmıştır. Çok Değişkenlik Kaynaklı Rasch Modeliyle ise birey, puanlayıcı, ve okul yüzeylerine ilişkin farklı güvenirlik hesaplamaları yapılmıştır. Hesaplama işlemi “Genellenebilirlik Kuramı” için EduG 6.1 (Swiss Society for Research IN Education Working

Group, 2010) bilgisayar programı, ÇDKRM analizleri ise FACETS (Linacre, 2007) programı kullanılmıştır.

Bu araştırma sonrası elde edilen sonuçlara göre Genellenebilirlik Kuramına göre İngilizce yeterliliğinin G ve Phi katsayılarının alternatif karar çalışmalarında nispeten de olsa gerçek duruma göre farklı görülmüştür. Puanlayıcılar arası uyumun belirlenmesinde farklı değerler gösteren puanlayıcıların verdikleri puanlar arasında anlamlı bir ilişki olduğu sonucuna varılmıştır. Genellenebilirlik Kuramı doğrultusunda İngilizce yeterliliğinin ölçülmesiyle elde edilen puanların birinci puanlayıcının G katsayısı 0.85 iken ikinci ve üçüncü puanlayıcının 0.88 düzeyindedir. Birinci puanlayıcının Phi katsayısı 0.79, ikinci puanlayıcının Phi katsayısı 0.72 ve üçüncü puanlayıcının Phi katsayısı 0.82 olarak hesaplanmıştır. Bu sonuçlara göre, 2022 yılında uygulanan İngilizce yeterliliğinin ölçülmesi sonucunda puanlayıcılar arasında farklılık gösterdiği görülmüştür. İngilizce yeterliliğini ölçme adına yapılan çalışmada üç puanlayıcının puanları ortalamaları arasında fark olduğu görülmüş olup birbirleriyle uyumlu puanlama yaptıkları belirlenmiştir. Ayrıca madde sayısını artırmanın puanlayıcı sayısı artırmadan işgücü zaman ve maliyet açısından daha faydalı olacağı sonucuna ulaşılmıştır.

Yapılan bu çalışmada İngilizce yeterliliğini ölçmede kullanılan kuramların arasında iki kuramında birbirine yakın sonuç verdiği görülmüştür. Bu iki kuramında detaylı inceleme yapılması amacıyla kullanılabilirliğin faydalı olacağı öngörülmektedir. Ayrıca puanlayıcıların ölçekli ve ölçeksiz puanlamaları sonrası ölçek kullanımının ortaya çıkarmış olduğu farkı görmek adına literatüre katkı sağlayacağı düşünülmüştür.

Anahtar Sözcükler: *Genellenebilirlik kuramı, Rasch modeli, Puanlayıcılar arası güvenilirlik, Bütünsel dereceleme anahtarı, Performansa dayalı durum belirleme*

ABSTRACT

DETERMINATION OF INTER-RATER RELIABILITY OF SECONDARY EDUCATION ENGLISH PROFICIENCY USING GENERALIZATION THEORY AND RASCH MODEL

Tuzan, Aykut

Master's Degree, Department of Measurement and Evaluation in Education

Thesis Advisor: Assoc. Dr. Hakan KOGAR

June 2023, 98 pages

In this study; As a result of scoring the English proficiency scores of 12th grade secondary school students in the 2022-2023 academic year by different raters with and without a holistic scoring key, it was aimed to determine their reliability and consistency with the "Generalizability Theory" and "Multivariability-Based Rasch Model". In Generalizability Theory, the validity and comparison of decision studies were examined by considering cross-over design. The data was handled with the multivariability model through the Rasch model. While using these theories, school, gender and scaled and non-scaled evaluation variables were taken into account and the analyzes were carried out entirely on real data.

In order to evaluate the English proficiency that is the subject of the research, it was applied to 124 students in the 12th grade in 4 different secondary schools, which were planned to be evaluated by 3 different experts in order to score the composition writing skill at the 12th grade level. The answers given by these students were scored by 3 different raters, one with a scale and the other two without a scale. In the next stage of the research, the reliability of the scores obtained was examined according to "Generalizability Theory" and "Multivariability Rasch Model Theory". In Generalizability Theory, reliability coefficients were calculated using the bxp (b: individual, p: rater) crossed design. Separate reliability calculations were made for individual, rater and school dimensions with the Multivariability Rasch Model. EduG 6.1 (Swiss Society for Research IN Education Working Group, 2010) computer program was used for the calculation process, and FACETS (Linacre, 2007) program was used for the MLRM analysis.

According to the results obtained after this research, according to the Generalizability Theory, the G and Phi coefficients of English proficiency were seen to be relatively different compared to the real situation in alternative decision studies. It was concluded that there was a significant relationship between the scores given by the raters, showing varying values in determining inter-rater agreement. According to the Generalizability Theory, the G coefficient of the first rater of the scores obtained by measuring English proficiency is 0.85, while the second and third rater's are 0.88. The Phi coefficient of the first rater was calculated as 0.79, the Phi coefficient of the second rater was calculated as 0.72, and the Phi coefficient of the third rater was calculated as 0.82. According to these results, it was observed that there were differences between raters as a result of measuring English proficiency applied in 2022. In the study conducted to measure English proficiency, it was determined that although there was a difference between the average scores of the three raters, they scored compatible with each other. Additionally, it was concluded that increasing the number of items would be more beneficial in terms of labor time and cost than increasing the number of raters.

In this research, it was seen that among the theories used to measure English proficiency, two theories gave similar results. In this research, it was seen that among the theories used to measure English proficiency, two theories gave similar results. In addition, it was thought that it would contribute to the literature in order to see the difference made by the use of the scale after the raters scored with and without a scale.

Key words: *Generalizability theory, Rasch model, Inter-rater reliability, Holistic rubric, Performance-based status determination.*

İÇİNDEKİLER

TEŞEKKÜR	i
ÖZET	ii
ABSTRACT	iv
TABLO VE ŞEKİLLER DİZİNİ.....	viii
KISALTMALAR LİSTESİ.....	ix

BÖLÜM I

GİRİŞ

1.1.Problem Durumu	1
1.3.Araştırmanın Amacı	4
1.4.Araştırmanın Önemi	4
1.5. Sayıtlılar	5
1.6. Sınırlılıklar	5

BÖLÜM II

KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

2.1.Kuramsal Çerçeve	7
2.1.1.İngilizce Öğretiminde Mevcut Değerlendirme Sistemi	8
2.1.2.Tamamlayıcı Ölçme ve Değerlendirme Teknikleri	10
2.2.Araştırmada Kullanılan Kuramlar	22
2.2.1Madde Tepki Kuramı	22
2.2.2.Genellenebilirlik Kuramı	26
2.2.İlgili Araştırmalar	31
2.2.1.Yurt İçindeki Araştırmalar.....	32
2.2.2. Yurt Dışındaki Araştırmalar	35

BÖLÜM III

YÖNTEM

3.1.Araştırmanın Modeli.....	43
3.2.Araştırma Grubu	43
3.3.Verİ Toplama Araçları	43
3.4.Verilerin Toplanması	44

3.5.Verilerin Analizi	44
-----------------------------	----

BÖLÜM IV

BULGULAR

4.1. Genellenebilirlik Analizi	47
4.2. Çok Değişkenlik Kaynaklı Rasch Modeli Analizi	57

BÖLÜM V

SONUÇLAR VE ÖNERİLER

5.1.Sonuçlar	62
5.2.Öneriler	66
5.2.1.Araştırmadan Çıkarılacak Öneriler	67
5.2.2.İleride Yapılacak Araştırmalar İçin Öneriler	67

KAYNAKLAR	68
------------------------	-----------

EKLER.....	75
-------------------	-----------

Ek 1: İNGLİZCE YETERLİLİĞİ PERFORMANS TESTİ.....	75
--	----

Ek 2: BÜTÜNSEL PUANLAMA ANAHTARI (WRITING ASSESSMENT CRITERIA (PARAGRAPH)).....	76
---	----

Ek 3: ÖLÇEKLE PUANLANDIRILMIŞ PERFORMANS TESTİ ÖRNEĞİ.....	77
--	----

Ek 4: ÖLÇEKSİZ PUANLAMA YÖNTEMİ İLE PUANLANAN ÖĞRENCİ TESTLERİ ÖRNEKLERİ.....	78
---	----

ÖZGEÇMİŞ	81
-----------------------	-----------

İNTİHAL RAPORU	83
-----------------------------	-----------

TABLO VE ŞEKİLLER DİZİNİ

Tablo:2.1. Yabancı Dil Yeterliliği için Ölçme Değerlendirme Biçim ve İçerikleri (Brown, 2004:83).....	9
Tablo 2. 2. Analitik Puanlama Örneği	17
Tablo 4.1. 2022-2023 Öğretim Yılı 12. Sınıf Öğrencilerinden Elde Edilen Örneklemen İngilizce Performans Değerlendirmesi Sonucu Alınan Puanlara İlişkin Betimsel İstatistikler.....	45
Tablo 4.2.Örnekleme Ait Tek Değişkenli G Çalışması Sonucunda Anadolu Lisesinin Varyans Değerleri.....	46
Tablo 4.3. Anadolu Lisesinin Örneklemen Çok Değişkenli Model K Çalışması İle Farklı Sayıdaki Puanlayıcı Sayıları Senaryolarına Göre G Ve Phi Katsayıları.....	47
Tablo 4.4. Puanlayıcıların Anadolu Lisesine göre G Ve Phi Katsayıları.....	48
Tablo 4.5. G Çalışması Sonucu Fen Lisesinin Varyans Değerleri.....	48
Tablo 4.6. Fen Lisesi Örneklemen Çok Değişkenli Model K Çalışması İle Farklı Sayıdaki Okul Ve Puanlayıcı Sayıları Senaryolarına Göre G Ve Phi Katsayıları.....	49
Tablo 4.7. Fen Lisesini puanlayıcıların G ve Phi katsayıları.....	50
Tablo 4.8. Örnekleme ait tek değişkenli G Çalışması sonucunda özel okulunun varyans Değerleri.....	51
Tablo 4.9. Özel Okul Örneklemen Çok Değişkenli Model K Çalışması İle Farklı Sayıdaki Puanlayıcı Sayıları Senaryolarına Göre G Ve Phi Katsayıları.....	52
Tablo 4.10. Özel Okulu puanlayan Puanlayıcıların G Ve Phi Katsayıları.....	53
Tablo 4.11. Örnekleme Ait Tek Değişkenli G Çalışması Sonucunda Çok Programlı Lisenin Varyans Değerleri.....	53
Tablo 4.12. Çok Programlı Lisenin Örneklemen Çok Değişkenli Model K Çalışması İle Farklı Sayıdaki Puanlayıcı Sayıları Senaryolarına Göre G Ve Phi Katsayıları.....	55
Tablo 4.13.Çok Programlı Lisenin Puanlayıcıların G Ve Phi Katsayıları.....	56
Tablo 4.14.Bireylerin Ölçütlerine İlişkin Ölçüm Raporu.....	58
Tablo 4.15. Puanlayıcı Ölçüm Raporu.....	59
Tablo 4.16. Okulların Ölçüm Raporu.....	60
Şekil 2.1. Performans Örneği	16
Şekil 4.1. Çok Değişken Kaynaklı Rasch Modeli İle Çalışmanın 3 Yüzeyli Logit Cetveli.....	57

KISALTMALAR LİSTESİ

MEB	: Milli Eğitim Bakanlığı
MTK	: Madde Tepki Kuramı
G Kuramı	: Genellenebilirlik Kuramı
G Çalışması	: Genellenebilirlik Çalışması
K Çalışması	: Karar Çalışması
G Katsayısı (Ep2)	: Genellenebilirlik katsayısı
Phi Katsayısı	: Güvenirlik katsayısı
ÇDKRM	: Çok Değişken Kaynaklı Rasch Modeli
b	: birey
p	: puanlayıcı
o	: okul
bxp	: birey x puanlayıcı ortak etkisi
dzy	:düzey
evr	:evren
snsz	:sonsuz

BÖLÜM I

GİRİŞ

Bu bölümde, problem durumu, araştırmanın amacı, araştırmanın önemi, varsayımlar, sınırlılıklar ve tanımlar başlıklar altında açıklamalara yer verilmiştir.

1.1. Problem Durumu

Eğitimde, bilgi ve teknolojinin gelişimiyle öğrendiği bilgileri gerçek yaşamına uyarlayabilecek ve bu bilgilere yenisini ekleyerek kullanabilecek bireylerin yetiştirilmesi ana hedef haline gelmiştir. Bu bireylerin bilgi ve becerilerini ne derece kazanıldığını niteliklerine göre seçilmesinde ve öğrenci seviyelerini belirlemede uygun ölçme ve ölçme süreci önemli bir yere sahiptir.

Tarihimize baktığımızda da ölçme ve değerlendirmede bilimsel çalışmaları 19. yüzyılın başlarından itibaren önemli bir durum olarak görülüp bir tanım haline gelmiş gibi görölse de eski Türk devletlerinden itibaren bireyin bir işe alınması gibi amaçlarla seçme yapıldığı bilinmektedir. Ölçme ve değerlendirme Selçuklu ve Osmanlı da, özellikle Fatih döneminde medreselere öğretmen alımında kullanıldığı, yine sonrasında devlet hizmeti olarak görev yapacak kişilere yapılan seçme sınavlarında uygulandığı bilinmektedir (Turgut ve Baykul, 2010). Özellikle cumhuriyet döneminden itibaren bilimsel olarak ölçme ve değerlendirme alanı hız kazanmıştır. Ülkemizde ilk çalışma olarak bilinen İbrahim Alaettin Gövsa'nın Binet-Simon zeka testini Türkçeye çevirmesiyle ölçme ve değerlendirme ile ilgili araştırmaların başladığı görölmektedir (Öner, 1994). Cumhuriyet döneminden itibaren günümüze kadar bakıldığında ölçme ve değerlendirmeye ilgili akademik çalışmalar artarak devam etmiştir. Günümüzde ölçme ve değerlendirmeye hemen hemen her alanda ihtiyaç duyulmaktadır. Özellikle ülkemizde öğrenci seçme ve yerleştirmede tercih edilen yöntem ve teknikler her ne kadar tartışma konusu olsa da ölçme ve değerlendirme, bireyi doğru yere yönlendirmek adına vazgeçilmez bir unsurdur.

Ölçme kavramı, amaca göre belirli bir özelliğin gözlemlenip verilerin toplanarak sonuçlarının nicel ve sembolik olarak ifade edilme işlemiyken; değerlendirme ise bu sonuçlar doğrultusunda belli bir ölçüte göre karar verme aşaması olarak tanımlanmaktadır (Baykul,

2000). Ölçme ayrıca bireylerdeki devam eden değişimi gözlemlemek bireyler arası farklılığı da belirtmek için yapılması gereken bir durumdur. Ölçme yapılırken ölçmenin özelliğinin belirtilmesi, ölçme sonuçlarının belirlemede sayı ve sembollerin belirlenmesi ve bu sayı ve sembollerin derece veya miktar olarak tanımlanması gerekmektedir (Doğan, 2020). Bu bakımdan doğru sonuca ulaşmak adına farklı ölçülmesi gereken özellikler uygun ölçme aracıyla belirlenmesinin yanında ölçmeyi etkileyen bazı değişkenlerinin olduğunu da göz ardı etmemek gerekir. Bu değişkenlerden biri de puanlayıcılarıdır. Birden fazla puanlayıcının yer aldığı ölçme durumlarında aynı ölçme durumuna karşı puanlamaların yakın veya uzak olması puanlayıcılar arası güvenirlikle ilgili bir durumdur. Bu durumda bahsedilmesi gereken hata kaynağı puanlayıcılarıdır. Ayrıca puanlayıcı seçimi ve seçilen puanlayıcının uzman olması, puanlayıcı hatalarının ya da puanlama yanlışlığı adına etkili olması söz konusudur (Arsan, 2012). Ölçme sürecindeki puanlayıcıların güvenirlik çalışmalarını analiz ederken birkaç hata kaynağının birlikte değerlendirilmesi gerekmektedir. Çünkü birden fazla değişkenliğin bir arada değerlendirilmesi daha güvenilir ve kapsamlı bir değerlendirmeyi sağlamaktadır (Atılğan, 2004).

Ortaöğretim İngilizce öğretimi okulların genel özelliğine bağlı olarak haftalık ders süresi farklılık göstermektedir. Ancak yapılan dönem içi başarı testlerinin biçimi birbirine benzerlik göstermektedir. Yabancı dil veya ikinci bir dilin öğreniminde teorik bilginin yanında uygulama yapmanın da önemli olduğu açıktır. Bu öğrenimde teorik veya uygulama bölümünü değerlendirmek adına performans testleri önem arz etmektedir. Değerlendirme sürecinin değerlendirilen adayın ya da bireyin bilgisi olması ve belirli bir yönergeler dahilinde değerlendirmelerin yapılması puanlayıcının işini kolaylaştırmaktadır. Performans testlerinin geçerli olması büyük önem taşırken bu testlerinin değerlendirilmesi sonrası sonuçların da güvenilir olması da istenilen bir durumdur (Yavuz-Kırık, 2008) .

Güvenirlik, ölçme neticelerinin tesadüfi hatadan soyutlanması olarak dile getirilmektedir (Turgut ve Baykul, 2010). Ölçme ve değerlendirmede güvenirliği etkileyen değişkenlerden bir puanlayıcılarıdır. Puanlayıcı sayısı birden fazla olma durumunda bir özellik için verilen puanların uyumluluğu puanlayıcı güvenirliği olarak nitelenmektedir (Nalbantoğlu, 2009). Ölçme aşamasında önemli bir yer edinen puanlayıcıların serbest puanlama dediğimiz alanında uzman olan puanlayıcıların herhangi bir ölçüde bağlı olmadan kendince yapmış olduğu puanlama stiline göre puanlama yapması hata payının da artmasına sebep olmaktadır (Damlar Demirci, 2019). Ayrıca tek bir değişkenlik üzerinden değerlendirilmesi birçok sınırlılığın yanında hata kaynağını birlikte değerlendirilmemesi güvenirlik anlamında bazı

sorunlara yol açmaktadır. Bu hata kaynaklarının birlikte değerlendirilmesi ve güvenilirliğin incelenmesi adına klasik yaklaşımların yerini bu durumları birlikte değerlendiren G Kuramı ve Çok Değişkenlik Kaynaklı Rasch Modeli almaktadır (Atılğan, 2004).

G kuramı, klasik kuramın tek seferde birden fazla hata kaynağını ele alamayışı gibi sınırlılıklarını yok ederek daha geniş kapsamlı bir güvenilirlik sonucu elde etmemize yardımcı olmaktadır. Özellikle puanlayıcılar arası uyum ve güvenilirliği incelemede genellenebilirlik ve rasch modeli son yıllarda kullanım açısından da kolaylık içeren uygulamalarla elde edilebilen sonuçlar vermesi nedeniyle kullanılmaya başlanmıştır (Atılğan, 2004).

Bu araştırmada gerçek ölçme durumlarından elde edilen verilerin ölçekli ve ölçeksiz olarak birden fazla puanlayıcılar tarafından puanlanmasıyla elde edilen sonuçların G kuramı ve ÇDKRM'yle detaylı olarak incelenmesi amaçlanmıştır. Bu doğrultuda aşağıda yer alan problem cümlesi ve alt problem cümlelerine yanıt aranmaktadır.

Bu çalışmada ölçme sürecinde farklı ortaöğretim kurumlarında yer alan 12. sınıf İngilizce dersi için elde edilen performans değerlendirme sonuçlarının 3 farklı puanlayıcı tarafından ölçekli ve ölçeksiz olarak puanlaması durumunda güvenilirliğin Genellenebilirlik Kuramı ve Çok Değişkenlik Kaynaklı Rasch Modeli ile elde edilen veriler analiz edilmiştir.

Belirlenen amaçlar doğrultusunda, bu çalışmada şu sorulara yanıt aranmaktadır:

- Ortaöğretim İngilizce yeterliliğinin okullara ait betimsel istatistikleri nasıldır?
- Puanlama anahtarı verilerek yapılan değerlendirmede durum nedir?
- Puanlama anahtarı verilmeden yapılan değerlendirmelerin uyum veya güvenilirlik ilişkisi nasıldır?
- Puanlama anahtarı kullanılarak ve puanlama anahtarı kullanılmadan yapılan değerlendirmelerin arasındaki ilişki nasıldır?
- G Kuramının tek değişkenli modeli ile gerçek durum için hesaplanan G ve Phi Katsayıları, senaryolarla kestirilen G ve Phi katsayılarıyla tutarlı mıdır?
- Ortaöğretim İngilizce yeterliliğinin performans sonuçlarının Çok Değişken Kaynaklı Rasch Modeli analizleri nasıldır?

1.3. Araştırmanın Amacı

Araştırmanın genel amacı, birden fazla çok değişkenlik kaynağını beraber değerlendiren G Kuramı ve Çok Değişken Kaynaklı Rasch Modeli ile gerçek veriler kullanılarak kestirilen parametreler incelenerek, iki yaklaşım üzerinden puanlayıcıların güvenilirliğinin incelenmesidir.

Buna göre araştırmada G kuramı kapsamında G ve K çalışmalarıyla ölçekli ve ölçeksiz olarak farklı okullardan yer alan bireylerin performansları puanlayıcıların puanlaması istenerek puanlayıcıların güvenilirliğini kuramlar üzerinden incelemesi yapılmıştır. Aynı veriler üzerinden G Kuramı ve Çok Değişkenlik Rasch Modelinden elde edilen verilerin karşılaştırması sonrasında bu iki yaklaşımdan elde edilen bilgilerin analizleri yapılarak elde edilen veriler incelenmektedir.

Bu amaçla; G kuramıyla, 12. Sınıf İngilizce yeterlilik performansının farklı değişkenler üzerinden puanlayan puanlayıcıların sonuçlarının G Kuramı tek değişkenli modelleri ile kestirilen Genellenebilirlik ve Karar çalışmaları parametrelerinin karşılaştırılarak ve modellerin birbiriyle tutarlılıkları incelenmek istenmiştir. Bununla beraber tek değişkenli model ile alternatif K çalışmalarıyla elde edilen farklı puanlayıcıların ve gerçek uygulamalar sonucunda kestirilen puanlayıcılardan elde edilen G ve Phi katsayılarını incelenmesi ve yapılan performans değerlendirme sonucunda puanlayıcılara verilen ölçek sonucunda ortaya çıkan verilerin G kuramı ve Çok Değişkenli Kaynaklı Rasch Modeli tutarlılıklarını belirlenmesi yine amaçlar arasındadır.

Ayrıca yapılan performans değerlendirme sonucunda puanlayıcıların ölçeksiz puanladıkları verilerden elde edilen sonuçların G Kuramı ve Çok Değişkenli Kaynaklı Rasch Modeli üzerinden karşılaştırılması ve tutarlılıkların belirlenmesiyle beraber elde edilen parametrelerin ölçekli puanlama ve ölçeksiz puanlama sonrasında ortaya çıkan sonuçların G Kuramı ve Çok Değişkenli Rasch Modeliyle analiz edilmesi diğer amaçlardan biridir.

1.4. Araştırmanın Önemi

Son dönemlerde ölçme değerlendirmenin hız kazanması ve ölçme yöntemlerinin çeşitliliği göstermesi ile beraber ölçülecek özelliği en hassas şekilde ölçmek adına geçerlik ve güvenilirlik de önemli boyuta ulaşmıştır. Geçerlik ve güvenilirliği yüksek tutmak adına birden fazla hatayı tek bir seferde ele alan “G Kuramı” ve MTK’ nın bir kolu olan “Çok Değişkenlik Kaynaklı Rasch Modeli” nin kullanımı uygun görülmüş ve sonuçları sonraki çalışmalara

yardımcısı olması açısından irdelenmek istenmiştir. Bunun yanında birden fazla puanlayıcı ve değişken ile yapılan ölçme durumlarının birbirine ne derece tutarlı sonuçlar vereceği hakkında bilgi edinilmek istenmiş olup bu puanlayıcıların ölçme durumlarında kullandıkları puanlama anahtarının da sağlamış olduğu güvenilirlik durumu incelenmiştir. Bu araştırmayla Genellenebilirlik Kuramı ve Çok Değişken Kaynaklı Rasch Modeli yaklaşımlarının gerçek veriler kullanılmış olup birbirleriyle ve kendi içlerinde tutarlılıkları test edilerek çeşitli okullardaki öğrencilerden elde edilen puanların farklı puanlayıcıların puanlama yönergesiyle puanlama ve ölçeksiz puanlama ile değerlendirilmesi sonucu elde edilen veriler ele alınmıştır. Böylelikle farklı nitelikteki okul ve puanlayıcı değişkenlerinin bu kuramlar çerçevesinde gerçek hayat şartlarında benzer verilerle test edilerek uygulanabilirliği, olumlu ve olumsuz yanları belirlenerek literatürün gelişimine ve tartışmalara katkı sağlaması beklenmektedir. Bununla birlikte yapılan bu uygulamanın güvenilirliğinin bu iki yaklaşım açısından ele alınarak puanlayıcılara verilen ölçeğin ne kadar fayda sağladığını görmek ve puanlayıcıların puanlamalarının güvenilirliklerini incelemek adına büyük önem taşımaktadır. Birçok değişkenlik kaynağı bulunan ölçme durumları için tek bir analizle kapsamlı bir güvenilirliğin hesaplanmasına yönelik çalışmalardan biri olup sonraki araştırmalara öncülük etmesi, yaklaşımların geliştirilebilmesi ve kullanımının artırılması adına diğer tartışmalara katkı sağlayabilir.

1.5. Sayıtlar

Araştırmada kullanılan Antalya ili Elmalı ve Kumluca ilçelerindeki Fen Lisesi, Anadolu Lisesi, Çok Programlı Lise ve kolej olmak üzere özel ve devlet ortaöğretim kurumlarının 12. sınıf öğrencilerinin İngilizce yeterliliği verilerini, puanlama işlemini yapan alanında uzman 3 puanlayıcı tarafından puanlamaları birbirinden bağımsız olarak yapmış oldukları varsayılmıştır. Yapılan performans testini yanıtlamaya katılan öğrencilerin soruyu cevaplandırarak hazırbulunuşlukta olduğu ve yalnızca kendi bilgi birikimiyle cevapladıkları varsayılmıştır.

1.6. Sınırlılıklar

Araştırma 2022-2023 yılları arasında yalnızca Antalya ili Elmalı ve Kumluca ilçesinden veri toplanarak sadece ortaöğretim 12. sınıf öğrencilerinin İngilizce yeterlilik performansları 3 puanlayıcı, 4 okul ve her okulda 31 birey için toplamda 124 öğrenci ile sınırlıdır.

Arařtırmaya konu olan puanlayıcıların gvenirlięinin incelenmesi yalnızca İngilizce yeterlięi ile sınırlı kalmıřtır. Zaman ve maliyet aısından yalnızca yazılı anlatım becerisi zerinden istatistikler elde edilmiřtir. Arařtırmaya sadece 12. sınıf ęrencilerinin verileri analize dahil edilmiřtir.



BÖLÜM II

KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

2.1. Kuramsal Çerçeve

Toplumların gelişiminde etkili olan bilimin birçok tanımı bulunmaktadır. Bu tanımlar bilimin kuramsal ve deneysel bir tarafının olduğu konusunda mutabık kalmışlardır. Bu kuramların ve deneylerin de doğruluğunu kanıtlamak için bazı testlerden geçmesi gerekir. Bunu da en iyi ölçme dediğimiz kavram karşılamaktadır. Ölçme, yapılan bu araştırmaların ileriki araştırmalara da kaynaklık etmesi açısından önemli bir yere sahip olmasının yanında ayrı bir bilim dalı olarak da bilimin bir parçasıdır (Atılğan, 2004).

Ölçme işleminde bilime katkı bulunmak adına geçerli ve güvenilir bir sonuç elde etmek doğrudan bir amaç olarak karşımıza çıkmaktadır. Bununla birlikte ölçmenin de farklı durumlarda değerlendirme yapılmasına olanak sağlayan kuramlar ortaya çıkmıştır. Bu kuramlar Klasik Test Kuramı (KTK), Madde Tepki Kuramı (MTK) ve Genellenebilirlik Kuramı (G)'dir. Bu üç kuramda da ölçme sonuçlarının hatasız olması amaçlanmaktadır. Ölçülen sonuçların gerçek sonuca olabildiğince yakın yani hatasız olması güvenilirlikle ilgili bir durumdur. Güvenirlik hata kaynaklarına göre çeşitlilik göstermektedir. Bunlardan biri de puanlayıcılar arası güvenirliliktir. Puanlayıcılar arası güvenirlilik birden fazla puanlayanın aynı özelliğe verdikleri puanlamanın tutarlılığı olarak kabul edilmektedir (Atılğan, 2004). Burada hata kaynağı puanlayıcılar olarak kabul edilmektedir.

Birden fazla puanlayıcı hatası bulunmaktadır. Ancak en fazla rastlanan puanlayıcı hataları Saal, Downey ve Lahey, (1980) tarafından belirtilen cömertlik/katılık, merkezi eğilim, ranj daralması ve halo etkisidir. Herhangi bir testte yapılan değerlendirmeler sonucunun düşük veya yüksek olması katılık/cömertlik durumunun var olduğuyula açıklanabilmektedir. Yani bireyin gerçek sonucundan aşağıda olması puanlayıcının katı değerlendirme yaptığını tam tersi gerçek sonuçtan yukarıda puanlama sonucu da puanlayıcının cömert puanlama yaptığını gösterebilmektedir. Cömertlik/katılık, hata durumunu puanlamanın çarpıklık yönüne göre bilmek mümkündür. Sağa çarpıklık, puanlayıcının cömertliğini gösterirken sola çarpıklık puanlayıcının katılığını göstermektedir (Arsan, 2012). Ranj daralması dediğimiz hata türü ise puanlamayı daha az ayırt etmeye sebep olan puanların bir bölüme yığılması olarak açıklanmaktadır (Wilson ve Case, 2000) Yığılma tam orta noktada oluşuyorsa merkezi eğilim olarak ifade edilmekte orta noktadan yüksek veya alçak bölgede yığılması yine ranj

daralmasının bir nedeni olarak görüldüğünü göstermektedir. Ranj daralmasını standart sapmanın sonucuna göre tespit etmek mümkündür. Standart sapmanın küçük olması ranj daralmasının olduğunu göstermektedir (Saal, Downy ve Lahey, 1980). Halo etkisi puanlayıcının puanlama yaptığı bireye ait özelliklerden etkilenmesi olarak yorumlanabilir. Örneğin; bireyin yazısının güzel olması ve puanlayıcının yazıdan etkilenerek fazla puan vermesi olumlu halo etkisi olarak açıklanmaktadır. Bunların dışında puanlayıcıdan kaynaklanan birçok hata bulunmaktayken ölçme aracı, ölçme aracındaki madde sayısı, ölçme ortamı ve zamanı, ölçme yöntemi de hataya sebep olabilir (Atılğan, 2004).

Ölçme işleminde güvenilir sonuçlarla doğru yorumlar yapabilmek adına hata kaynaklarının minimuma indirmek gerekmektedir. Bu sebeple kullandığımız KTK ve MTK modeller tek bir değişkenlik kaynağından gelen hatayı dikkate alırken Genellenabilirlik Kuramında birden çok hata kaynağını tek bir analizle ele alabilmektedir. Bu durumun önemi özellikle birden fazla puanlayıcının bir ölçmeyi puanlama esnasında ne kadar dikkate alınması gerektiğini ortaya çıkarmaktadır.

Puanlayıcılar arası uyum ve güven araştırmacılar ve uygulayıcılar tarafından çoğu zaman aynı anlam taşımaktadır (Goodwin, 2001). Bu uyum ve güveni belirlemede kullanılan birden fazla yöntem ve teknik bulunmaktadır (Atılğan, 2001). Bu yöntemlerden en yenileri olarak kabul edeceğimiz hem puanlayıcılardan gelen hataları hem de puanlayıcıları değerlendiren G kuramı ve Çok Değişkenlik Kaynaklı Rasch Modeli puanlayıcılar arası güvenilirliği hesaplamaktadır (Atılğan, 2004).

2.1.1. İngilizce Öğretiminde Mevcut Değerlendirme Sistemi

Birçok bilim dalında olduğu gibi eğitimde de bireylerin zeka, yetenek ve öğrenme gibi birbirinden farklılık gösteren özelliklerin incelenmesi veya ölçülmesinde ölçme araçları kullanılması gerekmektedir (Baykul, 2000). Bir anlamda eğitimde girdi ve süreçlerde ölçme ve değerlendirme olmazsa olmazdır. Kullanılan değerlendirme sistemi, ülkeye ve eğitim seviyesine göre değişebilir.

Dil öğrenme ve öğretme sürecinde öğretimin ne kadarının gerçekleştiğini anlamak adına ölçme ve değerlendirmenin yapılması gerekli görülmektedir. Bu nedenle yapılacak ölçme ve değerlendirme aracının sözcük, anlama ve dil becerileri bakımından birbirini tamamlayacak özellikte olmalıdır. Kısaca kullanılan ölçmenin yalnızca bilişsel alanda değil duyuşsal ve

psikomotor alan bakımından da değerlendirmeye esas alınması öğretimin niteliğini ortaya çıkarma bakımından gerekli olduğu düşünülmektedir (Gürses, 2006).

Ülkemizde genel olarak İngilizce öğretimi ikinci bir dil olarak kullanmayı amaçlayan bir sistem üzerine kuruludur. Bu sebeple ölçme ve değerlendirme boyutu İngilizce öğretimi yöntem, içerik ve uygulamalara göre şekil almaktadır (Yavuz-Kırık, 2008). İngilizce öğretiminde kullanılan değerlendirme yöntemleri genelde çoktan seçmeli testlerdir (Karlı, 2003). Bu testlerde amaç öğrencinin programına ve seviyesine uygun yeterlilikte olduğunu görmek bu doğrultuda öğrenci başarısı ve gelişimini sağlamaktır. Üniversite aşamasında, hazırlık sınıfındaki öğrencilere her üniversitenin kendine ait dönem sonunda dört beceriden oluşan okuma, yazma, dinleme ve konuşma becerilerini ölçmek için sınavlar kullanılır. Bu sınavlar da genellikle çoktan seçmeli, karşılaştırma, sıraya koyma, boşluk doldurma veya doğru-yanlış sorulardan oluşur (Cesur, 2008). Bunun yanında eğitim düzeyine göre öğrencilerin İngilizce becerilerini uygulamaları ve geliştirmeleri için proje ve performans ödevleri verilir. Bu ödevler, öğrencilerin araştırma yapmalarını, sunum hazırlamalarını ve yazma becerilerini geliştirmelerini sağlar. Konuşma becerilerini ölçmek için öğrencilere konuşma sınavları yapılır. Bu sınavlar, öğrencilerin İngilizce dilinde rahatça iletişim kurabilme yeteneklerini ölçer.

Öğrencilerin İngilizce becerilerinin gelişimini izlemek için öğretmenler, öğrenci performanslarını gözlemleyebilirler. Bu gözlemler, öğrencilerin öğrendikleri İngilizce dil bilgisini kullanma, okuma, yazma ve konuşma becerilerini gösterir. Ayrıca öğrencilerin İngilizce becerilerini değerlendirmek için portfolyo kullanılabilir. Portfolyo, öğrencinin İngilizce becerilerinin gelişimini gösteren bir koleksiyondur. Bu koleksiyonda öğrencinin yazıları, proje ödevleri, konuşma sınavları gibi çalışmalar yer alır.

Yabancı dil öğretiminde öğrencilerin ihtiyaçlarına ve dil becerilerine göre geliştirilen ölçme araçlarının başarıyı yordamak adına başlangıç için önemli bir adım olarak görülmektedir. İngilizce eğitiminde kullanılan değerlendirme yöntemleri, öğrencilerin İngilizce becerilerinin gelişimini takip etmelerine, zayıf yönlerini belirlemelerine ve bu zayıf yönleri geliştirmelerine yardımcı olur. Genel olarak yapılan dil yeterlilik sınavlarında dil becerilerini tümüyle kapsam içine alan belirli kazanımları ölçmek zorunululuğu olmalıdır. Bu dil becerilerini belirlemek için yeterliliğin değerlendirilmesi adına Swain(1990) tarafından belirlenen aşağıdaki tablo ortaya çıkmıştır (Brown, 2004:83).

Tablo:2.1. Yabancı Dil Yeterliliği İçin Ölçme Değerlendirme Biçim ve İçerikleri (Brown, 2004:83)

<i>Dilbilimsel Özellikler</i>	<i>Dilbilgisi</i>	<i>Söylem</i>	<i>Sosyo-Dilbilim</i>
Yöntem	(Odakta) Dilbilgisel Doğruluk	(Odakta) Metinsel Bağdaşıklık ve Bağlılık	(Odakta)Dil kullanımında sosyal uygunluk
Sözlü	(Yönlendirmeli) Sözlü	Öykü Anlatımı; Tartışma/İkna	Konuşmada rol alışı: Talep, şikayet öneri ve teklifler
Yazılı	Sözdizimi, Önek, Sözcük-biçim bilgisi doğruluğunun ölçümü	Saptama, Mantıklı Sıralama ve Bağdaşıklığın ölçümü	Kurallı ve günlük dil kullanımının ayırımını yapma becerisini ölçme
Çoktan Seçmeli	Tümce düzeyinde ‘Doğru yapıyı seç’ alıştırmaları, Fiil Sözdizimi, Önek ve diğer maddeler	Paragraf düzeyinde ‘Uygun Tümceyi Seç’ alıştırmaları	Formalkurallı-ve günlük dil kullanımının ayırımını yapma becerisini ölçme
Yazılı Kompozisyon	Öykü /Anlatı ve İkna Mektupları	Öykü/Anlatı ve İkna Mektupları	Resmi talep istek mektupları günlük dil kullanımı

Ülkemizde İngilizce öğretimi MEB’in 2004 yılında güncellediği ve sonraki eğitim öğretim yılından itibaren uygulanmaya başlayan yapılandırıcı eğitim sistemi programına doğrultusunda belirlenerek oluşturulmuştur. Ölçme ve değerlendirme öğrenme amaçları doğrultusunda materyallere ve görevlere hizmet etmektedir. Değerlendirme olumsuz pekiştirme aracı olmaktan ziyade, ingilizce öğrenimini geliştirmek ve bunun yanında kontrol etmek amaçlı kullanılmaktadır. Programa göre ölçme ve değerlendirme dilin gerçek kullanımını sağlamak adına geleneksel değerlendirme durumlarının yerine portfolyo, proje ve iletişim gerektiren etkinlikleri ön planda tutmaktadır. Bunun yanında öz değerlendirmeye bireylerin kendini değerlendirme uygulamasının da önemli olduğuna dikkat çekilmiştir (MEB, 2018).

2.1.2. Tamamlayıcı Ölçme ve Değerlendirme Teknikleri

Alanyazında alternatif ölçme ve değerlendirme teknikleri olarak da adlandırılan tamamlayıcı ölçme ve tamamlayıcı değerlendirme, eğitim sürecinin öğrenci merkezli anlayışa yaklaşmasıyla birlikte değerlendirmenin de öğrenci merkezli değerlendirme anlamını taşımaktadır. Bu ölçme değerlendirme tekniklerini tamamlayıcı olarak tanımlamanın amacı klasik ölçme değerlendirme tekniklerinin eksiklerini tamamlayıcı olmasından kaynaklıdır (Doğan, 2020). Performans değerlendirme, portfolyo değerlendirme, yapılandırılmış grid, öz

ve grup değerlendirme gibi değerlendirmeler tamamlayıcı ölçme değerlendirme tekniklerine örnektir.

Geleneksel ölçme ve değerlendirme tekniklerinden farklı olarak tamamlayıcı ölçme ve değerlendirme tekniklerinde amaç öğrencinin bilgi düzeyini direk sorgulamaktan ziyade öğrenme süreci boyunca öğrencilerin gözlemlenerek süreç değerlendirilmesi ve sonuçta meydana gelen ürünü de değerlendirmektir. Bunun yanında öğrenci kendini ve akranlarını da değerlendirmektedir. Öğrenciler bu süreçte öğrendiklerini nasıl öğrendiğinin de farkına varmaktadır.

Tamamlayıcı ölçme ve değerlendirme tekniklerinin geleneksel değerlendirme tekniklerine göre sınırlılıkları da vardır. Geleneksel ölçme ve değerlendirme tekniklerine göre ölçmede hata oranı daha fazladır. Puanlayıcıların birden fazla olmaması da güvenilirliği düşürecektir (Doğan, 2020).

2.1.2.1. Performansa Dayalı Durum Belirleme

Büyüköztürk (2007)'e göre performans bir öğrenme durumuna karşı gösterilen gayret ve bu gayretin sonucunda ortaya çıkan bir ürünken performans değerlendirme ise bireyin ortaya koymuş olduğu ürünün yeterliliğini veya noksanlığını belirlenmesini sağlayan bir değerlendirme biçimi olarak ortaya çıkmaktadır. Performans kavramı eğitim alanında genel olarak öğrencilerin kazanmış veya öğrenmiş olduğu bilgi ve becerilerden yola çıkarak ortaya çıkan ürün adına göstermiş olduğu çaba olarak da tanımlanmaktadır (Kutlu, Doğan ve Karakaya, 2017). TDK' ya göre performans, başarımlar olarak Türkçeleştirilmiştir.

Eğitimin olmazsa olmazlarından olan öğrencilerin ders veya alan kapsamındaki yeterliliklerini ve programdaki kazanımların ne kadarını gerçekleştirebileceklerini ölçmek adına performans değerlendirme önemli bir yere sahiptir. Özellikle üst düzey zihinsel beceri gerektiren karmaşık öğrenme hedeflerini yerine getirmede geleneksel ölçme teknikleri yetersiz kalması sebebiyle performansa dayalı durum değerlendirmenin kullanımı ön plana çıkmaktadır (Doğan, 2020). Bununla birlikte yalnızca soruları yanıtlamadan ziyade gelecek yaşama da fayda sağlayacak bir yaşantı sunar. Bu soru yazılı ve sözlü olarak açık uçlu veya bir deney sorusundan oluşabilmektedir. Bu performansların da güvenilir ve amaca uygun değerlendirme yapmak adına önceden belirlenmiş olan ölçütlere göre amaca uygun ölçme araçları kullanılmalıdır.

Eğitimde birçok ölçme araçları bulunmaktadır. Bunların arasından en çok tercih edilen ölçme araçları çoktan seçmeli testler ile beraber performans ölçümleridir (Alharby, 2006). Özellikle çok sayıda bireyi değerlendirmede çoktan seçmeli testler uygulanabilirliği ve değerlendirilmesi açısından en pratik ölçme araçlarından biridir. Ancak ölçme durumunda şans başarısının yüksek olması, gereken kazanımların kazanılmasını tam olarak ölçmemesi ve Bloom taksonomisinin üst becerilerinin ölçülmesinin kısıtlı olması çoktan seçmeli testlerin sınırlılıklarındandır. Özellikle görsel, fiziksel ve dil becerileri gibi performans gerektiren durumlarda çoktan seçmeli testleri uygulamak uygun olmamaktadır (Güler, 2008). Bununla beraber süreçten ziyade sonuca yönelik değerlendirmelerde bulunan klasik testler bireyin durumu hakkında daha kısıtlı sonuçlar vermektedir (Kutlu, Doğan ve Karakaya, 2017). Bu durumda çoktan seçmeli testlerin yerini performans değerlendirme almaktadır. Performans değerlendirmede tek bir doğru olmamasının yanında hedefi temsil edebilen ürün başarılı sayılabilmektedir. Performansa dayalı durum değerlendirmede birçok sorudan oluşan çoktan seçmeli sorular yerine tek bir soruyla birden fazla üst düzey zihinsel beceriyi ortaya çıkaracak soru sorulabilir.

Performans ölçmenin en büyük avantajı davranışları somut olarak ölçmesi ve üst düzey becerileri sınamasıdır (Güler, 2008). Performans ölçme aynı zamanda öğrencilerden yaratıcı davranışlar ortaya koymasını beklemektedir. Bu sebeple performans dayalı durum değerlendirme daha geçerli ölçme verilerini vermektedir. (Alharby, 2006). Örneğin bir bireyin bir müzik aletini çalabilme becerisi üzerinden alanı uzman olan puanlayıcılar tarafından değerlendirilmesi performans değerlendirmeye örnektir. Performans değerlendirme hem bilişsel hem de psikomotor alanda da kullanılabilir. Performans değerlendirme sonuçtan ziyade sürece odaklıdır.

Nitelikli bir performansa dayalı durum belirlemede bazı dikkat edilmesi gereken hususlar bulunmaktadır (Kutlu, Doğan ve Karakaya, 2017). Bu hususlar aşağıda verilmiştir:

- Performans görevleri, öğrencileri üst düzey zihinsel yapıları ortaya koyabilecek nitelikte olmalıdır.
- Performans görevi anlaşılır, ilgi çekici ve yazıldığı sınıf düzeyine uygun olmalıdır.
- Performans görevinin yönergesi açık, net ve öğrenciye çalışma hakkında rehberlik edecek nitelikte olmalıdır.

Performansa dayalı durum değerlendirmenin bazı faydaları ve sınırlılıkları bulunmaktadır. Performansa dayalı durum değerlendirmenin en önemli yararlarından biri farklı

yöntemlerle ölçülemeyen özellikleri ölçmede daha kullanışlı olmasıdır. Bununla beraber ölçülen özelliğin sonucunun yanı sıra ölçme sürecinde bireyin gelişim süreci ve performans sürecini de analiz edebilmektedir. Ayrıca durum değerlendirme öğrenciye veya bireye ait performansın gerçek yaşama da yansıtabilmesini sağlar. Diğer yandan performans durum değerlendirmenin bazı sınırlılıkları da mevcuttur. Bunlardan en önemlisi de puanlamadaki güvenilirlik sorunudur. Her ne kadar analitik rubrikler kullanılsa da istenilen anlamda güvenilirlik sağlanamayabilir. Ayrıca uzun süreli görevlerde başkalarından yardım alma durumu da güvenilirlik anlamında bazı sıkıntılara yol açacaktır. Bununla birlikte performansa dayalı değerlendirmenin uygulanış bakımından zaman alıcı olması, planlama, puanlamanın zaman alıcı olması ve etkili bir şekilde kullanılmamasından kaynaklı olumsuz durumlara sahiptir (Kutlu, Doğan ve Karakaya, 2017).

2.1.2.2. Performansa Dayalı Durum Belirlemede Puanlama Yöntemleri

Performansa dayalı durum belirlemede puanlayıcıların kullandıkları yöntem, geçerliği ve güvenilirliği belirlemede önemli rol oynamaktadır. Kullanılan yöntemlerden bazıları dereceli puanlama anahtarı olan bütünsel ve analitik olarak ikiye ayrılmaktadır (Mertler, 2001).

Standart bir puanlama olmayan genel izlenim puanlama durumunda puanlayıcı etkisi oldukça fazla ve bireyden bireye puan yanlılığı değişkenlik gösterecektir. Bu durumda öğrenci de performansına ait değerlendirmenin noksan olan durumları hakkında bilgi sahibi olması mümkün olmaz. Ölçmenin puanlama güvenilirliğinin de düşmesine neden olur (Damlar-Demirci, 2019). Ancak testin ilgi alanına göre uzmanlar tarafından hazırlanan puanlama anahtarı puanlama yanlılığını azaltacağı için güvenilirliği yüksek değere ulaşmasını sağlayacaktır. Bu puanlama yöntemleri farklı puanlayıcıların puanlaması durumunda puanlama uyumluluğunu artıracığı gibi değerlendirmenin de amacına bağlı yürütülmesini sağlamaktadır (Atmaz, 2009). Ayrıca hata oranını en aza indirmek için değerlendirmenin uygun bir ölçekle yapılması büyük önem arz etmektedir. Puanlama anahtarlarının mevcut göreve göre farklı kullanım alanları bulunmaktadır (Atılgan, Kan ve Doğan, 2017).

2.1.2.3. Dereceli Puanlama Anahtarı (Rubrik)

Dereceli puanlama anahtarı, değerlendirilen bireylerin performansların belirli bir ölçüte dayalı olarak analizini yapabilmek için uzmanlar ya da puanlayıcılar tarafından oluşturulan puanlama sistemidir. Dereceli puanlama anahtarı puanlayıcıların değerlendirmeyi daha

güvenilir ve yansız yapabilmesi için puanlama öncesinde belirli bir kategori oluşturularak puanlayıcılar adına yol gösterici niteliğindedir (Güler, 2008). Dereceli puanlama anahtarıyla önceden belirlenmiş ölçütler sayesinde farklı puanlayıcıdan puanlayıcıya değişmeden nesnel ve standart değerlendirme yaparak bireyin performansı veya başarısı hakkında daha net ve güvenilir sonuç vermesini sağlamaktadır (Atılğan, Kan ve Doğan, 2017).

Dereceli puanlama anahtarlarının kullanımı konu veya sınıf düzeyinden ziyade yapılacak olan ölçmenin amacına bağlı olarak belirlenmelidir (Moskal, 2000). Beceri temelli değerlendirmelerde rubriklerin kullanılması hem performans süreci hem de sonraki değerlendirme aşamalarında ölçülecek bireylere rehberlik edebilmektedir. Böylece birey almış olduğu sonuç doğrultusunda yapmış olduğu hatalar ve eksiklikler hakkında bilgi sahibi olabilmektedir. Dereceli puanlama anahtarları bütünsel (holistik) ve analitik olmak üzere iki farklı yapısal özellikten oluşmaktadır. (Mertler, 2001).

2.1.2.4. Dereceli Puanlama Anahtarı Türleri

2.1.2.4.1. Bütünsel (Holistik, Bütüncül) Dereceli Puanlama Anahtarı

Bütünsel dereceli puanlama anahtarları 20. yüzyılın ortalarından itibaren ortaya çıkmış olup bireylerin göstermiş olduğu performansların unsurlarını sorgulamadan tek bir bütün olarak puanlama yöntemi olarak tanımlanmıştır (Nitko, 2001). Bu puanlama türü genel izlenimle puanlama olarak da tanımlanmaktadır (Mertler, 2001).

Bütünsel puanlama anahtarı puanlama yapılacak olan performansın doğru yanıtı bakımından tek bir doğru olmadığında ya da öğrencinin ortaya çıkarmış olan ürünün özgünlüğü ön planda olduğu durumlarda kullanılmasının daha uygun olacağı ifade edilmektedir. Bu durumda puanlayıcı puanlama yapacağı öğrenciye ait cevabın veya çözüm yolunun bütün olarak değerlendirmesini esas almalıdır. Ancak puanlama öğrenciye ait becerinin niteliğini net olarak gösterememektedir (Ayverdi 2023).

Bütünsel puanlama türünde puanlama daha yüzeysel yapılmakta ve bazı hatalar göz ardı edilmesinin yanında zaman bakımından hızlı ve çabuk olacağından kullanılması fayda sağlayacaktır. Bütünsel dereceli puanlama sisteminde temel hedef performansı belli bileşenlere ayırmadan bütünsel olarak değerlendirilmek amaçlı kullanılmaktadır. Doğan (2020)'a göre

bütünsel puanlamanın en önemli yanı ekonomik ve etkili olmasıdır. Kısacası değerlendirme ölçeğini tercih ederken amaca yönelik olarak bütünsel veya analitik olarak seçmede yarar vardır.

2.1.2.4.2. Analitik Dereceli Puanlama Anahtarı

Analitik dereceli puanlama anahtarlarında bireyin her bir madde veya görev için göstermiş olduğu performansın bütünü parçalara ayrılarak puanlama yapılır. Bu puanlama anahtarların hangisinin kullanımının tercih edilmesi alana, yapılan sınava ve ölçme yapılacak birey sayısına göre farklılık gösterebilmektedir. Bu puanlama anahtarının analitik değerini artırmak için kategori ve ölçütlerin artırılması gerektiğini ancak bu durumun puanlama anahtarın hazırlanmasını ve puanlanmasını zorlaştırdığı öne sürülmektedir (Yıldıztekin, 2014). Analitik dereceli puanlama anahtarı her aşamayı bir bir puanlaması nedeniyle bireyin hangi aşamada eksiği olduğunu net bir şekilde yansıtabilmektedir. Böylece birey eksik olduğu yönleri net bir şekilde görerek düzeltme durumu oluşmaktadır.

Doğan (2020)'a göre analitik puanlama anahtarının en önemli yanı tanılayıcı bilgi sunması olup bunun da puanlayıcıya her adım için ayrıntılı bilgi vererek anlamlı bir dönüt sağlamış olur. Analitik puanlama anahtarları ayrıntılı düzenlenen ölçütlerden meydana gelmesiyle bütünsel puanlama anahtarına göre daha daha geçerli ve güvenilir sonuçlara ulaşmaktadır. Genel olarak eğitimin ilk basamaklarında temel beceriler ölçüldüğü için bütünsel kullanımının daha uygun olacağı sınıf seviyesinin ilerlemesiyle beraber öğrencilerden beklenen bilgi ve beceriler daha üst seviyede olacağı için analitik dereceli puanlama anahtarının kullanımının fazla olmasının daha doğru olacağı belirtilmektedir (Atılğan, Kan ve Doğan, 2017).

2.1.4.3. Dereceli Puanlama Anahtarının Faydaları ve Sınırlılıkları

Bir performansı ölçmek için birden fazla durumu düşünerekten analitik veya bütünsel dereceli puanlama anahtarından hangisinin kullanılacağını öğretmen veya uzmanlar tarafından iyi ayırt edilmesi gerekmektedir. Örneğin bir ürünü süreç içinde değerlendirilmesi halinde alt bileşenleriyle beraber ölçülmesi aşamasında analitik dereceli puanlama kullanılırken, daha yüzeysel ve performansın sonucuna odaklanılmış bir ürün için bütünsel dereceli puanlama anahtarı daha uygun olacaktır. Dereceli puanlama anahtarı türünde çok sayıda bireyin

performanslarını ölçmede bütünsel puanlama anahtarı tercih edilmesi daha doğru olacaktır. Bunun yanında performansın amacının da bu iki türden birinin seçimi için önemlidir. Fakat analitik dereceli puanlama anahtarı bütünsel anahtarına göre daha zaman alıcı ve teferruatlı olmaktadır. (Boring, 2002). Ayrıca bir özelliği ölçülen bireyin dereceli puanlama anahtarına göre puanlanacağını bilmesi değerlendirme aşamasında da bilgi sahibi olmasını sağlamaktadır (Wren, 2009). Dereceli puanlama anahtarının kullanım durumunda uzman veya öğretmenlerin bu anahtarların faydalarını ve sınırlılıklarını hakkında yeterince bilgi sahibi olarak amaca göre bu anahtarlardan birini seçmesi gerekmektedir. Bu durum güvenilirlik adına önemli bir katkı sağlayacaktır (Atılğan, Kan ve Doğan, 2017).



Performance Task

Content	Grade	Learning outcomes	Assessment
English Chapter (Unit):Our Future	Primary 7th grades	• Creativity • Writing Skill • Speaking skill	Analytical rubric

While you were reading newspaper the door bell rang. When you opened the door you saw nobody but a small box covered with coloured paper. There was a glasses in the box and a small sheet of paper with a note on it.



This is a magic glasses. When you put on it, you can see the future. You can arrange the time by pressing the small buttum on glasses.

You arranged the time 2100 and put on the glasses.

- What did you see in the future? Write at least 2 pages considering technology, life family relations, fashion, eating habits in 2100.
- Draw the Picture of the İstanbul in 2100
- Present your work orally in the classroom (maximum 5 minutes)

Instruction

- You have to hand in your Project in 3 class hours.
- You should use both simple past tense and future tense while doing your project
- Make outline and plan your work in advance
- You have to use your imaginative ideaswhile doing your Project
- Keep in mind. If you do your project by yourself, you can get higher points also you can discover your abilities.
- While doing your Project ask your teachers feedback very often, this will help you in difficult time.

Şekil 2.1. Performans Örneği (Kutlu Ö., Doğan C. ve Karakaya İ. 2017: 182)

Tablo 2.2. Analitik Puanlama Performans Örneği (Kutlu Ö., Doğan C. ve Karakaya İ. 2017:183)

Criteria	Achievement Leves			Achievement score	Teacher's feedback
	Need to be				
	Improved	Good	Very Good		
	(1)	(2)	(3)		
Time Usage	Students hands in his/her Project after deadline	—	Students hands in his/her Project in time		
Grammar	Students makes More than Five errors in grammar	Student makes 2-5 errors in grammar.	Student makes almost no error on grammar		
Prediction	Student's Predictions about the future life is not logical and ordinary.	Student's predictions about the future life is partly logical and unique.	Student's predictions about the future life is logical and unique.		
Spelling	Student makes more than errors in spelling	Student makes 2-5 errors in spelling .	Student makes almost no error on spelling		
Presentation (Fluency)	Student often pauses during presentation and doesnt speak fluently	Student rarely pauses during presentation and speaks fluently.	Student never pauses during presentation and speaks very fluently		
Visual Aids	Very few visual Aids are used in the Project.	Some visual aid is used in the Project. But not enough.	Enough visual aid is used in the project		
Originalitiy	Some part of the work (more than %50) is copied from a source.	The Work is partly original and unique (more than %50)	The work is original and unique.		
Content	More than half of the student's work don't match with what it is asked in the project	Student's work partly match with what itiş asked in teh project	Students work completely match with what it is asked in the Project.		
TOTAL SCORE					
Teacher Comment:					

2.1.3.3. Performans Ölçmede Geçerlik ve Güvenirlik İlişkisi

Geçerlik kavramı ölçülmesi gereken özelliğin dikkate alınarak uygun ölçme aracılığıyla yalnızca gerçek yeteneği hakkında ölçme durumudur. Ölçme aracının kullanım amacına göre geçerlik durumu çeşitlilik göstermektedir (Baykul, 2000). Değer olarak +1'e yaklaşan ölçmenin geçerliğinin yüksek olduğu anlamına gelmektedir. Fakat hiçbir ölçme işlemi +1 değerine ulaşamaz. Bunun nedeni ölçmenin mutlaka hata payının olmasından kaynaklıdır. Geçerliğin belirlenmesinde bazı özelliklerin dikkate alınması gerekmektedir. Ölçmenin geçerliği güvenilirliğine bağlıdır. Yani geçerliliği sağlamak için amacına göre minimum farklılık göstermesi gereken bir durumdur (Linn ve Gronlund, 2000).

Geçerlik yapı geçerliği, kapsam geçerliği, ölçüt (kriter) geçerliği uygunluk ve görünüş geçerliği gibi kavramdan oluşmaktadır. (Baykul, 2010). Yapı geçerliği tam olarak ölçülmek istenen özelliğin gerçekte ölçülebilmek derecesini, kapsam geçerliği ölçme aracının ölçülmesi amaçlanan kapsam alanını temsil edebilme derecesini, ölçüt geçerliği ise ölçme sonuçlarının diğer bir ölçüt ile olan ilişkisinin derecesini ifade etmektedir (Güler, 2008). Görünüş geçerliği ise ölçme aracının görünüş bakımından neyi ölçtüğünü betimler (Tan, 2014). Genel olarak geçerlik testin herhangi bir özelliğinden ziyade testten alınan puanlara ait bir özellik olarak kabul görmektedir (Seheryeli, 2018).

Güvenirlik ölçme sonuçlarının tesadüfi hatalarından ayrıştırılmış olması anlamına gelmektedir (Baykul, 2000). Bir ölçme güvenilir olduğu sürece farklı zamanlarda uygulanmasıyla elde edilen sonuçların aynı olma ihtimali fazladır. Ölçmenin tam olarak güvenilir olması mümkün olmamaktadır. Gözlenen puan varyansı gerçek puan varyansına ne kadar yakınsa o kadar güvenilir denilebilir (Arşan, 2012). Performans içerikli değerlendirmelerde puanlama esnasında puanlayıcı yanlılığı olabileceğinden puanlayıcı güvenilirliği kavramını ön planda tutmayı gerektirmektedir. Puanlama yanlılığında hata azaldıkça bireyde ölçülmüş olan özelliğin doğruluk payı yüksektir.

2.1.3.4. Puanlayıcılar Arası Güvenirlik

Puanlayıcılar arası güvenirlik ölçmede birbirlerinden bağımsız olarak aynı zamanda veya farklı zamanlarda bir bireyin performansını puanlaması sonrasında ortaya çıkan tutarlılık durumudur (Anastasi ve Urbina, 1997). Ölçmede birden fazla puanlayıcı, eğitimin dışında farklı bilim dallarında da kullanılmaktadır. Bu nedenle puanlayıcılar arası güvenirlikte önemli bir yere sahiptir (Deliceoğlu, 2009). Özellikle performans alanındaki güvenirlik puanlayıcıların birbiriyle olan uyumuyla ilgilidir. Puanlayıcılar arası güvenirlik G kuramı ve Çok değişkenlik kaynaklı Rasch modeliyle hesaplanabilmektedir (Brennan, 2001).

Aynı durumu puanlayan birden fazla puanlayıcının iki kez puanlaması durumunda puanlayıcıların verdikleri puanların korelasyon değeri uyum indeksi olarak adlandırılmaktadır (Deliceoğlu, 2009). Puanlayıcıların tam anlamıyla tesadüfi olarak puanlama yapmaları durumunda uyum düzeyi beklenildiği gibi çıkmamama durumu olağandır. (Goodwin, 2001).

Puanlanan sonuçların güvenilir olması için uyum indeksi değerinin %75'in üzerinde olması gerekir. Bunun altında kalan değerler değerlendirmede puanlayıcıların önemli ölçüde farklı düşündükleri anlamına gelmektedir. (Şencan, 2005).

Şans başarısı üzerinden uyum indeksini kestirmeyi sağlayan Cohen kappa istatistiği birden fazla puanlayıcının yaptığı değerlendirmeler arasındaki uyum indeksini belirlemeye yarayan bir istatistiksel kavramdır (Helms, 1999). Kendall uyum katsayısı (W) ise çok sayıda puanlayıcının vermiş olduğu puanlar arasındaki uyum düzeyini belirlemeye yaramaktadır (Akt Tavşancıl, 2002). Kendall'ın uyum katsayısı üzerinden yapılan değerlendirmelerde puanlayıcılar arasındaki uyuşmanın en aşağı 0,80 seviyesinde olması istenilen değer olarak kabul edilmektedir. (Szymanski ve Linkowski, 1993).

İki puanlayıcının bireylere, objelere göre ölçülen özellik kapsamına verdikleri puanlar arasındaki ilişki durumu veya tutarlılığı olarak tanımlanan pearson çarpım-moment korelasyon analizi, güvenirlik hesaplamasında iki puanlayıcının puanlarının doğrusal ilişkisini gösterir (Balcı, 2000). İki puanlayıcının değerlendirmeleri eşit aralıklı ölçek verilerine dayanıyorsa ve verilerin büyüklük sırası önemsiz görülüp veri sayısı 30'dan büyük olduğunda ilişkiyi belirlemek amacıyla kullanılmaktadır. Fakat Pearson çarpım-moment korelasyon analizi iki puan değeri arasındaki ilişkiyi belirler. Puanlayıcıların uyuşma yüzdeliği hakkında bilgi

vermemektedir. Bu nedenle puanlayıcılar arasındaki varyansı dikkate almamasından dolayı puanlayıcılar arasındaki değişkenliğe karşı herhangi bir veri sağlamamaktadır (Şencan, 2005).

Hartman (1977)'e göre Pearson korelasyon katsayısının hesaplandığı grubun büyüklüğünden doğrudan doğruya etkilenmesi olumsuz bir durum oluşturmaktadır (Akt: Goodwin 2001). Puanlayıcılar arası güvenilirliğin hesaplanmasında Pearson korelasyonunun kullanılmasının diğer bir olumsuz yanı ise, ortalamaya bağlı olmamasıdır. Bu da, güvenilirlik hesaplamada iki puanlayıcının puanlama değerlerine ilişkin benzerlik ve farklılıklarını gösterememesine neden olmaktadır. Diğer bir şekilde, Pearson çarpım-moment korelasyon katsayısı iki puanlayıcının verdikleri puanların doğrusal ilişkisi yani bu iki değişkenin birlikte değişiminin bir ölçüsüdür (Baykul, 2000).

Pearson korelasyon katsayısının puanlayıcılar arası uyumun ya da güvenilirliğin belirlenmesindeki olumsuz tarafını ortadan kaldırmak amacıyla varyans analizine bağlı olarak gruplar içi korelasyonu hesaplanmalıdır. Varyans analizi yoluyla hesaplanabilen bu güvenilirlik katsayısı 0 ile + 1 arasında değerler alır. Bu katsayı varyans analizi literatüründe grup-içi korelasyon (Intraclass Corelation Coefficient) denilen özel bir korelasyon katsayısıdır (Goodwin, 2001).

Puanlayıcılar tarafından elde edilen ölçümlerin sonrasında ortaya çıkan ölçme sonuçların aynı veya yakın olması durumunda ölçmenin güvenilir olması muhtemeldir. Herhangi bir ölçme aracıyla yapılan ölçmenin sonucunda yapılan değerlendirmelerde puanlayıcılar arasındaki uyuşmanın minimum 0,80 değerinde olması beklenmektedir. Eğer ölçme yalnızca gözlemde bulunmak maksadıyla yapılmışsa puanlayıcılar arasındaki uyum indeksinin 0,70 değerinde olmasının yeterli olacağı belirtilmiştir (Szymanski ve Linkowski, 1993).

2.2. Arařtırmada Kullanılan Kuramlar

2.2.1. Madde Tepki Kuramı

Madde Tepki Kuramı (MTK), ilk olarak psikometride yeteneęi ölçmek adına Danimarkalı bir matematikçi olan George Rasch tarafından ortaya çıkarılarak 20. yüzyılın sonlarına doğru gelişmeye başlamış bireyin gizil becerileri ile madde arasındaki baęı matematiksel modellerle ortaya koyan bir kuramdır (Atılğan, 2004). Madde tepki kuramı gizil özellik veya yetenek denilen bileşen ile bireylerin göstermiş olduęu performansın açıklanabilirlięi ve belli bir örüntü olarak artan madde karakteristik eğrisi vasıtasıyla bireyin madde performansı ile ilgili maddenin altında yatan özellik veya yetenek kümesi arasındaki ilişki tanımlanabilmesi olarak iki farklı ana varsayıma baęlıdır (Koęar, 2022).

MTK testteki maddenin özelliklerini ayrıntılı olarak vermekte ve bireyde var olan gizil özelliklerini ve gerçek yeteneklerini ortaya çıkararak çeşitli becerilere sahip bireyleri seçemeyen maddelere ait sorunları tespit etmeyi saęlayan bir kuramdır (Arşan 2012). Madde tepki kuramında dikkate alınan çeşitli deęişkenlik kaynakları bulunmaktadır. Özellikle performansa dayalı ölçümlerde birey, puanlayıcı, görev ve madde gibi deęişkenlik kaynakları ölçmede güvenilirlięi etkilemektedir (Güler, 2008).

Puanlayıcılar arası güvenilirlięin veya deęişkenlik kaynaęının fazla olduęu durumlarda güvenilirlik hesaplanmasında kullanılan bir model olup Rasch modelinden türemiştir (2004, Atılğan). Rasch Modeli puanlayıcılar arası güvenilirlięin deęerlendirilmesinde kullanılabilen bir modeldir. Ayrıca Madde Tepki Kuramı kişinin örtük becerileri ölçme aracındaki maddeler arasındaki ilişkiyi istatiksel olarak açıklamayla ilgilenirken dięer kuramlar gözlenebilen puanlara dayalı analizlerle bunu saęlamaktadır (Güler, 2008). MTK'nın bazı dezavantajlı durumları da bulunmaktadır. Kuramın uygulanabilmesi için genelde daha büyük örneklemelere ihtiyaç vardır. Bunun yanında karmaşıklıęı sebebiyle yorumlamak oldukça zor ve bazı yazılımların vasıtasıyla ortaya çıkarmak gerekmektedir. (Hambleton ve Swaminathan, 1985). Madde tepki kuramında puanlayıcılar arası güvenilirlięi belirlemek ya da bireyin yeteneęi ile madde güçlüęü arasındaki ilişkiyi belirlemede ve performans deęerlendirmede de kullanılabilen model, Rasch modeli olarak bilinmektedir (Güler, 2008).

2.2.1.1. Çok Değişkenlik Kaynaklı Rasch Modeli

Rasch modeli bireylerin yetenek düzeylerini ve madde güçlük düzeylerini birlikte ele alarak çalışan modeldir. Bu iki kavram çerçevesinde kurulmuş olup Linacre (1989) tarafından geliştirilen ÇDKRM'nin gelişmiş bir uzantısıdır. (Linacre, 1989). ÇDKRM yalnızca bir değişkenlik kaynağını değerlendirmeden ziyade birden fazla değişkenlik kaynağını aynı anda değerlendirmeyi mümkün kılmaktadır (Sudweeks, Reeve ve Bradshaw, 2005). Bu modelin eşitlik formülü aşağıdaki gibidir. (Linacre, 1993).

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

P_{nijk} : n. bireyin i. maddede j. Puanlayıcı tarafından k olarak puanlandırma olasılığı

P_{nijk-1} : n. Bireyin i maddede j. Puanlayıcı tarafından k-1 olarak puanlandırma olasılığı

β_n : n. bireyin yeteneği

δ_i : i maddenin güçlük düzeyi

λ_j : j. Puanlayıcı katılığı

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

Performans ölçümlerinde birden fazla değişkenlik kaynakları bulunmaktadır. Bu değişkenlerin başlıcaları birey, görev, puanlayıcılar, puanlanan maddeler ve puanlama ölçeğidir (Lunz ve Wright, 1997). Bu beş değişkenlik kaynağının birbirleriyle olan ilişkisini incelemek adına Çok Değişkenlik Rasch Modeli istatistiksel olarak bir veri sağlamaktadır. Bu model sayesinde puanlayıcı katılığı veya cömertliği gibi bireyin gizli özelliğini kestirmeyi etkileyen durumları da hesaplamak mümkün hale gelmektedir. (Linacre, 1989). Bu modelde farklı değişkenlik kaynakları eklenerek de değerlendirme yapılabilmektedir. Performans ölçümlerinde puanlayıcıların farklı yapılarda olması nedeniyle puanlamayı etkileyen önemli bir değişken konumundadır. Bununla beraber diğer önemli bir değişkenlik kaynağı olan puanlama ölçeği de puanlamayı yapacak puanlayıcının ne denli objektif puanlayacağını belirlemek adına oldukça önemli bir yere sahiptir. Çok değişkenlik Rasch modelinde puanlayıcı katılıkları ve bireyin yetenekleri gibi durumlar puanlayıcıların her bir madde için verdiği puanlamadan yola çıkılarak sonuca varılmaktadır (Atılman, 2004).

Çok Değişkenlik Kaynaklı Rasch Modelinin Genellenebilirlik Kuramından ayrılan bazı özellikleri vardır. Çok Değişkenlik Rasch Modeli her bir değişkenlik kaynağının kendisiyle beraber, diğer değişkenlik kaynağına ilişkin örneklemden bağımsız olarak parametre kestiriminde bulunmayı sağlar (Güler, 2008). Genellenebilirlik Kuramında ölçme niteliği bir grup veya bütünsel ele alınırken Çok Değişkenlik Kaynaklı Rasch Modelinde değişkenlerin her biri ayrıntılı olarak ele alınmaktadır (Alharby, 2006). Ayrıca Genellenebilirlik Kuramında bir ölçme için güvenilirlik katsayısı bulunurken, Çok Değişkenlik Kaynaklı Rasch Modelinde her bir değişkenlik kaynağı için ayrı ayrı güvenilirlik katsayıları bulunabilmektedir. Çok Değişkenlik Kaynaklı Rasch Modeli bütün değişkenlik kaynaklarını içine alan aynı zamanda bu değişkenlik kaynaklarının beraber araştırılmasına uygun olan bir desendir (Güler, 2008).

ÇDKRM analizi Lineacre'nin geliştirmiş olduğu FACETS adlı yazılım ile yapılmaktadır. Özellikle performans değerlendirme sonuçlarında birden fazla bileşeni ele alma, bireyin yeteneğini, madde güçlüğü ve punalayıcı katılığı/cömertliği gibi değişkenleri belirlemede model-veri uyumluluğunu test etmede kolaylık sağlamaktadır (Arsan, 2012).

Çok Değişkenlik Kaynaklı Rasch Modelinde elde edilen istatistiklerden biri olan uyum indeksleri bütüncül olarak ya da bireysel olarak değişkenlik kaynaklarının ÇDKRM ile hesaplanan gözlenen veriler ile istatistiksel modellerin arasında olan uyumunu içermektedir. Uyum istatistiği değeri beklenen puanların gözlenen puanlardan farkıyla hesaplanan değerlerdir. Veriler modelle uyumlu olduğunda her bir artık değeri sıfır olması beklenmektedir. Veri modele yaklaştıkça standartlaşmış artığın kestirimi, ortalaması 0 ve varyansı 1 olacak şekilde minimum veya maksimum değerde normal dağılım göstermektedir. Buna göre verinin modele uyumu için kullanışlı bir ölçüt olduğundan standartlaştırılmış artıkların yaklaşık normal dağılım gösterip göstermediğine bakılarak yorum yapılmaktadır (Hetherman, 2004).

ÇDKRM'deki uyum istatistikleri dış uyum (outfit) ve iç uyum (infit) istatistikleri olarak iki bakımdan vermektedir. Uyum istatistikleri birey ile diğer değişkenler arasındaki 0 ile sonsuz değer arasında değer alabilmekte ancak değer 1 olması beklenmektedir. Uyum istatistiklerinin 1'den küçük olması durumunda veri anlamsız ve sınırlıdır. 1'den büyük olması durumunda ise verilerin tutarsız ve uç değerlerde olduğunu kabul edilmektedir (Smith ve Kulikowich, 2004). Her bir puanlayıcının puanlamadaki katılığı veya cömertliği, maddede yer alan diğer değişkenlik kaynaklarından bağımsız olarak kestirilir. Bu kestirimle ortaya çıkan yüksek değerler puanlayıcının ortak ölçekte kendisinin altında kalan puanlayıcılardan daha katı olduğunun bir göstergesidir. Her bir puanlayıcının katılığının düzeyinin kestiriminden sonra, bu kestirimlerin verilere uygun olup olmadığı değerlendirilir. Benzer şekilde her bir bireyin

yeteneğinin kestirimi, her bir maddenin güçlük düzeyinin kestirimi ya da modeldeki diğer tüm değişkenlik kaynaklarının kestirimleri için yapılabilir (Hetherman, 2004).

Alan yazında, hem iç hem dış uyum istatistiği için genel olarak kabul görebilen değerlerin 0.5 ile 1.5 arasında olması önerilmektedir (Turner, 2003). Değişkenlik kaynağı içindeki her bir elemana ilişkin iç uyum veya dış uyum istatistik değerinin 1.5'den büyük olması, anlamlı bir uyumsuzluk durumunun varlığına işaret etmektedir. Uyum istatistiği 1'den büyüdükçe verilerin tutarsızlığının arttığını, 1'den küçüldükçe verilerdeki bağımlılığın ve kısıtlılığın arttığını göstermektedir (Smith ve Kulikowich, 2004).

Ayırt etme indeksi olarak bilinen değer Rasch modelinde, verinin modele ne kadar uyumlu olduğunu belirlemeye yarayan bir diğer istatistik kavramıdır. Ayırt etme indeksi, değişkenlik kaynağında mevcut bulunan nesnelerin birbirinden ne kadar farklılaştığını gösteren bir ölçü durumudur. Bireyler veya maddeler için bu ölçünün büyük, puanlayıcılar için ise sıfıra yakın değer çıkması beklenmektedir. Değerin büyük olduğu bir ayırma indeksi, puanlayıcıların birbirlerinden ne kadar ayrışma göstererek puanlama yaptıklarının işaretidir. Puanlayıcı değişkenlik kaynağı için elde edilen düşük ayırma indeks değerleri, istenen bir durum olarak katılık veya cömertlik açısından küçük farklılıklar olduğunun bir göstergesidir. Ayırma indeksi G ayrıca her bir değişkenlik kaynağı için, değişkenlik kaynağında alan elemanların (Lee ve Kantor, 2003;Hetherman, 2004).

Cronbach alfa katsayısına benzer bir şekilde ayırt etme indeksi güvenilirliği de 0 ile 1 arasında farklı değerler almaktadır. Bu değer değişkenlik boyutundaki elemanların birbirlerinden ne kadar güvenilir bir şekilde ayrıldıklarını gösterir. Birey, madde ve görev gibi değişkenlik kaynakları için ayırt etme indeksi güvenilirliğinin bire yakın çıkması istenen bir durumdur. Puanlayıcı değişkenlik kaynağı açısından katılık veya cömertlik açısından birbirlerinden farklı olduğunu gösterdiği için istenen bir durum değildir.

ÇDKRM'de ki-kare değeri puanlayıcılar arasındaki tutarlılığı göstermektedir. Bu değer birden fazla puanlayıcının katılık/cömertlik anlamında farkını ortaya koymaktadır. Böylece puanlayıcılar arası farkın bu şekilde önemini ortaya çıkarmaktadır (Wright ve Lineacre, 1994).

Bu bilgilere bakarak ÇDKRM, performans değerlendirme açısından daha güvenilir veriler elde edilmekte ve puanlayıcı etkisini daha net açıklamaktadır. Elde edilen bilgilerin detaylı olması durumu ise özellikle puanlayıcının daha doğru ve geçerlik güvenilirlik indeksi yüksek değerlendirme yapmasına olanak sağlamaktadır.

2.2.2.Genellenebilirlik Kuramı

Her ölçme sonuçları bir kısım hata barınmaktadır. Ölçmeyi gerçekleştiren kişi bu durumda hata oranını azaltmak için birtakım çalışmalar yapmaktadır. Örneğin; bir testin ölçme sonuçlarının hata payını azaltmak adına birden fazla puanlayıcı tarafından puanlama yaparak bu puanların ortalaması alınabilir. Veya farklı zamanlarda aynı testin puanlaması yapılarak yine test puanların ortalaması alınarak hata oranı azaltılabilir. Bu durum tam olarak testin güvenilir sonucunu vermese de ideal kestirimi olarak kabul edilmektedir. Testin güvenilirliğini artırmak için o teste özel bir ölçme aracı da kullanılabilir. Ancak bu durumda ölçmenin genellenebilirliği kısıtlanmış olmaktadır (Güler, Kaya-Uyanık ve Taşdelen-Teker, 2012).

Puanlayıcılar arası güvenilirliğin değerlendirilmesinde kullanılan yaklaşımlardan biri olan genellenebilirlik kuramı birey, madde, puanlayıcı, gibi benzer değişkenlik kaynaklarını birlikte ele alan, davranış ölçümlerinde güvenilirliğin değerlendirilmesinin yanında güvenilir gözlemlerin oluşturulmasını ve tanımlanmasını sağlayan varyans analizine (ANOVA) bağlı güçlü bir istatistiksel bir kuramdır (Atılğan, 2004).

Bu kuram Klasik Test Kuramının kısıtlı durumlarını gidermek adına Cronbach ve arkadaşları (1963-1965) tarafından ortaya atılmış gibi görünse de öncesinde güvenilirlik kestiriminde Burt(1936) , Hoyt, Jackson ve Ferguson (1941) tarafından varyans kesitimi tartışılmaya başlanmıştır. (Güler, Kaya-Uyanık ve Taşdelen-Teker, 2012). Klasik test kuramında ölçme neticelerine karışmış olan hatalar yalnızca bir değişkenlik kaynağından gelen hatalar olarak dikkate alırken genellenebilirlik kuramında sistematik ve sistematik olmayan hatalar birlikte ele alınarak eş zamanlı hesaplamalar yapabilmektedir. Genellenebilirlik kuramının klasik test kuramından üstün kılan bir diğer durumu ise klasik test kuramı güvenilirlik hesaplamada yalnızca bağıl değerlendirmeyi esas alırken genellenebilirlik kuramı hem bağıl hem mutlak değerlendirme için katsayı oluşturabilmektedir (Brennan, 2001). Yine klasik test kuramı farklı değişkenlik kaynaklarından gelen hataları birlikte değerlendiremezken genellenebilirlik kuramı görev, madde, puanlayıcı, test ve zaman gibi birden fazla değişkenlerden kaynaklı hataları beraber değerlendirerek kapsamlı bir biçimde tek güvenilirlik katsayısı hesaplanmasını sağlamaktadır (Shavelson ve Webb, 1991). Bu durum daha doğru hata analizi yapmasını sağlayarak genellenebilirlik kuramını diğer kuramlardan daha güçlü bir istatistiksel kuram haline getirmektedir.

G kuramına göre ayrıca evrende birden fazla varyans kaynağı bulunabilmektedir. Bu varyans kaynakları değişkenlik kaynağı (facet) olarak adlandırılmaktadır (Brennan, 2001). Bu değişkenlik kaynaklarının da farklı düzeyleri koşul (condition) olarak tanımlanmaktadır.

Genellenebilirlik kuramında değişkenlik kaynağının başka bir değişkenlik kaynağının koşullarıyla çaprazlandığı yani değişkenlik kaynaklarının tüm koşulları diğer değişkenlik kaynaklarının tüm koşullarını etkiliyorsa çaprazlanmış (crossed) denilmektedir (Shavelson ve Webb, 1991). Bu desende değişkenlik kaynakları arasında “x” işareti kullanılır. Örneğin her puanlayıcı her öğrenciyi aynı maddeler için puanladığında çaprazlanmış desen $p \times \ddot{o} \times g$ oluşturur (p:puanlayıcı, $\ddot{o}$:öğrenci, g: görev).

Değişkenlik kaynaklarının başka bir değişkenlik kaynaklarıyla çaprazlanmadığı durumlarda yani bir değişkenlik kaynağının koşulları yer alan bazı değişkenlik kaynaklarını gözlemliyorsa yuvalanmış (nested) olarak tasarlanmalıdır. Yuvalanmış desende değişkenler arasında “:” işareti kullanılır (Shevelson ve Webb, 1991). Örneğin; farklı okullarda okuyan öğrencilerin birden fazla puanlayıcıların puanladığı durumlarda $(o: \ddot{o}) \times p$ yuvalanmış desen oluşmaktadır. (o:okul, $\ddot{o}$: öğrenci, p: puanlayıcı) Bunların dışında hem çaprazlanmış hem de yuvalanmış desenin bulunduğu desenler karma desen olarak adlandırılmaktadır.

G kuramında bağlı (G katsayısı) ve mutlak (Phi katsayısından) sayıdan oluşan iki güvenilirlik katsayısı yer almaktadır. Bireyler arası durumu odak olarak alarak hesaplanan katsayı g katsayıyken bireyin performans da ne kadar başarılı olduğunu ölçmek adına phi katsayı hesaplanmaktadır. Bu durumda birey bağımsız değerlendirilmekte diğer bireylerle ilişkisine bakılmamaktadır.

G Kuramında, genellenebilirlik (G) çalışmaları (Generalizability (G) Studies) ve Karar (K) çalışmaları (Decision (D) Studies) olmak üzere güvenilirliğin araştırılmasında iki farklı çalışma türü bulunmaktadır (Goodwin, 2001).

2.2.2.1 G Çalışması

G çalışması ölçmelerden kaynaklı farklı türdeki ölçmeye karışan hataları tespit etmek ve değişkenlik kaynaklarına ait varyans bileşenlerini öngörme adına veri elde etmektir. (Rentz, 1987). Kısacası bir örneklem üzerinden yola çıkarak evrene genellemek olarak da açıklanmaktadır (Güler, Kaya Uyanık ve Taşdelen Teker, 2012).

G çalışmasını düzenlerken ilk olarak madde, görev, zaman, birey gibi olabilecek değişkenlik kaynaklarının ve evrenin belirlenmesi gerekmektedir (Atılğan, 2004). G çalışması evrendeki tanımlanan puanların varyansına katkısını ve faktörlerin gözlenen puanlardaki etkisini bulmaya yardımcı olur. Aynı zamanda eldeki ölçme sonuçlarına göre güvenilirlik durumunu ortaya koymaktadır (Atılğan, 2008). G çalışmalarında facet denilen puanlayıcı, madde, okul gibi değişkenlik kaynakları koşul olarak adlandırılan değişkenler yer almaktadır. G çalışması da bu değişkenlik kaynaklarının varyansına etkisini araştırmakla yükümlüdür. G çalışmasında yer alan hata kaynaklarına göre farklı desenler ortaya çıkmaktadır. Ortaya çıkan desenler hata kaynağı sayısına ve türüne göre çeşitlilik oluşturmaktadır. G kuramının desenleri buna göre ifade edilmektedir. Tek yüzeyli bir desende ölçme objesi, değişkenlik kaynağı ve ölçme objesi, değişkenlik kaynağı ile artık varyansdan oluşan 3 farklı değişkenlik kaynağı bulunmaktadır. G çalışmasında ölçme objesinden kaynaklanan varyans değeri evren puanı olarak kabul edildiğinden olabildiğince yüksek çıkması istenmektedir (Yılmaz-Koğar ve Koğar, 2022). Birden fazla değişkenlik kaynağının yer almasıyla değişkenlik kaynağının sayısının da doğrusal olmayan biçimde artmaktadır. Örneğin; bireyi ölçme objesi olarak kabul ettiğimiz tek facetli bir desende değişkenlik kaynağı yalnızca maddeyse birey, madde ve birey-madde ve artık varyansdan oluşan 3 farklı değişkenlik kaynağı yer alırken değişkenlik kaynağına puanlayıcıyı da eklediğimizde iki facetli bir desen oluşmakta ve birey, madde, puanlayıcı, birey-madde, birey-puanlayıcı, madde-puanlayıcı ve birey-madde-puanlayıcı ile artık varyans etkileşiminden oluşan 7 farklı değişkenlik kaynağı oluşmaktadır.

G kuramında değişkenlik kaynakları sabit ve rastgele olmak üzere ikiye ayrılmaktadır. Güler vd. (2012)'de örneklem büyüklüğünün evren büyüklüğünden küçük ve örneklem evrenden tesadüfi olarak seçilerek ya da aynı evrenden seçilecek başka bir örneklem ile değiştirilebilmesi mümkünse bu durum rastgele yüzey olarak adlandırılmaktadır. Başka bir deyişle eğer çalışmanın amacı örneklemin ötesinde bir genelleme yapmak değil ise sabit olarak adlandırılmaktadır. Bu durumun da ölçmenin genellenebilirliği üzerinde etkili olmaktadır. Yani sabit yüzeyli desende hata varyansı azalır ve g ve phi katsayıların artmasına neden olmaktadır. Minimum bir yüzeyin sabit olması hata varyansını düşürerek güvenilirlik katsayısını artırırken elde edilen sonuçların yorumlanmasını da kısıtlamaktadır (Brennan, 2011).

2.2.2.2. Karar Çalışması (D Study)

K (Karar) çalışması G çalışmasından elde edilen bilgilerden yararlanarak hataları en aza indirmenin yollarını araştırıp karar vermek adına en uygun duruma ulaşılmasını sağlamaktır. (2004, Atılğan). K çalışmasında amaç yapılan G çalışmasına göre elde edilen bilgilerden yola çıkıp hataları minimuma indirerek daha fazla güvenilirlik elde etmeye yöneliktir. Bu durum aslında G kuramını Klasik Test Kuramından ayırtıran önemli bir faktördür. G kuramı, öznel ya da nesnel kararların alınması durumunda güvenilirliği etkileyen tüm hata kaynaklarını ele almaktadır. Bu karar çalışmalarıyla elde edilen gerçek puanlayıcı-madde sayılarının artırılıp azaltılmasıyla çeşitli durumlarda ortaya çıkabilecek olan sonuçlar yorumlanabilmektedir. Sonuçların benzerlik göstermesi durumunda ölçeğin kullanılabileceği sonucuna varılmaktadır (Dimitrov, 2002). Ayrıca daha sonra yapılacak olan çalışmaların nasıl uygulanacağı ve daha iyi hale getirilmesi konusunda öngörude bulunulmasını sağlamaktadır.

Karar çalışmasında Genellenebilirlik katsayısıyla beraber Phi katsayısının da yer aldığı iki farklı güvenilirlik katsayısı hesaplanabilmektedir. Hata varyansına göre hesaplanmış olan bu iki güvenilirlik katsayıları mutlak ve bağıl olmak üzere iki durumdan oluşmaktadır. G çalışmasında temel amaç kestirilen varyans bileşenlerin evrendeki toplam varyans üzerindeki etkisini açıklamak iken K çalışmasının amacı maksimum güvenilirliği sağlamak amacıyla, G çalışmalarında toplanmış olan verileri yorumlayarak araştırmacının genellemesine en uygun ölçme durumu hakkında karar vermesidir. (Güler, Uyanık ve Teker, 2012). G katsayılarında bulunan hata varyanslarını minimuma indirmek adına farklı senaryolarda K çalışmaları oluşturulabilmektedir.

2.2.2.3.Genellenebilirlik (G) Katsayısı

Genellenebilirlik katsayısı (G), Cronbach (1972) tarafından belirlenmiş olan bir ölçüm aracının güvenilirliğini ölçmek için kullanılan bir istatistiksel kavramdır. Genellenebilirlik katsayısı, ölçüm aracının ne kadar istikrarlı olduğunu ve aynı ölçümü farklı zamanlarda yaptığında ne kadar tutarlı sonuçlar verdiğini gösterir. Kısacası bağıl (görelî) güvenilirliği kestirmek için kullanılan evren puanı varyansının gözlenen puan varyansına oranı olarak da tanımlanmaktadır (Shavelson ve Webb, 1981).

G katsayısı değeri, 0 ile 1 arasında değişen bir sayıdır. Ancak bazı durumlarda 0'ın altında değerler alabilmektedir. Bu değerlerin negatif olarak büyük olması hatalı olduğu anlamına gelmektedir. Bu durumda negatif değer alan hata varyansını 0 olarak kabul etmek daha doğru olacaktır. G katsayısı 1'e ne kadar yakınsa, ölçüm aracının o kadar güvenilir olduğu kabul edilir. Genellikle, 0.70 ve üzeri bir G katsayı değeri, ölçüm aracının kabul edilebilir bir güvenilirlik seviyesine sahip olduğunu gösterir.

G Katsayısı değeri, ölçülen özellik veya değişkenin doğasına ve kullanılan ölçüm aracına bağlı olarak değişebilir. Örneğin, bir kişinin matematik başarısını ölçmeye yarayan ölçme aracı farklıyken boy uzunluğunu ölçmeye yarayan ölçme aracı farklıdır.

Genellenebilirlik katsayısı, araştırmacıların ölçüm araçlarının güvenilirliğini değerlendirmelerine ve ölçülen özelliklerin gerçek değerleri maddeline daha doğru bir şekilde yaklaşmalarına yardımcı olur. Örneğin belli sayıda bir bireyin 10 maddelik test başarısını puanlamada 2 puanlayıcı olduğunu varsayalım. Bu çalışmanın genellenebilirlik katsayısını artırmak için madde sayısını ve puanlayıcı sayısını artırmak gerekecektir. Ancak tek yüzeyle G katsayısı yalnızca hata kaynağı dikkate alınacağı için Cronbach alfa sayısına eşittir (Yılmaz-Koğar ve Koğar, 2022).

2.2.2.4. Phi Katsayısı

Phi katsayısı (ϕ), iki kategorik değişken arasındaki ilişkiyi ölçmek için kullanılan ve güvenilirlik indeksi olarak da adlandırılan bir terimdir. Phi katsayısı, 2x2 lik tablolarda kullanılır ve mutlak güvenilirliği incelemek adına kullanılır.

Phi katsayısı, Pearson korelasyon katsayısı ile benzer bir değerlendirme yapan ve mutlak kararlarda kullanılan güvenilirlik katsayısı olarak bilinmektedir (Brennan, 2001). Ancak, Pearson korelasyon katsayısı, iki sürekli değişken arasındaki ilişkiyi ölçerken, Phi katsayısı iki kategorik değişken arasındaki ilişkiyi ölçer.

Phi katsayısı, -1 ile 1 arasında değişken olarak değer almaktadır. İki değişken arasında tam ters orantılı bir ilişki varsa -1 değer alırken, iki değişken arasında doğru orantılı bir ilişki varsa da +1 değer gösterir. İki değişken arasında herhangi bir ilişkinin yer almaması durumunda 0 değer alır.

Phi katsayısı, özellikle tıp, sosyal bilimler, psikoloji gibi alanlarda kullanılmaktadır. Örneğin, bir araştırmacı, kardeş sayısıyla başarı arasındaki ilişkiyi ölçmek istediğinde, Phi katsayısı kullanarak iki değişken arasındaki ilişkiyi değerlendirebilir. Genellenebilirlik katsayısında olduğu gibi puanlayıcı ve madde sayısındaki artış phi katsayısının artışı sağlamaktadır (Güler, Uyanık ve Teker, 2012).

2.2.2.5. Tek değişkenlik kaynaklı evrenler (bxm)

G kuramında tek değişkenlik kaynaklı evrenlerde ölçmeye dahil olan hata tek bir kaynaktan gelmektedir. (Güler, Uyanık ve Teker, 2012). Örneğin 5 soruluk 20 bireye uygulanan bir testte her soruyu tüm bireyler cevapladığı takdirde değişkenlik kaynağı maddenin zorluk derecesine göre tek yüzeyi maddeler ve ölçme objesi birey olmaktadır. Bu durumda çaprazlanmış desen bxm olmaktadır. Başka bir şekilde yine aynı sayıda bireylerin aynı sayıda maddelere ayrı ayrı cevap vermiş olduğunda yani 5 birey eşit bir şekilde 4 er madde cevapladığında bu kez yuvalanmış desen b:m olmaktadır.

2.2.2.6. İki değişkenlik kaynaklı evrenler (bxmxd)

İki farklı yüzeyin yer aldığı karmaşık yapılardan oluşur. Örneğin birey, madde ve puanlayıcıdan oluşan değişkenlerin iki yüzeyi hata kaynağı madde ve puanlayıcılar oluşturmaktadır.

2.2.2.7. İkiden Fazla Değişkenlik Kaynaklı Evrenler

İki veya daha fazla değişkenlik kaynağının üç veya daha fazla değişkenlik kaynaklı evrenler genellenmesiyle meydana gelir. Örneğin farklı okullarda yer alan bireylerin birden fazla puanlayıcının puanladığı 10 soruluk bir testte okul, madde ve puanlayıcı dan oluşan üç yüzey bulunmaktadır. Gözlemlenen evrenin genişlemesiyle örneklemden evrene yapılan genellemede de hata yapma ihtimali artma eğilimindedir (Güler, Uyanık ve Teker, 2012).

2.2.İlgili Araştırmalar

Bu başlık altında, araştırmada kullanılan çok değişkenli kaynağını birlikte değerlendirebilen iki yaklaşım olan Genellenebilirlik Kuramı ve Çok Değişkenlik Kaynaklı Rasch Modeli ilişkin hem kendi aralarındaki hem de farklı kuramlarla olan ilişkisini incelemek amacıyla araştırmalar özetlenmiştir. İlgili araştırmalar özetlenirken Genellenebilirlik Kuramı

ve Çok Değişkenlik Kaynaklı Rasch modelinin yurt içinde ve yurt dışındaki araştırmalar sınıflanarak sunulmuştur.

2.2.1.Yurt İçindeki Araştırmalar

Atılğan (2004), iki yıl art arda yıllarda yapılmış olan Müzik üzerine özel yetenek seçme sınavları verilerini referans alarak Genellenebilirlik Kuramı ve Çok Değişken Kaynaklı Rasch Modeli analizleri ile gerçek verilerden çıkan istatistikleri karşılaştırılmıştır. 7 boyut ile bu boyutlara bağlı olarak 28 görevden elde edilen verilerin analizi sonucunda; bu iki kuramın istatistikleri karşılaştırılmış ve Genellenebilirlik kuramının çeşitli senaryolar üzerine yapılan çalışmalarının gerçek durumla ilişkilerine bakılmıştır. Bununla birlikte Genellenebilirlik Kuramının tek değişkenli ve çok değişkenli modellerinin aynı ölçme durumu için farklı sonuçlar verdiği ancak özellikle puanlayıcı ve görev bakımından her iki yaklaşımının da uyumlu sonuçlar verdiği görülmüştür. G ve Phi katsayılarının modelden etkilenecek alternatif karar çalışmalarıyla farklı puanlayıcı sayıları senaryoları karşısında elde edilen G ve Phi katsayılarının gerçek durumda elde edilen farklı olduğu belirlenmiştir. Fakat özellikle puanlayıcı ve görev değişkenlik kaynağı hem genellenebilirlik kuramı ve çok değişken kaynaklı Rasch modeli ile değişkenlik kaynakları için bulunmuş olan sonuçlarının büyük oranda tutarlı sonuçlar verdiği görülmüştür.

Yelboğa (2007), iki yıl art arda iş performansı ölçeği üzerinden Klasik Test Kuramı ve Genellenebilirlik Kuramına göre ortaya çıkan güvenilirlik katsayıları karşılaştırılmıştır. Klasik test kuramında Cronbach alfa ve test tekrar test güvenilirlik katsayıları genellenebilirlik kuramında ise G ve Phi katsayıları karşılaştırılmıştır. Araştırma 11 ayrı birimde çalışan 176 bireye uygulanmıştır. Bunun sonucunda güvenilirlik katsayıların birbiriyle uyumlu sonuçlar verdiği görülmüştür. Ayrıca birden fazla puanlayıcının puanlama yaptığı durumlarda genellenebilirlik kuramının kullanılmasının daha doğru olacağı belirtilmiştir.

Güler (2008), Klasik Test Kuramı, Rasch modeli ve Genellenebilirlik Kuramını birlikte ele aldığı araştırmada matematik başarısını birden fazla puanlayıcı tarafından ölçülmesinden çıkan sonuçların güvenilirliklerini bularak karşılaştırmıştır. Bu araştırmada 8. Ve 9. Sınıf öğrencilerinden oluşan 203 bireye 1999 TIMS sorularından oluşan açık uçlu maddeler bütünsel puanlama anahtarıyla 4 puanlayıcı tarafından puanlanarak elde edilen sonuçlar karşılaştırılmıştır. Araştırma sonucunda matematik başarısını ölçmek için kullanılan ölçme aracının güvenilir sonuçlar verdiği ve puanlayıcıların birbiriyle tutarlı puan verdiği

görülmüştür. Bu sebeple puanlayıcı sayısını artırmak yerine madde sayısını arttırmanın daha iyi sonuçlar vereceği sonucuna ulaşılmıştır. Bunun yanında araştırma puanların güvenilirliğinin belirlenmesinde minimum iki kuram üzerinden faydalanılmasının doğru olacağı sonucuna ulaşılmıştır.

Yelboğa (2008), otomotiv alanında personel seçme aşamasında yapılandırılmış mülakat formu kullanarak G ve Phi katsayılarını hesaplamış ve alternatif karar çalışması yapmıştır. Çalışmanın sonucunda puanlayıcıların birbirinden farklı değerlendirme yaptıkları gibi sonuçlar bulunarak puanlayıcı artırma yerine madde sayısını arttırmanın G ve Phi katsayılarını arttırmada daha geçerli olacağını belirtmiştir.

Nalbantoğlu (2009), yapmış olduğu “Performans Ölçümlerinde Genellenebilirlik Kuramıyla Farklı Desenlerin Karşılaştırılması” adlı çalışmada OSCE (objektif yapılandırılmış klinik sınavı) bireylerin performanslarını birden fazla puanlayıcı tarafından puanlanması sonucunda ortaya çıkan desenlerin G ve K çalışma sonuçlarını karşılaştırmıştır. 48 öğrenci ve 3 puanlayıcının yer aldığı çalışma sonuçlarına bakıldığında çaprazlanmış ve yuvalanmış desenlere ilişkin kestirilen varyans değerlerin birbiriyle uyumlu olduğu görülmektedir. Aynı zamanda her iki desenle yapılan karar çalışmaları neticesinde G ve Phi katsayıları birbirine yakın çıkmaktadır.

Güler ve Çetin (2010), tarafından yapılan çalışmada oluşturulan çoktan seçmeli bir testte farklı sayıda gözetmen kullanılması durumunda tek değişkenlik boyutunda bireylerden elde edilen puanların güvenilirlik katsayılarını klasik test kuramı ve genellenebilirlik kuramına göre hesaplaması yapılmıştır. Bu araştırmaya göre G ve Phi katsayıları, iki değişkenlik kaynağının ele alındığı çalışmada tek değişkenlik kaynaklı çalışmalardan elde edilen değerlerden daha büyük çıkmıştır.

Güler ve Gelbal (2010), Klasik Test Kuramı ve Çok değişkenlik kaynaklı rasch modeli ile ilgili yapmış olduğu çalışmada matematik başarısında puanlayıcılar arası güvenilirliğin belirlenmesinde bu yaklaşımlar karşılaştırılmıştır. Elde edilen değerler puanlayıcılar arası uyumun var olduğunu göstermekte, ayırma indeksinin yüksek çıkması sebebiyle puanlayıcıların katılık/cömertlik bakımından birbirinden farklı değerlere sahip olduğu görülmüştür.

Arsan (2012), buz pateni hakem değerlendirmesi üzerine yapmış olduğu çalışmada GK ve ÇDKRM ile analiz edilen güvenilirlik sonuçlarını inceleyerek karşılaştırması yapılmıştır. 397

birey 5 kategoriden oluşan değerlendirme üzerinden 9 hakemin puanladığı çalışmada her iki kuramında birbirinin tamamlayıcısı olduğu ortaya çıkmıştır.

Büyükkıdık (2012), yaptığı çalışmada matematik problemi çözme yeterliliğinin değerlendirilmesine yönelik hazırlanmış olan performans görevlerinin ortaokul öğrencilerinden oluşan 132 öğrenciye uygulanarak analitik ve bütünsel olmak üzere 4 puanlayıcı tarafından puanlanarak KTK ve G Kuramı modelleriyle güvenilirlik katsayıları hesaplanmıştır. Buna göre her iki kurama göre de güvenilirlikler yüksek çıkmıştır.

Bağcı (2015), öğrencilerin matematiksel becerilerinin ölçülmesi üzerine yaptığı araştırmada KTK ve GK'nin çaprazlanmış ve yuvalanmış desenlerin uygulanması neticesinde katsayılar karşılaştırılmıştır. 187 kişiden oluşan 7. Sınıf öğrencilerinin performansları 3 puanlayıcı tarafından puanlanmış ve katsayılar tutarlı sonuçlar verdiği görülmüştür. Ayrıca genellenebilirlik kuramının daha detaylı sonuç verdiği ve çaprazlanmış desenin yuvalanmış desenden elde edilen katsayılara göre daha yüksek olduğu ortaya çıkmıştır.

İlhan (2016), açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyle rasch modeliyle hesaplanmış olan yetenek kestirimlerinin karşılaştırılması isimli çalışmada 8. Sınıf öğrencilerinden oluşan 100 öğrencinin 6 açık uçlu matematik sorusundan oluşan başarı testini cevaplamalarını istemiştir. Cevaplar 4 puanlayıcı tarafından puanlanarak ÇDKRM ve KTK na göre analizleri yapılmıştır. Araştırma sonucunda bu kuramlar arası görel uyumun çok yüksek olduğu ve yetenek kestirimlerinin ÇDKRM nin KTK na göre daha yüksek çıktığı sonucuna varılmıştır.

Anadol (2017), puanlama anahtarlarının güvenilirliğinin farklı deneyim yıllarına sahip puanlayıcıların kullanıldığı durumlarda incelenmesi üzerine yapmış olduğu araştırmada İngilizce yazılı anlatım becerisinin 3 grup puanlayıcı tarafından puanlanması sonucu elde edilen verilerin G çalışması sonuçları karşılaştırılmıştır. Buna göre G ve Phi katsayıların paralellik gösterdiği, puanlayıcıların farklı çalışma yıllarına sahip olmasının dereceli puanlama anahtarının güvenilirliğini etkilemediği sonucuna varılmıştır.

Atılğan (2019), İzmir'deki üç ilçeden seçilen 443 8. Sınıf öğrencileri üzerinden yapmış olduğu kompozisyon değerlendirme çalışmada güvenilirliği ve genellenebilirliği incelemek adına genellenebilirlik kuramını kullanmıştır. Bu çalışmada en uygun puanlayıcı sayısı ile puanlayıcıları seçmede dikkat edilmesi gereken durumlar belirlenmeye çalışılmıştır. Alanında uzman 4 Türkçe öğretmeni tarafından değerlendirilen araştırmada birey(p), madde(i) ve puanlayıcı (r) olmak üzere üç farklı değişken üzerinden çapraz desen kullanılarak analizler elde

edilmiştir. Bu sonuçlara göre birey varyansının en büyük olduğu puanlayıcıların birbiriyle tutarlı puan verdikleri görülmüştür. Bu duruma göre puanlayan puanlayıcıların iki tane olmasının yeterli olacağı puanlamada ekonomiklik sağlanması amacıyla yuvalanmış desenin kullanılmasının daha uygun olacağı sonucuna varılmıştır.

Deniz ve Ilıcan (2021) yapmış oldukları çalışmada gerçek durumdan elde edilen G ve Phi katsayıları ile genellenbilirlik kuramı karar çalışmasıyla elde edilen G ve Phi katsayıları karşılaştırılmışlardır. Farklı illerden 80 tane 7. Sınıf öğrencisinin yer aldığı araştırmada öğrencilerin 12 görevden oluşan yazılı anlatım becerilerinin 4 Türkçe öğretmeni tarafından puanlamışlardır. Çapraz desen üzerinden analiz yapılan araştırmada madde güçlükleri farklı maddeler eklenerek karar çalışmaları yapılmıştır. Buna göre ortaya çıkan G ve Phi katsayı değerlerinin birbirine oldukça yakın olduğu görülmüştür.

Yurt içi çalışmalara göre genellenebilirlik kuramının diğer kuramlarla karşılaştırılmasına yönelik çalışmaların çoğunlukta olduğu genel olarak KTK, ÇDKRM ve G kuramının birbiriyle tutarlı sonuçlar verse de Genellenebilirlik kuramının sonraki yapılacak çalışmalara daha net bilgiler verdiği görülmüştür. Birden fazla puanlayıcı kullanılması durumunda puanlayıcı sayısının artmasıyla güvenilirliğin yüksek çıktığı ancak belli bir puanlayıcı sayısından sonra iş yükü ve çalışmanın ekonomik boyutu artacağından puanlayıcı sayısı artırmak yerine madde sayısının artırmanın daha doğru bir yol olacağı ortaya çıkmaktadır. Ayrıca yapılan çalışmalarda puanlama anahtarı ile yapılan puanlamanın güvenilirliği artırdığı sonucuna ulaşılmıştır.

2.2.2. Yurt Dışındaki Araştırmalar

Bell (1985), araştırmasında genellenebilirlik kuramında dengelenmemiş ve karma desen uygulamalarında kullanılan programların sınırlılıklarına çözüm aranmıştır. Yapılan çalışmada (p), okul(s) ve soru (q) ile belirtilerek çeşitli okullardan gelen öğrencilerin kullanıldığı (p:s) x q deseni ile dengelenmiş (balanced) ve dengelenmemiş (unbalanced) desenler için analiz yapılmıştır. Yapılan araştırma sonrasında MIVQUE yönteminin bazı sıkıntıları ortadan kaldırdığını ve dengelenmiş desenlerde SAS veya GENOVA programlarını kullanmanın daha uygun olacağı sonucuna varılmıştır.

Lane ve Sabers (1989), yapmış oldukları öğrencilerin yazma yetenekleriyle ilgili çalışmada değerlendirme sonrasında puanlama tekniğinin güvenilirliği ölçmek için genellenebilirlik kuramından yararlanılmıştır. 15 öğrencinin verilen bir konu hakkında

kompozisyon yazmaları istenerek 8 uzman tarafından puanlanması yapılmış ve değerlendirmeye alınmıştır. Araştırmada öğrenci (s), puanlayıcı (r) ve puanlama sistemindeki kategori (c) ile belirtilerek kategori değişkeni sabit olarak belirlenerek bütün değişkenlerin çaprazlandığı $s \times r \times c$ deseniyle Genellenebilirlik ve Karar çalışmaları yapılmıştır. Araştırma sonuçlarına göre puanlayıcı sayısının azaldığı durumlarda mutlak ve bağıl hata varyanslarının artarken, G ve Phi katsayılarının azaldığı görülmüştür.

Brown, W. L., O’Gorman, K. Ve Du, Y. (1996), matematik performansın ölçülmesine yönelik geçerlik ve güvenilirliğini incelemiş oldukları araştırmada resmi okullarda uygulanmış olan matematik problem çözme performansını ölçmede psikometrik özellikler araştırılmış ve ölçmenin geçerliğinin yanında puanlayıcılar arası güvenilirliği ve puanlama güvenilirliğine yoğunlaşmıştır. 5. Sınıf öğrencilerinden oluşan araştırmada 1300 öğrencinin kağıdını FACETS modeliyle analiz edilmiştir. Araştırma sonucunda matematikte problem çözme becerisinin ölçülmesi geçerlik ve güvenilirliği doğrulamaktadır.

Engelhard (1996), performans ölçümünde çok değişkenlik rasch modeli kullanarak puanlayıcıların doğruluğunu değerlendirmek adına yeni bir yöntem açıklamıştır. Çalışmada FACETS yöntemiyle 20 puanlayıcının 373 testin puanlarla örneklendirilmiştir. Araştırma sonucunda puanlayıcı doğruluğunda manidar bir fark olduğunu ortaya çıkarmıştır.

Sun ve Valiga (1997), akademik danışmanların öğrenciler tarafından puanlanmasının güvenilirliğini değerlendirmek için Cronbach Alpha güvenilirlik katsayısının sınırlılıklarından hareketle G kuramını kullanarak bir araştırma yapmışlardır. Araştırmada, akademik Danışma Anketinin (Survey of Academic Advising) (SAA) üçüncü bölümü olan ve 18 maddeden oluşan beşli likert tipi “akademik danışma gereksinimleri” bölümünün tesadüfi olarak seçilen 15 enstitü, 150 akademik danışman ve 1500 öğrenci örnekleminde elde edilen veriler kullanılmıştır. Değişkenlik kaynağı olarak; öğrenciler (s), danışmanlar (a), üniversite (c) ve madde (i) tanımlanmış ve $(s:a:c) \times i$ deseniyle G ve K çalışmaları yapılmıştır. Araştırmada, Cronbach Alpha ve G kuramının güvenilirlik yaklaşımları karşılaştırılmış, G kuramı ile daha uygun ve daha doğru olarak güvenilirliğin kestirebildiği sonucuna varılmıştır.

Lynch ve McNamara (1998), Avustralya’da göçmen olarak gelmek isteyen bireylerin dil becerilerine ait performans ölçmek amaçlı kullanılan İkinci Dil Performans Testinin değişkenlik kaynaklarının hangisinin adayların testteki başarısını etkilediğini G kuramı ve Çok Değişken Kaynaklı Rasch Modelini kullanarak araştırmışlardır. Araştırmada her iki yaklaşımın performans temelli değerlendirme işlemlerinin geliştirilmesindeki potansiyel rolleri belirlenmeye çalışılmıştır. Bu amaçla araştırmacılar, 1992 yılında testin konuşma becerileri

bölümünde 83 adaya uygulanan 7 görevin 23 maddenin, 13 uzman puanlayıcı arasından seçilen dört puanlayıcı tarafından altılı derecelenme ölçeğinin kullanılması ile verilen puanlar kullanılarak G kuramı için GENOVA ve Çok Değişken Kaynaklı Rasch Modeli için ise FACETS bilgisayar programlarını kullanmışlar ve her iki programla kestirilen değerleri incelemişlerdir. Araştırma sonucunda iki yaklaşımında farklı kullanılabilir bilgiler sağladığını belirtmişlerdir. K çalışması ile elde edilen verilere dayanarak testin daha kısa olması ve puanlayıcı sayısının ikiye indirilmesi durumunda güvenilirlikte kayıp olup olmadığı belirlenerek testin daha ekonomik ve daha az zaman harcanarak kullanılabileceği sonucuna varmışlardır. FACETS analizi ile de puanlayıcı, görev, madde ve bireylerin tek tek ve birbirleri ile etkileşimlerinin incelenebileceğini vurgulamışlardır.

Banerji (1999), 110 kişiden oluşan bir örnekleme 19 matematik performansı görevinden oluşan bir test vererek elde ettiği verileri, ÇDKRM ve klasik kuramı birlikte kullanarak analiz ederek bir araştırma yapmıştır. Araştırma sonucunda tüm ölçekten ve alt alanlarından elde edilen puanların betimleyici istatistiklerinden, becerinin yaşla arttığı standart başarı testinin matematik bataryasıyla elde edilen ölçüt geçerliklerinin .22 ve .57 arasında, puanlayıcılar arası güvenilirlik katsayılarının ortancasının .80 ve toplam yetenek puanlarının iç tutarlılık güvenilirliklerinin .90 ve .92 olduğunu bulunmuştur. Araştırmada, görevlerin klasik madde istatistikleri ve rasch logit güçlükleri, sıralanmış ölçek yapısında uygulanmış, fakat sekiz maddenin Rasch ölçütlerine uymadığı sonucuna varılmıştır.

Sthulmann ve diğerleri (1999), 23'ü kurallı puanlama ölçeği (rubric) boyutların yorumlamada eğitilmiş ve 17'si eğitilmemiş olan toplam 40 anaokulu ve birinci sınıf öğretmeninden oluşan bir öğretmen grubunun birinci sınıf öğrencilerinin hazırladıkları 20 yazma örneğini bir kurallı puanlama ölçeğiyle (rubric) derecelerinden elde edilen veriler üzerinden öğrencilerin yazımlarının kurallı puanlama ölçeği (rubric) kullanılarak puanlanmasında öğretmenlere eğitim verilmesinin güvenilirliğe etkisini G kuramı kullanarak araştırmışlardır. Bu amaçla araştırmada kısmi-yuvalanmış desen kullanılarak kurallı puanlama ölçeğinin puanlama boyutu ve toplam ham puanlarına dayalı her bir okul için (eğitilmiş ve eğitimsizler) varyanslar kestirilerek genellenebilirlik katsayıları hesaplanmıştır. Araştırma bulgularından öğretmenlerin kurallı puanlama ölçeklerinin puanlanması konusunda eğitilmiş olmalarının güvenilirliği artırmadığı sonucuna varılmıştır.

Tobar ve ark. (1999), 6 alt ölçek ve 65 maddeden üzerinden duygu durum profili ölçeğinin güvenilirliğini G kuramı ile araştırmışlardır. Araştırma sonucunda tesadüfi desende

madde sayısının arttırıldığında G katsayısının arttığı, alt ölçeklerin sabit değişken kaynağı olarak tanımlanması durumunda G katsayısının yükseldiği görülmüştür.

Johnson ve ark., (2000), dereceli puanlama anahtarını (rubric) kullanarak yapmış oldukları çalışma puanların çözümü ve puanlayıcılar arasındaki tutarlılığıyla beraber iki puanlayıcının puanlanması ve puanlayıcı uyumu, asıl puanlayıcıların puanları yerine bir uzmanın puanlarının kullanılması ve iki puanlayıcı ile uzman puanlayıcının puanların kullanılması gibi seçenekler üzerinden ortaya çıkan puanların kararlılığı incelenilmiştir. Georgia Lisesi yazma testinden alınarak kullanılan 120 kompozisyon araştırma verileri, birden fazla puanlayıcının ve bir çift uzmanın farklı alan kriterlerini değerlendirerek vermiş oldukları puanlarla puanlayıcılar arası güvenilirlik uyum yüzdesi, Spearman sıra farkları korelasyonu ve G kuramının Phi katsayısı ile hesaplanmıştır. Bunun sonucunda puanlayıcıların belli bir eğitimin geçmesi gerektiği ve puanlayıcıları arttırmanın puanların güvenilirliğini olumlu yönde etkilediği görülmüştür.

MacMillian (2000), 4930 öğrencinin verdikleri yanıtların 70 puanlayıcı havuzundan seçilen 3 puanlayıcı tarafından dokuzlu dereceleme ölçeğinde verdikleri puanları kullanarak bir araştırma yapmıştır. Araştırmada klasik test kuramı (KTK), genellenebilirlik kuramı (GK) ve Çok Değişken Kaynaklı Rasch Modeli (ÇDKRM) yaklaşımlarını kullanarak puanlayıcılar arasındaki değişmeyi ve bu değişimin düzeltilmesini araştırarak karşılaştırmıştır. KTK ve ÇDKRM yaklaşımları puanlayıcılar arasında farklılığın önemli boyutta olduğu, buna karşılık GK’da puanlayıcı varyans bileşenleri ve Rasch histogramları oldukça az puanlayıcı değişimi izlenimi vermiştir. KTK ve ÇDKRM düzeltme işlemleri bireylerin yarısından daha fazlası için farklı puanlar üretmiş fakat her bir düzeltmeden sonra bireylerin %75’i için aynı sonuçlar elde edilmiştir. Macmillan (2000) KTK, G Kuramı ve ÇDKRM’ ni büyük ölçekli performans değerlendirme durumlarında puanlayıcı katılığı, tutarlılığı ve uyumu bakımından karşılaştırdığı çalışmasında 3 farklı göreve katılan 4930 birey ve 70 puanlayıcı havuzundan seçilen 3 puanlayıcı tarafından puanlanmıştır. Araştırma sonucunda bu yaklaşımlarının her durumda aynı sonucu vermediği duruma göre yaklaşımların seçilmesi gerektiği sonucu ortaya çıkmıştır.

Lee ve ark, (2002), yeni TOEFL sınav sisteminin yazma ve konuşma bölümlerinin puanlarının güvenilirliği üzerine yapmış olduğu bir çalışmada G kuramı ile araştırmışlardır. Üç aşamada yapmış olduğu çalışmanın ilk bölümünde 488 bireye sorulan okuma-yazma ve dinleme-yazma becerileri 27 kişiden oluşan puanlayıcı tarafından 5 kategorili bir ölçek üzerinden puanlanmıştır. İkinci bölümde üçüncü alt grubun, altı göreve ilişkin yazdıkları yazı her biri aynı 6 puanlayıcı tarafından puanlanarak G kuramında birey, görev ve puanlayıcıdan

oluşan b x g x p çapraz desenle incelenmiştir. Üçüncü bölümde ise konuşma yeteneklerine bağlı olarak 261 adaya yöneltilmiş ve bireylerin tepkileri 5 kategorili bir ölçekte 22 kişilik bir puanlayıcı arasından bağımsız iki puanlayıcı tarafından puanlanmıştır. Analizlerde çıkan sonuca göre her bir görev için yazma ve konuşma bölümlerinde puanlayıcıların artırılmasının ekonomik açıdan zor olacağı, bundan dolayı görev sayısının artırılmasının puanların güvenilirliğinin artırmada daha etkili olduğu sonucuna varmıştır.

Nakamura (2002), öğrenciler ile beraber öğretmenlerin İngilizce dilinde sözel sunum yeteneklerini ölçmek adına yapmış olduğu çalışmada puanların güvenilirliğini çok değişkenlik Rasch modeli vasıtasıyla incelenmiştir. Araştırma sonuçlara göre her bir değişkenlik kaynağında uyumsuzluk gösteren elemanların puanlamadan çıkarılıp yeniden yapılacak puanlamalarda ölçmenin daha güvenilir sonuçlar verecek şekilde nasıl geliştirilebileceğine yönelik açıklamalara değinmiştir.

Monahan ve Schumacker (2003), Genel Eğitimsel Gelişme testi ile ÇDKRM ve G kuramı yaklaşımlarını kullanarak bir araştırma yapmışlardır. Araştırmada, üç farklı yargıcının (bir deneyimli GED öğretmeni, bir özel eğitim öğretmeni ve bir deneyimsiz GED öğretmeni) 15-20 yaşlar arasındaki öğrenme bozukluğu olan dokuz erkek gencin, yirmişer dakikalık iki oturumda yazdıkları beş paragraflık GED kompozisyonlarının 1-4 ölçeğinde puanlaması sonucunda elde edilen 27 GED kompozisyon puanını veri olarak kullanılmıştır. Araştırmada ÇDKRM ve G kuramı kullanılarak “üç farklı öğretmenin GED öğrencilerinin kompozisyonlarını farklı puanlayıp puanlamadıkları” araştırılmıştır. Analizlerle kestirilen ÇDKRM ve G kuramının parametrelerinin karşılaştırılması sonucunda, ÇDKRM’nin öğrencilerin GED kompozisyonlarının puanlanmasında eğitimli yargıcılar için daha tutarlı ve daha az yanlılık olduğu ve de yararlı bilgiler sağladığı belirtilmiştir.

Lee ve diğerleri (2003), deneme (kompozisyon) yazma sınavı üzerinde puanlayıcılar, görevler ve bireylerin birbirleriyle etkileşimlerini / yanlılıklarını ÇDKRM ile analiz ederek tamamlayıcı ve bağımsız görevlere bireylerin verdikleri tepkilerin elde edilen puanlarındaki puanlayıcılarla ilgili potansiyel hataları araştırmışlardır. Araştırmada 6 akademik yazma görevi için 162 bireyden elde edilen 970 yazı örneğinin, 6 eğitimli puanlayıcı tarafından beş kategorili bir ölçekte puanlaması ile elde edilen veriler kullanılmıştır. Puanlayıcıların, görevlerin, bireyleri ayrı ayrı olduğu kadar, birbirleriyle etkileşimlerinin tüm olası kombinasyonlarının yanlılık analizleri FACETS bilgisayar programı kullanılarak yapılmıştır. Analizler sonucunda; birey yetenek ölçümlerinin yayılımının, görev güçlüğü ölçümlerinden ve puanlayıcı katılıklarından daha geniş, değerlendirme hatasını n en büyük kaynağını puanlayıcılardan çok

görevlerle ilişkili olduğu ve de görevlerde puanlayıcıların tutarlılıklarını korumada bazı güçlerin olduğu görülmüştür.

Turner (2003), öğrencilerin resim yeteneklerinin ölçülmesi adına yaptığı portfolyo değerlendirmesinin güvenilirliğini test etmek için çok değişkenlik kaynaklı rasch modelini kullanmıştır. Araştırma için 12 portfolyo projesini 15 öğrenci tarafından tamamlanmasıyla ortaya çıkan verileri 10 puanlayıcı resim yeteneği tahminin üstünde, 3 öğrencinin resim yeteneği de tahminin altında bulunmuştur.

Hetherman (2004), İngilizce kompozisyon yazma testinin etkinliğini rasch ölçme modeli kullanarak araştırmak için çok değişkenlik rasch ölçme modelini kullanarak 230 örneklem üzerinden elde ettiği sonuçlarda bireylerdeki dil yeteneği düzeyinin 0 ve 1 olmak üzere iki farklı seviyede olduğunu görmüştür. Bununla beraber dil yeteneği güçlük seviyesinin -1.68 ile 1.33 arasında değiştiği sonucuna ulaşmıştır. Ayrıca puanlayıcılar arası güvenilirliğin yüksek ve z puanlarının da -1.48 ile 3.5 arasında olduğu görülmüştür.

Kozaki (2004), Japonca'dan İngilizce'ye tercüme yapma sertifikalandırılması için kullanılan performans değerlendirmelerinde ÇDKRM ve G Kuramını karşılaştırmışlardır. Buna göre puanlayıcı özellikleri ile ilgili ÇDKRM, G Kuramına göre bireysel seviyede puanlayıcı etkilerinin araştırılmasına imkan tanıdığı için daha uygun bir teknik olduğu ortaya çıkmaktadır.

Smith ve Kulikowich (2004), karmaşık problem çözüme becerilerinin değerlendirilmesinden elde edilen puanları kullanarak genellenebilirlik ve ÇDKRM'ni karşılaştırmışlardır. 4. Sınıf öğrencilerinden 44 öğrenci ve 2 puanlayıcının katılımıyla gerçekleştirilen bu araştırma sonucunda değişkenlik kaynakları arasındaki değişimin göreceli büyüklüğünün karşılaştırılabilir olduğu ortaya çıkmıştır.

Sudweeks, Reeve ve Bradshaw (2005), üniversite öğrencilerinin yazılarıyla ilgili değerlendirmede G Kuramı ve ÇDKRM'ni karşılaştırmışlardır. Buna göre her iki kuramında kendine göre güçlü ve zayıf yönlerinin bulunduğu ve araştırma amacına göre her iki kuramından kullanımında farklı faydalar sağlayacağı belirtilmiştir.

Wang (2005), genellenebilirlik kuramı ve geleneksel yöntemiyle güvenilirlik hesaplamada bir karşılaştırma yapmıştır. Buna göre öğrencilerin yazma becerilerini ölçmek adına hazırlanan sınav verileri sonucunda öğrenci (p), madde (i) ve beceri (t) üzerinden $p \times (i: t)$ deseniyle G çalışması yapılmıştır.

Alharby (2006), arařtırmasında bütüncül ve analitik yaklaşımları karşılařtırdığı alışmasında G kuramı ve DKRM ile ölçümlerin güvenilirliğini analiz etmiştir. Buna göre analitik puanlama yöntemi bütüncül puanlamaya göre daha yüksek G katsayısına sahip olduėu, DKRM uygulandığında bütüncül yaklaşımın analitik yaklaşıma göre daha iyi uyum sağladığı görülmüřtür.

Lee ve Kantor (2007), TOEFL sınavı yazma yetilerinin ölçmek için alternatif deėerlendirme yöntemleri ve görev örnekleri belirlemek amacıyla genellenebilirlik kuramını kullanmışlardır. alışma sonucunda en yüksek puan güvenilirliği için sınavda kullanılan yazma görevlerinin sayısını artırmanın görev başına düşen puanlayıcı sayısını artırmaktan daha iyi sonuçlar vereceğini belirtmiştir. Bunun yanında dinleme-yazma görev sayısının, okuma-yazma görevleri sayısından daha çok olduğunda daha yüksek puan güvenilirliği elde edildiėi sonucuna ulaşmıştır.

Kim ve Wilson (2009), öğrencilerin yazma becerilerinin deėerlendirilmesinde Genellenebilirlik kuramı ve DKRM karşılaştırılmıştır. Bu alışmada her iki modelin de farklı deėişkenler hakkında detaylı analiz ürettiėi, arařtırmacının grupların alışmasına yönelik incelenmesi durumlarında g kuramının daha üstün durumda olduğuna ancak bireysel performansa yönelik alışmalarda DKRM kullanmanın faydalı olacağı sonucuna varılmaktadır.

Sebok (2010), bir üniversitenin yüksek lisans öğrenci seçme sürecini DKRM ve G Kuramı kullanarak deėerlendirmiştir. Her iki yöntemin de olumlu ve olumsuz yanların bulunduėu belirtilmektedir. Küçük örneklem durumunda DKRM olumlu sonuçlar verdiği uygun örneklem büyüklüėü ve uyum istatistikleri gibi durumlarda bazı sınırlılıkların bulunduėu belirtilmiştir. Genellenebilirlik kuramı ise her bir bileşenleri ayrı ayrı varyanslarda açıkladığı farklı desenler üzerinde karar alışmaları yaptığı sonucuna varmaktadır.

Yang (2010), sözel İngilizce yeterlik testinde yaptığı alışmanın deėerlendirmesinde puanlayıcıların bireylerin puanlarındaki katılık etkisini arařtırmıştır. Bu alışma sonucunda puanlayıcıların farklı katılık düzeyinde olduğuna fakat test puanlarındaki hakem katılığın etkisi çok küçük olduğuna görülmektedir.

Brennan (2011), KTK, ve GK karşılaştırılarak bu iki kuramın MTK ile kıyaslamaları üzerinde durulmuřtur. Bu üç kuramın da farklı güçlü yönlerinin ve zayıf yönlerinin olduğuna ortaya çıkmıştır.

Ure (2011), misyoner öğretime ilişkin performans deėerlendirme üzerine yapmış olduğuna arařtırmada iki farklı koşuldan oluşan puanlayıcı davranışını ve deėerlendirmenin güvenilirliğini

nasıl etkilediğini araştırdığı çalışmada ÇDKRM ve GK'nı güvenilirlik belirleme performanslarını karşılaştırmıştır. Her iki kuramın da farklı amaçlar doğrultusunda avantajlarının bulunduğu ancak daha detaylı bilgi vermesi nedeniyle GK'nın üstün konumda yer aldığı sonucuna varılmıştır.

Genel olarak yapılan araştırmalara bakıldığında G kuramının kullanıldığı araştırmaların, kuramsal özelliklerin incelenmesi ve diğer kuramlarla karşılaştırılması ile ilgili olduğu görülmektedir. Ayrıca G kuramının diğer kuramlara göre daha açıklayıcı sonuçlar verdiği görülse de G kuramının kullanılması ve yorumlanmasının zor olması kuramın bir handikapı olarak görülmüştür. Rasch modeli ise tüm değişkenlik kaynaklarının tek bir ölçek üzerinden değerlendirilmesi ve her bireye değişkenlik kaynağı açısından ayrı güvenilirlik hesaplamasının olanak dahilinde olması fakat verinin az sayıda olması durumunda kullanılmasının mümkün olmadığı ve yorumlamanın zor olduğu sonucuna ulaşılmaktadır. Sonuç olarak araştırmanın amacına bağlı olarak bu kuramların kullanılmasının geçerli ve uygun olacağı görüşü yaygındır.

BÖLÜM III

YÖNTEM

Bu başlık altında araştırmanın türü, çalışma grubu, veri toplama araçları, verilerin toplanması ve verilerin analizi ilgili açıklamalara yer verilmiştir.

3.1.Araştırmanın Modeli

Araştırmaya konu olan bu çalışmayı puanlayıcıların Genellenebilirlik (G) Kuramı ve Çok Değişkenlik Kaynaklı Rasch Modelinin (ÇDKRM) gerçek verilerle uygulanmasını, gerçek verilerle bu iki kuramsal yaklaşımın uygulanabilirliklerinin belirlenmesini sağlayarak puanlayıcı değişkenliğinin güvenilirliğini analiz etmeye dayalıdır.

Ortaöğretim 12. sınıf öğrencilerinin İngilizce yeterliliği üzerine yapılan ölçmenin farklı okulların, birden fazla puanlayıcıların puanlama yönergesi ve ölçeksiz değerlendirmede bu yaklaşımlara dayalı olarak güvenilirliğini test etmek, tutarlılık ve farklılıkları saptamak açısından betimsel bir araştırmadır.

3.2.Araştırma Grubu

Araştırmanın grubu; 2022-2023 eğitim-öğretim yılı dönemlerinde Antalya ili Elmalı ve Kumluca ilçelerindeki 4 farklı resmi ve özel ortaöğretim kurumlarından değerlendirilen her okuldan eşit sayıda öğrencinin katılımıyla 45 erkek, 79 kız öğrenci olmak üzere toplam 124 kişilik 12. sınıf öğrencileri ve puanlayıcı olarak 3 uzman İngilizce öğretmeninden oluşmaktadır. Bu bağlamda puanlayıcı olarak biri bütünsel puanlama anahtarıyla diğer ikisi ölçeksiz olmak üzere 3 farklı puanlayıcı 4 farklı okuldan eşit sayıda değerlendirmeye katılan toplam 124 öğrencisini puanlamıştır.

3.3.Verİ Toplama Araçları

Araştırmada farklı okullarda yer alan 12. sınıf öğrencilerine program kazanımları doğrultusunda açık uçlu bir performans testi verilerek çerçevesi belli bir konuda kompozisyon yazmaları istenmiştir.

Performans testini toplamda 200'e yakın öğrenci cevaplamıştır. Cevaplama yapan öğrenciler içinden her okuldan 31'er öğrencinin testi seçkisiz olarak seçilerek değerlendirmeye alınmıştır. Farklı okullardan 12. sınıf öğrencilerine uygulanan performans testinin sonuçları 1 tanesi ölçekli 2 tanesi ölçeksiz olmak üzere 3 uzman İngilizce öğretmenleri tarafından puanlanması istenmiştir. Değerlendirmede kullanılmış olan ölçek 4 aşama şeklinde uzman İngilizce öğretmenleri tarafından kazanımlara uygun olarak hazırlanmıştır. Hazırlanan bu ölçek bir ölçme ve değerlendirme alanında uzman ve bir İngilizce eğitiminde uzman akademisyen tarafından incelenerek değerlendirmede kullanılmaya uygun görülmüştür. (Ek 2).

3.4.Verilerin Toplanması

Araştırmanın amacına göre puanlama yönergesi öncelikle ölçülecek özelliğe bağlı olarak dereceli puanlama anahtarıyla değerlendirilmesine karar verilmiştir. Performans ölçütüne göre performansın gözlenebilir özellikleri açıkça belirtilmiş ve uzman görüşü alarak gözden geçirilmiştir (Airasian, 1991).

Belirlenen ölçek uygulanacak olan sınıf düzeyine uygun bir şekilde uzman öğretmenler tarafından hazırlanıp yerinde ve bir ders saati içinde öğrencilere uygulamayla ilgili yeterli bilgi verilmiştir. Derecelendirme ölçeği kullanacak puanlayıcı tarafından ölçek incelenmiş ve maddenin puanlanması sırasında dikkat edilmesi gereken hususlar belirlenmiştir. Toplanan verilerin ikisi ölçeksiz biri ölçekli olacak şekilde 3 uzman tarafından 0 ile 100 puan arasında ön plana çıkarılan öğrenme hedeflerinin doğrultusunda öğrencilerin performansı değerlendirilmiştir. Değerlendirme süreci tesadüfi hatayı en aza çekmek adına yaklaşık 1 aya dengeli bir şekilde dağıtılmış her uzman birbirinden bağımsız olacak şekilde aynı verileri puanlamıştır. Öğrencilerin kişisel bilgileri gizli tutularak veriler analiz edilmiştir.

3.5.Verilerin Analizi

Puanlayıcılar, araştırma içerisinde iki uzman ölçeksiz, bir uzman puanlama yönergesine göre aynı verileri birbirinden bağımsız puanlama yapmışlardır. Buna göre puanlama yönergesini kullanan ve kullanmayan puanlayıcıların verdikleri puanlar kuramlar bakımından güvenilirliği ve aynı zamanda manidar bir fark olup olmadığı tespit edilmeye çalışılmıştır. Verilerin elde edilmesinin sonrasında Genellenebilirlik kuramının yanında Çok yüzeyli rasch modeli ile de güvenilirliğe bakılarak aradaki ilişki hakkında bilgi edinilmeye çalışılmıştır.

Arařtırmada G Kuramı analizi kapsamında kullanılan veriler facets olarak adlandırılan deęiřkenlik kaynaklarının birey, puanlayıcı ve okul olarak belirlenmiřtir. Verilerin analizi sırasında G Kuramına iliřkin analizlerde G alıřması ve karar alıřmaları iin okulların ayrı ayrı istatistikleri Edu G 6.01 ile DKRM analizlerinde; FACETS programlarıyla model- veri uyumu, i ve dıř uyum ki-kare ve ayırt etme indeksi incelenmiř bununla beraber betimsel istatistik iin SPSS 26. programı kullanılmıřtır.



BÖLÜM IV

BULGULAR

Bu bölümde, G Kuramına ait çaprazlanmış desen kullanımına ait ve CDKRM puanlama desenine ait puanlayıcılar arası güvenilirliği incelemek amacıyla elde edilen ölçüm raporlar yer almış ve yorumlanmıştır.

Tablo 4.1. 2022-2023 Öğretim Yılı 12. Sınıf Öğrencilerinden Elde Edilen Örneklemen İngilizce Performans Değerlendirmesi Sonucu Alınan Puanlara İlişkin Betimsel İstatistikler

İstatistikler	Puanlayıcı 1 (ölçeksiz)	Puanlayıcı 2 (ölçeksiz)	Puanlayıcı 3 (ölçekli)
N	124	124	124
Erkek	45	45	45
Kız	79	79	79
Ortalama	60,56	49,25	46,33
Ortanca	65,0	50,0	50,0
Std sapma	27,04	22,09	27,97
Varyans	731,39	487,76	782,15
Çarpıklık	-0,66	-0,13	-0,12
Basıklık	-0,52	-0,33	-1,01

*Buna göre değerlendirmeye alınan öğrenci sayılarından %36,3 erkek iken %63,7' si kızdır.

Tablo 4.1'de görüldüğü üzere 124 öğrencinin bir performans yeterliliği üzerinden aldıkları puanlara ilişkin en yüksek ortalama 60,56 ile 1. puanlayıcıya aitken, en düşük ortalama ölçekle puanlama yapan 3. puanlayıcıya (46,33) aittir. En homojen puanlamayı ikinci puanlayıcı yapmıştır. Ortanca, her bir puanlayıcı için aritmetik ortalamadan büyüktür. Çarpıklık değerlerine bakıldığında negatif değer aldıkları için sola çarpık dağılım göstermektedir. Basıklık değeri 0'dan küçük değer alması nedeniyle normalden daha basık dağılım göstermektedir.

4.1. Genellenebilirlik Analizi

Bu bölümde tek değişkenli performansların değerlendirilmesi sonucu ortaya çıkan varyans değerlerini belirlemek adına 4 okul için ayrı ayrı G Kuramının bxp çapraz desen modelleri kullanılarak varyans kaynaklarının ve toplam varyansın belirlenmesi için G çalışması yapılmıştır. Değerlendirmeye alınan 124 öğrenci, 3 puanlayıcı ve 4 okuldan oluşan veri grubunun varyans ve toplam varyans açıklama yüzdeleri b ve p ana etkileri ile bp, ortak etkileri aşağıdaki tablo 4.2’ de verilmiştir.

Tablo 4.2. Örnekleme Ait Tek Değişkenli G Çalışması Sonucunda Anadolu Lisesi’nin Varyans Değerleri

Varyans kaynağı	Kareler toplamı SS	Sd	Kareler ortalaması MS	Bileşenler				
				Rastgele	Karışık	Varyans	%	SH
B	37419.570	30	1247.319	378.371	378.371	378.371	62.5	104.16
P	7351.634	2	3675.817	114.955	114.955	114.955	19.0	83.85
Bp	6732.366	60	112.206	112.206	112.206	112.206	18.5	20.15
Toplam	51503.570	92					100%	

Tablo 4.2.’de Anadolu Lisesi’nin tek değişkenli G çalışması sonrasında elde edilen varyanslar ve yüzdeleri yukarıda açıklanmıştır. Bu tabloya göre bireylere ait varyans bileşeni beklenen değer olarak en büyük değer aldığı görülmekte (378.37) ve toplam varyansın.%62.5’ini oluşturmaktadır. Bu değere göre bireylerin ölçülen özellik bakımından bakıldığında hepsi birbirinden yeterlilik düzeyi bakımından farklılık gösterdiği görülmektedir. Aynı zamanda toplam varyans bileşeninde birey varyansın çok büyük bir orana sahip olması evren puanı olarak değerlendirildiğinden beklenen bir durum olarak kabul edilir (Shavelson ve Webb, 1991; Brennan 2001). Bu aynı zamanda bireylerin ölçülen özellik bakımından birbirinden çeşitlilik gösterdiği anlamına gelmektedir.

Puanlayıcı varyans değeri ise (114.96) %19.0 olarak hesaplanmıştır. Bu da puanlayıcıların bireylere verdikleri puanların farklılık gösterdiği yani katılık veya cömertlik durumunun var olduğunu göstermektedir.

Birey x puanlayıcı (bxp) artık varyans etkisi (112.21) toplam varyansın %18.5’ini açıklamaktadır. Bu da tesadüfi hatanın yüksek olduğunun yani puanlayıcıların puanlamada bireyleri puanlamada çeşitli hata kaynaklarının bulunduğunu göstermektedir. Artık varyans en küçük varyans değeri olsa da oran olarak puanlayıcı varyansına yakın ve nispeten beklenenden yüksek hesaplanmıştır.

Genel olarak çok deęişkenli G alıřmalarıyla elde edilen varyans deęerleri ile varyans aıklama oranlarına bakıldığında varyans bileřenlerinin toplam varyans aıklama oranının farklılıklar gösterdiği görlmektedir.

G kuramı analizine gre ok deęişkenli (b x p) modeli ile farklı puanlayıcı sayıları senaryolarına gre yapılan K alıřmaları sonucunda kestirilen G ve Phi katsayıları, alt boyutlara gre ařaęıda tablo 4.3. deęerler verilmiřtir.

Tablo 4.3. Anadolu Lisesi'nin rneklemin ok Deęişkenli Model K alıřması İle Farklı Sayıdaki Puanlayıcı Sayıları Senaryolarına Gre G ve Phi Katsayıları

	G-alıřması		Senaryo 1		Senaryo 2		Senaryo 3		Senaryo 4		Senaryo 5	
	Dzy	Evr	Dzy	Evr	Dzy	Evr	Dzy	Evr	Dzy	Evr	Dzy	Evr
B	31	Snsz	31	Snsz	31	Snsz	31	Snsz	31	Snsz	31	Snsz
P	3	Snsz	1	Snsz	2	Snsz	4	Snsz	5	Snsz	6	Snsz
Gzlem	93		31		62		124		155		186	
G katsayısı	0.91		0.77		0.87		0.93		0.94		0.95	
Phi katsayısı	0.83		0.62		0.77		0.87		0.89		0.91	
Baęıl Hata Varyansı	37.402		112.206		56.103		28.052		22.441		18.701	
Rel. Std. Err. of M.	6.116		10.593		7.490		5.296		4.737		4.324	
Mutlak hata Varyansı	75.720		227.161		113.581		56.790		45.432		37.860	
Abs. Std. Err. of M.	8.702		15.072		10.657		7.536		6.740		6.153	
Dzy: dzey		Evr: evren		Snsz: sonsuz								

Tablo 4.3.'de farklı boyutlardaki ok deęişkenli K alıřması sonucunda rneklemdaki birey-puanlayıcı sayıları senaryolarına gre kestirilen G katsayıları aıklanmaya alıřılmıştır. rneklemdaki verilerden yola ıkarak puanlayıcı sayıların artırılıp azaltılması ve bireylerin aynı řekilde sabit tutulması senaryoları karřısında G katsayılarının deęiřiklięi incelenmeye alıřılmıştır.

Verilen tek soruluk performans deęerlendirme kompozisyon sorusunda her okuldaki 31 ğrenci alınarak toplamda 4 farklı okul ve 3 puanlayıcı iinden 1. okulun K alıřması ile elde edilen G katsayısı 0,91 hesaplanmıřtır. Bu boyut, iin, birey sayısı sabit tutularak puanlayıcı sayısı 2 azaltılması ($n_p=1$) durumunda kestirilen G katsayısı 0,77 birey sayısı sabit tutularak puanlayıcı sayısı 1 azaltılarak ($n_p=2$) durumunda ise G katsayısı 0,87 olarak hesaplanmıřtır. Birey sayısının sabit tutulmasıyla puanlayıcı sayısının bir arttırılmasıyla ($n_p=4$) G katsayısı 0.93 iki arttırılmasıyla ($n_p=5$) g katsayısı 0.94  arttırılmasıyla ($n_p=5$) g katsayısı 0.95 olarak bulunmuřtur. Bir kiřiden oluřması durumunda kestirilen G katsayısında 0.14 azalma olurken puanlayıcı sayısı altı olduęunda G katsayısındaki artıř 0,04 olmaktadır. Tablo 4.3'e gre

puanlayıcı sayısının azaltılıp veya artırılması durumunda G katsayısında fazla bir değişkenlik olmadığı görülmektedir.

bxp modelinin orijinal puanlayıcı ve birey sayılarına göre Phi katsayısı 0.83 olarak hesaplanmıştır. Farklı puanlayıcı senaryolarına göre yapılan K çalışmaları sonucunda elde edilen Phi katsayıları, yukarıdaki tabloda verilmiştir. Buna göre puanlayıcı sayılarının birer arttırılıp birer azaltılması sonucunda phi katsayıları değişkenlik göstermektedir. Puanlayıcı sayısını iki azaltıldığında ($n_p=1$) Phi katsayısı 0.62 iken, bir azaltıldığında ($n_p=2$) Phi katsayısı 0.77 olarak hesaplanmıştır. Bunun yanında puanlayıcı sayısı bir arttığında ($n_p=4$) Phi katsayısı 0.87, iki arttırıldığında ($n_p=5$) 0.89 ve üç arttırıldığında ($n_p=6$) 0.91 olarak artış olduğu görülmüştür. Bu tabloya göre puanlayıcı sayısını artırma veya azaltmanın Phi katsayısını çok fazla değiştirmemiştir. Bu duruma göre G katsayıların Phi katsayılardan daha yüksek çıktığı görülmekte aynı zamanda g katsayıların puanlayıcı senaryolarına göre Phi katsayıları da paralel olarak artmakta ya da azalmakta olduğu görülmektedir.

Tablo 4.4. Puanlayıcıların Anadolu Lisesi'ne göre G ve Phi Katsayıları

G-Facets analysis

Yüzey	Düzey	Bağıl G Katsayısı	Mutlak G Katsayısı
P	1	0.848	0.791
	2	0.882	0.722
	3	0.881	0.822

Tablo 4.4'e göre üç puanlayıcıdan birinci ve ikinci puanlayıcı ölçeğe puanlama yaparken üçüncü puanlayıcı ölçekli puanlama yapmıştır. Birinci puanlayıcının G katsayısı 0.85 iken ikinci ve üçüncü puanlayıcının 0.88 düzeyindedir. Birinci puanlayıcının Phi katsayısı 0.79, ikinci puanlayıcının Phi katsayısı 0.72 ve üçüncü puanlayıcının Phi katsayısı 0.82 olarak hesaplanmıştır.

Tablo 4.5. G Çalışması Sonucu Fen Lisesi'nin Varyans Değerleri

Varyans kaynağı	Kareler toplamı SS	Sd	Kareler ortalaması MS	Bileşenler				
				Rastgele	Karışık	Varyans	%	SH
B	7079.699	30	235.990	52.604	52.604	52.604	23.2	20.215
P	6101.312	2	3050.656	95.886	95.886	95.886	42.3	69.587
BP	4690.688	60	78.178	78.178	78.178	78.178	34.5	14.041
Toplam	17871.699	92					100%	

Tablo 4.5.'e Fen Lisesi'nin tek değişkenli G çalışması sonrasında elde edilen varyanslar ve yüzdeleri aşağıda açıklanmıştır. Buna göre bu tabloya göre bireylere ait varyans bileşeninin hem kendi içinde hem de diğer okulların birey varyansı ile karşılaştırıldığında en küçük değer aldığı görülmekte (52.60) ve toplam varyansın %23.2'sini oluşturmaktadır. Bu değere göre bireylerin ölçülen özellik bakımından bakıldığında birbirine yakın değerler gösterdiği görülmektedir. Aynı zamanda toplam varyans bileşeninde birey varyansın çok küçük bir orana sahip olması beklenen bir durum olarak görülmemekte ve bireylerin ölçülen özellik bakımından birbirine yakın değerler aldığı anlamına gelmektedir.

Puanlayıcı varyans değeri ise (95.89) %42.3 olarak en büyük değer olarak karşımıza çıkmaktadır. Bu da puanlayıcıların bireylere verdikleri puanın oldukça birbirinden farklılık gösterdiği yani katılık ve cömertlik durumunun büyük oranda var olduğunu göstermektedir.

Birey x puanlayıcı (bxp) artık varyans etkisi (78.178) toplam varyansın %34.5'ini açıklamaktadır. Bu da tesadüfi hatanın yüksek olduğunun yani puanlayıcıların puanlamada bireyler hakkında bilinmeyen hata kaynaklarının yüksek olduğunu göstermektedir.

G kuramı analizine göre çok değişkenli (b x p) modeli ile farklı okul ve puanlayıcı sayıları senaryolarına göre yapılan K çalışmaları sonucunda elde edilen G ve Phi katsayıları, alt boyutlara göre aşağıda tablo 4.6. değerler verilmiştir.

Tablo 4.6. Fen Lisesi Örneklemin Çok Değişkenli Model K Çalışması İle Farklı Sayıdaki Okul Ve Puanlayıcı Sayıları Senaryolarına Göre G ve Phi Katsayıları

	G-çalışması		Senaryo 1		Senaryo 2		Senaryo n 3		Senaryo 4		Senaryo 5	
	Dzy	Evr	Dzy	Evr	Dzy	Evr	Dzy	Evr	Dzy	Evr	Dzy	Evr
B	31	Snsz	31	Snsz	31	Snsz	31	Snsz	31	Snsz	31	Snsz
P	3	Snsz	1	Snsz	2	Snsz	4	Snsz	5	Snsz	6	Snsz
Gözlem	93		31		62		124		155		186	
G katsayısı	0.67		0.40		0.57		0.73		0.77		0.80	
Phi katsayısı	0.48		0.23		0.38		0.55		0.60		0.64	
Bağıl hata varyansı	26.059		78.178		39.089		19.545		15.636		13.030	
Rel. Std. Err. of M.	5.105		8.842		6.252		4.421		3.954		3.610	
Mutlak hata varyansı	58.022		174.065		87.032		43.516		34.813		29.011	
Abs. Std. Err. of M.	7.617		13.193		9.329		6.597		5.900		5.386	

Tablo 4.6' da farklı boyutlardaki çok değişkenli K çalışması sonucunda örneklemdaki birey-puanlayıcı sayıları senaryolarına göre kestirilen G katsayıları açıklanmaya çalışılmıştır. Örneklemdaki verilerden yola çıkarak puanlayıcı sayılarının artırılıp azaltılması ve bireylerin

aynı şekilde sabit tutulması senaryoları karşısında G katsayılarının değışikliğı incelenmeye çalışılmıştır.

Verilen tek soruluk performans değeriendirme kompozisyon sorusunda her okuldan 31 öğrenci alınarak toplamda 4 farklı okul ve 3 puanlayıcı içinden ikinci okulun K çalışması ile elde edilen G katsayısı 0,67 hesaplanmıştır. Bu boyut, için, birey sayısı sabit tutularak puanlayıcı sayısı 2 azaltılması ($n_p=1$) durumunda kestirilen G katsayısı 0,40 birey sayısı sabit tutularak puanlayıcı sayısı 1 azaltılarak ($n_p=2$) durumunda ise G katsayısı 0,57 olarak hesaplanmıştır. Birey sayısının sabit tutulmasıyla puanlayıcı sayısının bir arttırılmasıyla ($n_p=4$) G katsayısı 0.73 iki arttırılmasıyla ($n_p=5$) G katsayısı 0.77 üç arttırılmasıyla ($n_p=6$) g katsayısı 0.80 olarak bulunmuştur. Bir kişiden oluşması durumunda elde edilen G katsayısında 0.27 azalma olurken puanlayıcı sayısı altı olduğunda G katsayısındaki artış 0,13 olmaktadır. Tablo 4.6' ya göre puanlayıcı sayısının azaltılıp veya arttırılması durumunda G katsayısında fazla bir değışkenlik olmadığı görülmektedir.

bxp modelinin orijinal puanlayıcı ve birey sayılarına göre phi katsayısı 0.48 olarak hesaplanmıştır. Farklı puanlayıcı senaryolarına göre yapılan K çalışmaları sonucunda elde edilen phi katsayıları, yukarıdaki tabloda verilmiştir. Buna göre puanlayıcı sayılarının birer arttırılıp birer azaltılması sonucunda phi katsayıları değışkenlik göstermektedir. Puanlayıcı sayısını iki azaltıldığında ($n_p=1$) Phi katsayısı 0.23 iken, bir azaltıldığında ($n_p=2$) phi katsayısı 0.38 olarak hesaplanmıştır. Bunun yanında puanlayıcı sayısı bir arttığında ($n_p=4$) phi katsayısı 0.73, iki arttırıldığında ($n_p=5$) 0.77 ve üç arttırıldığında ($n_p=6$) 0.80 olarak artış olduğu görülmüştür. Bu tabloya göre puanlayıcı sayısını artırma veya azaltmanın Phi katsayısını çok fazla etkilemediğı G katsayılarının Phi katsayılardan büyük olduğu görülmektedir. Bunun yanında G katsayılarının, farklı puanlayıcı senaryolarına göre artışına ve azalışına paralel olarak Phi katsayıları da paralel olarak artmakta ya da azalmakta olduğu görülmektedir.

Tablo 4.7. Fen Lisesi'ne göre puanlayıcıların G ve Phi katsayıları

G-Facets analysis

Yüzey	düzeş	Bağıl G Katsayısı	Mutlak G Katsayısı
P	1	0.299	0.210
	2	0.351	0.297
	3	0.881	0.496

Tablo 4.7.'ye göre üç puanlayıcıdan birinci ve ikinci puanlayıcı ölçeksiz puanlama yaparken üçüncü puanlayıcı ölçekli puanlama yapmıştır. Birinci puanlayıcının G katsayısı 0.30

iken ikinci puanlayıcının G katsayısı 0.35 ve üçüncü puanlayıcının 0.88 düzeyindedir. Birinci puanlayıcının Phi katsayısı 0.21, ikinci puanlayıcının Phi katsayısı 0.30 ve üçüncü puanlayıcının Phi katsayısı 0.50 olarak hesaplanmıştır.

Tablo 4.8. Örneklemeye ait tek değişkenli G Çalışması sonucunda özel okulun varyans değerleri

Varyans kaynağı	Kareler toplamı SS	Sd	Kareler ortalaması MS	Bileşenler				
				Rastgele	Karışık	Varyans	%	SH
B	39083.656	30	1302.789	387.344	387.344	387.344	62.7	108.892
P	5813.957	2	2906.978	89.233	89.233	89.233	14.5	66.313
BP	8445.376	60	140.756	140.756	140.756	140.756	22.8	25.281
Toplam	53342.989	92					100%	

Tablo 4.8’de özel okulunun tek değişkenli G çalışması sonrasında elde edilen varyanslar ve yüzdeleri aşağıda açıklanmıştır.

Bu tabloya göre bireylere ait varyans bileşeninin en büyük değer aldığı görülmekte (387.34) ve toplam varyansın.%62.7’sini oluşturmaktadır. Bu değere göre bireylerin ölçülen özellik bakımından bakıldığında hepsi birbirinden farklılık gösterdiği görülmektedir. Aynı zamanda toplam varyans bileşeninde birey varyansın çok büyük bir orana sahip olması beklenen bir durum olarak bireylerin ölçülen özellik bakımından birbirinden çeşitlilik gösterdiği anlamına gelmektedir.

Puanlayıcı varyans değeri ise (89.23) %14.5 olarak en düşük varyans olarak çıkmaktadır. Bu da puanlayıcıların bireylere verdikleri puanın çok fazla farklılık göstermediği yani birbirlerine yakın puanlamalar yaptığını göstermektedir.

Birey x puanlayıcı (b_xp) artık varyans etkisi (140.76) toplam varyansın %22.8’ini açıklamaktadır. Bu da tesadüfi hatanın var olduğunun yani puanlayıcıların puanlamada bireyler hakkında ölçülemeyen varyans kaynaklarının bulunduğunu göstermektedir.

G kuramı analizine göre çok değişkenli (b x p) modeli ile farklı okul ve puanlayıcı sayıları senaryolarına göre yapılan K çalışmaları sonucunda kestirilen G ve Phi katsayıları, alt boyutlara göre aşağıda tablo 4.9.’da değerler verilmiştir.

Tablo 4.9. Puanlayıcı Sayıları Senaryolarına Göre G ve Phi Katsayıları

	G-çalışması		Senaryo 1		Senaryo 2		Senaryo 3		Senaryo 4		Senaryo 5	
	Dzy	Evr	Dzy	Evr	Dzy	Evr	Dzy	Evr	Dzy	Evr	Dzy	Evr
B	31	Snsz	31	Snsz	31	Snsz	31	Snsz	31	Snsz	31	Snsz
P	3	Snsz	1	Snsz	2	Snsz	4	Snsz	5	Snsz	6	Snsz
Gözlem	93		31		62		124		155		186	
G katsayısı	0.89		0.73		0.85		0.92		0.93		0.94	
Phi katsayısı	0.83		0.63		0.77		0.87		0.89		0.91	
Bağıl hata varyansı	46.919		140.756		70.378		35.189		28.151		23.459	
Rel. Std. Err. of M.	6.850		11.864		8.389		5.932		5.306		4.843	
Mutlak hata varyansı	76.663		229.989		114.995		57.497		45.998		38.332	
Abs. Std. Err. of M.	8.756		15.165		10.724		7.583		6.782		6.191	

Tablo 4.9’ da farklı boyutlardaki çok değişkenli K çalışması sonucunda örneklemdaki birey-puanlayıcı sayıları senaryolarına göre kestirilen G katsayıları açıklanmaya çalışılmıştır. Örneklemdaki verilerden yola çıkarak puanlayıcı sayıların artırılıp azaltılması ve bireylerin aynı şekilde sabit tutulması senaryoları karşısında G katsayılarının değişikliği incelenmeye çalışılmıştır.

Verilen tek soruluk performans değerlendirme kompozisyon sorusunda her okuldan 31 öğrenci alınarak toplamda 4 farklı okul ve 3 puanlayıcı içinden üçüncü okulun K çalışması ile elde edilen G katsayısı 0,89 hesaplanmıştır. Bu boyut, için, birey sayısı sabit tutularak puanlayıcı sayısı 2 azaltılması ($n_p=1$) durumunda kestirilen G katsayısı 0,73 birey sayısı sabit tutularak puanlayıcı sayısı 1 azaltılarak ($n_p=2$) durumunda ise G katsayısı 0,85 olarak hesaplanmıştır. Birey sayısının sabit tutulmasıyla puanlayıcı sayısının bir arttırılmasıyla ($n_p=4$) G katsayısı 0.92 iki arttırılmasıyla ($n_p=5$) G katsayısı 0.93 üç arttırılmasıyla ($n_p=6$) g katsayısı 0.94 olarak bulunmuştur. Bir kişiden oluşması durumunda kestirilen G katsayısında 0.16 azalma olurken puanlayıcı sayısı altı olduğunda G katsayısındaki artış 0,05 olmaktadır. Tablo 4.9’a göre puanlayıcı sayısının azaltılıp veya artırılması durumunda G katsayısında fazla bir değişkenlik olmadığı görülmektedir.

bxp modelinin orijinal puanlayıcı ve birey sayılarına göre Phi katsayısı 0.83 olarak hesaplanmıştır. Farklı puanlayıcı senaryolarına göre yapılan K çalışmaları sonucunda elde edilen Phi katsayıları, yukarıdaki tabloda verilmiştir. Buna göre puanlayıcı sayılarının birer arttırılıp birer azaltılması sonucunda phi katsayıları değişkenlik göstermektedir. Puanlayıcı sayısını iki azaltıldığında ($n_p=1$) Phi katsayısı 0.63 iken, bir azaltıldığında ($n_p=2$) Phi katsayısı 0.77 olarak hesaplanmıştır. Bunun yanında puanlayıcı sayısı bir arttığında ($n_p=4$) Phi katsayısı

0.87, iki arttırıldığında ($n_p=5$) 0.89 ve üç arttırıldığında ($n_p=6$) 0.91 olarak artış olduğu görülmüştür. Bu tabloya göre puanlayıcı sayısını artırma veya azaltmanın Phi katsayısını çok fazla etkilemediği ve G katsayıların Phi katsayılarından büyük olduğu görülmektedir. Bunun yanında G katsayılarının, farklı puanlayıcı senaryolarına göre artışına veya azalışına paralel olarak Phi katsayıları da paralel olarak artmakta veya azalmaktadır.

Tablo 4.10. Özel okula göre Puanlayıcıların G ve Phi Katsayıları

Yüzey	düzy	Bağıl G katsayısı	Mutlak G katsayısı
P	1	0.862	0.842
	2	0.847	0.721
	3	0.825	0.759

Tablo 4.10.'a göre üç puanlayıcıdan birinci ve ikinci puanlayıcı ölçeksiz puanlama yaparken üçüncü puanlayıcı ölçekli puanlama yapmıştır. Birinci puanlayıcının G katsayısı 0.86 iken ikinci puanlayıcının 0.85 ve üçüncü puanlayıcının 0.83 düzeyindedir. Birinci puanlayıcının Phi katsayısı 0.84, ikinci puanlayıcının Phi katsayısı 0.72 ve üçüncü puanlayıcının Phi katsayısı 0.76 olarak hesaplanmıştır.

Çok değişkenli ve G çalışması sonucunda performansların değerlendirilmesi sonucu ortaya çıkan varyans değerlerini belirlemek için G Kuramının bxp çapraz desen modelleri kullanılarak varyans kaynaklarının ve toplam varyansın belirlenmesi için G çalışması yapılmıştır. Değerlendirmeye alınan 124 öğrenci, 3 puanlayıcı ve 4 okuldan oluşan veri grubunun varyans açıklama yüzdeleri b ve p ana etkileri ile bp, ortak etkileri aşağıdaki tablo 4.11' de verilmiştir.

Tablo 4.11. Örnekleme Ait Tek Değişkenli G Çalışması Sonucunda Çok Programlı Lisenin Okulunun Varyans Değerleri

Varyans kaynağı	Kareler toplamı SS	Sd	Kareler ortalaması MS	Bileşenler				
				Rastgele	Karışık	Varyans	%	SH
B	72990.796	30	2433.027	796.513	796.513	796.513	92.9	202.769
P	1154.000	2	577.000	17.210	17.210	17.210	2.0	13.164
BP	2609.333	60	43.489	43.489	43.489	43.489	5.1	7.811
Toplam	76754.129	92					100 %	

Tablo 4.11.'e göre çok programlı lisenin tek değişkenli G çalışması sonrasında elde edilen varyanslar ve yüzdeleri aşağıda açıklanmıştır.

Bu tabloya göre bireylere ait varyans bileşeninin en büyük değer aldığı görülmekte (796.51) ve toplam varyansın.%92.9'unu oluşturmaktadır. Bu değere göre bireylerin ölçülen özellik bakımından bakıldığında hepsi birbirinden oldukça farklılık gösterdiği görülmektedir. Aynı zamanda toplam varyans bileşeninde birey varyansın çok büyük bir orana sahip olması beklenen bir durum olarak bireylerin ölçülen özellik bakımından birbirinden yine oldukça çeşitlilik gösterdiği anlamına gelmektedir.

Puanlayıcı varyans değeri ise (17.21) %2.0 olarak çıkmaktadır. Bu da puanlayıcıların bireylere verdikleri puanın birbirine çok yakın puanladıkları aralarında puanlama anlamında tutarlılık olduğu görülmektedir.

Birey x puanlayıcı (b_xp) artık varyans etkisi (43.49) toplam varyansın %5.1'ini açıklamaktadır. Bu da tesadüfi hatanın az olduğunun yani puanlayıcıların puanlamada bireyler hakkında tutarlı değerlendirmelerde bulunduğunu göstermektedir.

G kuramı analizine göre çok değişkenli (b x p) modeli ile farklı puanlayıcı sayıları senaryolarına göre yapılan K çalışmaları sonucunda kestirilen G ve Phi katsayıları, alt boyutlara göre aşağıda tablo 4.12' de değerler verilmiştir.

Tablo 4.12. Çok programlı lisenin Örneklemin Çok Değişkenli Model K Çalışması İle Farklı Sayıdaki Puanlayıcı Sayıları Senaryolarına Göre G ve Phi Katsayıları

	G çalışması		Senaryo 1		Senaryo 2		Senaryo 3		Senaryo 4		Senaryo 5	
	Dzy	Evr	Dzy	Evr	Dzy	Evr	Dzy	Evr	Dzy	Evr	Dzy	Evr
B	31	Snsz	31	Snsz	31	Snsz	31	Snsz	31	Snsz	31	Snsz
P	3	Snsz	1	Snsz	2	Snsz	4	Snsz	5	Snsz	6	Snsz
Gözlem	93		31		62		124		155		186	
G katsayısı	0.98		0.95		0.97		0.99		0.99		0.99	
Phi katsayısı	0.98		0.93		0.96		0.98		0.98		0.99	
Bağıl hata varyansı	14.496		43.489		21.744		10.872		8.698		7.248	
Rel. Std. Err. of M.	3.807		6.595		4.663		3.297		2.949		2.692	
Mutlak hata varyansı	20.233		60.699		30.349		15.175		12.140		10.116	
Abs. Std. Err. of M.	4.498		7.791		5.509		3.895		3.484		3.181	

Tablo 4.12.' de farklı boyutlardaki çok değişkenli K çalışması sonucunda örneklemdaki birey-puanlayıcı sayıları senaryolarına göre kestirilen G katsayıları açıklanmaya çalışılmıştır. Örneklemdaki verilerden yola çıkarak puanlayıcı sayıların artırılıp azaltılması ve bireylerin

aynı şekilde sabit tutulması senaryoları karşısında G katsayılarının değışikliğı incelenmeye çalışılmıştır.

Verilen tek soruluk performans değeriendirme kompozisyon sorusunda her okuldan 31 öğrenci alınarak toplamda 4 farklı okul ve 3 puanlayıcı içinden dördüncü okulun K çalışması ile kestirilen G katsayısı 0,98 hesaplanmıştır. Bu boyut, için, birey sayısı sabit tutularak puanlayıcı sayısı 2 azaltılması ($n_p=1$) durumunda kestirilen G katsayısı 0,95 birey sayısı sabit tutularak puanlayıcı sayısı 1 azaltılarak ($n_p=2$) durumunda ise G katsayısı 0,97 olarak hesaplanmıştır. Birey sayısının sabit tutulmasıyla puanlayıcı sayısının bir arttırılmasıyla ($n_p=4$) G katsayısı 0.99 iki arttırılmasıyla ($n_p=5$) G katsayısı 0.99 üç arttırılmasıyla ($n_p=6$) g katsayısı 0.99 olarak bulunmuştur. Bir kişiden oluşması durumunda kestirilen G katsayısında 0.03 azalma olurken puanlayıcı sayısı dört, beş ve altı olduğunda G katsayısındaki artış 0,01 olmaktadır. Tablo 4.12' ye göre puanlayıcı sayısının azaltılıp veya arttırılması durumunda G katsayısında neredeyse çok bir değışkenlik olmadığı görölmektedir.

bxp modelinin orijinal puanlayıcı ve birey sayılarına göre Phi katsayısı 0.98 olarak hesaplanmıştır. Farklı puanlayıcı senaryolarına göre yapılan K çalışmaları sonucunda elde edilen Phi katsayıları, yukarıdaki tabloda verilmiştir. Buna göre puanlayıcı sayılarının birer arttırılıp birer azaltılması sonucunda Phi katsayıları değışkenlik göstermektedir. Puanlayıcı sayısını iki azaltıldığında ($n_p=1$) Phi katsayısı 0.93 iken, bir azaltıldığında ($n_p=2$) Phi katsayısı 0.96 olarak hesaplanmıştır. Bunun yanında puanlayıcı sayısı bir arttığında ($n_p=4$) Phi katsayısı 0.98, iki arttırıldığında ($n_p=5$) 0.98 ve üç arttırıldığında ($n_p=6$) 0.99 olarak artış olduğu görölmüştür. Bu tabloya göre puanlayıcı sayısını artırma veya azaltmanın Phi katsayısını çok fazla etkilemediğı ve G ve Phi değeriilerinin de birbirine çok yakın olduğu görölmektedir.

Tablo 4.13. Çok Programlı Liseye göre puanlayıcıların G ve Phi Katsayıları

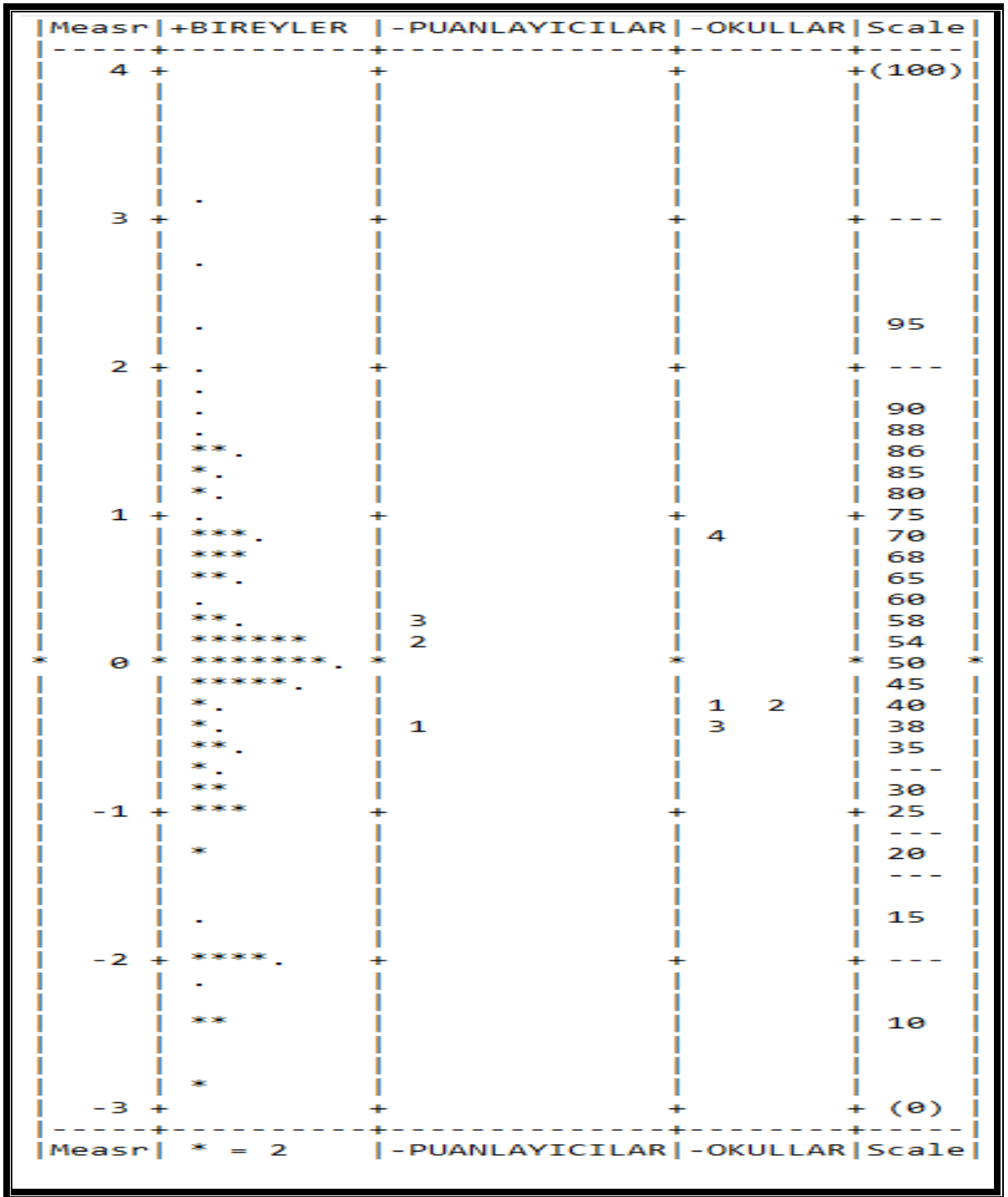
Yüzey	düzey	Bağıl G Katsayıları	Mutlak G Katsayıları
P	1	0.964	0.956
	2	0.979	0.960
	3	0.977	0.974

Tablo 4.13.'e göre üç puanlayıcıdan birinci ve ikinci puanlayıcı ölçeşsiz puanlama yaparken üçüncü puanlayıcı ölçekli puanlama yapmıştır. Birinci puanlayıcının G katsayısı 0.96 iken ikinci puanlayıcının G katsayısı 0.98 ve üçüncü puanlayıcının 0.98 düzeyindedir. Birinci puanlayıcının Phi katsayısı 0.96, ikinci puanlayıcının Phi katsayısı 0.96 ve üçüncü puanlayıcının Phi katsayısı 0.97 olarak hesaplanmıştır.

4.2. Çok Değişkenlik Kaynaklı Rasch Modeli Analizi

Puanlayıcıların güvenilirliğinin belirlenmesi adına yapılan bu çalışma için Çok Değişkenlik Kaynaklı Rasch Modeli (ÇDKRM) ile elde edilen model veri uygunluk istatistikleri, birey yetenek ölçüleri ve uygunluk istatistikleri, puanlayıcı katılık/cömertlikleri ve uygunluk istatistikleri, okul seviyeleri ve uygunluk istatistikleri, puanlayıcı x birey, okul x birey, puanlayıcı x okul ve puanlayıcı x birey x okul değişkenlik kaynaklarına uygunluk istatistikleri ve yanlılıkları aşağıda açıklanmış ve logit haritası Şekil 4.1.'de verilmiştir. Bu cetvele göre ölçütleri karşılama seviyeleri yukarıdan aşağıya doğru sıralanmıştır.

Cetvelin ikinci sütununda verilen bireylerin genelde değer olarak 0 da yoğunlaştığı maksimum 3-4 arası minimum (-2)-(-3) arasında değiştiği görülmektedir. Üçüncü sütunda ise puanlayıcıların katılık/cömertlik düzeylerin yüksek veya düşük sıralaması yapılmıştır. Buna göre en katı ölçekli olarak puanlama yapan 3. puanlayıcıyken en cömert 1. puanlayıcı olarak görülmektedir. Dördüncü sütunda okulların düzeyleri verilmiş olup başarı ortalaması en yüksek Anadolu Lisesi olurken, başarı ortalaması en düşük çok programlı lise olarak ortaya çıkmıştır. Logit cetvelinin dışında yüzeylerle ilgili ayrıntılı bilgiler ayrı ayrı aşağıdaki tabloda verilmiştir.



Şekil 4.1. Çok Değişken Kaynaklı Rasch Modeli İle Çalışmanın 3 Yüzeyli Logit Cetveli

Analize dahil olan 4 farklı okuldan toplam 124 öğrencinin İngilizce yeterlilik seviyelerinin 3 farklı puanlayıcının puanlaması sonucunda logit değerleri 3,14 ve -2,90 arasında farklılık göstermektedir. Bu değerlere göre logit ortalaması -0,08 ve standart sapması 1,09 ‘dur. Ayırma indeksi 3,26 ve güvenilirlik 0,91 olarak çıkmaktadır. Güvenirlik değerinin 1’e yakın olması bireylerin yeteri kadar ayırt edici olarak puanlandığını gösterir. Cetvelin ikinci sütununda yer alan puanlayıcılara ait düzeyler yer almakta, en katı olanı yüksek puan alırken

en cömert olanı düşük puan alarak birbirlerinden ayrılabilirdikleri görülmektedir. Buna göre üçüncü puanlayıcı en katı puanlayıcı olurken birinci puanlayıcı en cömert puanlayıcı olmaktadır. Cetvelin üçüncü sütununda okul değişkeni yer almaktadır. Okullar da cetveldeki yerlerine bakarak birbirlerinden farklılık göstermektedir. Logit cetvel yer alan değişkenlerin bir tablo içinde değerlendirilip yorumlanmasını sağlarken daha detaylı ölçüm raporları ayrı ayrı aşağıdaki tablolarda verilmiştir.

Tablo 4.14. Bireylerin Ölçütlerine İlişkin Ölçüm Raporu

			Birey düzeyi		Uygunluk İçi		Uygunluk Dışı	
	Birey Ort.	Birey Toplam r	Logit Ölçüsü	Std Hata.	Kareler Ort.	Z Std.	Kareler Ort.	Z Std
Ortalama	52.05	49.15	-0.08	0.29	0.94	-0.1	0.98	0
SS	24.17	23.67	1.14	0.17	0.83	1.1	0.91	1.1
RMSE(Model) = 1.09		SS = 3.26	Ayırma indeksi= 4.68		Güvenirlik: 0.91			
Tamamı aynı Kay-Kare =1405		Sd= 123	p= 0.00					

Tablo 4.14.'deki değerlere bakıldığında bireylerin ortalaması 52,05 ve standart sapmanın 24,17 olduğu görülmektedir. Uygunluk içi kareler ortalamaları (infit mean square) 0,94 ve uygunluk içi Z puanı ortalaması -0.1 çıkmıştır. Uyum dışı istatistiği 0.98 ve uyum dışı Z puanı 0 olarak bulunmuştur. Uyum dışı istatistiği, uyum içi istatistiğine göre beklenmedik değerlere karşı daha duyarlı olduğu ve model-veri uyumunun olması gereken değerleri 0.6 ile 1.5 arasında olması gerektiği kabul görmektedir(Lunz, Wright ve Linacre, 1990). Uyum indeksinin değer olarak 1'e yakın olması model ve veri arasında uyumun oldukça yüksek olduğunu göstermektedir. (Güler,2014; Hetherman, 2004). Tabloya göre ayırma indeksi 4.68, güvenirlik 0.91 olarak hesaplanmıştır. Bu değerlere bakıldığında öğrencilerin yeterliğinin birbirinden farklılık gösterdiği sonucuna ulaşılabilir. RMSE değeri ise 1.09 olarak bulunmuştur.

Tablo 4.15. Puanlayıcı Ölçüm Raporu

Puanl. No	Puanl. Ort.	Puanl Top. r	Puanlayıcı katılığı		Uygunluk İçi		Uygunluk Dışı	
			Logit Ölçüsü	Std Hata.	Kareler Ort.	Z Std.	Kareler Ort.	Z Std
1	60.56	58.68	-0,40	0.04	0.71	-1.8	0.70	-1.9
2	49.25	41.71	0.15	0.03	0.91	-0.6	0.96	-0.2
3	46.33	39.43	0.24	0.03	1.33	2.2	1.25	1.8
1,2,3	52.05	46.61	0.00	0.03	0.98	-0.2	0.98	-0.1
Ortlma	6.14	8.58	0.28	0.0	0.26	1.9	0.22	1.6
SS	7.52	10.51	0.35	0.0	0.32	2.3	0.27	1.6
RMSE(Model) = 0.28			SS = 8.03		Ayırma indeksi = 11.04		Güvenirlilik: 0.98	
Tamamı aynı Kay-Kare =186.5			Sd= 2		p= 0.00			

Tablo 4.15'e göre üç puanlayıcının katılık logit değerleri -0.40 ile 0.24 aralığında değişmektedir. (birinci puanlayıcı; -0,40, ikinci puanlayıcı; 0,15 ve üçüncü puanlayıcı 0,24 logit). Puanlayıcıların logit ortalaması 0.0 ve standart sapması 0.35'dir. Puanlamada birinci puanlayıcı (-0.40) diğer puanlayıcılara göre oldukça cömert davrandığı görülmektedir. Puanlayıcı ayırma indeksi 11.04 ve güvenirlilik 0.98'dir. Güvenirlilik değerinin bu kadar yüksek olması yani 1'e yakın olması puanlamada puanlayıcıların birbirlerinden farklı şekilde puanlama yaptığını göstermektedir. Aynı şekilde ayırma indeksinin yüksek olması da puanlayıcıların puanlamada farklılık gösterdiğini ifade etmektedir. Bu da yine puanlayıcıların katılık/cömertlik bakımından farklılık gösterdiğinin göstergesidir. Kay-kare testi 186.5, serbestlik derecesi 2 ve $p=0$ çıkmaktadır. Ayrıca puanlayıcıların katılık cömertlikte çok olmasa da nispeten puanlamada farklılık olduğu görülmektedir. Özellikle ölçekli puanlama yapan üçüncü puanlayıcı diğer puanlayıcılara göre puanlamada daha katı davranmıştır. Ayrıca puanlayıcılar arası belli bir farklılık olması sebebiyle birbirlerinden bağımsız puanlama yaptıkları görülmektedir.

Uygunluk (infit) kareler ortalaması istatistiğinin ortalaması 0.26 ve standart sapması 0.32 olarak hesaplanırken standartlaştırılmış Z puanları sırasıyla -1,8, -0,6 ve 2,2'dir. Literatürde uygunluk içi ve uygunluk dışı istatistiklerinin 0.5 ve 1.5 değerleri arasında olması istenir (Turner, 2003). Çalışmanın analizlerine bakıldığında puanların modelle uyumlu olduğunu göstermektedir. Ayırma indeksinin 2.00'den büyük olması istenmeyen bir durum olarak görülmekte yani puanlamayı yapan puanlayıcılardan kaynaklı hataların var olduğunu göstermektedir. Bununla birlikte katılık/cömertlik durumunun puanlayıcılar arasında farklılık gösterdiğini belirtmektedir (Karakaya 2018). RMSE değeri 0.28 bulunmuştur.

Tablo 4.16. Okulların Ölçüm Raporu

Okul no	Okul Ort.	Okul Toplam r	Okul güçlüğü		Uygunluk İçi		Uygunluk Dışı	
			Logit Ölçüsü	Std Hata.	Kareler Ort.	Z Std.	Kareler Ort.	Z Std
1	57.30	53.53	-0.23	0.04	0.93	-0.4	1.12	0.8
2	61.78	56.50	-0.32	0.03	1.06	0.4	0.95	-0.3
3	59.01	57.58	-0.36	0.04	1.16	1.0	1.15	1.12
4	30.10	25.14	0.92	0.05	0.66	-1.8	0.70	-1.9
Ortalama	52.05	48.19	0.00	0.04	0.95	-0.2	0.98	-0.1
SS	12.77	13.39	0.53	0.01	0.19	1.1	0.18	1.1
RMSE(Model) = 0.53		SS = 12.52	Ayırma indeksi= 17.03		Güvenirlilik: 0.99			
Tamamı aynı kay kare=449.5		Sd= 3	p= 0.00					
Normal kay-kare =3.00		Sd=2	P=0.22					

Her okuldan 31 öğrencinin yer aldığı çalışmada toplam 4 farklı düzeydeki okulların logit değerleri -0,23 ile 0,92 arasında dağıldığı görülmektedir. Buna göre ortalamasının 0,00 olduğu, standart sapmasının 0,53 olduğu görülmektedir. Okullar arasından çok programlı lisedeki bireylerinin (0,92) en az yeterlilikte olduğu görülmekte en fazla yeterliliğe sahip okulun bireyleri ise Anadolu Lisesinde (-0,23) olduğu hesaplanmıştır. Okul ayırma indeksi 17,03 ve ayırma güvenirliliği 0.99'dur. Ayırma indeksinin oldukça yüksek çıkması okulların birbirinden fazlasıyla farklılık gösterdiğini göstermektedir. Bu ayırma indeksi ve ayırma güvenirliliği test edildiğinde sabit etki (fixed effect) hipotezi, Kay-Kare testiyle; χ^2 :449,5 ve serbestlik derecesinde ($p=0,00$)<0,05 olarak kabul edilmektedir. Bu sonuca göre okul seviyelerinde istatistiksel olarak anlamlı fark olduğunu göstermektedir.

ÇDKRM'de 0 ile sonsuz değer arasında değerler alan dış uyum ve iç uyum istatistiğinin 1 olması istenilen durumdur. Değer 1'den küçük değerlere doğru azaldıkça değerlerin kısıtlanması veya sıkıştırılmasına neden olurken, 1 den büyük değerler aldıkça beklenmeyen değişkenliği veya tutarsızlığı belirtmektedir. Okulların uygunluk (infit) kareler ortalaması 1,16 ile 0,66 arasında değerler almakta, uygunluk standart Z puanları 1 ile -1,8 arasında değişmektedir. İç uyum ve dış uyum kareler ortalaması en çok tutarlılığı Anadolu Lisesinde okulda görülürken tutarsızlığın en fazla çok programlı lisede olduğu okulda görülmektedir.

BÖLÜM V

SONUÇ, TARTIŞMA VE ÖNERİLER

Araştırmanın bu bölümünde bulgulardan yola çıkarak elde edilen sonuçlar ve bu sonuçlara göre çıkarılacak sonraki çalışmalara yardımcı olacak önerilere yer verilmektedir.

5.1. Sonuçlar

Araştırmadan elde edilen bulgulara ait sonuçlar ve alt problem sonuçları sırayla verilmiştir. Buna göre 2022-2023 eğitim – öğretim yılına ait farklı ortaöğretim kurumlarında yer alan 12. sınıfların İngilizce yeterliğine ait değerlendirme sonuçlarının bir ölçekli diğer ikisi ölçeksiz olmak üzere 3 farklı puanlayıcı tarafından puanlanması sonucu Genellenebilirlik ve Çok Değişkenli Rasch Modeliyle güvenilirliği ölçülmek istenmiştir. Çıkan sonuçlara göre ayrı ayrı modeller üzerinden bakıldığında benzerliklerin yanında bazı farklılıkların da ortaya çıktığı görülmektedir.

Örneklem üzerinden yapılan analizler sonucu çıkan betimsel istatistiksel verilerde okulların puan ortalaması birinci puanlayıcıya göre üçüncü puanlayıcının puanlaması sonucunda okul puan ortalamaları arasındaki fark artmış olduğu ve puanların ortalamaların genel olarak düştüğü görülmektedir. Bunun da sebebi olarak üçüncü puanlayıcının puanlamayı yaparken bütünsel dereceli puanlama anahtarıyla yaptığı ve bu durumun diğer puanlayıcıların serbest puanlamasından anlamlı derecede fark oluşturduğu görülmektedir. Hem G Kuramı hem de ÇDKRM analizlerinde okullar arasında anlamlı bir farklılık olduğu görülmektedir. Puanlayıcıların her okuldan eşit sayıda öğrenciyi puanladıkları analiz sonucunda en yüksek ortalama ya bağlı okul Fen Lisesi (61,78) olurken en düşük ortalama ya sahip okul çok programlı lise (30,10) olarak görülmektedir.

Kompozisyon performans testini puanlayan puanlayıcıların Anadolu Lisesindeki bireyleri puanlayan puanlayıcıların ölçeksiz puanlayan ilk puanlayıcının G çalışması sonucunda G katsayısı 0.85 ikinci ölçeksiz puanlayan puanlayıcının 0.88 ve ölçekli puanlayan diğer puanlayıcının, 0.88 gibi değerle birbirlerine yakın puanlamalar yaptığı görülmüştür. Puanlayıcı sayısını birer azaltıp artırılmasıyla G katsayılarında doğru orantıda azalma ve artma görülmüştür. Phi katsayılarına bakıldığında birinci puanlayıcının Phi katsayısı 0.79, ikinci puanlayıcının Phi katsayısı 0.72 ve üçüncü puanlayıcının Phi katsayısı 0.82 olarak

hesaplanmıştır. Karar çalışması doğrultusunda puanlayıcı sayılarının artırılmasıyla ortaya çıkan Phi katsayılarında azalma olurken puanlayıcıların artışında Phi katsayıları da artmaktadır.

Puanlayıcıların Fen Lisesi bireyleri puanlaması sonucu yapılan bxp tek değişkenli G çalışmaları sonuçlarındaki birinci puanlayıcının G katsayısı 0.30 ikinci ölçeksiz puanlayan puanlayıcının 0.35 ve ölçekli puanlayan diğer puanlayıcının, 0.88 gibi değerler hesaplanmıştır. İkinci okulu puanlamada ölçeksiz puanlayanlar birbirine oldukça yakın değerler alırken ölçekle puanlayan puanlayıcının G katsayısı oldukça farklı ve büyük hesaplanmıştır. Puanlayıcı sayısını birer azaltıp artırılmasıyla G katsayılarında doğru orantıda azalma ve artma görülmüştür. Phi katsayılarına bakıldığında birinci puanlayıcının Phi katsayısı 0.21, ikinci puanlayıcının Phi katsayısı 0.30 ve üçüncü puanlayıcının Phi katsayısı 0.50 olarak hesaplanmıştır. Karar çalışması doğrultusunda puanlayıcı sayılarının artırılmasıyla ortaya çıkan Phi katsayılarında azalma olurken puanlayıcıların artışında Phi katsayıları da artmaktadır.

Özel okulu puanlamada bxp tek değişkenli sonuçlara bakıldığında birinci puanlayıcının G katsayısı 0.86 ikinci ölçeksiz puanlayan puanlayıcının 0.85 ve ölçekli puanlayan diğer puanlayıcının, 0.82 gibi değerler hesaplanmıştır. İkinci okulu puanlamada ölçeksiz puanlayanlar birbirine oldukça yakın değerler alırken ölçekle puanlayan puanlayıcının G katsayısı daha küçük hesaplanmıştır. Puanlayıcı sayısını birer azaltıp artırılmasıyla G katsayılarında doğru orantıda azalma ve artma görülmüştür. Phi katsayılarına bakıldığında birinci puanlayıcının Phi katsayısı 0.84, ikinci puanlayıcının Phi katsayısı 0.82 ve üçüncü puanlayıcının Phi katsayısı 0.76 olarak hesaplanmıştır. Karar çalışması doğrultusunda puanlayıcı sayılarının artırılmasıyla ortaya çıkan Phi katsayılarında azalma olurken puanlayıcıların artışında Phi katsayıları da artmaktadır.

Tek değişkenli bxp sonuçlarında çok programlı lisenin değerleri birinci puanlayıcının G katsayısı 0.96 ikinci ölçeksiz puanlayan puanlayıcının 0.98 ve ölçekli puanlayan diğer puanlayıcının, 0.98 gibi değerler hesaplanmıştır. Dördüncü okulu puanlamada birinci puanlayıcı en düşük çıkarken diğer ikinci ölçeksiz puanlayan puanlayıcı ile ölçekli puanlayan puanlayıcının değerleri aynı hesaplanmıştır. Puanlayıcı sayısını birer azaltıp artırılmasıyla G katsayılarında doğru orantıda azalma ve artma görülmüştür. Phi katsayılarına bakıldığında birinci puanlayıcının Phi katsayısı 0.96, ikinci puanlayıcının Phi katsayısı 0.96 ve üçüncü puanlayıcının Phi katsayısı 0.92 olarak hesaplanmıştır. Karar çalışması doğrultusunda puanlayıcı sayılarının artırılmasıyla ortaya çıkan Phi katsayılarında azalma olurken puanlayıcı sayısının arttığı durumda Phi katsayıları da artmaktadır.

Hem Genellenebilirlik hem de ÇDKRM analizlerinin Klasik Test Kuramından ayıran en önemli özelliği birey, okul ve puanlayıcı gibi birden fazla değişkenleri aynı anda ele alarak güvenilirlik hesaplaması yapabilmesidir. İki veya daha fazla puanlayıcının yapmış olduğu puanlamada tüm hata kaynaklarını kestirebilmesi neticesinde puanlayıcılar arasındaki uyumu ve farkı güvenilirlik katsayılarıyla beraber hesaplamakta ve alt problemleri sırasına uygun olarak vermektedir. Bu çalışmada da iki puanlayıcının ölçeksiz, bir puanlayıcının ölçekli puanlamasıyla elde edilen sonuçlara bakıldığında puanlayıcılar arasında tutarlılık olduğu ancak birbirlerinden farklılık gösterdiği de görülmektedir. Özellikle bütünsel dereceli puanlama anahtarıyla puanlama yapan üçüncü puanlayıcının diğer puanlayıcılara göre daha katı puanlama yaptığı görülmektedir. Ölçeksiz puanlayan diğer iki puanlayıcıların öğrencilerin yazı güzelliği, mesleki uzmanlık gibi farklılıklardan kaynaklanabileceği sebeplerle daha cömert veya katı puanlama yapmış olabileceği şeklinde yorumlanabilir.

Bu araştırma sonucunda ortaöğretim öğrencilerinin ölçekli ve ölçeksiz puanlama yapan 3 farklı puanlayıcının puanlaması sonucunda çıkan sonuçlar ÇDKRM ile analiz edilerek güvenilirlikleri incelenmiştir. Buna göre bireylerin 3 farklı puanlayıcı tarafından ayırt edilir bir şekilde puanlanmış ve performans yeterlilikleri birbirinden farklı ancak model veri uyumu tutarlı yani uygunluk istatistikleri istenilen değerler arasında çıkmıştır. Çalışmada yer alan puanlayıcıların katılık/cömertlik değerlerinin birbirinden farklı çıktığı, özellikle ölçekle puanlama yapan puanlayıcının daha katı puanlama yaptığı görülmüştür. Ölçeksiz puanlama yapan puanlayıcılar arasında da katılık/cömertlik değerleri arasında farklılık olduğu görülmekte ve bunun nedeninin hizmet yılı, çalıştığı kurum veya farklı değişkenlerin de etkili olabileceği düşünülmektedir.

Analizler sonrasında üçüncü bir değişken olan okullara bakıldığında yine 4 okulun arasında anlamlı bir fark olduğu, okul düzeylerinin 3 puanlayıcı tarafından doğru orantıda sıralandığı ancak okul uygunluk değerlerinin yine aynı şekilde istenen değerde olduğu görülmüştür.

Eldeki bulgulara göre G kuramı analizlerinde toplam varyans açıklamada bireyler arası varyansların okullara göre farklılık gösterdiği ancak genelde birey varyansının yüksek olduğu ortaya çıkmıştır. Genellenebilirlik analizlerine bakıldığında ölçme objesi olan birey ana etkisi %92.9 ile %23.2 arasında değişmekte bir okulun analizi dışında en büyük yüzdeliği elde etmektedir. G (bağıl) katsayılarına bakıldığında 0.67 ile 0.98 arasında değişmektedir. Phi (mutlak) katsayılarına bakıldığında 0.48 ile 0.98 arasında hesaplanmıştır. Genel olarak yüksek değer sonucu güvenilir sonuçlara ulaşıldığı görülmektedir. ÇDKRM’de ise ayırma indeksi ve güvenilirlik katsayılarına bakıldığında sırasıyla birey bazında 4.68 ile 0.91 çıkmıştır. Bu

değerlere göre bireylerin birbirinden ayrıldığı güvenilirliğin oldukça yüksek çıktığı görülmüştür. Aynı zamanda kay-kare testine bakarak da anlamlı bulunmuştur. Bu sonuçların genellenebilirlik sonuçlarıyla tutarlı olduğu görülmektedir.

Araştırmadaki diğer bir değişken olan puanlayıcıların genellenebilirlik sonuçlarına göre puanlayıcı ana varyans etkisinin %2 ile %42.5 arasında değişkenlik gösterdiği görülmektedir. Çok programlı lisedeki varyans etkisinin oranı düşük çıkarken diğer okullarda oldukça yüksek çıkmaktadır. Bu da çok programlı liseyi puanlayan puanlayıcıların birbiriyle uyum içinde puanlama yaptıkları diğer okulların öğrencilerini puanlamada birbirinden farklılık gösterdiği görülmektedir. ÇDKRM’de ise puanlayıcıların öğrencileri puanlamada gösterdiği uyumu veren uyum istatistikleri 0.71 ile 1.33 arasında çıkarak 1’e oldukça yakın bulunmuştur. Bunun da puanlamadaki puanlayıcıların birbiriyle uyum içinde olduklarını göstermektedir. Bununla birlikte ayırma indeksi ve güvenilirlik katsayısının da oldukça yüksek olduğu görülmektedir.

Genellenebilirlik kuramında güvenilirlik bakımından g ve phi katsayıları güvenilirlik katsayısı olarak kabul edilmiştir. ÇDKRM’ de ise ayırma indeksi ve güvenilirlik katsayısı güvenilirlik katsayısı olarak hesaplanmaktadır. Bu iki analize bakarak güvenilirlikle ilgili sonuçlar bakımından benzer sonuçlar çıkmaktadır.

Çok değişkenli Rasch Modeline göre model-veri uyumunu sağlandığı görülmektedir. ÇDKRM bakımından yapılan yorumlamalarda değişkenlik kaynağı olarak birey söz konusuysa ayırma indeksi ve güvenilirliğin yüksek olması beklenmektedir. Puanlayıcılar değişkenlerine bakıldığında ayırma indeksi düşük çıkması beklenmekte ancak ölçekli ve ölçeksiz puanlamalar göz önüne alındığında burada da ayırma indeksinin yüksek çıkabileceği yorumuna ulaşılabilir. Her üç puanlayıcının bütün bireylere yönelik yaptığı puanlamada birbirine yakın katılık/cömertlikte puanlama yapsa da en katı ölçekli puanlayan puanlayıcı üçüncü puanlayıcıdır. En cömert ölçeksiz puanlayan birinci puanlayıcı olarak görülmektedir. Aynı şekilde üç puanlayıcının ÇDKRM ye göre güvenilir indeksinin yüksek çıktığı görülmektedir. Üç puanlayıcının da tüm bireylere verdikleri puanların uygun olduğu ortaya çıkmıştır.

Bireylerin yetenekleri G Kuramı ile kestirilen varyans ile ÇDKRM ile elde edilen parametrelere birlikte bakıldığında genel anlamda tutarlı olduğu ve puanlayıcıların yapmış olduğu puanlama sonrasında katılık ve cömertlik olarak birbiri ile örtüştüğü birey x puanlayıcı değişkenlik kaynağı için G Kuramı ile elde edilen varyans yüzdeleriyle ÇDKRM ile elde edilen yanlılık yüzdelerinin birbirinden farklılık gösterdiği ancak sonuç olarak hem G Kuramıyla hem de ÇDKRM analizleriyle beraber güvenilirliğin birbirine yakın sonuçlar verdiği görülmektedir.

Araştırmada elde edilen bulgulardan yola çıkarak hem Genellenebilirlik Kuramı hem de ÇDKRM üzerinden yapılan performans güvenilirlik testinde birden fazla puanlayıcının hem ölçekli hem ölçeksiz olarak yapmış olduğu ölçümlerde bireyler arası farklılıkların uyumlu bulunduğu bu farklılıkları görebilmeye iki kuramın da etkili olduğu görülmüştür. Bireyler arası ve okullar arası farklılaşmayı verebilmesi adına güvenilirlik yüksek çıkmış bu da çalışmanın amacı doğrultusunda net bilgi sağlamıştır.

Sonuç olarak iki modelde de bireylerin performans testinin ayırt edilebildiği, iki modelde de güvenilirliğin tutarlı sonuçlar verdiği, puanlayıcıların katılık ve cömertlik düzeylerinin birbirinden farklılık gösterdiği fakat model-veri uyumunun sağlandığı görülmektedir. Değişkenlik kaynakları istatistiklerini farklı şekilde hesaplamalarına rağmen her iki modelde de güvenilirlik sonuçları genel anlamda uyum göstermiş farklı durumlarda birbirleri ile arasındaki üstünlük durumu değişerek genel anlamda birden fazla puanlayıcı güvenilirliği açısından ikisinin kullanımının da uygun olacağı sonucuna ulaşılmıştır.

5.2.Tartışma

Bireylerin performans yetenekleri ve uygunluk istatistikleri ile beraber puanlayıcıların katılık/cömertlik durumunu görmek amacıyla tamamı aynı kay-kare, normal kay kare değerleri, ayırt etme indeksi ve güvenilirlik değerleri incelenmiştir. Sıfır ile bir değerler arasında yer alan ayırt etme indeksi ve ayırma oranı sıfır ile sonsuz değerler arasında ÇDKRM'nin iki farklı güvenilirlik katsayısı olarak kabul edilmektedir (Bond ve Fox, 2007). Değişkenlik kaynağı olarak bireylerin istatistiği yapılacağı zaman gerçek varyansın gözlenen varyansa oranını belirten ayırt etme indeksi ve ayırma oranının büyük olması beklenmektedir. Bu değerlerin küçük olması merkezi eğilim hatasını belirtmekte puanlayıcıların bireylere ait performans sonuçlarının ayırt edemediğini ortaya koymaktadır. (Mford ve Wolfe, 2003; Sudweeks, Reeve ve Bradshaw, 2005).

5.3. Öneriler

Bu başlıkla beraber yapılan çalışma sonrası elde edilen analiz ve bulgular sonucunda hem genellenebilirlik kuramı hem de rasch modeli üzerinden çıkarılacak önerilere ve sonraki araştırmalara yönelik önerilere yer verilmiştir.

5.2.1.Araştırmadan Çıkarılacak Öneriler

1. Araştırmalarda, birden fazla hata kaynağı göz önünde bulundurarak puanlayıcılar arası güvenirlik hesaplanmasında Genellenebilirlik Kuramı ve Rasch modelinin yararlanılmasının uygun olacağı sonucu ortaya çıkmaktadır.
2. Genellenebilirlik kuramını kullanmanın farklı sayıdaki değişkenler üzerinden oluşturulacak senaryoları görmek adına uygun olacağı görülmüştür.
3. Çok değişkenlik kaynaklı rasch modelinde birden fazla değişken için hem puanlayıcı hem de okul değişkenlerine yönelik güvenirlik ve uyum durumunu test etmek adına yararlanabileceğinin uygun olduğu görülmüştür.
4. Puanlayıcı sayısının artırma veya azaltma yerine zaman ve maliyet açısından madde sayısı artırılmasının denenebileceği önerilmektedir.

5.2.2. İleride Yapılacak Araştırmalar İçin Öneriler

1. Araştırmaya konu olmuş bu çalışmada hem G Kuramı hem de ÇDKRM'nin güvenirliği belirlemede pratik ve etkili bir sonuç verdiği görülmektedir. Yapılan bu çalışma ve önceki çalışmalar ışığında madde veya görev sayısı arttırılarak farklı desenlerde benzer çalışmalar yapılabilir.

2.Farklı sınav metotları veya yaş grubunda uygulamalar yapılarak kuramlar arasında karşılaştırmalar yapılabilir. Mutlak ve bağıl değerlendirmeler yapılarak güvenirlik sonuçlarında bakılabilir.

3.G kuramı açısından farklı desen uygulamaları yapılarak G ve Phi katsayılarıyla beraber varyans yüzdelikleri karşılaştırılabilir.

4.Bütünsel dereceli puanlamanın yanı sıra analitik dereceli puanlama anahtarı da kullanılarak farklı alanlarda farklı değişkenler üzerinden çalışmalar yapılabilir.

KAYNAKLAR

- Airasian, P. W. (1991). *Classroom assessment*. New York: McGraw-Hill
- Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Prentice Hall/Pearson Education.
- Alharby, E. (2006). *A comparasion between two scoring methods, holistic vs analytic, using two measurement models, the generalizability theory and many-faced rasch measurement within the context of performance assesment* [Yayınlanmamış Doktora Tezi]. The Pennyslvania State University.
- Anadol, H. (2017). *Dereceli puanlama anahtarının güvenilirliğinin farklı deneyim yıllarına sahip puanlayıcıların kullandığı durumda incelenmesi*. [Yüksek Lisans Tezi]. Ankara Üniversitesi, Ankara.
- Arsan, N. (2012). *Buz pateninde hakem değerlendirmelerinin genellenebilirlik kuramı ve rasch modeli ile incelenmesi*. [Yayınlanmamış Yüksek Lisans Tezi]. Hacettepe Üniveristesi, Ankara.
- Atılğan, H. (2004). *Genellenebilirlik kuramı ve çok değişkenlik kaynaklı rasch modelinin karşılaştırılmasına ilişkin bir araştırma*. Hacettepe Üniversitesi. [Doktora Tezi], Ankara.
- Atılğan, H., Kan, A. Ve Doğan, N. (2006). *Eğitimde ölçme ve değerlendirme*. Ankara: Anı Yayıncılık.
- Atılğan, H. (2019). Reliability of essay ratings: A study on generalizability theory. *Eurasian Journal of Educational Research*, 19 (80), 133-150. Retrieved from <https://dergipark.org.tr/en/pub/ejer/issue/43338/548921>
- Atmaz, G. (2009). *Puanlama yönergesi (rubrik) kullanılması durumunda puanlayıcı güvenilirliğinin incelenmesi*. [Yüksek Lisans Tezi]. Mersin Üniversitesi
- Balcı, A. (2022). *Sosyal bilimlerde araştırma: Yöntem, teknik ve ilkeler*. Ankara: Pegem Akademi yayıncılık.
- Banerjee, M. Capozzoli M., Mcweeney, L. ve Sinha, D. (1999). Beyond kappa: A reiew of interrater agreement measures. *The Canadian Journal of Statistics*, 27(1), 3-23.
- Baykul, Y. (2021). *Eğitimde ve psikolojide ölçme: Klasik Test Teorisi ve Uygulaması*. Ankara: Pegem Akademi
- Bağcı, V. (2015). *Matematiksel muhakeme becerisinin ölçülmesinde klasik test kuramı ile*

- genellenebilirlik kuramındaki farklı desenlerin karşılaştırılması* [Yüksek Lisans Tezi].
Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.
- Bell, J. F. (1985). Generalizability theory: *The Software Problem*. *Journal of Educational Statistics*, 10(1), 19-29.
- Brennan, R.L. (2001). *Generalizability theory*. Springer- Verlag, New York.
- Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24, s.1-21.
- Bond, T. G. ve Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Boring, R. l. (2002). *Human and computerized essay assessment: A comparative analysis of holistic, analytic and latent semantic methods*. [Yayınlanmamış doktora tezi]. New Mexico State University, New Mexico.
- Brown, H. D. (2004). *Language assessment: Principles and Classroom Practices*. New York: Longman.
- Brown, W. L. Gorman, K. ve Du, Y. (1996). The Reliability and Validity of Mathematics Performance Assessment. *Paper for Presentation of The Annual Conference American Educational Research Association*, New York.
- Büyükkıdık, S. (2012). *Problem çözme becerisinin değerlendirilmesinde puanlayıcılar arası güvenilirliğin klasik test kuramı ve genellenebilirlik kuramına göre karşılaştırılması*. [Yüksek Lisans Tezi] Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.
- Büyüköztürk, Ş. (2007). Performansa Dayalı Durum Belirleme nedir? *İlköğretmen Dergisi*, 8, 28-32
- Cesur, K. (2008). “Öğrenci ve Öğretmenlerin, Üniversite Seviyesinde Dil Becerilerini Değerlendirmede Kullanılan Test Tekniklerine İlişkin Algulamaları”, Çanakkale On sekiz Mart Üniversitesi, Sosyal Bilimler Enstitüsü, İngilizce Öğretmenliği Bölümü, Yüksek Lisans Tezi, Çanakkale
- Damlar Demirci, P. (2019). *Açık Uçlu Soruların Puanlama Yöntemlerinin Genellenebilirlik Kuramına Göre İncelenmesi* [Yüksek Lisans Tezi]. Ege Üniversitesi.
- Deliceoğlu, G. (2009). *Futbol Yetilerine İlişkin Dereceleme Ölçeğinin Genellenebilirlik Ve Klasik Test Kuramına Dayalı Güvenirliklerinin Karşılaştırılması*. Ankara Üniversitesi Eğitim Bilimleri Enstitüsü Eğitimde Psikolojik Hizmetler Anabilim Dalı Ölçme Ve Değerlendirme Bilim Dalı [Doktora Tezi], Ankara.
- Deniz, K. Z., & Ilıcan, E. (2021). Comparison of G and Phi coefficients estimated in generalizability theory with real cases. *International Journal of Assessment Tools in*

- Education*, 8(3), (p.583-595).
- Dimitrov, M. D. (2002) Relability: Arguments for Multiple Perspectives and Potential Problems with Generalization Across Studies. *Educational and Psychological Measurement*, 62; (p.783-801).
- Doğan, N (2020). *Eğitimde ölçme ve değerlendirme (2. Baskı)* Ankara: Pegem Akademi (s.182-215).
- Engelhard, G. (1996). Evaluating rater accuracy in performance assessment. *Journal of Educational Measurement*, 33 (1), 56-70.
- Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Psychological Education and Exercises Science*, 5 (1), 13-14.
- Güler, N. (2008). *Klasik test kuramı genellenebilirlik kuramı ve rasch model üzerine bir araştırma*. [Yayımlanmamış Doktora Tezi]. Ankara: Hacettepe Üniversitesi.
- Güler, N. Gelbal, S. (2010) *Açık uçlu matematik sorularının güvenilirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi*, Kuram ve Uygulamada Eğitim Bilimleri.
- Güler, N. Kaya-Uyanık ve G. Taşdelen-Teker, G. (2012) *Genellenebilirlik Kuramı*. Ankara, Pegem Akademi,
- Gürses, R. (2006). Gürses, R. (2006). Dil öğretiminde ölçme ve değerlendirmenin yeri. *Türk dili araştırmaları yaylığı - Belleten*, 54 (2006/1), 131-154. Retrieved from <https://dergipark.org.tr/tr/pub/belleten/issue/45390/568981>
- Hambleton, R.K. & Swamitnathan, H. (1985). *Item response theory: Principles and Applications*. Kluwer-Nijhoff Publishing, Boston
- Helms, L. S. (1999). Basic concepts in classical test theory: Tests aren't reliable, the nature of alpha, and reliability generalization as meta-analytic method. (*ERIC Document Reproduction Service* No. ED 427 083).
- Hetherman, S. C. (2004). *An application of multi-faceted Rasch measurement to monitor effectiveness of the written composition in English in the New York City Department of Education*. Teachers College, Columbia University.
- Johnson, R. L., Penny, J. , Gordon, B. (2000). The Relation Between Score Resolution Methods And Interrater Reliability: An Emprical Study Of Analytic Scoring Rubric. *Applied Measurement In Education*, 13(2), 121-138.
- İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 31(2), 346-368.

- Karslı, M.D. (2003). *Öğretmenlik Mesleğine Giriş (Alternatif Yaklaşım)*, Ankara: Pegem A Yayınları.
- Karakaya Özzer, K. (2018). Açıköğretim sistemindeki açık uçlu soruların çok yüzeyli Rasch modeli analizi yöntemiyle puanlanması. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, 4(1), 61-77.
- Kim, S.C. ve Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generalizability theory and many-facet rasch measurement. *Journal of Applied Measurement*, 10(4), 408-422.
- Koşar H. (2022). *R ile geçerlilik ve güvenilirlik analizleri (3. Baskı)* Ankara, Pegem Akademi
- Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese to English. *Language Testing*, 21 (1), 1-27.
- Kutlu Ö. Doğan C. ve Karakaya İ. (2017). *Ölçme ve değerlendirme performans ve portfolyaya dayalı durum değerlendirme (5. Baskı)* Ankara Pegem Akademi Yayıncılık
- Lane, S. & Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoring system for sample essays. *Applied measurement in education*, 2(3), 195-205.
- Lee, Y. Kantor, R. ve Mollaun, P. (2002). score dependability of the writing and speaking section of new TOEFL. *Educational Testing Service*. 94
- Lee, Y. W. ve Kantor, R. (2003). Investigating differential rater functioning for academic writing samples: an mfrm approach. *Paper to be presented at the annual meeting of National Council On Measurement in Education* held in Chicago, IL.
- Lee, Y., W., Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for a new esl writing test through g theory. *International Journal of Testing*, 7(4), 353-385.
- Linacre, J. M. (1989). Many-faceted Rasch measurement [Doctoral dissertation, The University of Chicago]. Linacre, J. M.i, "A Facet Model for Judgmental Scoring". MESA Memo 61.199
- Linacre, J. M. (1993). *Generalizability theory and many facet rasch measurement*. Paper presented at the Annual Meeting Of The American Educational Research Association, Atlanta.
- Linn, R. L. ve Gronlund, N.E. (2000). *Measurement and assessment in teaching*. Prentice Hall, Columbus, Ohio, USA.
- Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied measurement in education*, 3(4), 331-345.

- Lynch, B. K. & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15(2), 158-180.
- MacMillan, P. D. (2000). Classical, generalizability and multifaceted Rasch detection of interrater variability in large, sparse data sets. *The Journal of Experimental Education*, 68(2), 167-190.
- MEB. (2018). *Ortaöğretim İngilizce dersi öğretim programı* (9.10.11 ve 12 sınıflar). Ankara: Millî Eğitim Bakanlığı.
- Mertler, C.A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment Research and Evaluation*, 7(25).
- Moskal, B. M. (2000). Scoring rubrics: What, when, how? Practical assessment. *Research, and Evaluation*, 7(3), 1-11.
- Nitko, A. J. (2001). *Educational assessment of students* (3. Baskı). Upper Saddle River, NJ: Merrill.
- Nalbantoğlu, F. (2009). *Performans Ölçümlerinde Genellenebilirlik Kuramı ile Farklı Desenlerin Karşılaştırılması*, [Yüksek Lisans Tezi], Hacettepe Üniversitesi, Ankara.
- Öner, N. (1994). *Türkiye’de kullanılan psikolojik testler. Bir başvuru kaynağı*. İstanbul: Boğaziçi Üniversitesi Matbaası.
- Seheryeli M. (2018). *Yazılı anlatım becerisi puanlama anahtarının güvenilirliğinin klasik test, genellenebilirlik ve madde tepki kuramlarına göre incelenmesi* [Yüksek Lisans Tezi]. Ankara: Gazi Üniversitesi
- Saal, F., Downey, R. ve Lahey, M. A. (1980). Rating the ratings: assessing the quality of rating data. *Psychological Bulletin*, 88(2), 413-428.
- Sebok, S. (2010). “*Pick me, pick me, I want to be a counsellor*” assessment of med. counselling application selection process using Rasch analysis and generalizability theory [Yayınlanmamış Yüksek Lisans Tezi]. University of Northern British Columbia: USA.
- Shavelson, J. R. & Webb, N. M. (1991). *Generalizability theory: a primer*. Newbury Park: CA: Sage Publications.
- Smith Jr, E. V. & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement*, 64(4), (s.617-639).
- Sudweeks, R. R., Reeve, S. ve Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9, (s.239-261)

- Sun, A. ve Valiga, M.J.. “Using Generalizability Theory to Asses the reliability of student Ratings of Academic Advising”, *Journal of Experimental Education*, Vol. 65, Issue 4, (p.367), Summer 1997.
- Szymanski, E. M. ve Linkowski, D. C. (1993). Human Resource Development: An Examination of Perceived Training Needs of Certified Rehabilitation Counselors. *Rehabilitation Counseling Bulletin*, 37 (2), 163-176.
- Şencan, H. (2005). *Sosyal ve davranışsal ölçmelerde güvenirlik ve geçerlik*. Ankara: Sözkese Matbaacılık
- Tan, Ş. (2014). *Öğretimde ölçme ve değerlendirme KPSS el kitabı*. Ankara: Pegem Akademi.
- Tanverdi, A. (2023). *Yabancı dil olarak türkçe öğretiminde c1 düzeyi konuşma becerisi analitik dereceli puanlama anahtarı geliştirilmesi* [Yüksek Lisans Tezi]. Ankara: Gazi Üniveritesi
- Tavşancıl, E. (2002). *Tutumların ölçülmesi ve spss ile veri analizi*. Ankara: Nobel Yayın Dağıtım Ltd. Şti.
- Tobar, D. A., Stegner, A. J. ve Kane, M. T. (1999). The use of generalizability theory in examining the dependability of scores on the profil of mood state. *Measurement in Physical Education And Execise Science*. 3 (3), 141-156.
- Turner, J. (2003). *Examining on art portfolio assesment using amany facet rasch measurement model*. [Yayınlanmamış doktora tezi]. Boston College, Boston.
- Turgut, M. F. ve Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme (2. baskı)*. Ankara: Pegem Akademi Yayıncılık.
- Ure, A. C. (2011). *The effect of raters and rating conditions on the reliability of the missionary teaching assessment*. Master Thesis, University of Brigham Young, USA.
- Wang, Z. (2005). *Estimating reliability under a generalizability theory model for writing scores in c-base*. Master Thesis, University of Missouri, Columbia.
- Wren, D. G. (2009). *Performance assesment: a key component of a balanced assesment system*. Research Brief: Report From Department Of Research, Evaluation and Assesment, 2.
- Wright, B. D. ve Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8 (3), 370.

- Wilson, M. ve Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. In M. Wilson, ve G. Engelhard (Eds.), *Objectivemeasurement: Theory into practice* (Vol. 5, pp. 113-134). Stamford, CT: Ablex
- Yang, R. (2010). *A many- facet rasch analysis of rater effecets on an oral english proficiency test* [Yayımlanmamış doktora tezi]. Purdue University, USA.
- Yavuz-Kırık, M. (2008). *Yabancı dil olarak İngilizce öğretmenlerinin ölçme değerlendirme bağlamında tutum ve yaklaşımları*. [Yayınlanmamış doktora tezi]. İstanbul Üniversitesi.
- Yelboğa, A. (2007). *Klasik test ve genellenebilirlik kuramına göre güvenirliğin bir iş performansı ölççeği üzerinde incelenmesi* [Yayımlanmamış Doktora Tezi]. Ankara Üniversitesi.
- Yelboğa, A. (2008). *Genellenebilirlik kuramı ile bir personel seçme yönteminde güvenirliğin değerlendirilmesi*. I. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi: 14-16 Mayıs 2008- Ankara.
- Yıldıztekin, B. (2014). *Klasik test kuramı ve genellenebilirlik kuramından puanlayıcılar arası tutarlılığın farklı yöntemlere göre karşılaştırılması* [Yüksek Lisans Tezi]. Hacettepe Üniversitesi
- Yılmaz-Koğar E. ve Koğar H. (2022). *M Plus ve R ile ileri düzey psikometri uygulamaları*. Ankara: Pegem Akademi Yayıncılık.

EKLER

EK 1: İNGLİZCE YETERLİLİĞİ PERFORMANS TESTİ

Name / Surname :

Class:

School :

TOPIC :WHAT ARE THE ADVANTAGES OF LIVING IN A SMALL TOWN ?

SUGGESTED IDEA FOR THIS TOPIC:

- ☐ Fresh food
- ☐ Clean air
- ☐ Safe environment
- ☐ Less traffic

(Use minimum 200 words)

EK 2: BÜTÜNSSEL PUANLAMA ANAHTARI (WRITING ASSESSMENT CRITERIA (PARAGRAPH))

	CONTENT (5)	ORGANIZATION(5)	VOCABULARY(4)	GRAMMAR (4)	MECHANICS(2)
0	• No relevance to the given topic or task not attempted • No development of ideas	• Completely disorganized • No unity and coherence	• No level-appropriate vocabulary • Poor knowledge of words and word forms	• Serious lack of language • No attempt to use learned structures	• No control over spelling and punctuation, capitalization, and paragraph indentation etc...
1	• Considerably irrelevant to the given topic • Shows little knowledge of subject	• Disorganized (no sentence with main idea or supporting sentences) • Serious disorganization. • Mostly incoherent; ideas and examples are disconnected / almost no signal or linking words	• Uses limited range of level-appropriate vocabulary. • Most of them are inaccurate or irrelevant • Translation-based errors	• The language is rarely clear • Significant defects in the use of even basic structures • Frequent errors in agreement, number, tense, negation, word order, articles, pronouns, prepositions.	• Occasional errors in spelling, punctuation, capitalization, paragraph indentation, etc., which do not interfere with meaning
2	• Inadequate coverage of the topic with inappropriate and/or insufficient ideas and examples	• Inadequately organized (badly expressed or misplaced sentences, insufficient supporting ideas/very short details/ no sentence which concludes the paragraph. • Considerably incoherent; ideas are not well connected	• Shows some use of varied word choice that is relevant to the content • Frequent word errors; inappropriate choice and usage	• Few examples of learned structures • Occasional mistakes in using appropriate grammatical structures, which cause problems in conveying the message.	• Almost no errors in conventions of spelling, punctuation, capitalization, paragraph indentation, etc
3	• Adequate coverage of the topic, but addresses the topic using somewhat appropriate and/or sufficient ideas	• Adequately organized (i.e: a weak sentence (containing the main idea) with correct placement, unbalanced development of supporting ideas or a weak concluding sentence) • Mostly coherent and fluent passage; ideas flow meaningfully and logically, but may contain some redundancy and some unclear connections	• Adequate range of level-appropriate vocabulary that is relevant to the content • Some errors of word/idiom choice, but effective transmission of meaning.	• No major difficulties in the use of appropriate grammatical structures • Some problems in the use of complex constructions.	
4	• Good coverage of the topic with relevant, clear, appropriate, and sufficient ideas and examples	• Generally well-organized (sufficiently developed sentences) • Coherent and fluent passage; ideas generally flow meaningfully and logically	• A wide range of level-appropriate vocabulary that is relevant to the content is used effectively and accurately • Very few mistakes in the word choice	• Use of a wide range of learned structures effectively and accurately. • Hardly any grammatical mistakes causing problem in conveying the message.	
5	• Very good coverage of the topic • Fully developed paragraph or paragraphs with very good justification	• Well-organized (a good sentence which contains the main idea, fully developed supporting ideas, and a good sentence which concludes the paragraph) • Very coherent and fluent passage with appropriate linking/signal words.			

EK 3: ÖLÇEKLE PUANLANDIRILMIŞ PERFORMANS TESTİ ÖRNEĞİ

TOPIC : WHAT ARE THE ADVANTAGES OF LIVING IN A SMALL TOWN ?

SUGGESTED IDEA FOR THIS TOPIC:

- Fresh food
- Clean air
- Safe environment
- Less traffic

Content = 0
Organization = 10
Vocabulary = 10
Grammar = 15
Mechanics = 0
+ 35

(Use minimum 200 words)

Living in a small city gives people confidence. Small town is intimacy, more
milbrity. However, there are many negative aspects of living in small cities. Big
ies have a lot of energy, a quiet life in small cities can make the person
g there sluggish both physically and mentally after a while. Calmness leads to
art, comfort leads to laziness. A character who reads and produces with
energy becomes withdrawn in a small city or falls into depressive states. In
cities, life flows slowly. Then this slowness becomes a behavior. Cities, towns
villages also get smaller towards the east in Anatolia. I love living in
mall city

EK 4: ÖLÇEKSİZ PUANLAMA YÖNTEMİ İLE PUANLANAN ÖĞRENCİ TESTLERİ ÖRNEKLERİ

Class: 12-B

School: Muhammed Hakkı Yılmaz Anadolu Lisesi

TOPIC : WHAT ARE THE ADVANTAGES OF LIVING IN A SMALL TOWN ?

SUGGESTED IDEA FOR THIS TOPIC:

- Fresh food
- Clean air
- Safe environment
- Less traffic

(Use minimum 200 words)

The air in the village is cleaner but in the city is more polluted. City life is exciting because there are plenty of things you can do. This is why living in the village can be boring. Cities are bigger than villages and more crowded because more people live there. People in the city may feel better because they don't have time for each other. In the village there is less crime so people feel safer. In the city life is more dangerous. The cost of living is more expensive in the city. There is a strong sense of community in the village and neighbours are usually more helpful. People are more friendly in the village. Life in the city is faster - in the village life moves a bit slower. People seem to be busier in the city because they are always working and running around. I love email. Looking at the sports is fun. People are very happy. There are not many queues in the markets. Everything is cheap. Small cities are cute places. Eat fresh fruits and vegetables in small cities. Small city has very good transportation. No boring traffic. It's hot in summers and it's in winters. Address directions are very easy.

TOPIC : WHAT ARE THE ADVANTAGES OF LIVING IN A SMALL TOWN ?

SUGGESTED IDEA FOR THIS TOPIC:

- Fresh food
- Clean air
- Safe environment
- Less traffic

(Use minimum 200 words)

Everyone knows Antalya, İzmir, Muğla, Ankara carry high cost of living. Rent in those cities is among the highest in the nation, and home ownership is out-of-reach for many.

But you don't need to live in a major metropolitan area along one of the coasts to experience the budget-busting impact of high costs. Inland cities like Konya, Erzurum and Kars, offer relatively high cost-of-living as compared to many small towns dotted across the nation's landscape.

Far away from the hustle and bustle of a big city, the slower, more relaxed pace of small towns can be a welcome change of pace.

Class: 12A

School: K. J. Somaiya M.T.A.L

109

TOPIC : WHAT ARE THE ADVANTAGES OF LIVING IN A SMALL TOWN ?

SUGGESTED IDEAS FOR THIS TOPIC:

- Fresh food
- Clean air
- Safe environment
- Less traffic

(Use minimum 200 words)

You can save money ✓

You can breathe fresh air. ✓

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı : Aykut TÜZAN

Doğum Yeri ve Tarihi :

Eğitim Durumu

Lisans Öğrenimi : Anadolu Üniversitesi
Eğitim Fakültesi İlköğretim Sınıf Öğretmenliği (2011-2015)

Yüksek Lisans Öğrenimi :

Bildiği Yabancı Diller :

Bilimsel Faaliyetleri :

İş Deneyimi

Çalıştığı Kurumlar :

- Şırnak Silopi Çiftlik İlkokulu (2015-2018)
- İstanbul Esenyurt Sultan Alparslan İlkokulu (2018-2019)
- Antalya Elmalı Atatürk İlkokulu (2019 – Halen)

İletişim

E-Posta Adresi:

Tarih :02.10.2023

BİLDİRİM

Hazırladığım tezin/raporun tamamen kendi çalışmam olduğunu ve her alıntıya kaynak gösterdiğimi taahhüt eder, tezimin/raporumun kâğıt ve elektronik kopyalarının Akdeniz Üniversitesi Eğitim Bilimleri Enstitüsü arşivlerinde aşağıda belirttiğim koşullarda saklanmasına izin verdiğimi onaylarım:

- ☐ Tezimin/Raporumun tamamı her yerden erişime açılabilir.
- ☐ Tezim/Raporum sadece Akdeniz Üniversitesi yerleşkelerinden erişime açılabilir. Tezimin/Raporumun yıl süreyle erişime açılmasını istemiyorum. Bu sürenin sonunda uzatma için başvuruda bulunmadığım takdirde, tezimin/raporumun tamamı her yerden erişime açılabilir.

24.10.2023

Aykut TÜZAN

İNTİHAL RAPORU

ORJİNALLIK RAPORU

% 17	% 17	% 3	%
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	acikbilim.yok.gov.tr	% 7
2	acikarsiv.ankara.edu.tr	% 3
3	ydy.ikcu.edu.tr	% 2
4	www.openaccess.hacettepe.edu.tr:8080	% 1
5	docplayer.biz.tr	% 1
6	www.ulusaltezmerkezi.net	% 1
7	dspace.akdeniz.edu.tr:8080	% 1
8	egitimbilim.akdeniz.edu.tr	<% 1
9	acikerisim.akdeniz.edu.tr:8080	<% 1

10	www.hku.edu.tr İnternet Kaynağı	<% 1
11	erl.ucc.edu.gh:8080 İnternet Kaynağı	<% 1
12	abakus.inonu.edu.tr İnternet Kaynağı	<% 1
13	dergipark.org.tr İnternet Kaynağı	<% 1

Alıntıları çıkart üzerinde Eşleşmeleri çıkar < 30 words
Bibliyografyayı Çıkart üzerinde



AKDENİZ ÜNİVERSİTESİ REKTÖRLÜĞÜ
Sosyal ve Beşeri Bilimler Bilimsel Araştırma ve Yayın Etiği Kurulu
KURUL KARARI

TOPLANTI TARİHİ : 24.10.2022

TOPLANTI SAYISI : 18

KARAR SAYISI : 371

Üniversitemiz Eğitim Fakültesi Eğitim Bilimleri Bölümü öğretim üyesi **Doç. Dr. Hakan KOĞAR**'ın yürütücülüğünü, **Aykut TÜZAN**'ın araştırmacılığını üstlendiği, *"Ortaöğretim İngilizce Yeterliğinin Puanlayıcılar Arası Güvenirliğinin Genellenebilirlik Kuramı ve Rasch Modeli ile Belirlenmesi"* konulu çalışmanın, fikri hukuki ve telif hakları bakımından metot ve ölçeğine ilişkin sorumluluğun başvurucuya ait olmak üzere, proje süresince uygulanmasının etik olarak **uygun olduğuna** oy birliği ile karar verilmiştir.

Prof. Dr. Hilmi DEMİRKAYA
Kurul Başkanı

Başkan
Prof. Dr.
Hilmi DEMİRKAYA

Başkan Yrd.
Prof. Dr.
Sibel MEHTER AYKIN
(İzinli)

Üye
Prof. Dr.
Ebru İÇİGEN

Üye
Prof. Dr.
Nurşen ADAK

Üye
Prof. Dr.
Sibel PAŞAOĞLU YÖNDEM

Üye
Prof. Dr.
Taner KORKUT

Üye
Prof. Dr.
Gökhan AKYÜZ