MATEMATİKSEL MUHAKEME BECERİSİNİN ÖLÇÜLMESİNDE KLASİK TEST KURAMI İLE GENELLENEBİLİRLİK KURAMINDAKİ FARKLI DESENLERİN KARŞILAŞTIRILMASI

VİLDAN BAĞCI

YÜKSEK LİSANS TEZİ

EĞİTİM BİLİMLERİ ANABİLİM DALI

EĞİTİMDE ÖLÇME VE DEĞERLENDİRME BİLİM DALI

GAZİ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

EYLÜL, 2015

TELİF HAKKI ve TEZ FOTOKOPİ İZİN FORMU

YAZARIN

Adı: Vildan

Soyadı: Bağcı

Bölümü: Eğitim Bilimleri

İmza:

Teslim tarihi: 01.06.2016

TEZİN

Türkçe adı: Matematiksel Muhakeme Becerisinin Ölçülmesinde Klasik Test Kuramı İle Genellenebilirlik Kuramındaki Farklı Desenlerin Karşılaştırılması.

İngilizce adı: Comparison of Different Designs in Generalizability Theory with Classical Test Theory in the Measurement of Mathematical Reasoning Ability.

ETİK İLKELERE UYGUNLUK BEYANI

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

Yazar Adı Soyadı: Vildan Bağcı

İmza:

Jüri Onay Sayfası

Vildan Bağcı tarafından hazırlanan "Matematiksel Muhakeme Becerisinin Ölçülmesinde Klasik Test Kuramı ile Genellenebilirlik Kuramındaki Farklı Desenlerin Karşılaştırılması" adlı tez çalışması aşağıdaki jüri tarafından oy birliği/oy çokluğu ile Gazi Üniversitesi Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda Yüksek Lisans tezi olarak kabul edilmiştir.

Danışman: Doç. Dr. Şeref TAN

Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Gazi Üniversitesi

Başkan: Prof. Dr. Mehtap ÇAKAN

Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Gazi Üniversitesi

Üye: Yrd. Doç. Dr. Deniz GÜLLEROĞLU

Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Hacettepe Üniversitesi

Tez Savunma Tarihi: 10.09.2015

Bu tezin Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda Yüksek Lisans tezi olması için şartları yerine getirdiğini onaylıyorum.

Prof. Dr. Servet KARABAĞ

Eğitim Bilimleri Enstitüsü Müdürü

Biricik Aileme,

TEŞEKKÜR

Her şeyden önce tez sürecimde yanımda bulunan ve benden hiçbir yardımını esirgemeyen destekleyici, saygıdeğer tez danışmanım Doç. Dr. Şeref Tan'a; lisansüstü eğitimim boyunca yetersizliğimi fark etmemi sağlayan, öğretileriyle ufkumu genişleten çok sevgili hocalarıma, bölümümüzün neşeli, hoş sohbetli, her anlamda yardımsever tavırlarıyla günüme emek ve güzellik katan asistan arkadaşlarıma; maddiyatın maalesef önemli rol oynadığı şu hayatta, eksikliğini hissettirmeyen TÜBİTAK'a; ve tabi ki mutlu zamanlarımın yanı sıra içler acısı her halime şahit olan, zamanından erken ayrılmamak için kontrat yaptığım, kuzen yoldaşım Şahide Altıncaba'ya; taa uzaklarda demeye gerek duymadığım, çünkü hemen Ankara'nın dibinde (Konya'da) ikamet eden başka bir kardeş dosta, birbirimizi kamçılayarak başladığımız ders çalışma yolunda desteklerini esirgemeyen Aybegüm Albay'a; kendi canımdan kanımdan olduğu çok bariz belli olan, birbirimizi ancak bu kadar anlarız dediğim sayın psikolog Ayşe Mirza'ya, içimdeki umudu yükselten, biriciğim ananecime; bana yaşamın iyi-kötü tecrübelerle şekilleneceğini, inandığın olguların ancak hayaller diyarında gerçekleşeceğini ve gerçek dünyanın ancak realist düşünceyle var olduğunu tecrübe ettiren Sayın B'ye; ve vee onlar benim her şeyim dediğim, her birini ayrı ayrı söylemeden geçemeyeceğim babacığım, annem, ablam ve kardeşim, bana öyle güzel bir ömür biçtiğiniz, hep beraber güldüğümüz, ağladığımız ve siz olduğunuz için sonsuz teşekkürlerimi sunuyorum. Ayrıca her imkânda benimle gurur duyduklarını dile getiren ve beni mahcup eden aileme, nihayet fırsatını bulmuşken şunu söylemek istiyorum: sizinle gurur duyuyorum…

MATEMATİKSEL MUHAKEME BECERİSİNİN ÖLÇÜLMESİNDE KLASİK TEST KURAMI İLE GENELLENEBİLİRLİK KURAMINDAKİ FARKLI DESENLERİN KARŞILAŞTIRILMASI

(YÜKSEK LİSANS TEZİ)

VİLDAN BAĞCI

GAZİ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ

EYLÜL, 2015

ÖZ

Bu çalışmada ilköğretim yedinci sınıf öğrencilerine yönelik matematiksel muhakeme performansının belirlenmesinde kullanılan ölçekten elde edilen ölçümlerin güvenirliğinin incelenmesi amaçlanmıştır. Bu amaçla ölçeğin, üç bağımsız puanlayıcı tarafından puanlanmasıyla elde edilen ölçümlerin güvenirliği; Klasik Test Kuramı ile Genellenebilirlik (G) kuramının çaprazlanmış ve yuvalanmış desenlerinde karşılaştırılmıştır. Her iki kuramda yapılan güvenirlik analizleri sonucu elde edilen güvenirlik katsayıları karşılaştırılarak, aralarındaki farklılıkların manidarlığı test edilmiştir. Ayrıca elde edilen bulgulara dayalı olarak kuramların birbirine göre avantajları tartışılmıştır.

Araştırmanın çalışma gurubunu, Konya ilinde bulunan, 2014-2015 eğitim-öğretim yılında yedinci sınıfta öğrenim gören 187 kişilik öğrenci grubu oluşturmuştur. Öğrencilerin

matematiksel muhakeme seviyelerini belirleyen ölçek uygulanmış ve öğrenci cevapları 3 bağımsız puanlayıcı tarafından analitik puanlama anahtarı ile puanlanmıştır.

Genellenebilirlik kuramı için iki farklı senaryo kullanmak üzere iki desen tasarlanmıştır. Bu desenlerden birincisi, öğrenci (ö), soru (s) ve puanlayıcı (p) değişkenleri olmak üzere, öğrencilerin aynı sorular üzerinden puanlayıcıların her biri tarafından puanlandığı Ö X S X P çapraz desenidir. İkinci desen ise, her bir puanlayıcının soruların sadece bir kısmını puanlamasıyla oluşan, puanlayıcı ve soru değişkenlerinin yuvalanmış, öğrencilerin ise bu değişkenlerle çaprazlanmış olduğu Ö X (S:P) desenidir.

Verilerin analizi 3 aşamada gerçekleşmiştir. Birinci aşamada genellenebilirlik kuramı kapsamında Ö X S X P ve Ö X (S:P) desenlerinde ayrı ayrı G çalışmaları yapılarak ana ve ortak etkiler için varyans değerlerinin kestirimine yönelik analizler yapılmış, ardından yapılan Karar çalışmaları ile de farklı senaryolar oluşturularak kabul edilebilir güvenirlik katsayıları kestirilmiştir. İlk iki aşamada yapılan analizlerde EduG6.1e programından yararlanılmıştır. Son aşamada ise performans görevinden elde edilen puanların klasik test kuramında güvenirlik analizleri yapılmıştır.

Araştırma sonucunda her iki kuramdan kestirilen güvenirlik katsayıları da kabul edilebilir düzeyde bulunmuştur. Ö X (S:P) deseninde G çalışması sonucu kestirilen G ve Phi katsayıları Ö X S X P deseninden daha yüksek bulunmuştur. Klasik test kuramında her 3 puanlayıcı için ayrı ayrı hesaplanarak elde edilen Cronbach alfa katsayıları ise, her iki desende bağıl ölçme için kestirilen G katsayıları ile oldukça paraleldir. Ayrıca Genellenebilirlik kuramında yapılan karar çalışma ile de yüzeylerin sayılarının mutlak ve bağıl hata varyanslarına etkisi belirlenmiştir. Dolayısı ile G kuramı ile yapılan analizlerin KTK'ya göre daha detaylı bilgi verdiği görülmüştür.

Bilim Kodu: 10211

Anahtar Kelimeler: klasik test kuramı, genellenebilirlik kuramı, güvenirlik, G çalışması, K çalışması.

Sayfa Adedi: 116

Danışman: Doç. Dr. Şeref TAN

THE COMPARISON OF DIFFERENT DESIGNS IN GENERALIZABILITY THEORY WITH CLASSICAL TEST THEORY IN THE MEASUREMENT OF MATHEMATICAL REASONING ABILITY

(M.S THESIS)

VİLDAN BAĞCI

GAZI UNIVERSITY

GRADUATE SCHOOL OF EDUCATIONAL SCIENCES

SEPTEMBER, 2015

ABSTRACT

The purpose of this study is to examine the reliability of measurements obtained "Mathematical Reasoning Measurement Scale" for seventh grade students. For this purpose, the reliability of the measurements obtained by the scoring by three independent raters were compared by using Classical Test Theory and Generalizability Theory which has crossed and nested designs. The reliability coefficients obtained by reliability analyses of both theories were compared with each other and the significant test was made for the difference between them. Also, the advantages of theories were discussed, based on the findings.

This study has been conducted with totally 187 students in the seventh grade in the spring term of 2014-2015 academic year in Konya. "Mathematical Reasoning Measurement Scale" was applied to mentioned students and the student responses were scored by three independent raters with analytical rubric.

Two designs of Generalizability Theory were deliberated for the study. The first design is a fully crossed design S X I X R (student x item x rater) which all of the students answered all of the items and scored by all of the raters. The second design is a partially nested design S X (I:R) which students anwered all off the items by all of the raters, but the items were nested in raters.

Data analysis occurred in three stages. Firstly, Generalizability study which is enabled to identfy which sources of error variances have the greatest influence on the measurement results were carried out for both designs and then Decision study allowed the effects of different designs to contributions of measurement error. EduG6.1e was used to carry out analyses so far. At the last step, the reliability of the scores obtained from the scale were analyzed in Classical Test Theory.

Consequently, the reliability coefficients were estimated of both theory have been found acceptable. The reliability coefficients obtained from S X (I:R) design are relatively higher than the ones obtained from S X I X R design.

The Cronbach's alpha coefficients obtained by estimated for each of three raters in classical test theory and G coefficients for relative measurements in both designs is quite parallel.In addition, the impact of the number of facets to absolute and relative error variance was examined in decision studies. Therefore, analysis by G theory was found to give more detailed information than Classical Test Theory.

Science Code: 10211

Key Words: classical test theory, generalizability theory, reliability, G Study, D Study.

Page Number: 116

Supervisor: Doç. Dr. Şeref TAN

İÇİNDEKİLER

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

SİMGELER VE KISALTMALAR LİSTESİ

KTK	Klasik Test Kuramı
GK	Genellenebilirlik Kuramı
NAEP	National Assessment of Educational Progress
TIMMS	Trends in International Mathematics and Science Study
MTK	Madde Tepki Kuramı

BÖLÜM I

GİRİŞ

Bu bölümde, araştırmaya ilişkin problem durumu, araştırmanın amacı ve önemi, problem cümlesi, alt problemler, sayıtlılar ve sınırlılıklar yer almaktadır.

Problem Durumu

Matematik, insanı doğadaki diğer canlılardan ayıran en temel özelliği, "düşünebilmeyi" geliştiren önemli araçlardan birisidir (Tural, 2005). Öyle ki insanların yapılar arasında ilişki kurabilmesi, çözümleyebilmesi, anlam çıkarabilmesi gibi zihinsel becerilerinin tamamı düşünebilme yetisinde saklıdır. Bu nedenle matematik eğitiminin temel eğitimin önemli bir kısmını oluşturduğu söylenebilir.

Matematik, sadece sayıları, basit işlemleri öğretmekle kalmamakta; düşünme, olaylar arasında bağ kurma, akıl yürütme, tahminlerde bulunma, problem çözebilme gibi zihinsel süreçleri de kapsamaktadır (Umay, 2003). Aynı zamanda matematik bir düşünme alışkanlığı ya da düşünme biçimi olarak ifade edilmektedir (Baki, Güven ve Karataş, 2002). O halde matematik eğitiminin en önemli amaçlarından birisinin, bireyin matematiksel düşünme ve muhakeme yeteneğinin gelişmesine katkı sağlamak olduğu söylenebilir. Matematiksel düşünce ve muhakeme yeteneği gelişmiş olan bireylerin, yukarıda sayılan akıl yürütme, problem çözme gibi tüm zihinsel süreçlerin öğreniminde başarı göstermeleri kaçınılmazdır. Peki, bu zihinsel süreçlerin kazanılmasında önemli rolleri olan matematiksel düşünce ve matematiksel muhakeme kavramlarından ne anlaşılmaktadır?

Matematiksel Düşünme ve Muhakeme

Matematiksel düşünme, "tahmin edebilme, tümevarım, tümdengelim, betimleme, genelleme, örnekleme, biçimsel ve biçimsel olmayan usa vurma, doğrulama ve benzeri karmaşık süreçlerin bir birleşim kümesi olarak tanımlanmaktadır (Liu Po-Hung, 2003)". Söz konusu kavrama göre matematiksel düşünmenin bireyin çevresindeki nesneleri algılama ve onlar arasındaki ilişkiyi anlamlı kılma çabasına girdiği an oluşmaya başladığı söylenebilir (Tall, 1995).

Matematiksel muhakeme kavramının açıklamasına geçmeden önce muhakemenin ne olduğunun anlaşılması gerekmektedir. "Muhakeme; sonuçlardan, yargılardan, gerekçelerden ya da önermelerden bir sonuç çıkarma işlemi; önermeleri, yargıları bir kalıba bağlamak ve bunlardan emin olmaktır (Altıparmak ve Öziş, 2005)". O halde muhakeme, çeşitli düşünme tarzlarını içeren bir etkinliktir (Peresini ve Webb, 1999). Bu çeşitli düşünce tarzlarından kasıt eleştirel ve yaratıcı düşünmedir. Bir başka deyişle muhakeme, düşünmenin ileri basamaklarında ortaya çıkan bir beceridir (Umay, 2003). Bu açıdan, insanın görüş ve düşüncelerini mantıksal gerekçelere dayandırdığı bir bilişsel süreç olarak da tanımlanabilir.

Keşfetme, merak gibi duyguların tetiklediği neden, niçin soruları bireyin dünyaya gelmesiyle başlamaktadır. Bebekler etraflarını inceleyerek, gözlemleyerek; çocuklar sorular sorarak birtakım ilişkileri öğrenmeye çabalamaktadırlar. O halde, bireylerin doğumla birlikte birtakım zihinsel aktivitelerin içinde bulundukları (düşünmek, olaylar arasındaki ilişkileri keşfetmek, muhakeme etmek) söylenebilir. Muhakemenin doğuştan gelen bir yetenek olduğunun bilinmesinin yanı sıra; çevrenin, özellikle eğitim ve öğretim kurumlarının etkisiyle geliştirilebilir olduğu kabul görmektedir. Eğitim ve öğretimde öğrencilere olayları/durumları nedenleriyle açıklayabilme yaklaşımı söz konusudur. Bu yaklaşım muhakeme yapısının gelişiminin sağlanması ile örtüşür (Altıparmak ve Öziş, 2005; Çoban, 2010). Dolayısı ile muhakeme yapabilmenin bir yetenek olduğu, fakat çevresel ve eğitim yollarıyla geliştirilebilir olduğu kabul görülmekte ve bu alan araştırmacıların ilgi konusu olmaya devam etmektedir (Umay, 2003; Altıparmak ve Öziş, 2005; Çoban, 2010).

Matematik Eğitiminde Matematiksel Muhakemenin Yeri

Muhakemenin en yoğun kullanıldığı alanlardan birisi şüphesiz matematiktir. Matematik sayıları, cebiri, geometriyi, alan hesaplamayı, problem çözmeyi ve bunun gibi birçok konuyu öğretirken öğrencinin gerekçeli düşünmesini, akıl yürütmesini, tahminde bulunmasını,

sorgulamasını ve sonuca ulaşmasını da öğretir. Dolayısı ile matematiksel muhakemenin, matematiğin doğası gereği matematik öğretiminin temelini oluşturduğu söylenebilir (Umay, 2003).

Alanyazındaki çalışmalarda matematiksel muhakemenin basit anlamda, problem çözme yeteneği olduğu belirtilmiştir. Fakat bir hesap makinesinin de bu işlevi gördüğü düşünülürse muhakemenin problem çözme becerisinden ince bir çizgiyle ayrıldığı söylenebilir (Krulik ve Rudnick, 1993). O halde, matematiksel muhakemeyi problem çözme becerisinden farklı kılan özellikleri nelerdir?

Matematiksel muhakeme sistemi ilk olarak sabit şablonlara bağlı kalmayan problem tiplerini çözebilen ve çözümleri ifadelendiren bir yapı gerektirmektedir. Bu takdirde muhakeme sisteminin ifade gücü yüksek olan bir dizi temsile dayandığı söylenebilir. İkincisi, bilinen örtük yapının açık (anlaşılır) hale getirilmesi, yani her bilgi için eşdeğer temsiller oluşturmaktır. Örneğin "Kaplanlar tehlikeli hayvanlardır" ve "Bu bir kaplandır" bilgilerinden yararlanarak "O halde bu hayvan tehlikelidir" yargısına varabilmek birinci ve ikinci bilginin altındaki örtük bilgiyi açığa çıkarma işlemidir. İfadelerden oluşan bir problemi formüle dönüştürmek yine aynı bilgi için eşdeğer temsilciler oluşturmaya örnektir. Üçüncü olarak muhakeme sisteminin, problem çözme aşamasında bir çözüme ulaşıldığında ya da çabaların sonuçsuz olduğu durumda bile uygulanan dönüşümleri kontrol eden bir yapısının olması gerekmektedir. Matematiksel muhakeme sistemini oluşturan yukarıdaki her bir yapı için, makul ölçüde sayısal yeterliliğe sahip olmanın gerekliliği kuşkusuzdur (Krulik ve Rudnick, 1993).

Öğrenci, muhakeme ile üst düzey düşünmenin temel bileşenlerini kullanır ve muhakeme süreci sonundaki değerlendirmelerine bakarak mevcut bilgilerini yeniden yapılandırabilir. Bir problemi çözmeye başlamadan önce problemi mümkün olduğunca inceler, soruları anlamaya çalışır, çözüm sırasında da öncelikle dayanakça ve gerekçeleri gösterir. Benzer şekilde bir probleme farklı çözüm önerilerinin sunulması da o problemin matematiksel açıdan neyi ortaya koyduğunu bilen, bahsedilen süreçlerden geçen öğrencilerin varlığını göstermektedir. Öğrencilerin kendi fikirlerini ifade etmeleri, doğruluğunu ispatlamak için tartışmaları, düşüncelerinin eksik kalan kısımlarını fark etmeleri ve diğer öğrencilerin düşüncelerini eleştirebilmeleri, ancak matematiksel muhakemenin öğrenildiği bir sınıfta gerçekleşir (Pilten, 2008; Altıparmak ve Özişş, 2005).

Yukarıda sayılan özelliklere bakarak denilebilir ki matematik eğitimi muhakeme yeteneğinin geliştirilmesinde büyük bir paya sahiptir. Şöyle ki bir sınıftaki öğrencilerin problem çözme

durumunda kullandıkları stratejileri ve kuralları sırasıyla açıklama eğiliminde olmaları, üretilen çözümler hakkında tartışmaları ve daha iyi bir hale getirmeye çalışmaları, o sınıftaki muhakeme sistemini oluşturan matematiksel yapılardır (Umay, 2003; Back ve Wright, 1999). O halde muhakeme becerisinin gelişmesi için belirtilen davranışların üzerinde durulduğu sınıfların olması gerektiği açıktır. Öğrencilerin böyle sınıflarda muhakeme becerilerini geliştirebilmeleri, üst bilişsel çözüm stratejilerini gerektiren sorular için de bir basamak oluşturmaktadır. Ayrıca öğrencilerin matematiksel muhakemelerini ölçen çeşitli ölçek ve sınavlarla bu yetilerinin düzeyi belirlenmelidir.

Matematiksel Muhakeme Becerisinin Ölçülmesi

Öğrencilerin herhangi bir beceri ya da konuda, o beceri\konu için gerekli olan ön koşul bilgi ve davranışlara sahip olup olmadıklarını, öğrenme eksikliklerini, yanlışlıklarını ve düzeylerini belirlemek amacıyla ölçme değerlendirme çalışmalarına başvurulur. Ölçme en bilinen tanımıyla belli bir niteliğin gözlemlenerek, gözlem sonuçlarının (ölçümlerin) sayı ya da sembollerle ifade edilmesi olarak karşımıza çıkarken (Turgut, Baykul, 2010, s. 103); değerlendirme ölçme sonuçlarını belli bir ölçütle karşılaştırarak karar vermektir (Turgut, 1997).

Ölçmenin uygun araçlarla yapılmasının önemli olması gibi, yapılan ölçümlerin geçerli ve güvenilir olması da büyük önem taşımaktadır. Geçerlilik klasik anlamda ölçme aracının istenen amaca hizmet etme derecesi, başka değişkenlerle karıştırılmaması olarak tanımlanırken; güvenirlik, ölçme yapan kişinin mümkün olduğunca hatasız ölçümler elde etmesi ile ilgilidir (Baykul, 2000). Ölçmelerin güvenilir olması, ölçmeye karışan sistematik, sabit ve tesadüfî hata kaynaklarının düşük olmasını gerektirir. Bu hata kaynaklarından sabit ve sistematik hatalar araştırmacı tarafından ölçme esnasında minimize edilebilir ya da uzaklaştırılabilinirken, tesadüfî hata kaynakları için bu olanaksızdır. Özellikle bilgi düzeyi ötesinde düşünme gerektiren, karmaşık yapıdaki görevlerin ölçülmesinde puanlayıcıdan kaynaklı tesadüfî hatanın minimize edilmesi ve kabul edilebilir bir güvenirlik katsayısının elde edilmesi için puanlayıcı sayısının arttırılması yoluna gidilebilmektedir (Turgut ve Baykul, 2010). Bunun yanı sıra ölçme aracının; ölçmenin konusuna, kazanım düzeylerine göre belirlenmesi ölçmenin güvenilir ve geçerli olmasını etkileyen önemli bir faktördür. O halde matematiksel muhakeme gibi bilgi düzeyi ötesinde süreçleri gerektiren ölçmelerde, performansa dayalı durum belirleme öğrencilerin başarıları hakkında karar vermede uygun bir yöntem olacaktır. Bu bağlamda öncelikle performans kavramının açıklanmasında yarar vardır.

Performans Kutlu ve diğerleri (2009) tarafından üst düzey zihinsel süreç (kavrama, uygulama basamakları) gerektiren görev, soru ya da etkinliklerin yerine getirilirken ortaya konan çaba ve ürün olarak açıklanmıştır. Öğrencinin alt düzey düşünme süreçlerinden çok, üst düzey düşünme gerektiren görevlere odaklanılması istenmiştir. O halde öğrenci bilgi düzeyini aşmalı ve yeni bilgiler üretme aşamasına gelmelidir (Kutlu, Doğan ve Karakaya, 2009).

Performans değerlendirme ise, öğrencilerin öğrendiklerini gerçek yaşam problemleri üzerinde uygulayabilmeleri ile ilgilidir (Acar ve Anıl, 2009). Klasik değerlendirme yöntemlerinden en temel farkı budur. Bir diğer farkı ise öğrencide bilginin var olup olmamasını sorgulamaktan ziyade, öğrencinin o bilgiyi kullanırken gösterdiği performansı ya da gelişimi izlemesidir. Ayrıca öğretmenlerin öğrencilerini belli bir alanda bilgi ve yetilerini sergilediklerinde, bir yanıtı yapılandırdıklarında gözlemleyerek başarıları hakkında karar verebilmeleri performansa dayalı değerlendirme ile sağlanır (Kutlu, vd., 2009). Büyüköztürk'e (2007) göre performans değerlendirmenin amacı, öğrencilerin uzun süreli öğrenmelerinin bir fonksiyonu olarak tanımlanabilen yeteneklerin değerlendirilmesidir.

NAEP (National Assessment of Educational Progress); (2002) matematiksel muhakeme becerilerini problem çözme becerisi içerisinde ele almaktadır. Aynı şekilde bu beceri matematiksel tahminleri oluşturma, matematiksel tartışmaları geliştirme ve matematiksel bilgileri çeşitli şekillerde sunma gibi çeşitli üst düzey (bilgi düzeyi ötesi) performansları içermektedir (Pilten, 2008). Dolayısı ile öğrencilerin bir problem üzerinde düşünerek içeriği ile ilgili karar vermeleri, çözüm için gerekçeler sunmaları, buna uygun bir plan seçmeleri ve çözümü yorumlamaları gibi üst düzey becerileri ölçmeye en çok imkân tanıyan madde türünün açık uçlu maddeler olduğu söylenebilir.

Bu çalışmada alanyazında yer alan araştırmalarda kullanılan veri toplama araçları ve kurumlar tarafından ortaya konulmuş değerlendirme ile ilgili kriterler (yukarıdaki bilgiler doğrultusunda) göz önüne alınarak; öğrencilerin muhakeme performanslarını ölçmeye yönelik açık uçlu sorulardan oluşan bir ölçek kullanılmıştır. Öğrencilerin açık uçlu sorulara verdiği yanıtların güvenilir bir şekilde puanlanması için beklenilen davranışlar önceden belirlenmiş ve bu davranışlara göre dereceli puanlama anahtarı hazırlanmıştır. Bu bağlamda dereceli puanlama anahtarları ve çeşitleri hakkındaki bilgilere sırası ile değinilecektir.

Dereceli puanlama anahtarları, öğrencilerin sorulara verdikleri yanıtları önceden belirlenmiş kriterlere göre puanlamada kullanılan kılavuzlardır (Turgut ve Baykul, 2010). Puanlama yöntemine göre iki farklı dereceli puanlama anahtarı bulunur: Bunlar bütüncül ve analitik puanlama anahtarlarıdır.

Bütüncül puanlama anahtarlarında performans; genel olarak, öğelerine ayrılmadan bir bütün olarak puanlanmaktadır (Haladyna, 1997). Analitik puanlama anahtarları ise performansı öğelere ayırır ve her bir öğe için ayrı bir bütüncül anahtar geliştirilir. Analitik puanlama anahtarları performansın her bir alt boyutu için bilgi verdiği için daha detaylı ve iyi tanımlanmış anahtarlardır (Haladyna, 1997; Moskal, 2000). Bu nedenle bu çalışmada matematiksel muhakemenin belirlenmesinde kullanılan ölçeğin puanlanmasında, muhakeme becerisinin her bir alt boyutunu dikkate alarak hazırlanmış analitik puanlama anahtarı kullanılmıştır.

Performansın belirlenmesinde her ne kadar puanlama anahtarları kullanılsa da, puanlayıcılar arasındaki görüş ayrılıkları, objektif puanlayamama ya da çevresel değişiklikler gibi hatalar ölçmeyi olumsuz yönde etkiler. Özellikle puanlamanın birden fazla puanlayıcı tarafından yapıldığı ölçmelerde, puanlayıcılar da bir hata kaynağı olarak karşımıza çıkmaktadır. Puanlayıcının puanlama deneyiminin eksikliği, yaşı, cinsiyeti, kişisel özellikleri gibi pek çok sebepten ötürü puanlayıcılar arası tutarlılığın düştüğü söylenebilir. Bu sebeplerden dolayı öğrenci yanıtlarını değerlendirmeden önce yapılan ölçmenin güvenirliğinin incelenmesi gerekmektedir. Bu bağlamda bu çalışmada, yedinci sınıf öğrencilerinin matematiksel muhakeme becerileri, analitik puanlama anahtarı ile üç puanlayıcı tarafından puanlanırken, diğer hata kaynaklarının var olup olmadığı, hata kaynakları varsa bunların etkisinin belirlenebilmesine yönelik güvenirlik analizlerinin yapılması amaçlanmıştır. Bu amaç doğrultusunda:

Güvenirliğin kestirilmesinde hem değişkenlik kaynaklarını hem de bunlar arasındaki etkileşimleri dikkate alan bir yöntem olan Genellenebilirlik Kuramı çalışmada kullanılmıştır. Ölçümlerin güvenirliğini kestirmede Genellenebilirlik Kuramı ile karşılaştırılan diğer yöntem ise gerçek puan modeline dayanan Klasik Test Kuramı olmuştur.

Böylece öğrencilerin matematiksel muhakeme becerilerinin güvenilir ve geçerli bir şekilde ölçülmesi sağlanarak, ölçmeyi etkileyen olumlu-olumsuz değişkenlik kaynakları belirlenebilmiştir. En uygun güvenirlik indekslerinin hangi kuram ve hangi ölçme senaryoları ile elde edildiği saptanmıştır.

Çalışmanın kuramsal çerçevesi kapsamında sırası ile Genellenebilirlik Kuramı ve Klasik Test Kuramı sunulmuştur.

Genellenebilirlik Kuramı

Genellenebilirlik kuramı, ya da G kuramı, özellikle farklı hata kaynaklarını konu edinen ölçmelerde, bu hata kaynakları ile bunların etkileşiminden kaynaklı hataların kestirimini sağlayan ve temelinde varyans analizine (ANOVA) dayanan istatistiksel bir kuramdır (Shavelson ve Webb, 1991; Brennan, 2001a). G kuramı Klasik Test kuramının bir uzantısıdır. Klasik Test kuramı güvenirliğin sadece bir hata kaynağına bağlı kestirimine izin veren gerçek puan modeline dayalı bir kuramdır. G kuramı ise Klasik Test kuramının en açık sınırlılıklarından biri olan tek hata kaynağı içerme durumuna tepki olarak geliştirilmiştir (Güler, Uyanık ve Teker, 2012).

G kuramı sadece yapılan ölçümlerin güvenirliği hakkında tahmin yapmakla kalmaz, aynı zamanda gelecekteki uygulamalarda ölçme işlemleri geliştirmek için kullanılacak hata kaynakları hakkında bilgi sağlar.

Shavelson ve Webb (1991), Genellenebilirlik kuramının Klasik Test kuramının genişletilmiş bir uzantısı olduğunu dört maddeyle belirtmişlerdir.

1. Genellenebilirlik kuramı tek bir analizle birçok hata kaynağını kestirebilmektedir.
2. Değişkenlik kaynaklarının her birinin büyüklüğünü belirleyebilir.
3. Bireylerin performanslarına yönelik hem bağıl hem de mutlak kararlar alınabilir ve buna bağlı olarak iki farklı güvenirlik katsayısı hesaplanabilir.
4. İstenilen ölçme durumlarında, en uygun güvenirlik katsayısının elde edilebileceği Karar çalışmaları yapılabilir.

Genellenebilirlik kuramı varyans analizi (ANOVA) ve Klasik Test kuramının bir uzantısı olarak görülse de yukarıdaki maddelerden de anlaşılacağı üzere KTK'nın genişletilmiş hali olup matematiksel modeli itibari ile de varyans analizine benzemektedir. Veri setindeki toplam varyansı potansiyel varyans kaynaklarına bölmesi varyans analizi temelinde olduğunun göstergesidir.

Varyans analizinde toplam varyans, varyans bileşenlerine ayrılarak, bireylerin gözlenen puanlarının evren puanlarına genellemesi sağlanmaktadır (Brennan, 2001). Varyans analizinde "faktör" olarak adlandırdığımız bu hata kaynakları, Guttman tarafından değişkenlik kaynağı ya da yüzey (facet) olarak ifade edilmiştir. Değişkenlik kaynağı Güler (2012) tarafından, benzerlik gösteren ölçme durumları olarak tanımlanmıştır. G kuramında bu değişkenlik kaynakları, madde, puanlayıcı, zaman vb. olabilmekte ve bunlar ölçme hatasının olası kaynakları olarak görülmektedir. Dolayısı ile değişkenlik kaynaklarından gelen

varyansların olabildiğince küçük olması beklenmektedir (Alkan, 2013). Değişkenlik kaynaklarının ya da yüzeylerin (facet) düzeyleri bulunmaktadır. Bu düzeylere ise koşul (condition) adı verilir. Örneğin puanlayıcılar ve maddeler çalışmadaki yüzeyler ise her bir puanlayıcı ve madde birer koşuldur (Güler vd., 2012). G kuramında bir yüzeyin olası koşullarının genelde sonsuz sayıda olduğu varsayılmaktadır. Bu durumda G kuramı için önemli iki kavramı daha tanımlamak gerekir. Araştırmada alınabilecek olası tüm koşullardan elde edilen sonuçların evrenine "kabul edilebilir gözlemlerin evreni (the universe of admissible observation)" adı verilir. Araştırmacının genellemek istediği koşulların tamamı ya da kullanılan yüzeylere bağlı ölçme sonuçlarının oluşturduğu evrene ise "Genellenebilirlik evreni (the universe of generalization)" adı verilir (Shavelson, Webb ve Rowley, 1989).

Araştırmaların pek çoğunda bireyler ya da öğrenciler istenilen kararların alınacağı ölçme hedefi durumundadırlar. Bu nedenle G kuramında genellikle bireyler ölçme objesi (the object of measurement) olarak ele alınırlar. Bireyler (ölçme objesi) arası farklılıklar doğal ve muhtemel olduğu için, bireyler hatanın değişkenlik kaynağı olarak ele alınmaz. Bununla birlikte maddelerin ya da diğer değişkenlerin ölçme objesi olduğu ölçme durumları da bulunmaktadır.

G kuramında ölçme objelerinin genelleme evrenindeki tüm koşullardan aldığı puanın ortalamasına evren puanı (universe score) denilmektedir. Evren puanı bireylerin (ölçme objesinin) genelleme evrenindeki ideal puanı olup Klasik Test kuramındaki gerçek puan kavramına benzerdir (Güler vd., 2012).

Genellenebilirlik kuramında bir araştırmadaki değişkenlik kaynakları örnekleme durumuna göre tesadüfî (random) ya da sabit (fixed) olabilir. Tesadüfî değişkenlik, koşulların evren ya da ilgili popülâsyondan tesadüfî olarak örneklenmesi demektir. Bir diğer ifadeyle araştırmacı değişkenlik kaynağını ilgili tüm durumlara genellemek istiyorsa bu değişkenlik kaynağı tesadüfî olacaktır. Bunun yanında araştırmacının genelleme yaptığı evren sonsuz büyüklükte ve değiştirilebilir nitelikte varsayılıyorsa, değişkenlik kaynağı tesadüfî kabul edilir. Sabit (fixed) değişkenlik kaynakları ise araştırmacının belirlediği ve bunun dışında genelleme yapmak istemediği veya çalışılan evrenin küçük olmasından kaynaklı o evrende çalıştığı durumlarda kullanılan yüzeylerdir (Cardinet, Johnson ve Pini, 2010). Sabit yüzeylerde bir örnekleme durumu gerçekleşmediği için bu durumdan kaynaklı varyans elenir ve hata varyansı azalır. Bu sebeple sabit yüzeylerdeki güvenirlik için hesaplanacak katsayılar tesadüfî yüzeylere göre daha yüksek değerler almaktadır (Güler vd., 2012).

Genellenebilirlik kuramında güvenirliğin incelenmesinde iki çalışma söz konusudur: 1. Genellenebilirlik çalışması (G Study, G-çalışması), 2. Karar çalışması (D Study, K-çalışması). Bu çalışmalar sırası ile aşağıda açıklanmaktadır.

Genellenebilirlik (G) Çalışması

G-çalışmasının amacı ölçmedeki çeşitli varyans kaynakları hakkında mümkün olabildiğince bilgi verebilmektir. Bu sebeple G-çalışmalarının deseni; potansiyel varyans kaynaklarını tanıtmalı ve içermelidir. Başka bir değişle G-çalışmaları kabul edilebilir gözlemler evrenini olabildiğince geniş tanımlamalıdır (Shavelson ve Webb, 1991).

G-çalışmalarında puanların değişkenliğinde rol oynayan tüm varyans bileşenleri ve bunlar arasındaki etkileşimler tek bir analizle (ANOVA) kestirilir (Güler ve Gelbal, 2010). Bu varyans değişkenleri tek bir madde ya da tek bir puan üzerinden kestirilen değerlerdir. Amaç gözlenen puanlar ile evren puanları arasındaki ilişkileri incelemektir. Bu kestirimler de ölçme durumlarının daha operasyonel olabilmesi ya da karar çalışmalarında ölçme objeleri lehine uygun kararlar verilebilmesi için kullanılır.

Karar (K) Çalışması

K-çalışmaları G-çalışmalarından elde edilen bilgilerden yararlanarak araştırmacının yaptığı ölçmede belli bir amaç için en uygun tasarıyı gerçekleştirebilmesini sağlar. Başka bir değişle bu amaçla yaptığı ölçmedeki hataları minimize edecek sonuçları ortaya koyar.

Karar çalışmalarını planlarken takip edilmesi gereken adımlar Shavelson ve Webb (1991) tarafından şu şekilde özetlenmiştir:

a. Genelleme evreni tanımlanır, araştırmacının üzerinde genellemek istediği yüzeylerin sayısı ve genişliğini belirlenir.
b. Ölçmenin amacına uygun değerlendirme türü belirlenir, mutlak ya da göreceli değerlendirme durumuna göre ölçme hatası tanımlanır ve buna bağlı güvenirlik katsayıları hesaplanır.
c. G çalışmasından elde edilen değişkenlik kaynaklarının büyüklüklerini kullanarak minimum hata ve maksimum güvenirlik elde edebilecek çalışmalar düzenlenir.

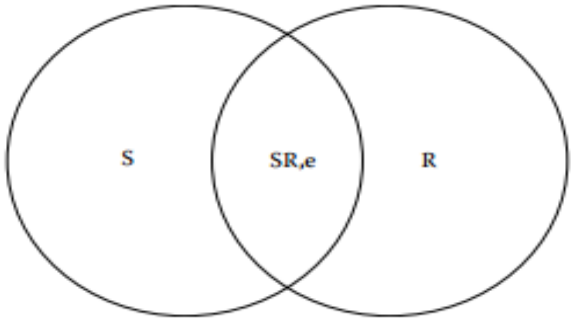
Kısacası G çalışmaları, değişkenlik kaynaklarını kestirerek o ölçmedeki zayıf ve güçlü yanları sunmakta iken, K çalışmaları bu zayıflıkların iyileştirilmesi ya da güçlü yanların

yorumlanmasını sağlamaktadır. Ayrıca K çalışmasında yüzeylerin koşul sayısının değişiminde güvenirliğin ne olacağı sorusunun cevabı verilebilmekte olup en uygun tasarıyı sağlamaktadır (Cardinet vd., 2010). Bu nedenle K-çalışmaları daha yüksek güvenirlik elde etmek için ölçme işlem veya koşullarını iyileştirme çalışmaları olarak da adlandırılır.

Çaprazlanmış (Crossed) ve Yuvalanmış (Nested) Desen

Genellenebilirlik kuramında, verilerin düzenlenme biçimine göre iki farklı desenden söz edilebilmektedir: çaprazlanmış desen, yuvalanmış desen. Çaprazlanmış desen bir değişkenlik kaynağının tüm koşullarının diğer bir değişkenlik kaynağının tüm koşullarında gözlenmesi durumudur ve bu durumda iki yüzey arasına "X" işareti konulur. Örneğin belli bir performans ölçme durumunda her öğrenciyi her puanlayıcı puanlıyorsa veriler çaprazlanmıştır ve S X I X R (öğrenci: S, madde: I, puanlayıcı: R) şeklinde gösterilir.

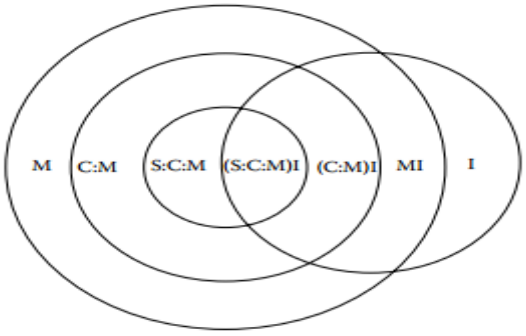
S öğrencileri R puanlayıcıları göstermek üzere S X R çapraz desenine ait varyans bileşenleri şekildeki gibidir:



Şekil 1. S X R çapraz desenine ait varyans bileşenleri

Yuvalanmış desen ise bir değişkenlik kaynağının tüm koşullarının başka bir değişkenlik kaynağının sadece bazı koşullarında gözlenmesi durumudur ve bu durumda yüzeyler arasına ":" işareti konulur. Örnek olarak aynı ölçme durumunda her öğrencinin performansı farklı bir puanlayıcı tarafından puanlanırsa desen yuvalanmış olur ve I X (S: R) şeklinde gösterilir (Yelboğa, 2007).

S öğrenciyi, M yöntemi, I maddeyi göstermek üzere (S:C:M) X I yuvalanmış desenine ait varyans bileşenleri şekildeki gibidir:



Şekil 2. (S:C:M) X I yuvalanmış desenine ait varyans bileşenleri

Araştırmalarda genellikle bütün varyans bileşenlerini hesaplamaya olanak tanıyan çaprazlanmış desen kullanımı tercih edilmektedir. Ancak yuvalanmış desen de bazı varyans bileşenlerinin hesaplanmasında serbestlik derecesini arttırdığı için faydalı olabilmektedir.

Öğrencilerin belli bir performans için her bir puanlayıcı tarafından değerlendirildiği çaprazlanmış S X I X R (öğrenci: S, madde: I, puanlayıcı: R) deseninde gözlenen puan şu şekilde gösterilir:

X_{sir} =

μ	genel ortalama
$+ \mu s - \mu$	öğrenci etkisi
$+ \mu i - \mu$	madde etkisi
$+ \mu r - \mu$	puanlayıcı etkisi
$+ \mu si - \mu s - \mu i + \mu$	öğrenci x madde ortak etkisi
$+ \mu sr - \mu s - \mu r + \mu$	öğrenci x puanlayıcı ortak etkisi
$+ \mu ir - \mu i - \mu r + \mu$	madde x puanlayıcı ortak etkisi
$+ Xsir - \mu s - \mu i - \mu r + \mu si + \mu sr + \mu ir - \mu$	artık etkisi

(Eşitlik 1)

G kuramında araştırmanın amacına bağlı olarak iki farklı ölçme katsayısı hesaplanabilmektedir: bağıl (relative) ölçme katsayısı, mutlak (absolute) ölçme katsayısı. Bunlardan birincisi bireylerin ya da ölçme objelerinin diğerlerine göre dağılımının ya da sıralamanın önemli olduğu çalışmalarda kullanılırken, mutlak ölçme katsayısı ise her bireyin diğer bireylerden bağımsız olarak ölçme aracındaki yerini kesin olarak belirlenmesinde hesaplanmaktadır. Dolayısı ile her iki katsayının hesaplanmasında kullanılan hata terimleri birbirinden farklıdır (Cardinet vd., 2010).

Genellenebilirlik (G) ve Phi(Φ) Katsayıları

Genellenebilirlik kuramı ile hesaplanan güvenirlik katsayıları bağıl ve mutlak ölçmelere göre ayrı ayrı hesaplanmaktadır.

Bağıl genellenebilirlik (G) katsayısı bağıl hata varyansı ile hesaplanan ve bağıl ölçmeler için uygun olan güvenirlik katsayısıdır. Bağıl hata varyansı, araştırmadaki ölçme objesini içeren ortak etkili varyans bileşenlerinin toplamıdır. B X M X P deseni için bağıl hata varyansı "δ" (Yunan alfabesindeki küçük delta harfi) ile gösterilmek üzere (Eşitlik 2 de madde ve puanlayıcı yüzeyinin büyük harfler ile gösterilmesi bu değerlerin ortalamalar üzerinden alındığını belirtmek içindir);

$$\sigma^2(\delta) = \sigma^2(\text{ bM }) + \sigma^2(\text{ bP }) + \sigma^2(\text{ bMP }) \text{ şeklindedir.} \qquad \text{(Eşitlik 2)}$$

"σ^2b" birey puanlarının evren değerinin varyansı Klasik Test kuramında gerçek puan varyansına karşılık geldiği için genellenebilirlik katsayısı şu şekilde hesaplanmaktadır:

$$G = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\delta^2} \qquad \text{(Eşitlik 3)}$$

Phi katsayısı, mutlak hata varyansı ile hesaplanan ve mutlak ölçmeler için kullanılan güvenirlik katsayısıdır. Mutlak hata varyansı bireylerin ya da ölçme objesinin gözlenen ve evren puanları arasındaki farkın varyansıdır. Mutlak hata "Δ" (Yunan alfabesindeki büyük delta harfi) ile gösterilmek üzere aşağıdaki eşitlikle kestirilir:

$$\sigma^2(\Delta) = \sigma^2(M) + \sigma^2(P) + \sigma^2(bM) + \sigma^2(bP) + \sigma^2(mP) + \sigma^2(bMP) \qquad \text{(Eşitlik 4)}$$

Phi(Φ) katsayısı ise aşağıdaki formülle elde edilir (Güler vd., 2012):

$$\varphi = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\Delta^2} \qquad \text{(Eşitlik 5)}$$

Mutlak ölçmelerde, madde yüzeyinin olduğu ölçme durumlarında, test formunun güçlüğü ölçme objesinin puanını etkilemekte ve bundan evren puanı da etkilenmektedir. Bağıl ölçmelerde ise test formunun etkisi herkes için sabit olup, sıralamada bir fark yaratmamaktadır. Bu sebeple madde ana etkisi bağıl ölçmelerde yer almazken mutlak ölçmelerde kararlarda etkili olduğu için yer almaktadır. Bu durum mutlak ölçmeler için hesaplanan Phi (Φ) katsayısının, G katsayısından daha küçük değerler almasına sebep olmaktadır (Alkan, 2013; Güler, Uyanık ve Teker, 2012).

Bağıl ölçmelerdeki G katsayısı ölçme objelerinin ne derecede iyi farklılaştığını gösterirken mutlak ölçmelerdeki Phi katsayısı ölçme işleminin; ölçme nesnelerini ölçeğe ne derece iyi

yerleştiğinin göstergesidir. Genel olarak mutlak ölçmelerdeki katsayı daha düşüktür. Çünkü mutlak ölçmede hata varyansının potansiyel kaynağı daha çoktur.

Alanyazında yer alan bu iki güvenirlik katsayılarından hariç kriter referanslı ölçmeler (criterion-referenced measurement) için hesaplanan Φ(λ) katsayısı bulunmaktadır. Bu katsayı bireysel puanların, λ kesme puanına ya da kriterine uygun olarak, hatasız bir şekilde ölçekte yerini belirlemek amacıyla kullanılır (Cardinet vd., 2010). Örneğin 0-100 arası puanlanan bir test için kesme puanı 60 olarak ele alınırsa, Φ(60); 60 puanın altında başarılı olanlarla 60 puan ve üzerinde başarılı olanların ne derece güvenilir olarak (mutlak anlamda) belirlendiğini ifade etmektedir.

Klasik Test Kuramı

Ölçme alanındaki ilk kuram olma özelliğini taşıyan klasik test kuramı (KTK), gerçek puanın gözlenen puanlar yardımıyla kestirilebileceğini ileri sürmektedir. Bu varsayım, gözlenen puan ile gerçek puan arasındaki doğrusal bir ilişki ile açıklanmaktadır. Bu nedenle KTK, gerçek puan modeli (true score model) olarak da karşımıza çıkar (Baykul, 2000, s. 97). Bir ölçme durumundaki gözlenen puan (X), gerçek puan (T) ve hata puanı (E) olmak üzere, bazı sayıltılar altında Klasik Test Kuramının modeli X = T + E olarak ifade edilir.

Klasik test kuramının Algina (1986) tarafından sayıltı da denilen, temel prensipleri şunlardır:

1. Hata puanlarının evrendeki dağılımının ortalaması sıfırdır (μ_E= 0).
2. Gerçek puanlar ile hata puanları arasındaki korelâsyon sıfırdır (ρ_{TE} = 0).
3. Ayrık hata puanları arasındaki korelâsyon sıfırdır (ρ_{E1E2} = 0).

Bu üç sayıltı, gerçek puanlar ile hata puanlarının temel prensiplerini tanımlayarak test puanlarının güvenirliğinin KTK ile incelenmesine rehber olmaktadır. Klasik Test Kuramında bu sayıtlılardan yola çıkarak güvenirlik katsayısı (ρ), gerçek puan varyansının gözlenen puan varyansına oranı olarak açıklanır (Crocker ve Algina, 1986).

$$\sigma^2_{\text{gözlenen}} = \sigma^2_{\text{gerçek}} + \sigma^2_{\text{hata}} \qquad \text{(Eşitlik 6)}$$

$$\rho = \sigma^2_{\text{gerçek}} / \sigma^2_{\text{gözlenen}} \qquad \text{(Eşitlik 7)}$$

Buradaki gözlenen puan varyansı Eşitlik 6 da görüldüğü gibi gerçek puan varyansı ile hata puan varyansından oluşur. Gerçek puan varyansı dışındaki varyansların farklı hata kaynaklarından gelebileceği düşünülür ve bu hata kaynaklarına bağlı olarak da güvenirlik farklı isimlerle ifade edilir. Ayrıca uygulamalarda gerçek değerin bilinememesi sebebiyle

katsayının bu şekilde hesabı teoride kalmaktadır. Bu nedenle güvenirlik katsayısını hata kaynaklarını dikkate alarak dolaylı yoldan hesaplayacak yöntemler geliştirilmiştir (Ercan ve Kan, 2004). Bu araştırmada Pilten (2008) tarafından geliştirilen ölçeğin bir kez uygulanması ve üç bağımsız puanlayıcı tarafından puanlanması sonucu elde edilen veriler ile güvenirlik kestirimi yapılmıştır. Dolayısı ile tek uygulamaya yönelik iç tutarlılık anlamındaki güvenirlik için Cronbach Alfa katsayısı ve birden fazla puanlayıcının bulunduğu ölçme durumlarında, hata kaynağı olan puanlayıcıların ölçümleri arasındaki uyum için sınıf içi ilişki katsayısı sırası ile incelenmiştir.

Cronbach Alfa Katsayısı

Cronbach tarafından 1951 yılında geliştirilen alfa katsayısı, Kuder-Richardson 20 formülünün genel bir hali olarak şu özelliklerle tanımlanmaktadır:

a. Tüm olası iki yarı güvenirlik (split-half) katsayılarının ortalamasıdır.
b. Verilen ilişkili testlerdeki madde havuzundan alınan iki rastgele örneklemin beklenen değeridir (Cronbach, 1951).
c. Güvenirlik katsayıları içinde en alt sınır olarak kabul edilebilir (Tekindal, 2014).
d. Madde kovaryanslarının bir fonksiyonudur ve maddeler arasındaki bu kovaryans genel bir faktör değil, bir sonuç olduğu için, bu katsayı tek boyutluluğun ölçüsü olarak düşünülmemelidir. Crocker ve Algina (1986), Alfa katsayısını genel bir faktörle açıklanamayan test puanlarındaki varyansın bir bölümü, alt sınır olarak yorumlanabileceğini ifade etmiştir.

Cronbach Alfa yönteminin KR 20 yönteminden farkı, çoklu puanlanabilen maddelerden oluşan testlere uygulanabilmesidir. Cronbach Alfa eşitliği aşağıdaki gibidir:

$$\alpha = \frac{K}{K-1}\left[1 - \frac{\sum S_j^2}{S_x^2}\right] \qquad \text{(Eşitlik 8)}$$

K: Testte bulunan madde sayısı

S_x^2 : Test puanları dağılımı varyansı

$\sum S_j^2$: Madde varyanslarının toplamı

Alfa katsayısı, diğer güvenirlik katsayıları gibi 0-1 aralığında değer almaktadır. 1'e yaklaştıkça güvenirlik düzeyi artmakta iken Kaplan ve Saccuzzo, (1982); Murphy ve

Davidshoper (1988); Nunnally, (1978) uygulama araştırmalarında yüksek düzeyde güvenirlik elde etmek için 0,90 ve üzeri Alfa katsayısını öngörmüşlerdir (Aktaran Yurdugül, 2010).

Sınıf İçi İlişki Katsayısı (Intraclass Correlation Coefficient – ICC)

Performans değerlendirmede, puanlayıcılara bağlı hata kaynağının hesaplanmasında KTK'ya dayalı pek çok yöntem bulunmaktadır. Uyum yüzdesi, sınıf içi ilişki katsayısı, Cohen'in kappası, Kendall'ın uyuşum katsayısı, Krippendorff alfa katsayısı bunlardan bazılarıdır. Bu çalışmadaki en temel amaç, aynı bireyler üzerinden üç puanlayıcının ölçümleri arasındaki uyumu belirlemektir. Bu bağlamda, çoklu puanlayıcılı ve ölçümlerin sürekli olduğu ölçme durumlarında kullanılabilen, varyans analizini temel alan sınıf içi ilişki katsayısı kullanılmıştır (Ateş, Öztuna ve Genç, 2009).

Shrout ve Fleiss (1979), KTK'ya dayalı pek çok güvenirlik indeksinin sınıf içi ilişki katsayısının versiyonu olarak gösterilebileceğini ifade etmişlerdir. Çünkü KTK'nın temel güvenirlik tanımında olduğu gibi sınıf içi ilişki katsayısı; ilgilenilen varyansın; ilgilenilen varyans ve hata varyansının toplamına oranı olarak ifade edilir:

$$\text{ICC}=\frac{\sigma_{PA}^2}{\sigma_{PA}^2+\sigma_{P\dot{I}}^2} \qquad \text{(Eşitlik 9)}$$

Eşitlik 9'daki σ_{PA}^2 puanlayıcılar arası varyansı; $\sigma_{P\dot{I}}^2$ puanlayıcılar içi varyansı ifade etmektedir. Bu değerler varyans hesaplamasındaki kareler ortalaması ile elde edilmektedir. Bireylerden elde edilen çoklu ölçüm, aynı puanlayıcının tekrarlı ölçümleri olabileceği gibi, iki ya da daha fazla puanlayıcının ölçümleri de olabilir. Bu durumda iki farklı sınıf içi ilişki katsayısından söz edilir: ilk durum için puanlayıcılar içi (intra-rater), ikinci durumda için ise puanlayıcılar arası (inter-rater) uyum ilişki katsayıları (Ateş, vd., 2009).

Farklı kriterlere göre pek çok sınıf içi ilişki katsayısı bulunmaktadır. Çalışma durumuna uygun olan sınıf içi ilişki katsayısını belirlemede üç önemli husus bulunmaktadır (Shrout ve Fleiss, 1979):

i. Güvenirlik analizi için tek yönlü rastgele etki modeli mi, çift yönlü rastgele etki modeli mi uygundur?
ii. Çalışmanın amacına bağlı olarak mutlak uyum mu, tutarlılık mı ön plandadır?
iii. Ölçümlerin elde edilme biçiminde tek ölçüm, ortalama puan ya da toplam puan mı alınmıştır?

Yukarıdaki ölçütler eşliğinde sınıf içi ilişki katsayı çeşitleri için 3 durumdan bahsedilebilir (Kılıç, 2009: 36):

Durum1: Değerlendiriciler her değerlendirilen birim için rastgele seçilmektedir.

Durum2: Aynı değerlendiriciler her birimi değerlendirir. Bunlar rastgele örneklemdir.

Durum3: Aynı değerlendiriciler her birimi değerlendirir. Bunlar özel seçimli değerlendiricilerdir.

Bu araştırmadaki ölçme durumu için, sadece birimler değil, puanlayıcılar da rastgele etki kaynağı olarak alındığından, ikinci durumdaki iki yönlü rastgele etki modeli uygun bulunmuştur. Bu durumda sınıf içi ilişki katsayısı aşağıdaki gibi hesaplanır (Shrout ve Fleiss, 1979):

$$\text{ICC}(2,1) = \frac{BMS - EMS}{BMS + (k-1)EMS + k(JMS - EMS)/n} \qquad \text{Eşitlik 10}$$

BMS: Gruplar arası kareler ortalaması

EMS: Hata kareleri ortalaması

JMS: Puanlayıcılar arası kareler ortalaması

k: Puanlayıcı sayısı

n: Birey sayısı

Eşitlik 10'da verilen ICC(2,1) katsayısı; puanlayıcıların rastgele çekildiği ve her bir bireyden ölçüm aldığı ölçme durumu için hesaplanan sınıf içi ilişki katsayısına işaret etmektedir. Bu araştırmada toplam puana göre puanlayıcılar arası uyuma bakılmıştır.

Problem Cümlesi

Öğrencilerin matematiksel muhakeme becerisine yönelik performanslarının, üç farklı puanlayıcı tarafından puanlanması sonucunda elde edilen ölçümlerin, Genellenebilirlik Kuramının farklı desenlerinden ve Klasik Test Kuramından elde edilen güvenirlik katsayıları nelerdir?

Araştırmanın Amacı

Bu araştırmanın amacı ilköğretim 7. sınıf öğrencilerine yönelik matematiksel muhakeme performansının belirlenmesinde kullanılan ölçeğin birden fazla puanlayıcı tarafından, çaprazlanmış ve yuvalanmış desene göre puanlanmasıyla elde edilen ölçümlerin güvenirliğini Genellenebilirlik (G) kuramı ve Klasik Test Kuramına dayalı olarak karşılaştırmaktır. Bu genel amaç doğrultusunda şu sorulara yanıt aranacaktır:

1. Öğrenci (ö), matematiksel muhakemeyi belirleme ölçeğindeki sorular (s) ve puanlayıcı (p) yüzeylerinin çaprazlandığı Ö X S X P deseninin Genellenebilirlik (G) çalışması sonuçlarının;

   1.a. Kestirilen varyansları ve toplam varyansları açıklama yüzdeleri nelerdir?

   1.b. Puanlayıcı ve soru sayısının arttırılması ve azaltılması sonucunda K çalışmasında kestirilen G ve Phi katsayıları nasıl değişmektedir?

2. Soru (s) ve puanlayıcı (p) yüzeylerinin yuvalanmış, öğrenci (ö) yüzeyinin ise çaprazlanmış olduğu Ö X (S:P) deseninin Genellenebilirlik (G) çalışması sonuçlarının;

   2.a. Kestirilen varyansları ve toplam varyansları açıklama yüzdeleri nelerdir?

   2.b. Puanlayıcı ve soru sayısının arttırılıp azaltılması sonucunda K çalışmasında kestirilen G ve Phi katsayıları nasıl değişmektedir?

3. Ölçeğin Ö X S X P ve Ö X (S:P) desenlerinden elde edilen G çalışması parametrelerinin değişimi nasıldır?

4. Ölçeğin Ö X S X P ve Ö X (S:P) desenlerinde puanlayıcı ve soru sayılarının arttırılıp azaltılmasıyla yapılan Karar çalışmaları parametrelerinin değişimi nasıldır?

   4.a. Ö X S X P ve Ö X (S:P) desenlerinde puanlayıcı ve soru sayılarının arttırılıp azaltılmasıyla yapılan Karar çalışmalarında mutlak ve bağıl hata varyanslarının değişimi nasıldır?

   4.b. Ö X S X P ve Ö X (S:P) desenlerinde puanlayıcı ve soru sayısının arttırılıp azaltılmasıyla yapılan Karar çalışmalarında elde edilen G ve Phi katsayılarının değişimi nasıldır?

5. Matematiksel muhakemenin belirlenmesinde kullanılan ölçek için her iki desende de (Ö X S X P ve Ö X (S:P)) kabul edilebilir bir düzeyde genellenebilirlik katsayısı elde etmek için gerekli minimum soru ve puanlayıcı sayısı nedir?

   5.a. Ö X S X P deseninde kabul edilebilir bir düzeyde genellenebilirlik katsayısı elde etmek için gerekli minimum soru ve puanlayıcı sayısı nedir?

   5.b. Ö X (S:P) deseninde kabul edilebilir bir düzeyde genellenebilirlik katsayısı elde etmek için gerekli minimum soru ve puanlayıcı sayısı nedir?

6. Klasik test kuramına göre; aşamalı puanlama anahtarı ile puanlanan matematiksel muhakemeyi belirleme ölçeğinden elde edilen puanların Cronbach Alfa ve Tabakalanmış Alfa güvenirlik katsayıları kaçtır?
7. Matematiksel muhakemeyi belirleme ölçeğinden elde edilen puanların Genellenebilirlik ve Klasik Test Kuramına dayalı güvenirlik katsayıları arasında manidar farklılık var mıdır?

   7.a. Ö X S X P deseninden elde edilen puanların güvenirlik katsayıları ile manidar farklılık var mıdır?

   7.b. Ö X (S:P) deseninden elde edilen puanların güvenirlik katsayıları ile manidar farklılık var mıdır?

Araştırmanın Önemi

Matematiksel muhakeme, öğrencilerin bir problem üzerinde düşünerek içeriği ile ilgili karar vermeleri, çözüm için gerekçeler sunmaları, buna uygun bir plan seçmeleri ve çözümü yorumlamaları gibi pek çok üst düzey becerileri içermektedir. Dolayısıyla bu beceriyi ölçmeye en çok imkân tanıyan madde türünün açık uçlu maddeler olduğu söylenebilir. Bu durum beraberinde öğrencilerin açık uçlu maddelere verdiği yanıtların güvenilir bir şekilde değerlendirilmesini gerektirmektedir. Bunun için beklenilen davranışlar önceden belirlenmeli ve bu davranışlara göre dereceli puanlama anahtarı hazırlanmalıdır.

Birden fazla puanlayıcının yer aldığı ölçme durumlarında, her ne kadar puanlama anahtarları kullanılsa da, puanlayıcılar arasındaki görüş ayrılıkları, objektif değerlendirmeme ya da çevresel değişiklikler gibi hatalar ölçmeyi olumsuz yönde etkilemektedir. Puanlayıcının puanlama deneyiminin eksikliği, yaşı, cinsiyeti, kişisel özellikleri gibi pek çok sebep bu

duruma örnek gösterilebilir. Bu sebeplerden dolayı öğrenci yanıtlarını değerlendirmeden önce yapılan ölçmenin güvenirliğinin incelenmesi gerekmektedir.

Bu araştırmayla, açık uçlu soruların puanlanmasındaki güvenirlik kestirimi, Klasik Test Kuramı'na ve Genellenebilirlik Kuramı'nın iki farklı desenine dayalı olarak, kuramların her iki desende de birbiriyle ve kendi içlerinde tutarlılıkları ele alınmıştır. Aynı zamanda iki kuramdan elde edilen güvenirlik kestirimleri karşılaştırılmıştır. Böylece aynı ölçme durumu için oluşturulmuş farklı desenlerden hangisinin, benzer ölçme durumlarında hangi kuramın daha uygun olacağının belirlenmesinin alana katkı sağlayacağı düşünülmektedir. Ayrıca matematik dersinde özellikle bilgi düzeyi üstü becerilerin ölçülmesinde, açık uçlu sorular gibi öznel değerlendirme araçlarının ne kadar güvenilir olduğunun belirlenmesi ve bu becerilerin ölçülmesinde etkili olan değişkenlik kaynaklarının ortaya çıkarılmasının özellikle matematik eğitimcileri için aydınlatıcı olacağı düşünülmektedir.

Sayıltılar

Puanlayıcılar öğrenci cevaplarını ciddiyetle puanlamıştır.

Uygulamaya katılan öğrenciler ölçekte yer alan soruları ciddiyetle cevaplamıştır.

Sınırlılıklar

Araştırma Konya ilinde yedinci sınıfta öğrenim gören 187 öğrenci ile sınırlıdır.

Araştırma gönüllü puanlayıcı olan 3 matematik eğitimcisi ile sınırlıdır.

İlgili Araştırmalar

Alanyazında yapılan çalışmalar yurt içinde yapılan araştırmalar ve yurt dışında yapılan araştırmalar olmak üzere iki başlık altında ve yayınlanma yılına göre sırası ile sunulmuştur.

Yurt İçinde Yapılan Araştırmalar

Atılgan (2004) araştırmasında, 2003 ve 2004 yıllarında yapılan Müzik öğretmenliği özel yetenek seçme sınavları verilerine Genellenebilirlik Kuramı ve Çok Değişkenlik Kaynaklı Rasch Modelini uygulamıştır. Her bir birey (b), her bir görev (g) için, her puanlayıcı (p) tarafından bağımsız olarak puanlanmıştır. Analiz sonucu elde edilen verilerle genellenebilirlik kuramının tek değişkenli ve çok değişkenli modellerinin G analizi sonuçlarında kestirilen

varyans bileşenleri farklı çıkmıştır. İki sınav için yapılan K çalışmalarında, alt testler ve birleşik testler için yapılan çok değişkenli modelin kestirilen G ve Phi katsayıları, tek değişkenli modelle kestirilenlerden büyük bulunmuştur. Ayrıca Genellenebilirlik Kuramı ile Çok Değişkenlik Kaynaklı Rasch Modeli (ÇDKRM) istatistikleri karşılaştırılmıştır. Araştırma sonunda iki kuramın değişkenlik kaynakları için kestirilen varyans bileşenlerinden birey, görev ve puanlayıcıya ait olanları örtüşürken; puanlayıcı x birey değişkenlik kaynağının G kuramı ile kestirilen varyans bileşeninin ÇDKRM ile elde edilen yanlılık yüzdelerinin tutarlı olmadığı görülmüştür. Görev-birey ile görev-birey-puanlayıcı değişkenlik kaynaklarının G kuramında kestirilen varyans bileşenlerinin ÇDKRM ile elde edilen yanlılık yüzdelerinden büyük olduğu ve puanlayıcı-görev değişkenlik kaynağına ait varyans bileşeninin ise daha küçük olduğu görülmüştür.

Yelboğa (2007), 2005 ve 2006 yıllarında uygulanan iş performansı ölçeğinin güvenirliğini Genellenebilirlik kuramı ve Klasik test kuramına göre karşılaştırmıştır. Araştırmada çalışma grubunu hizmet sektöründeki bir iş yerinin 11 farklı biriminden 176 personel oluşturmaktadır. Üç farklı yönetici (değerlendirici), 176 personeli (birey) birbirinden bağımsız olarak iş performansı ölçeği ile puanlamış, elde edilen veriler ile Klasik test kuramına göre test tekrar test ve Cronbach alfa güvenirlik katsayıları; Genellenebilirlik kuramına göre ise çok değişkenli model ile G ve Phi katsayıları hesaplanmıştır. Araştırma sonucunda, her 3 değerlendiricinin 2005-2006 yıllarına ilişkin test tekrar test uygulaması sonucunda elde edilen Pearson momentler çarpım korelâsyon katsayısının 0,85'ten yüksek olduğu görülmüştür. İş performansı ölçeğinin her iki yıl içinde kestirilen Cronbach Alfa değerlerinin ise 0,90'ın üzerinde olduğu görülmüştür. Çalışmada KTK çerçevesinde puanlayıcılar arasındaki tutarlılığın göstergesi olarak kabul edilen Kendall'ın uyuşum katsayısı ise her iki yıl için de 0,95'in üzerinde hesaplanmıştır. Genellenebilirlik kuramına göre G çalışması ile analiz edilen verilerde ise her iki yıla ait G ve Phi katsayılarının 0,90'ın üzerinde olduğu görülmüştür. Dolayısı ile her iki kurama göre elde edilen güvenirlik katsayılarının birbirleriyle uyumlu olduğu sonucuna ulaşılmıştır.

Güler (2008) araştırmasında, TIMMS-1999 da yer alan açık uçlu matematik sorularının ölçülmesinde klasik test kuramı, genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch ölçme modeli uygulayarak güvenirlik indekslerini karşılaştırmıştır. Öğrencilerin verdikleri cevaplar 4 puanlayıcı tarafından holistik rubrik kullanılarak puanlanmıştır. Verilerin güvenirlik analizinde, klasik test kuramında Cronbach Alfa katsayısı, puanlayıcılar arası uyumun belirlenmesinde Kendall'ın konkordans katsayısı ile puanlayıcılar arası korelâsyon

katsayısı hesaplanmıştır. Genellenebilirlik kuramında B X G X P tümüyle çaprazlanmış desen kullanılarak güvenirlik ve genellenebilirlik katsayıları hesaplanmıştır. Çok değişkenlik kaynaklı Rasch modeli ile de birey, madde ve puanlayıcı boyutlarına göre ayrı ayrı hesaplamalar yapılmıştır. Araştırma sonunda, 4 puanlayıcının da birbirleriyle uyumlu puanlamalar yapıldığı görülmüş ve matematik başarısının ölçülmesindeki güvenirliği test ederken en az iki kuramdan yararlanılmasının daha faydalı olacağı sonucuna ulaşılmıştır.

Nalbantoğlu (2009), Tıp Fakültesi üçüncü sınıf öğrencilerinin iletişim becerilerini, birden fazla puanlayıcı tarafından birlikte ve dönüşümlü olarak puanlanmasıyla oluşturulan desenlerden elde edilen G ve K çalışmaları sonuçlarını karşılaştırmıştır. Araştırmada 3 puanlayıcı öğrencileri, iletişim becerileri değerlendirme formuyla 15 görev doğrultusunda birlikte ve dönüşümlü olarak puanlamış olup, Ö X G X P ve (Ö:P) X G desenleri (ö: öğrenci, g: görev, p: puanlayıcı) için ayrı ayrı G ve K çalışması yapılmıştır. Araştırma sonucunda, puanlayıcıların puanlamaları arasında bir fark olmadığı görülmüştür. Yapılan G çalışmaları sonucunda kestirilen varyans bileşenlerinin birbirleriyle paralellik gösterdiği sonucuna ulaşılmıştır. Yalnızca görev değişkenlik kaynağına ait kestirilen varyans bileşeni (Ö:P) X G deseninde daha yüksek bulunmuştur. Bununla birlikte her iki desende de artık varyans bileşeninin yüksek olmasına rağmen yuvalanmış desende biraz daha fazla bulunmuştur. Bu durumun (Ö:P) X G deseninde, öğrenci-görev etkisine ait varyansın, artık varyansa dâhil olmasından kaynaklandığı düşünülmüştür. Karar çalışmaları sonucunda ise her iki desendeki G ve Phi katsayıları arasında çok fark olmamakla birlikte (Ö:P) X G deseninin katsayılarının daha büyük çıkma eğiliminde olduğu görülmüştür.

Deliceoğlu (2009), futbolcuların teknik yetilerinin ölçülmesi amacıyla 56 maddeden oluşan Futbol Yetilerine İlişkin Dereceleme Ölçeği'nden elde edilen ölçümlerin güvenirliklerini Klasik Test Kuramı ve Genellenebilirlik Kuramına dayalı olarak incelemiştir. Ölçekten elde edilen veriler ile, Klasik Test Kuramına dayalı olarak, ilk puanlama ve ikinci puanlama arasındaki tutarlılığı ölçmeye yönelik "Pearson momentler çarpımı" korelâsyon katsayısı, maddelerin iç tutarlılık güvenirliği için Cronbach α (alfa) katsayısı, puanlayıcılar arasındaki tutarlılık için Kendall uyuşum katsayıları hesaplanmıştır. Ölçeğin alt boyutlarından ve toplamından elde edilen puanların güvenirliği için ise Genellenebilirlik Kuramı'na dayalı olarak, puanlayıcı sayılarının 3, 4 ve 5 olduğu ve madde sayılarının bir arttırılıp, bir azaltıldığı koşullarda ana ve ortak etkilerin varyans bileşenlerin kestirilmesi için MGENOVA paket programında G ve Phi katsayıları bulunmuştur. Araştırma sonunda, ölçeğin puanlanmasından elde edilen güvenirlik parametreleri iç ölçütlere göre incelendiğinde G katsayısı ve Cronbach

Alfa katsayılarının beklenen değerlerinden yüksek; Phi katsayısı ile Kendall W Güvenirlik katsayılarının ise beklenen değerlerinden düşük olduğu ortaya çıkmıştır. Ayrıca Futbol Yetilerine İlişkin Dereceleme Ölçeği'nin alt boyutlarının güvenirliklerinin yüksek olması ölçeğin güvenilir bir ölçek olduğu sonucunu ortaya çıkarmıştır.

Çakıcı (2011), 9. sınıf düzeyindeki öğrencilere yönelik, PISA sorularından yararlanarak oluşturulan 15 maddelik ölçme aracının puanlanmasında ortaya çıkan puanlayıcı tutarlığını genellenebilirlik kuramı ve lojistik regresyon analizinden yararlanarak belirlemiş ve karşılaştırmıştır. Öğrenci cevapları üç puanlayıcı tarafından dereceli puanlama anahtarı kullanılarak puanlanmış ve veri seti G kuramına göre Ö X P (ö: öğrenci, p: puanlayıcı) deseninde analiz edilmiştir. Lojistik regresyon analizinde ise öğrenciler; her bir puanlayıcıdan elde ettikleri toplam puanların ortalamasına dayalı olarak başarılı-başarısız şeklinde sınıflandırılmıştır. Testin tamamına yönelik yapılan analizlerde G çalışması Ö X M X P (ö: öğrenci, m: madde, p:puanlayıcı) desenine göre gerçekleştirilirken, lojistik regresyon analizi başarılı-başarısız sınıflamasına dayalı olarak öğrencilerin her bir puanlayıcıdan elde ettikleri toplam puana göre yapılmıştır. Araştırma sonuçlarına göre puanlayıcılar arası tutarlılığı belirlemede genellenebilirlik kuramı ve regresyon analizinin benzer sonuçlar vermesi ile birlikte G kuramının daha hassas çıktılar verdiği sonucuna ulaşılmıştır.

Kaya (2011), üniversite 3. sınıf düzeyindeki öğrencilere uygulanan "madde" konulu iki farklı doldurma kavram haritasının değerlendirilmesinde genellenebilirlik kuramını kullanmıştır. Farklı teknik ve uygulama sırasıyla uygulanmış doldurma kavram haritasının değerlendirilmesinde, 10 maddeye tümüyle çaprazlanmış desende yapılan G çalışması sonucunda; birey (b) ana etkisi için kestirilen varyans bileşeninin toplam varyansın büyük bir oranını açıkladığı görülmüştür. G çalışmasının ardından yapılan karar çalışması sonucunda doldurma kavram haritası uygulamalarının değerlendirmesinde daha yüksek genellenebilirlik ve güvenirliğe ulaşmak için, harita tekniği sayısını arttırmanın daha ekonomik ve kullanışlı olacağı sonucuna ulaşılmıştır.

Büyükkıdık (2012) araştırmasında, matematik problemi çözme becerisinin dört puanlayıcı tarafından analitik ve bütünsel dereceli puanlama anahtarıyla puanlanmasından elde edilen veriler ile klasik test kuramı ve genellenebilirlik kuramı kullanarak puanlayıcılar arası güvenirlik sınaması yapmıştır. Klasik test kuramına göre puanlayıcılar arası güvenirlik analizi için sınıf içi ilişki katsayısı ve puanlayıcılar arası ilişki katsayısı ile hesaplanmıştır. Genellenebilirlik kuramına göre güvenirlik analizinde ise her iki puanlama anahtarı için birey görev ve puanlayıcı değişkenlerinin çaprazlandığı B X G X P ve anahtar çeşidinin değişkenlik

kaynağı olarak alındığı B X G X P X A tümüyle çaprazlanmış desenleri kullanılmıştır. Araştırma sonucunda, her iki dereceli puanlama anahtarı için genellenebilirlik kuramından elde edilen G ve Phi katsayıları 0,90'ın üzerinde bulunmuşken; KTK'daki Cronbach Alfa katsayısı 0,84 bulunmuştur. B X G X P deseninde puanlayıcı değişkenlik kaynağı, B X G X P X A deseninde olduğundan kısmen daha büyük oranda toplam varyansı açıklamaktadır. Bu sonucun B X G X P X A deseninde anahtar değişkenlik kaynağının etkisinin toplam varyansın paylaşılmasında rol oynayabileceğinden kaynaklı olduğu düşünülmüştür.

Özberk (2012) araştırmasında, iki farklı varyans bileşeni belirleme yöntemine göre kestirilen standart hata değerleri ile güvenirlik ve geçerlik üzerinde karşılaştırmalar yapmıştır. Veriler simülâsyon ile tek değişkenlik kaynaklı desene göre üretilmiştir. İlk aşamada B X M (b:birey, m:madde) desenine uygun birey-madde madde matrisi, 60 x 5 şeklinde veri seti simülâsyon ile R yazılımı kullanılarak üretilmiştir. İkinci aşamada B X M desenine uygun olarak değişkenlik kaynakları S PLUS yazılımında 1000 kere yeniden örneklenmiştir. Standart hatalar, varyans bileşenleri, mutlak ve bağıl hatalar ANOVA ve bootstrap yöntemleri kullanılarak kestirilmiştir. Araştırma nihayetinde, *boot−m* prosedürünün geçerlik hakkında daha fazla bilgi verdiği, *boot–b* prosedürünün de G Kuramı çalışmalarında evren puanlarını belirlemede daha kesin kestirimler yaptığı sonucuna varılmıştır.

Anıl ve Büyükkıdık (2012), çalışmalarında 6, 7 ve 8. sınıfta öğrenim görmekte olan 132 öğrencinin problem çözme becerisine yönelik hazırlanan 2 performans görevinde sergiledikleri performansların 4 puanlayıcı tarafından puanlanması ile elde edilen veriler üzerinde Genellenebilirlik kuramı uygulamıştır. Ölçmenin nesnesi olan farklı sınıflardaki bireyler (B:S); görev, ölçüt ve puanlayıcı yüzeyleri ile çaprazlanarak (B:S) X G X Ö X P deseni oluşturulmuştur. Araştırma sonunda G çalışması ile kestirilen en büyük varyans bileşeninin farklı sınıflardaki bireyler olduğu görülmüştür. Bu bulgu problem çözme becerisi açısından grubun heterojen olduğu ile açıklanmıştır. Ayrıca çalışmada sınıf düzeyine ait varyans bileşeninin önemli olmadığı görülmüştür. Bu durumun, üst sınıfların güdülenme düzeyinin alt sınıflardan daha düşük olmasından ya da uygulanan sınıflarda üst düzey düşünme becerilerini sergileyen benzer oranda öğrencinin varlığından kaynaklanabileceği ifade edilmiştir. Yapılan karar çalışmaları ile ölçümlerin güvenirliğini arttırmada, puanlayıcı sayısının, görev sayısına göre daha etkili olduğu sonucuna ulaşılmıştır.

Aktaş (2013)'ın araştırmasında, öğrencilerin hikâye yazma becerileri farklı puanlayıcılar tarafından; kontrol listesi, dereceleme ölçeği ve analitik rubrik araçlarıyla puanlanmış ve elde edilen puanların güvenirliği Genellenebilirlik Kuramı çerçevesinde incelenmiştir.

Öğrencilerin yazdıklarından seçilen 6 hikâye, 45 puanlayıcıya üç farklı puanlama anahtarı ile 10-15 gün aralıklarla puanlattırılmıştır. Araştırmaya katılan 45 puanlayıcı içerisinden 2, 3, 5 ve 10 puanlayıcılı 100'er örneklem seçilerek elde edilen 400 örneklem için Genellenebilirlik Kuramı çerçevesinde puanlayıcılar arası güvenirlikler hesaplanmıştır. Bulguların her 100 örneklemi için ortanca ve standart sapma değerleri hesaplanmıştır. Araştırmanın sonunda Genellenebilirlik Kuramı'ndan elde edilen standart hataların, puanlayıcı sayısı arttıkça azaldığı gözlenmiştir. Ayrıca puanlayıcı sayısının 5 ve kategori sayısının 2 olduğu durumda, güvenirlik katsayısı en yüksek değere ulaşmıştır.

Alkan (2013), PISA 2009 Okuma Becerileri performansını ölçen açık uçlu soruların, birden fazla puanlayıcı tarafından birlikte ve dönüşümlü olarak puanlanmasıyla elde edilen farklı desenleri Genellenebilirlik Kuramına göre karşılaştırmıştır. Kullanılan ilk desen, öğrenci (ö), soru (s) ve puanlayıcı (p) değişkenleri olmak üzere, Ö X S X P çapraz desenidir. İkinci desen ise öğrenci ve puanlayıcı değişkenlerinin yuvalanmış olduğu, soruların ise bu değişkenlerle çaprazlanmış olduğu (Ö:P) X S desenidir. Araştırma sonucunda, her iki desende de G çalışmaları ile kestirilen varyans bileşenleri öğrencilerin okuma becerileri açısından birbirinden farklılaştığını göstermiştir. Ayrıca her iki desende de puanlayıcıya ait varyans bileşeni oldukça küçük bulunmuş dolayısı ile puanlayıcıların öğrencileri tutarlı puanladığı sonucuna ulaşılmıştır. K çalışmaları ile (Ö:P) X S deseninde kestirilen mutlak ve bağıl hata varyanslarının Ö X S X P desenine göre daha küçük olduğu görülmüştür. Buna bağlı olarak da G ve Phi katsayılarının, (Ö:P) X S deseninde daha büyük değerler aldığı sonucuna ulaşılmıştır.

Yurt Dışında Yapılan Çalışmalar

Roebroeck, Harlaar ve Lankhorst (1993) çalışmalarında, izometrik kuvvet ölçümünün güvenirliğinin değerlendirilmesinde genellenebilirlik kuramını uygulamışlardır. Denek olarak 10 kadının bulunduğu çalışmada, deneklerin kas kuvvetleri birer hafta süre ile iki terapist tarafından standartlaştırılmış test protokollerine göre ölçülmüştür. Genellenebilirlik kuramına göre varyans kaynakları hesaplanmış ve en büyük hata kaynağının terapist x denek ortak etkisine ait olduğu görülmüştür. Denek x terapist varyans bileşeninin oldukça büyük olması, bazı deneklerin terapistten terapiste kasılma sürelerinin değiştiğinin göstergesidir. Terapist x tekrarlama durumuna ait varyans bileşenlerinin ise önemsiz olduğu yani deneklerin ortalama kasılma süresinin tekrarlama durumuna göre değişmediği görülmüştür. Araştırmada denek, terapist ve durum ana etkilerinin hata varyansına katkıları önemsiz derecede bulunmuş iken

bunlar arasındaki etkileşimlerden oluşan varyans bileşenlerinin ve artık varyansın önemli hata kaynakları olduğu sonucuna ulaşılmıştır.

Hoyt ve Melby (1999), çalışmalarında danışmanlık psikolojisindeki ölçümlerin güvenirliğini değerlendirmede Genellenebilirlik kuramının KTK'ya göre avantajlarını açıklamıştır. Klasik güvenirlik yaklaşımlarının aksine hatanın pek çok kaynağını eşzamanlı olarak kestirebilmeye olanak sağlaması, kestirilen varyans bileşenlerinin gelecekte yapılacak ilgili araştırmalar için yol gösterici olması G kuramının üstün yanları olarak vurgulanmıştır. Ayrıca özellikle danışmanlık psikolojisi alanı ve diğer alanlarda, GK'nın en yararlı özelliklerinden birisinin araştırmacının ihtiyaç ve istekleri doğrultusunda uygulanabilirliği olduğu vurgulanmıştır.

Goodwin (2001) araştırmasında, puanlayıcılar arasındaki uyum ve güvenirlik üzerine çalışmıştır. Araştırmada puanlayıcılar arasındaki tutarlılık ve güvenirlik; hipotetik veriler kullanılarak, basit yüzde ve korelasyon teknikleri, Kappa ve Genellenebilirlik kuramı analizleri ile hesaplanmış ve sonuçlar karşılaştırılmıştır. Araştırmada 10 çocuk 2 farklı puanlayıcı tarafından bağımsız olarak, 6 farklı günde, fiziksel aktiviteleri gözlemlenerek yarım saatlik periyotlarla puanlanmıştır. Genellenebilirlik kuramı kapsamında katılımcılara (çocuklar) ek olarak 2 yüzey daha desene katılmıştır: puanlayıcı ($n_p = 2$) ve zaman ($n_z = 6$). Değişkenlik yüzeylerinin tamamen çaprazlandığı Ç X P X Z deseni kullanılmış ve bu desene göre G ve K çalışmaları yapılmıştır. Yapılan G çalışmaları sonucunda çocuklara ait varyans bileşeninin toplam varyansı açıklama oranı % 36 iken; artık varyansın oranı % 42 bulunmuştur. Bu bağlamda çocukların ölçülen özellik açısından birbirlerinden farklılaştığı sonucuna ulaşılmıştır. Artık varyans bileşeninin yüksek olması; puanlayıcıların çocukları farklı zamanlarda tutarlı puanlayamamasından ya da puanları etkileyebilecek diğer varyans bileşenlerinin G çalışmasında olmamasından kaynaklanabilmektedir. Çalışma sonunda uyum indekslerinin güvenirlik kat sayılarına göre farklı ve daha dar anlamlar verdiği sonucuna ulaşılmıştır. Zira puanlayıcıların vermiş oldukları puanlar arasındaki Pearson korelâsyon katsayısı 0,90 bulunmuştur ki bu; puanlamaların birlikte değişiminin ölçüsünü verir. Yani 0,90 istatistiği çok genel anlamda bir puanlayıcı güvenirliği vermiştir. Bu teknikler arasında Genellenebilirlik kuramının tek bir analizle birçok hata kaynağını göstermesi ve daha kapsamlı olduğu düşüncesinden hareketle bu tür araştırmalarda daha avantajlı olduğu öne sürülmüştür.

Brown (2005), Genellenebilirlik ve Karar çalışmalarının neler olduğunu, bu çalışmaların nasıl ayrıldığını ve ne zaman kullanıldığını teorik açıdan ele almıştır. Bu bağlamda, Genellenebilirlik çalışmalarının, bir araştırmada ilgili değişkenlik yüzeylerinin kestirilmesine

yönelik ANOVA tekniklerinin kullanımına dayandığı ifade edilmiştir. Çalışmada bulunan bütün değişkenlik kaynakları ve bütün muhtemel etkileşimlerinden oluşan varyans bileşenlerinin kestirimi için ortalama kareler kullanılmaktadır. G çalışmalarını takiben yapılan Karar çalışmalarında ise önceden kestirilen bu varyans bileşenleri, çeşitli ölçme senaryolarında güvenirliğe etkilerini kestirmek için kullanılmaktadır. Alternatif senaryolarda hata varyanslarına ve güvenirlik katsayılarına bakılarak en uygun ölçme senaryosuna karar verilmektedir. Dolayısı ile araştırmacı G ve K çalışmalarının beraber ve sıralı (G çalışması ve ardından K çalışması) yapılmasının ilgili çalışmalar için anlamlı olacağını ifade etmiştir.

Chafolues, Christ, Tillman ve Chanese (2007), çalışmalarında anaokulu öğrencilerinin sosyal davranışlarını ölçmede doğrudan davranış puanlama ölçeğinin geçerlilik ve güvenirliklerini incelemişlerdir. Çalışmada öğrenciler üzerinde gözlemlenmesi hedeflenen iki sosyal davranış; çatışmaları çözmeye çalışma ve akranlarla işbirliği içinde bulunma davranışlarıdır. G kuramı kapsamında öğrenci, puanlayıcı, zaman, ortam ve skor (puan) olmak üzere 5 değişkenlik kaynağı, tamamıyla çaprazlanmıştır. Araştırma sonunda her iki davranış için de, öğrenci, puanlayıcı ve artık varyans bileşenleri yüksek bulunmuştur, diğer varyans bileşenlerinin katkısı ise %8'den azdır. Araştırmada güvenirlik katsayılarının; puanlayıcı sayısına ve davranış türüne bağlı olarak önemli ölçüde değişkenlik gösterdiği sonucuna ulaşılmıştır.

Briesch, Chafolueas ve Tillman (2010) tarafından yapılan araştırmada, akademik ilgiyi ölçen iki davranış değerlendirme metodu üzerinde genellenebilirlik ve güvenirliğe bakılmıştır. Çalışma grubunu 12 anaokulu öğrencisi oluşturmuştur. Çalışmanın bağımsız değişkeni olan akademik ilgi kavramı; aktif olarak öğrencilerin yazma, el kaldırma gibi becerileriyle; pasif olarak öğretmeni dinleme, sessizce okuma gibi becerilerle tanımlanmıştır. Genellenebilirlik kuramı için puanlayıcıların yöntemlere (Sistematik Doğrudan Gözlem- SDO ve Doğrudan Davranış Puanlama-DBR); gözlem periyotlarının zamana yuvalanmış; öğrencilerin ise bu yüzeylere çaprazlanmış olduğu P X (R:M) X (O:D) deseni tasarlanmıştır. Her yöntem için ayrı G çalışmaları yapılarak öğrenci, puanlayıcı, puanlama durumu, zaman ve bu değişkenlik kaynakları etkileşimlerinin etkisine bakılmıştır. Yapılan G çalışmaları ile SDO yöntemi için G ve Phi katsayıları sırası ile 0,98; 0,97 bulunmuş iken; DBR yöntemi için 0,82; 0,87 bulunmuştur. SDO ve DBR metotlarının toplam varyansı açıklayan en büyük değişkenlik kaynaklarının sırasıyla öğrenci x zaman ve puanlayıcı yüzeyleri olduğu görülmüştür.

Christ, Tillman, Chafouleas ve Boice (2010) çalışmalarında, iki adet doğrudan performans değerlendirme ölçeğinden gelen sonuçların genellenebilirlik ve güvenirliklerini incelemek üzere genellenebilirlik kuramı kullanmıştır. Çalışmada, yüz yirmi beş lisans öğrencisi,

öğrencilerin lego yapbozu göreviyle uğraştıkları video kayıtlarını izlemiş ve puanlamıştır. Bunun ardından değişik senaryoların olduğu (farklı sayıda puanlayıcı ve puanlama durumları) karar çalışmaları yapılmıştır. Çalışma sonucunda kısıtlı bir genelleme evreni ile her bir durumdaki eş zamanlı puanlayıcı sayısının, sonuçlar üzerinde çok az bir etkisi olduğu görülmüştür. Buna karşılık puanlayıcı yüzeyine ait genelleme evreninin sonsuz olduğu çalışmalarda puanlayıcı sayısının sonuçlar üzerine daha etkili olduğu sonucuna varılmıştır.

Brennan (2011) çalışmasında, Genellenebilirlik Kuramı ile Klasik Test Kuramını teorik açıdan ele almış ve karşılaştırmıştır. Güvenirlik kavramını her iki kurama göre tartışmış ve iki kuramın da Madde Tepki Kuramı ile güvenirlik ölçütünde kıyaslamalarını yapmıştır. Bu kıyaslamalar araştırmacılara tavsiye olabilecek şekilde çalışmada sunulmuştur: KTK da yalnızca bir hata teriminden (E) söz edilir, bu hatanın tek kaynağı olduğunu göstermez aksine bütün hata kaynakları tek bir "E" terimi ile gösterilir. G kuramında ise araştırmacının ilgilendiği pek çok hata kaynağı kestirilebilmektedir. Araştırmacı bu hata kaynaklarından hangisi ile ilgileneceğine karar verebilmeli ve hangilerinin ölçme yüzeylerini etkili biçimde tanımladığını bilmelidir. Bunlara ek olarak, kuramların varsayım kıyaslamaları yapıldığında KTK ve G kuramının aksine Madde Tepki kuramının varsayımlarının güçlü olduğu belirtilmiştir. Ayrıca her üç kuramın güçlü ve zayıf yönleri karşılaştırılmıştır: KTK'nın basit ve yaygın kullanımlı olması; G kuramının farklı hata kaynaklarını tek bir analizle kestirmeye olanak sağlaması, sabit ve sonsuz yüzeylere uygulanabilirliği; MTK'nın matematiksel karmaşık yapıda olan ölçme durumlarına uygun olması gibi. KTK'nın hata varyansını ayrıştıramaması, G kuramının kavramsal zorluğu, MTK'nın ise yalnızca sabit yüzeylere uygulanabilir olması ise bu kuramların zayıf yönleri olarak sıralanmıştır.

Volpe ve Briesch (2012) çalışmalarında, yapıcı ve yıkıcı davranışı ölçmek için geliştirilen tek maddeli ve çok maddeli doğrudan davranış derecelendirme ölçeğinin genellenebilirliği ve güvenirliği üzerinde çalışmışlardır. Çalışmada doktora derecesindeki iki puanlayıcı 7. Sınıf düzeyindeki 8 ortaokul öğrencisini her iki metoda göre, 3 farklı zamandaki, 10 dakikalık video kliplerini gözlemleyerek puanlamıştır. Genellenebilirlik kuramı kapsamında hem G hem de K çalışması yapılmıştır. Araştırmada G çalışması sonucunda çok maddeli ve tek maddeli puanlama ölçeklerinden elde edilen bireylere ait varyans bileşeninin toplam varyansı açıklama yüzdeleri sırası ile %67 ve %46 bulunmuştur. Bu bulgu yöntemden yönteme bireylerin farklılaştığını göstermektedir. Yöntem x zaman etkileşiminden kaynaklanan varyans bileşeni her iki yöntemde de ikinci büyük varyans bileşenini oluşturmaktadır. Zaman ve zaman x puanlayıcıya ait varyans bileşenlerinin ise toplam varyansa katkısı neredeyse

0'dır. Bu durum zaman karşısında öğrencilerin ölçülen davranışları açısından puanlayıcıların tutarlı olduğunun göstergesidir. Karar çalışmalarında ise çok maddeli puanlama ölçeklerinden elde edilen G ve Phi katsayıları daha yüksek bulunmuştur. Tek maddeli ölçeğin göreli ve mutlak kararlar verebilmek için kabul edilebilir güvenirlik parametrelerine ulaşması için zaman sayısının minimum 8 olması gerektiği sonucuna ulaşılmıştır.

Arterberry, Martens, Cadigan ve Smith (2012), çalışmalarında alkol kullanma güdüsünü ölçen bir ölçeğin güvenirliğini Genellenebilirlik kuramı ile değerlendirmişlerdir. Çalışmada 367 üniversite öğrencisi, üç zaman aralığında ölçeği tamamlamıştır. Genellenebilirlik çalışması kapsamında birey, madde ve durumların tamamıyla çaprazlandığı B X M X D deseni kullanılmıştır. G çalışması kapsamında ölçeğin 4 alt boyutu içinde ayrı ayrı varyans bileşenleri ve G katsayıları kestirilmiştir. Karar çalışmaları için ise durum ve madde sayısının artıp azalmasına yönelik senaryolar oluşturularak G katsayıları incelenmiştir. Araştırmada bireylerin birbirine göre durumları önemsendiği yani göreli karar söz konusu olduğu için G katsayıları önemsenmiştir. Analiz sonuçlarında 3 alt ölçek için varyans bileşenlerinin ve G katsayılarının benzer olduğu görülmüştür. Diğer alt ölçek için ise maddelerin sırasıyla çıkarılarak varyans bileşenlerinin kestirildiği Post Hoc testleri yapılmış ve diğer alt ölçekler ile paralel sonuçlara ulaşılmıştır.

Hill, Charalombous ve Kraft (2012) çalışmalarında, puanlayıcı tutarlılığının yeterli olmadığı durumlarda öğretmen gözlem sisteminin geliştirilmesine yönelik, G çalışmasının kullanıldığı bir uygulama yapmışlardır. Çalışmada sınıf içi gözlem süreçlerinin amaçlarına ulaşması için, güvenilir gözlem araçları geliştirmenin yanında gözlem sistemlerinin geliştirilmesi gerektiğine odaklanılmıştır. Bunun için öğretmenlerden oluşan deneysel bir örneklemden matematiksel gözlem aracı (matematiksel yönergelere uyum düzeyi) ile 9 puanlayıcı tarafından puanlanarak elde edilen veriler G kuramı ile analiz edilmiştir. G çalışması kapsamında derslerin öğretmenlere yuvalandığı ve bu iki yüzeyin puanlayıcılarla çaprazlandığı (D:Ö) X P deseni kullanılmıştır. Araştırma sonunda genel yargının aksine belli bir aracın güvenirliğinden bahsetmek yerine; güvenirliğin, ölçme araçlarının, puanlayıcı eğitiminin, spesifik bir puanlama sisteminin yani bir gözlem sisteminin birleşiminden etkilendiği belirtilmiştir. Ayrıca araştırmada olduğu gibi, özellikle karmaşık performansların ölçümünde puanlayıcı güvenirliliğinin tek bir kriteri olmaması gerektiği sonucuna ulaşılmıştır.

Huang (2012), çalışmasında iki farklı dil grubundaki öğrencilerin İngilizce yazma becerilerinin puanlanması sonucu elde edilen verilerin geçerliğini ve doğruluğunu (accuracy)

Genellenebilirlik kuramı ile incelemiştir, her iki grup için 3 yıl boyunca ve ayrı ayrı hesaplanan G katsayılarının arasındaki farlılıkların manidarlığını Feldt'in yöntemi ile test etmiştir. Araştırmada G çalışması için ayrı dil grubundaki öğrencilerin görevlerle ve puanlayıcılarla çaprazlandığı B X G X P deseni kullanılmıştır. Dil grupları sabit yüzey iken; puanlayıcı ve görev tesadüfî değişkenlik kaynaklarıdır. Her iki grup için yapılan G çalışmaları sonucu elde edilen G katsayıları birinci dil grubu için 3 yıl boyunca daha küçük bulunmuş ve katsayılar arasındaki farklılık her 3 sene için de 0,05 düzeyinde anlamlı bulunmuştur. Ayrıca puanlayıcı x birey değişkenlik kaynağına ait varyans bileşeni birinci dil grubundaki öğrencilerde (0,05 düzeyinde anlamlı derecede) daha büyük bulunmuştur. Bu durum, birinci dil grubundaki öğrencilerin puanlamalarının daha az tutarlı olduğunu göstermektedir.

Briesch, Swaminathan, Welsh ve Chafouleas (2014) çalışmalarında, Genellenebilirlik kuramı çalışma deseni oluşturma, uygulama ve yorumlamaya yönelik pratik bir rehber sunmuşlardır. Eş zamanlı olarak birden çok hata kaynağını kestirmeye ve sonuçların genellenebilirliği ve güvenirliği ile ilgili hem mutlak hem de bağıl kararlar almaya olanak sağlayan G kuramının; psikoloji ve eğitim alanlarında çok kullanılamamasının, kavramsal alt yapısının yeterli düzeyde anlaşılamamasından kaynaklı olduğu ifade etmişlerdir. Çalışmada şu amaçlara değinilmiştir: (1) G kuramının kullanımında kavramsal altyapıyı ve terminolojiyi göstermek. (2) Genellenebilirlik ve güvenirlik çalışmaları için tasarlanan desenlerin uygulanması ve yorumlanmasına katkı sağlamak. Bu bağlamda G ve K çalışmaları adımları sırası ile sunulmuştur. Çalışma deseni oluşturulurken desenin amaca uygunluğunun ve yüzeylerin açıklanabilmesi; ölçme modelinin; yüzeylerini, koşullarını ve etkileşimlerini göz önünde bulundurarak tanımlanabilmesi önerilerinde bulunulmuştur.

Genel olarak yapılan araştırmalara bakıldığında, Genellenebilirlik kuramıyla ilgili olarak farklı desenlerdeki güvenirlik katsayılarının karşılaştırılmasına, farklı senaryolardaki K çalışmalarında en uygun güvenirlik katsayısını bulmaya yönelik çalışmalara odaklanıldığı görülmektedir. Ayrıca belli bir alana yönelik performansın ölçülmesinde G kuramı ile Klasik Test Kuramının karşılaştırıldığı çalışmalar da mevcuttur. Yapılan araştırmalarda G kuramının birden fazla hata kaynağını ortaya çıkarabilmesi ve daha hassas güvenirlik parametreleri elde edebilmesi açısından daha çok tercih edildiği görülmüştür.

Bunun yanında Genellenebilirlik Kuramı ile Klasik Test Kuramı'nın karşılaştırıldığı çalışmalarda, ölçekle elde edilen puanların tek bir desendeki güvenirliği incelenmiştir, G kuramına göre farklı desenlerden elde edilen güvenirlik katsayılarının Klasik Test Kuramındaki güvenirlik katsayısı ile karşılaştırıldığı bir çalışmaya rastlanmamıştır.

BÖLÜM II

YÖNTEM

Bu bölümde araştırmanın modeli, çalışma grubu, araştırma verilerini toplama araçları ve verilerin analizine yer verilmiştir.

Araştırmanın Modeli

Araştırma, Matematiksel muhakeme performansının belirlenmesine yönelik kullanılan ölçeğin Genellenebilirlik Kuramı'na göre iki farklı deseninden elde edilen ölçümlere ait güvenirlik kestirimlerinin incelenmesi açısından betimsel araştırma özelliği taşırken, her iki desendeki güvenirlik parametrelerinin Klasik Test Kuramı parametreleri ile karşılaştırılmasına yönelik olması açısından temel bir araştırmadır.

Çalışma Grubu

Araştırmanın çalışma gurubunu, Konya ilinde bulunan, 2014-2015 eğitim-öğretim yılında yedinci sınıfta öğrenim gören 187 kişilik öğrenci grubu oluşturmuştur. Öğrencilerin okullara göre dağılımı Tablo 1'de sunulmuştur:

Tablo 1. Öğrencilerin Cinsiyete ve Okul Türlerine Göre Dağılımı

Okul İsimleri	Kız		Erkek		Toplam	
	Frekans (f)	Yüzde %	Frekans (f)	Yüzde %	Frekans (f)	Yüzde %
Hâkim Ömer Onsun Ortaokulu	18	56	14	44	32	17,1
Özel Meram Abdullah Aymaz Ortaokulu	42	55	35	45	77	41,2
Özel Tür-mak Ortaokulu	25	63	15	27	40	21,4
Özel Gündoğdu Koleji	18	47	20	53	38	20,3
Toplam					187	100

Çalışma grubu, kolay ulaşılabilir ve uygulama yapılabilir nitelikteki okullardan oluşmuştur. Bu sebeple bu çalışma grubundan toplanan verilerle yapılan güvenirlik analizi sonuçları Konya ilinde bulunan tüm okullara genelleme özelliği taşımamaktadır.

Veri Toplama Araçları

Araştırmada veri toplama aracı olarak, matematiksel muhakeme becerisinin belirlenmesine yönelik, Pilten (2008) tarafından geliştirilen ölçek kullanılmıştır.

Matematiksel muhakemenin belirlenmesinde kullanılan bu ölçek, "Üstbiliş Stratejileri Öğretiminin İlköğretim Beşinci Sınıf Öğrencilerinin Matematiksel Muhakeme Becerilerine Etkisi" konulu doktora tezinde uygulanmak üzere geliştirilmiştir. Ölçeğin geliştirilmesinde, Tracy ve Gibson (2005) tarafından ortaya konulan üç aşama kullanılmıştır.

İlk aşamada literatür taraması yoluyla matematiksel muhakemeyi değerlendirmeye yönelik araştırmalar incelenmiştir. Ölçeğin kavramsal temelleri literatür sonucu oturtulmuştur. İkinci aşamada, ölçek maddeleri; cevap formatları belirlenerek, madde havuzu oluşturularak, kapsam geçerliliği için her bir madde, madde ölçek boyutlarında yeniden değerlendirilerek ve nihayetinde geniş bir örneklem üzerinde denenerek son halini almıştır. Üçüncü aşamada veri analizleri; madde analizi, geçerlik ve güvenirlik çalışmaları kapsamında tamamlanmıştır (Pilten, 2008).

Ölçeğin geliştirilmesinde muhakemenin değerlendirilmesi ile ilgili çalışmalar incelenerek ölçeğin alt boyutları oluşturulmuştur. Bu takdirde literatürün ortaya koyduğu kuramsal temeller şu şekildedir:

1. NCTM (1989), muhakeme becerilerini; muhakeme, matematik gücü ve problem çözme gibi kategorilerle birlikte ele almaktadır.
2. NAEP (2002), muhakeme becerilerine problem çözme becerisi içerisinde yer vermektedir.
3. Bloom'un sınıflamasında muhakeme becerilerini; sentez ve değerlendirme gibi üst basamaklar karşılamaktadır (Suzuki, 1998).
4. TIMMS (2003), muhakeme becerilerini; araştırma - problem çözme ve matematiksel muhakeme kategorileri içerisinde tanımlamaktadır.
5. MEB (2005), muhakeme becerilerini akıl yürütme becerisi olarak ele almaktadır (Pilten, 2008).

Soru sayılarının ölçek boyutlarına göre dağılımı Tablo 2'de sunulmuştur.

Tablo 2. Soru Sayılarının Ölçek Boyutlarına Göre Dağılımı

Ölçek Boyutları	Literatürde Belirtilen Boyutta Yer Alan Toplam Beceri Sayısı	Literatürde Yer Alan Becerilerin Ağırlıklarına Göre Ölçek Boyutlarına Yansıtılacak Yüzdeler	Ölçek Boyutlarında Yer Alacak Soru Sayıları
Uygun Muhakemeyi Belirleme Ve Kullanma	6	% 32	4
Çözüme İlişkin Mantıklı Tartışmalar Geliştirme	3	% 16	5
Çözüm Yolu/ Sonucun Doğruluğuna Karar Verme	5	% 26	3
Genelleme Yapma	3	% 16	3
Rutin Olmayan Problemleri Çözme	2	% 10	3
TOPLAM	19	100	18

Kaynak: Pilten, P. (2008).

Tablo 2 literatürde yer alan matematiksel muhakeme becerileri ölçeğinin alt boyutlarını oluşturacak şekilde belli başlıklar altında toplanmıştır ve madde tipinin belirlenmesinde alanyazından yararlanılmıştır.

Ölçeğin madde tipi, belli kurumların değerlendirmeye yönelik verdiği kriterleri dikkate alarak öğrencilerin muhakeme becerilerini detaylıca ortaya koyabileceği açık uçlu sorular olarak belirlenmiştir.

Ölçeğin madde analizi, güvenirlik ve geçerlik çalışmaları, Pilten (2008); araştırmacı sırasına göre verilmiştir.

Madde Analizi

Madde analizi kapsamında sırasıyla, maddelerin ölçülen özellik bakımından bireyleri ayırmasının ölçüsü olan madde ayırıcılık gücü indeksi; madde zorluğunun ölçüsü ya da doğru cevaplama yüzdesi olarak da anılan madde güçlüğü indeksleri verilmiştir (Baykul, 2000).

Ölçekte yer alan açık uçlu soruların analizinde Pilten (2008), Öncü'nün (1999) belirttiği formülleri kullanmıştır.

Madde ayırt ediciliği için;

$$d_j = \frac{\sum ü - \sum a}{N\ .(maks - puan\ )}$$

d_j = j Maddesinin Ayırt Edicilik İndeksi

$\sum ü$ = Üst %25'in Puanlarının Toplamı

$\sum a$ = Alt %25'in Puanlarının Toplamı

N = Test Edilen Öğrencilerin %25'i

maks−puan = Sorudan Alınabilecek En Büyük Puan

Madde güçlüğü için;

$$d_j = \frac{\sum ü + \sum a}{2N\ .(maks - puan\ )}$$

p_j = j Maddesinin Güçlük İndeksi

$\sum ü$ = Üst %25'in Puanlarının Toplamı

$\sum a$ = Alt %25'in Puanlarının Toplamı

N = Test Edilen Öğrencilerin %25'i

maks−puan = Sorudan Alınabilecek En Büyük Puan

Pilten (2008) tarafından yapılan analizler sonucunda tüm maddelerin ayırt edicilik ve güçlük indekslerinin uygun olduğu ortaya çıkmıştır.

Araştırmacı ise, her 3 puanlayıcıdan elde edilen puanlar için ayrı ayrı madde analizleri gerçekleştirmiştir, her bir puanlayıcı için hesaplanan madde ayırt edicilik ve güçlük indeksleri sırasıyla Tablo 3 ve Tablo 4'te sunulmuştur.

Araştırmacı madde ayırt ediciliği için düzeltilmiş nokta çift serili korelâsyon katsayısını kullanmıştır. Hesaplanan korelâsyon katsayısı madde ayırıcılık gücü için özellikle madde sayısının az olduğu durumlarda uygun bir istatistik olmaktadır.

Tablo 3. Her 3 Puanlayıcı için Hesaplanan Madde Ayırt Edicilik İndeksleri

Maddeler	Birinci Puanlayıcı için Madde Ayırt Edicilikleri	İkinci Puanlayıcı için Madde Ayırt Edicilikleri	Üçüncü Puanlayıcı için Madde Ayırt Edicilikleri
1	0,67	0,71	0,64
2	0,69	0,66	0,59
3	0,69	0,45	0,40
4	0,66	0,66	0,61
5	0,54	0,62	0,51
6	0,56	0,66	0,62
7	0,54	0,49	0,54
8	0,43	0,52	0,44
9	0,43	0,53	0,44
10	0,35	0,31	0,48
11	0,60	0,73	0,59
12	0,50	0,59	0,54
13	0,55	0,65	0,57
14	0,45	0,32	0,33
15	0,58	0,52	0,63
16	0,62	0,50	0,46
17	0,71	0,68	0,75
18	0,50	0,49	0,57

Tablo 3 incelendiğinde puanlayıcılara ait madde ayırt edicilik indeksleri birbirlerine oldukça yakın görünmektedir. Madde ayırıcılık gücü indeksi 0,20'den küçük olan maddelerin ölçekten atılmasının ölçeğin güvenirliğini önemli ölçüde arttırdığı öngörülmüştür. Ayrıca madde ayırıcılık gücü indeksinin 0,40'tan büyük olması o maddenin iyi bir madde olduğunun göstergesidir. Bu bağlamda, tüm maddelerin ayırt edicilik gücü indekslerinin oldukça yüksek olduğu, dolayısı ile hiçbir maddenin ölçekten atılmasına gerek duyulmadığı görülmüştür.

Araştırmacı madde güçlüğü için ise; bütün ölçek puanlarının hesaplamaya dâhil edildiği;

$p_j = \frac{madde\ \ puanlar\ ının\ aritmetik\ \ ortalaması}{maddenin\ \ maksimum\ \ puanı}$ formülünü kullanmıştır. Bu bağlamda her bir puanlayıcı için hesaplanan madde güçlükleri Tablo 4'te sunulmuştur.

Tablo 4. Her 3 Puanlayıcıdan Elde Edilen Madde Güçlükleri

Maddeler	Birinci Puanlayıcı için Madde Güçlükleri	İkinci Puanlayıcı için Madde Güçlükleri	Üçüncü Puanlayıcı için Madde Güçlükleri
1	0,39	0,42	0,51
2	0,40	0,48	0,51
3	0,65	0,72	0,77
4	0,61	0,64	0,60
5	0,49	0,51	0,46
6	0,59	0,55	0,55
7	0,67	0,65	0,62
8	0,74	0,73	0,68
9	0,71	0,69	0,72
10	0,51	0,48	0,51
11	0,45	0,63	0,45
12	0,67	0,68	0,66
13	0,58	0,52	0,60
14	0,58	0,63	0,64
15	0,35	0,26	0,33
16	0,30	0,27	0,28
17	0,43	0,28	0,44
18	0,43	0,45	0,57
Testin Güçlüğü	0,53	0,53	0,55

Madde güçlüğünün 0,50 olması, madde güvenirliğini yükseltmesi bakımından istenen bir durumken, Kehoe (1995) iyi bir test için madde güçlüğünün 0,30-0,80 arasında olmasının yeterli olabileceğini belirtmiştir (Kehoe'den aktaran Tan, 2012). Tablo 4 incelendiğinde maddelerin hemen hemen hepsinin güçlüklerinin belirtilen aralıkta olduğu; 14, 15 ve 16 numaralı maddelerin ise bazı puanlayıcıları için madde güçlüklerinin 0,30'un altında olduğu

görülmektedir. Ancak, birbirilerine oldukça yakın değerler alan test güçlükleri, testin güçlük düzeyinin oldukça makul olduğunu göstermektedir.

Ölçümlerin Güvenirliği

Ölçeğin güvenirlik çalışmaları kapsamında Pilten (2008) tarafından iç tutarlılık katsayısı ve test tekrar test güvenirliği hesaplanmış, sırasıyla 0,87; 0,76 bulunmuştur. Testin iç tutarlılığı için kabul edilebilir katsayının en az 0,70 olması (Tezbaşaran, 1996; Tavşancıl, 2002) ve Guildford'a (1956) göre, test tekrar test ile belirlenecek bir testin güvenirlik katsayısının en az 0,70 olması görüşlerini öngören Pilten, Matematiksel muhakeme performansının belirlenmesine yönelik kullanılan ölçekten elde edilen ölçümlerin yüksek güvenirliğe sahip olduğunu ifade etmiştir.

Araştırmacı ise, ölçümlerin güvenirliğini Genellenebilirlik kuramı ve Klasik Test kuramı kapsamında analiz etmiştir.

Ölçümlerin Geçerliği

Araştırmacı ölçeğin yapı geçerliği analizinde, daha önce Pilten tarafından tanımlanmış ve sınırlandırılmış modelin doğrulanıp doğrulanmadığını test etmek amacında olduğu için doğrulayıcı faktör analizini kullanmıştır (Büyüköztürk, Çokluk ve Şekercioğlu, 2010).

Çalışmada 187 öğrencinin ölçek maddelerine vermiş oldukları cevaplar 3 farklı puanlayıcı tarafından puanlandığı için, her 3 puanlayıcının vermiş olduklara puanlara yönelik ayrı ayrı doğrulayıcı faktör analizi (DFA), LISREL 8.8 programı kullanılarak yapılmıştır. Tablo 5'te her üç puanlayıcı için ayrı hesaplanmış, maddelerin faktör yük değer aralıkları bulunmaktadır.

Tablo 5. Üç Puanlayıcıya ilişkin Faktör Yük Değerleri Aralıkları

	1. Puanlayıcı	2. Puanlayıcı	3. Puanlayıcı
Faktör 1	0,36-0,80	0,42-0,76	0,34-0,79
Faktör 2	0,53-0,68	0,57-0,74	0,47-0,71
Faktör 3	0,37-0,68	0,32-0,98	0,58-0,65
Faktör 4	0,35-0,69	0,32-0,77	0,44-0,67
Faktör 5	0,53-0,73	0,40-0,89	0,55-0,88

Tablo 5 incelendiğinde daha önce Pilten (2008) tarafından tanımlanan beş faktöründe yük değerlerinin 0,30'dan yüksek olduğu görülmektedir. Ayrıca DFA analizi sonucu t değerleri de incelenerek faktör yük değerlerinin anlamlı olduğu görülmüştür. Ölçek maddelerinin ilgili yapılarla (faktörlerle) uyumunun derecesini gösteren, uyum iyiliği indekslerinden GFI, CFI, NNFI değerleri her üç puanlayıcı için de 0,90 üzerinde bulunmuştur. GFI'nın düzenlenmiş bir türü olan AGFI değerleri ise 0,80-0,84 arasında değişmektedir. O halde yapılan faktör analizi sonucunda uyum iyiliği indekslerinin ve faktör yük değerlerinin kabul edilebilir aralıkta olduğu görülerek ölçek maddeleri ile ilgili faktörlerden kurulan modelin doğrulanmış olduğu söylenebilir.

Verilerin Toplanması

Çalışmada verilerin toplanması araştırma amaçlarına uygun olarak iki aşamada gerçekleşmiştir. Bu aşamalar verilerin toplanması başlığı altında sırasıyla açıklanmıştır.

Ön Uygulama Aşaması

Öğrencilerin matematiksel muhakeme performanslarını ölçmeden önce Pilten (2008) tarafından geliştirilen matematiksel muhakeme performansını belirleme ölçeği orta düzeyde başarı gösteren ilköğretim beşinci, altıncı ve yedinci sınıf düzeylerindeki 3 öğrenciye klinik mülakat yöntemi ile uygulanmıştır. Öğrencilerin bilişsel becerilerinin değerlendirebilmesini sağlayan ve esnek sorular sorma tekniği olarak da adlandırılan klinik mülakat tekniği ile ölçekte yer alan sorular her 3 öğrenciye de ayrı ayrı yönlendirilmiş ve öğrencilerin verdiği cevaplar araştırmacı tarafından gözlemlenerek not edilmiştir (Baki, Karataş ve Güven, 2002). Yapılan mülakat sonucu ölçekte yer alan soruların yedinci sınıf düzeyindeki öğrenciler için daha uygun olduğu görülmüştür. Bununla birlikte araştırmacı ölçekte yer alan soruların bilişsel seviyelerine yönelik, uygulamayı yapan 4 matematik öğretmeni ile de uzlaşmış, soruların yedinci sınıf öğrencilerine uygulamanın daha uygun olacağı sonucuna ulaşmıştır.

Matematiksel Muhakeme Performansının Belirlenmesi ve Puanlanması Aşaması

Araştırmada 187 ilköğretim yedinci sınıf öğrencisinin matematiksel muhakeme becerilerinin ölçülmesinde Pilten (2008) tarafından geliştirilen ölçek, araştırmacının gözleminde bir ders saati süresince uygulanmıştır. Ayrıca uygulamaya geçilmeden önce öğrencilere, ölçeğin

madde tipi ve buna bağlı cevaplama şekli hakkında netliği kazanmaları için araştırmacı tarafından gerekli bilgilendirme yapılmıştır.

Böylece 187 öğrencinin açık uçlu sorulara verdikleri cevaplar puanlayıcılar tarafından puanlanmak üzere kodlanarak hazır hale getirilmiştir. Uygulamanın ardından öğrenci grubunun ölçeğe verdikleri cevapları, biri araştırmacının kendisi olmak üzere, Ankara'da görev yapan matematik öğretmenliği lisans eğitimli 3 gönüllü uzman puanlamıştır. Ayrıca puanlama yapılmadan önce uzmanlar matematiksel muhakeme kavramının alt boyutları, soruların hangi alt boyutlara ait olduğu ve performansa dayalı durum belirleme ile alakalı bilgilendirilmişlerdir. Ardından gönüllü üç uzman ölçekte yer alan soruları EK. 2 de yer alan dereceli puanlama anahtarına göre puanlamışlardır.

Verilerin Analizi

Verilerin analizi alt problemlere göre sırasıyla dört aşamada gerçekleştirilmiştir. Birinci aşamada ilk alt problem için genellenebilirlik kuramı kapsamında Ö X S X P deseninde G ve K çalışması yapılarak ana ve ortak etkiler için varyans değerleri analizi yapılmıştır. İkinci aşamada ikinci alt problem için genellenebilirlik kuramı kapsamında Ö X (S:P) deseninde G ve K çalışması yapılarak ana ve ortak etkiler için varyans değerleri analizi yapılmıştır. İlk iki aşamada yapılan analizlerde EduG6.1e programından yararlanılmıştır. Üçüncü aşamada, performans görevinden elde edilen puanların Klasik Test kuramında geçerlik ve güvenirlik analizleri yapılmıştır.

Matematiksel muhakeme becerisini belirleme ölçeği verilerine ilişkin olarak, Genellenebilirlik kuramı için iki farklı senaryo kullanmak üzere iki desen tasarlanmıştır. Bu desenlerden birincisi, öğrenci (Ö), soru (S) ve puanlayıcı (P) değişkenleri olmak üzere, öğrencilerin aynı sorular üzerinden puanlayıcıların her biri tarafından puanlandığı Ö X S X P çapraz desendir. İkinci desen ise, her bir puanlayıcının soruların bir kısmını puanlamasıyla oluşan, puanlayıcı ve soru değişkenlerinin yuvalanmış, öğrencilerin ise bu değişkenlerle çaprazlanmış olduğu Ö X (S:P) desenidir.

Çapraz ve yuvalanmış desenin her biri için 7. sınıf düzeyindeki öğrencilerin oluşturduğu 187 kişilik örneklemden alınan veriler kullanılmış ve her bir desen için G ve K çalışmaları yapılmıştır. Karar çalışmalarında, 3 olan puanlayıcı sayısı artırıp azaltılarak G ve Phi katsayılarının değişimine bakılmıştır.

Soru sayısının etkisini araştırmak için 18 olan soru sayısı her iki desende de arttırıp azaltılmış ve kabul edilebilir bir genellenebilirlik katsayısı elde etmek için minimum soru sayısı belirlenmesi hedeflenmiştir.

Araştırmanın Klasik Test kuramı kapsamında ise, Genellenebilirlik kuramından elde edilen güvenirlik katsayıları ile kıyaslamak üzere Cronbach α ve Tabakalanmış α katsayıları hesaplanmıştır. Çaprazlanmış desen ile karşılaştırmak üzere her 3 puanlayıcı için Cronbach α değerleri hesaplanmış iken; yuvalanmış desen için çok faktörlü ölçeklerin güvenirliğini tek bir α katsayısı olarak hesaplayan, Cronbach, Schonemann ve Brennan (1965) tarafından önerilmiş, tabakalanmış alfa katsayısı kullanılmıştır (Aktaran Tan, 2009). Zira G kuramında Ö X (S:P) deseninde her bir puanlayıcı 6'şar soruyu puanlamıştır. Tabakalanmış alfa katsayısı için bu alt testlerin teorik olarak anlamlı ve iyi tanımlanmış olması istenirken; alanyazında yapılan çalışmalarda faktör analizi ile elde edilen faktörlerin hepsinin teorik olarak anlamlı olamayabileceği gösterilmiştir (Güloğlu ve Aydın, 2001). Tabakalanmış alfa katsayısı aşağıdaki formülle hesaplanmaktadır:

$$\text{Strat } \alpha\rho_{xx'} = \frac{\Sigma_1^c \ (1-\alpha\rho_{xjx\,j'}) \ \ \sigma_{xj}^2}{\sigma_{xtop}^2} \qquad \text{(Eşitlik 11)}$$

x_j : Alt test puanı

$\alpha\rho_{xjx\,j'}$: Alt testlere ait Cronbach α değerleri

c: Alt test sayısı

x_{top}: Toplam puan

Son aşamada ise her iki kurama göre elde edilen güvenirlik katsayıları arasındaki farklılığın manidarlığı test edilmiştir. Alanyazında iki güvenirlik katsayısının karşılaştırılmasına yönelik bazı çalışmalarda Fisher'ın z testi gibi korelâsyon karşılaştırma yöntemleri kullanılmıştır (Büyükturan ve Demirtaşlı, 2013; Şahin ve Gülleroğlu, 2013). Fakat Demirtaşlı ve Aybek'in (2014) de belirttiği gibi madde kovaryansına dayalı hesaplamalarla elde edilen güvenirlik katsayılarının karşılaştırılmasında Feldt'in (1969) F istatistiğinin daha uygun olacağı düşünülmüştür. Karşılaştırılmak istenen güvenirlik katsayıları α_1 ve α_2 olmak üzere, Eşitlik 13'teki gibi elde edilen W değeri, F (N-1, N-1) (N: testi alan öğrenci sayısı) serbestlik derecesinden elde edilen tablo değeri ile kıyaslanmıştır.

$$W = \frac{1-\alpha_2}{1-\alpha_{1,}} \qquad \text{(Eşitlik 12)}$$

Feldt (1969), bu formülü iki örneklemin birbirinden bağımsız olduğu durumlar için önerirken; K, ölçme aracındaki madde sayısı olmak üzere; (K-1)(N-1) çarpımının 1000'den büyük olması durumunda kullanılabileceğini öngörmüştür. Dolayısı ile son aşamada KTK ve G kuramından elde edilen Cronbach α, Tabakalanmış α ve G katsayıları aralarındaki farlılıkların manidarlığı Feldt'in yöntemi ile test edilmiştir.

BÖLÜM III

BULGULAR VE YORUMLAR

Bu bölümde araştırma alt problemlerine göre sırasıyla elde dilen bulgular ve yorumlar yer almaktadır.

Birinci Alt Probleme İlişkin Bulgular ve Yorumlar

1. **Öğrenci (ö), matematiksel muhakeme performansını belirleme ölçeğindeki sorular (s) ve puanlayıcı (p) yüzeylerinin çaprazlandığı Ö X S X P deseninin Genellenebilirlik (G) çalışması sonuçlarının;**

    1.1. Kestirilen varyansları ve toplam varyansları açıklama yüzdeleri nelerdir?

İlköğretim yedinci sınıfta öğrenim görmekte olan 187 öğrencinin matematiksel muhakeme performanslarını belirlemeye yönelik ölçeğe vermiş oldukları cevaplar 3 puanlayıcı tarafından puanlanmış ve elde edilen puanlarla Genellenebilirlik kuramı kapsamında Ö X S X P deseni oluşturularak G çalışması yapılmıştır. G çalışması sonucunda her bir değişkenin kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri Tablo 6'da sunulmuştur.

Tablo 6. Ö X S X P Deseni G Çalışması Sonucu Kestirilen Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri

Varyans kaynağı	Serbestlik Derecesi	Kareler Toplamı	Kareler Ortalaması	G Çalışması ile Kestirilen Varyans	Yüzde %
Ö	186	9099,04	48,92	0,79	27,5
S	17	63,05	3,71	-0,002	0,0
P	2	10,43	5,22	0,001	0,0
ÖS	3162	15937,69	5,04	1,52	53,2
ÖP	372	667,87	1,80	0,07	2,6
SP	34	9,49	0,28	-0,001	0,0
ÖSP	6324	3030,21	0,48	0,48	16,7
Toplam	10097	28817,78			100

Ö: öğrenci, P:puanlayıcı, S: soru

Tablo 6'daki varyans bileşenleri ve toplam varyansı açıklama yüzdeleri incelendiğinde, ölçme objesinin yani öğrenci varyans bileşeninin toplam varyansın % 28'ini açıkladığı görülmüştür. Bu varyans bileşeninin en büyük ikinci varyans bileşenine sahip olması ölçekte yer alan soruların, öğrencileri ayırt edebildiğine işaret etmektedir. O halde ölçek ile öğrenciler matematiksel muhakeme performansları açısından birbirlerinden farklılaşmaktadır. Bununla birlikte en yüksek varyans bileşenine sahip olan öğrenci x soru etkileşimi varyansı öğrencilerin sorudan soruya performanslarının büyük ölçüde değiştiğini göstermektedir. O halde soruların güçlüklerinin bireyden bireye farklılık gösterdiği söylenebilir.

Soru varyans bileşeni ile soru x puanlayıcı varyans bileşeninin negatif değere sahip olması şaşırtıcı görünebilir. Çünkü varyans bileşenleri teoride yalnızca pozitif değer ya da sıfır değer almaktadır. Fakat burada belli bir örneklem ve kareler ortalaması üzerinden varyans tahmini yapıldığı için gerçek varyans değerinin sıfıra oldukça yakın olduğu söylenebilir (Cardinet vd., 2010). Buradaki soru varyans bileşeninin değeri de sıfıra oldukça yakındır. EduG6.1e negatif varyans bileşenlerini 0 olarak analize katar, dolayısı ile sonuçlara bir etkisi olmamaktadır. Soru varyans bileşeninin sıfıra yakın bir değerde olması soruların güçlük düzeylerinin farklı olmadığını göstermektedir. Soru x puanlayıcı varyans bileşeni ise puanlayıcıların verdikleri puanların sorudan soruya farklılaşmadığını göstermektedir.

Puanlayıcıya ait varyans bileşeni (0,001) toplam varyansın % 0'ını açıklamaktadır. Bu varyans bileşeninin toplam hata varyansına katkısının olmaması puanlayıcıların birbirleriyle tutarlı puanlar verdiği şeklinde yorumlanır.

Öğrenci x puanlayıcı varyans bileşeni (0,07); toplam hata varyansının % 2,6'sını açıklamaktadır. Bu bileşen puanlayıcıların verdikleri puanların öğrenciden öğrenciye büyük ölçüde değişiklik göstermediğine işaret eder. O halde puanlayıcıların bazı öğrenciler için cömertlik-katılık açısından tutarsızlığının büyük ölçüde olmadığı söylenebilir.

1.2. Puanlayıcı ve soru sayısının arttırılması ve azaltılması sonucunda K çalışmasında kestirilen G ve Phi katsayıları nasıl değişmektedir?

Öğrenci, soru ve puanlayıcı değişkenlerinin tümünün çaprazlandığı Ö X S X P deseninde öğrencilerin ölçme objesi olarak alınıp, puanlayıcı ve soru sayısının arttırılıp azaltılmasıyla oluşturulan senaryolarla K çalışmaları yapılmıştır. Puanlayıcı ve soru sayılarının arttırılıp azaltıldığı her bir senaryo için kestirilen G ve Phi katsayıları, bağıl ve mutlak hata varyansları Tablo 7'de sunulmuştur.

Tablo 7. Ö X S X P Deseninde Puanlayıcı ve Soru Sayılarının Arttırılıp Azaltıldığı Her Bir Senaryo İçin Kestirilen G ve Phi Katsayıları, Bağıl ve Mutlak Hata Varyansları

n_p	n_s	G Katsayısı	Phi Katsayısı	Bağıl Hata Varyansı	Mutlak Hata Varyansı
1	15	**0,792**	**0,792**	**0,206**	**0,208**
1	18	0,810	0,810	0,184	0,185
1	21	0,824	0,823	0,168	0,170
1	24	0,834	0,834	0,156	0,158
1	27	0,843	0,842	0,147	0,149
1	30	0,849	0,848	0,140	0,141
2	15	0,837	0,836	0,154	0,154
2	18	0,854	0,854	0,134	0,135
2	21	0,868	0,867	0,120	0,121
2	24	0,878	0,877	0,110	0,110
2	27	0,886	0,885	0,102	0,102
2	30	0,892	0,892	0,095	0,096
3	15	0,853	0,852	0,136	0,137
3	**18**	**0,870**	**0,870**	**0,118**	**0,118**
3	21	0,883	0,883	0,104	0,105
3	24	0,893	0,893	0,094	0,095
3	27	0,901	0,901	0,087	0,087
3	30	0,908	0,907	0,080	0,081
4	15	0,861	0,860	0,128	0,128
4	18	0,878	0,878	0,109	0,110
4	21	0,891	0,891	0,096	0,097
4	24	0,901	0,901	0,087	0,087
4	27	0,909	0,909	0,079	0,079
4	30	0,915	0,915	0,073	0,073

n_p: puanlayıcı sayısı, n_s: soru sayısı

187 ilköğretim yedinci sınıf öğrencisine uygulanan matematiksel muhakeme performansını belirleme ölçeği, öğrencilerin, puanlayıcıların ve soruların çaprazlanmış olduğu Ö X S X P deseni dâhilinde 3 puanlayıcı tarafından birbirlerinden bağımsız olarak puanlanmıştır. Bu puanlama sonucunda elde edilen puanların G ve Phi katsayıları Tablo 7'de koyu rakamlarla sunulduğu gibi sırası ile 0,87 ve 0,8697 olarak bulunmuştur.

Tablo 7 incelendiğinde, puanlayıcı sayısının 3 olduğu çalışmada, soru sayısının 27 ve üstü olduğu senaryolarda G ve Phi katsayılarının 0,90'ın üzerinde olduğu görülmüştür.

Puanlayıcı sayısı, soru sayısı sabit iken ($\boldsymbol{n_m}$=18) 4'e yükseltildiğinde ise G ve Phi katsayıları sırasıyla 0,8781; 0,8779'a yükselmiştir. Puanlayıcı sayısının 1 ve 2'ye düştüğü senaryolarda ise G ve Phi parametreleri sırasıyla 0,8106; 0,8097 ile 0,8544; 0,8539'a düşmüştür. Dolayısı ile puanlayıcı sayısının artıp azalmasından etkilenen hata varyans değerlerine bağlı olarak güvenirlik parametrelerinin artıp azaldığı söylenebilir. Soru sayısının sabit olduğu bu senaryolarda puanlayıcı sayısının 1 olduğu durumda bile güvenirlik parametrelerinin 0,80'in

üzerinde olduğu, dolayısı ile kabul edilebilir düzeyde olduğu görülmüştür. Sadece puanlayıcı sayısının 1 ve soru sayısının 15 olduğu senaryoda, G ve Phi parametrelerinin 0,80'in altında olduğu görülmüştür.

Puanlayıcı sayısının sabit ($\boldsymbol{n_p}$=3) olduğu, soru sayılarının 3'er arttırıldığı senaryolarda G ve Phi katsayılarının sırasıyla 0,8830, 0,8827; 0,8930, 0,8927; 0,9010, 0,9006 ve 0,9071; 0,9075'e yükseldikleri görülmüştür. Puanlayıcı sayısının 3 ve soru sayısının 15 olduğu durumda ise G ve Phi katsayıları sırası ile 0,8525; 0,8521' e düşmüştür. O halde soru sayısındaki azalışa bağlı olarak mutlak ve bağıl hata varyansları yükselmiş ve dolayısı ile güvenirlik parametreleri düşmüştür denilebilir. Soru sayısının 15 olduğu bu senaryoda dahi güvenirlik parametreleri kabul edilebilir düzeydedir (Cardinet vd., 2010).

Tablo 8. Ö X S X P Deseninde Puanlayıcı Sayısının Sabit Olduğu ve Soru Sayısının Birer Arttırılıp Azaltıldığı Her Bir Senaryo İçin Kestirilen G ve Phi Katsayıları, Bağıl ve Mutlak Hata Varyansları

	Seçenek 1 Düzey	Seçenek 2 Düzey	Seçenek 3 Düzey	Seçenek 3 Düzey	Seçenek 4 Düzey
Öğrenci	187	187	**187**	187	187
Soru	16	17	**18**	19	20
Puanlayıcı	3	3	**3**	3	3
Gözlem	8976	9537	**10098**	10659	11220
G Katsayısı	0,859	0,865	**0,870**	0,875	0,879
Phi Katsayısı	0,859	0,864	**0,870**	0,874	0,879
Bağıl Hata Varyansı	0,129	0,123	**0,118**	0,113	0,108
Ölçmenin Standart Hatası	0,360	0,351	**0,343**	0,336	0,329
Mutlak Hata Varyansı	0,130	0,124	**0,118**	0,113	0,109
Ölçmenin Standart Hatası	0,360	0,352	**0,344**	0,336	0,330

(Tablo 8'de koyu rakamlarla belirtilen değerler araştırmada kullanılan senaryoya ait değerlerdir.)

Tablo 8'de puanlayıcı sayısının sabit olduğu ve soru sayısının birer arttırılıp azaltıldığı durumdaki G ve Phi katsayıları görülmektedir. Puanlayıcı sayısının sabit ve soru sayısının bir azaltıldığı durumda G ve Phi katsayıları 0,005 miktar azalmıştır. Puanlayıcı sayısının; soru sayısı sabit iken bir azaltıldığı durumda ise güvenirlik parametrelerinin 0,02'şer azaldığı görülmektedir. Soru ve puanlayıcı sayılarının birer arttırıldığı senaryolarda ise puanlayıcılar için güvenirlik indeksleri 0,008; sorular için 0,005 miktar artmıştır. Bu durumda puanlayıcı

varyans bileşeninin soru varyans bileşenine göre küçük bir farklada olsa G ve Phi katsayılarını arttırmaya yönelik daha fazla katkı sağladığı söylenebilir. G çalışması sonuçlarından anlaşılacağı üzere; iki varyans bileşeninin de toplam varyansa katkısının olmaması aradaki farkın önemsenmeyecek kadar küçük olduğunun göstergesidir.

Tablo 7 ve Tablo 8 genel olarak incelendiğinde, soru ve puanlayıcı sayılarının artması ile bağıl ve mutlak hata varyanslarının küçüldüğü, buna bağlı olarak G ve Phi katsayı değerlerinin büyüdüğü; soru ve puanlayıcı sayılarının azalması ile ise hata varyanslarının arttığı ve G ve Phi katsayılarının küçüldüğü söylenebilir.

İkinci Alt Probleme İlişkin Bulgular ve Yorumlar

2. **Soru (s) ve puanlayıcı (p) yüzeylerinin yuvalanmış, öğrenci (ö) yüzeyinin ise çaprazlanmış olduğu Ö X (S:P) deseninin Genellenebilirlik (G) çalışması sonuçlarının;**

2.1. Kestirilen varyansları ve toplam varyansları açıklama yüzdeleri nelerdir?

Araştırmanın ikinci alt problemi için ilk desende kullanılan 187 ilköğretim yedinci sınıf öğrencisinin verileri kullanılmıştır. Ö X (S:P) deseni için tüm öğrencilerden alınan verilerin 6'şar sorusu 3 puanlayıcıya randum olarak atanmış ve soruların puanlayıcılara yuvalandığı, öğrencilerin ise tüm değişkenlik yüzeyleri ile çaprazlandığı desen oluşturulmuştur. Öğrenciler bu desen için de ölçmenin objesi durumundadır. Ö X (S:P) deseni dâhilinde yapılan G çalışması sonucunda her bir değişkenin kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri Tablo 9 da sunulmuştur.

Tablo 9. Ö X (S:P) Deseni G Çalışması Sonucunda Her Bir Değişkenin Kestirilen Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri

Varyans kaynağı	Serbestlik Derecesi	Kareler Toplamı	Kareler Ortalaması	G Çalışması ile Kestirilen Varyans	%
Ö	186	3187,669	17,138	**0,895**	28,3
S:P	15	875,598	58,373	0,302	9,5
P	2	44,378	22,189	-0,031	0,0
ÖS:P	2790	5500,235	1,971	1,971	62,2
ÖP	372	382,288	1,028	-0,157	0,0
Toplam	3365	9990.16904			100

Ö: öğrenci, P:puanlayıcı, S: soru

Tablo 9 incelendiğinde, öğrenciye ait varyans bileşeninin toplam varyansın yaklaşık % 28'ini açıkladığı görülmüştür. Ölçmenin objesi olan öğrencilere ait varyans; bireylerin birbirinden ne derece sistematik bir şekilde ayrıldığını gösterdiği için bu varyans bileşeninin yüksek olması istenen bir durumdur. Bu durumda öğrenci varyans bileşeninin en büyük ikinci varyans bileşenine sahip olması yuvalanmış desen için de, öğrencilerin birbirlerinden matematiksel muhakeme becerileri açısından farklılaştığını yani ölçeğin öğrencileri ayırt etmede başarılı olduğunu göstermektedir.

Puanlayıcı ve öğrenci x puanlayıcı varyans bileşenlerine ait değerlerin toplam hata varyansına katkıda bulunmadığı görülmektedir. Bu varyans bileşenlerinin negatif ya da sıfıra oldukça yakın değerler alması ya da sıfır olarak kabul edilmesi puanlayıcı varyans bileşeni için, puanlamalar arasında bir değişkenlik olmadığının göstergesidir. Başka bir deyişle puanlayıcıların puanlamalarındaki tutarlılık mükemmeldir (Güler vd., 2012, s. 76). Öğrenci x puanlayıcı varyans bileşeni için ise puanlayıcıların bazı öğrencilere karşı cömertlik-katılık anlamında bir tutarsızlık yapmadığının göstergesidir.

Soru-puanlayıcı varyans bileşeni toplam hata varyansının % 10'unu açıklamaktadır. Bu varyans bileşeni soruların puanlayıcılara yuvalanmış etkisini bir başka deyişle her bir puanlayıcının puanladığı sorulardan alınan puanların farklılığını göstermektedir.

En yüksek varyans bileşenine sahip değişkenlik kaynağı ise öğrenci-soru-puanlayıcı etkileşiminden kaynaklanan ya da tesadüfi hataların varlığını gösteren artık varyanstır. Bu varyans bileşeni toplam varyansın % 62,2 sini açıklamaktadır. Artık varyansın yüksek olması istenmeyen bir sonuçtur. Bu durumda çalışmada ölçülemeyen değişkenlik kaynakları ya da öğrenci-soru-puanlayıcı etkileşimiyle ortaya çıkan etkilerin varlığı söz konusu olabilir.

2.2. Puanlayıcı ve soru sayısının arttırılıp azaltılması sonucunda K çalışmasında kestirilen G ve Phi katsayıları nasıl değişmektedir?

18 sorudan oluşan matematiksel muhakeme performansını belirleme ölçeği Ö X (S:P) deseni altında 187 ilköğretim yedinci sınıf öğrencisine uygulanmış ve öğrenci cevapları 3 puanlayıcı tarafından bağımsız bir şekilde puanlanmıştır. Bu desen dâhilinde elde edilen puanların G ve Phi katsayıları sırasıyla 0,89 ve 0,88 bulunmuştur. Tablo 10'da puanlayıcı sayılarının birer arttırılıp azaltıldığı; puanlayıcıların içine yuvalanmış soru sayılarının ise; her bir durum için birer arttırılıp azaltıldığı senaryolar için hesaplanmış G ve Phi katsayıları yer almaktadır.

Tablo 10. Ö X (S:P) Deseninde Puanlayıcı ve Soru Sayısının Arttırılıp Azaltıldığı Her Bir Senaryo İçin Kestirilen G ve Phi Katsayıları, Bağıl ve Mutlak Hata Varyansları

n_p	n_s: P	G Katsayısı	Phi Katsayısı	Bağıl Hata Varyansı	Mutlak Hata Varyansı
2	5	0,819	0,797	0,197	0,227
2	6	0,845	0,825	0,164	0,189
2	7	0,864	0,846	0,141	0,162
2	8	0,879	0,863	0,123	0,142
2	9	0,891	0,876	0,110	0,126
2	10	0,90	0,887	0,099	0,114
3	5	0,872	0,855	0,131	0,152
3	**6**	**0,891**	**0,87**	**0,110**	**0,126**
3	7	0,905	0,892	0,094	0,108
3	8	0,916	0,904	0,082	0,095
3	9	0,925	0,914	0,073	0,084
3	10	0,932	0,922	0,066	0,076
4	5	0,90	0,887	0,099	0,114
4	6	0,916	0,904	0,082	0,095
4	7	0,927	0,917	0,070	0,081
4	8	0,936	0,926	0,061	0,071
4	9	0,942	0,934	0,055	0,063
4	10	0,948	0,940	0,049	0,057

n_p: puanlayıcı sayısı, n_s: soru sayısı

Tablo 10 incelendiğinde puanlayıcı sayısı 3 iken soru sayısı 3 arttırıldığında (n_s= 21), G ve Phi katsayılarının sırasıyla 0,91 ve 0,89 olduğu; soru sayısının 24, 27 ve 30 olduğu durumlarda ise güvenirlik parametrelerinin her ikisinin de 0,90'ın üstünde olduğu görülmektedir.

Puanlayıcı sayısının bir azaltıldığı (n_p=2) ve soru sayısının sabit tutulduğu senaryoda güvenirlik katsayıları hata varyanslarındaki artış ile birlikte azalmıştır. Ancak puanlayıcı sayısı 2 iken soru sayısının 20'ye yükseltildiği durumda (n_s: $P = 10$) güvenirlik parametreleri 0,90'ın üzerinde elde edilmiştir. Bununla birlikte puanlayıcı sayısının bir azaltılmasıyla oluşturulan senaryoların tümü için G ve Phi katsayıları 0,80'in üzerindedir. Dolayısı ile güvenirlikleri kabul edilebilir düzeyde bulunmaktadır.

Puanlayıcı sayısını bir arttırılmasıyla ($\boldsymbol{n_p}$=4) oluşturulan senaryolar Tablo 10'dan incelenirse, bağıl ve mutlak hata varyanslarının oldukça küçüldüğü, G ve Phi katsayılarının arttığı; soru sayısının 20 ($\boldsymbol{n_s}: P = 5$) ve üstü olduğu tüm senaryolarda güvenirlik katsayılarının her ikisinin de 0,90'ın üzerinde olduğu dolayısı ile yüksek güvenirliğe ulaşıldığı görülmektedir.

Puanlayıcı sayısının ($\boldsymbol{n_p}$=3) sabit ve soru sayısının 1 azaldığı durumda (her bir puanlayıcı için birer azaltıldığı, $\boldsymbol{n_s}: P = 5$) mutlak ve bağıl hata varyanslarının arttığı ve buna bağlı olarak G ve Phi katsayılarının azalarak sırasıyla 0,872; 0,855 bulunduğu görülmüştür. Ancak bu durumda bile güvenirlik parametreleri 0,80'in üzerinde olduğundan, kabul edilebilir düzeydedir.

Soru sayısının etkisini belirlemek üzere, puanlayıcı sayısı sabit tutulup, soru sayısı 1 arttırıldığında, G ve Phi katsayılarında sırası ile 0,014; 0,022 artış sağlandığı görülmektedir. Puanlayıcı etkisini belirlemek için ise, soru sayısı sabit tutulup puanlayıcı sayısı 1 arttırıldığında, G ve Phi katsayıları sırası ile 0,025; 0,034 miktar artmıştır. Bu durum puanlayıcı yüzeyindeki artışın soru yüzeyindeki artışa göre az bir farkla daha etkili olduğunu göstermektedir.

Üçüncü Alt Probleme İlişkin Bulgu ve Yorumlar

3. Ölçeğin Ö X S X P ve Ö X (S:P) desenlerinden elde edilen G çalışması parametrelerinin değişimi nasıldır?

Matematiksel muhakeme performansının belirlenmesinde kullanılan ölçek 187 öğrenciye uygulanmış ve 3 puanlayıcı tarafından puanlanmıştır, ardından elde edilen puanlar ile Ö X S X P ve Ö X (S:P) desenleri oluşturularak G analizi yapılmış ve her iki desenden elde edilen varyans bileşenlerinin yüzdeleri, Tablo 11'de sunulmuştur.

Tablo 11. Ö X S X P ve Ö X (S:P) Desenlerinden Elde Edilen G Çalışması Parametreleri

Ö X S X P Deseni				Ö X (S:P) Deseni			
Varyans Kaynağı	Sd	G Çalışması ile Kesitilen Varyans	Yüzde %	Varyans Kaynağı	Sd	G Çalışması ile Kestirilen Varyans	Yüzde %
Öğrenci	186	0,79	27,5	Öğrenci	186	0,90	28,3
Soru	17	-0,002	0	S:P	15	0,30	9,5
Puanlayıcı	2	0	0	Puanlayıcı	2	-0,03	0
ÖS	3162	1,52	53,2	ÖP	372	-0,16	0
ÖP	372	0,07	2,6				
SP	34	-0,002	0				
ÖSP	6324	0,48	16,7	ÖS:P	2790	1,97	62,2
Toplam	10097		100		3365		100

Tablo 11 ile her iki desende yapılan G çalışmaları sonuçları incelendiğinde, çalışmanın ölçme objesi olan öğrenci değişkeni; Ö X S X P deseninde toplam varyansın % 27,5'ini açıklarken Ö X (S:P) deseninde toplam varyansın % 28,3'ünü açıklamaktadır. İki desendeki kestirilen varyans yüzdelerine bakılarak öğrenci ana etkisine ait bileşenin yuvalanmış desende küçük bir miktar fazla olmasıyla birlikte benzer düzeyde oldukları söylenebilir. Bu varyans bileşeni bireylerin matematiksel muhakeme becerilerine yönelik farklılıkları göstermektedir. Dolayısı ile her iki desen için de ölçeğin bireyleri ayırt etme yeterliliğinin benzer olduğu söylenebilir.

Sorulara ait varyans bileşeni incelendiğinde, çaprazlanmış desende toplam hata varyansını açıklama yüzdesinin 0 olduğu başka bir deyişle toplam hata varyansına hiçbir katkısının olmadığı görülmektedir. Yuvalanmış desende sorulara ait varyans bileşeni puanlayıcı değişkeninden bağımsız incelenememektedir. Bu desende sorulara ait varyans bileşeni toplam hata varyansının % 10'unu açıklamaktadır. O halde çaprazlanmış desende soruların zorluk dereceleri açısından bir farklılık bulunmazken; yuvalanmış desende her bir puanlayıcının puanladığı sorulardan alınan puanlar farklılık göstermektedir. Bu durum yuvalanmış desenin etkisini göstermektedir (Güler vd., 2012, s. 76).

Puanlayıcı ana etkisine ait varyans bileşeni her iki desende de 0 bulunmuştur. Bu durumda her iki desen içinde puanlayıcılar arası tutarlılığın mükemmel olduğu söylenebilir.

Öğrenci x puanlayıcı ortak etkisine ait varyans bileşenine bakıldığında, çaprazlanmış desende toplam hata varyansının % 2,6'sını açıkladığı; yuvalanmış desende ise toplam hata varyansına bir katkısının olmadığı görülmüştür. Her iki desen için de puanlayıcıların verdikleri puanların bireyden bireye büyük miktarda farklılık göstermediği söylenebilir. Yani bir puanlayıcının yüksek puan verdiği bireye; diğer puanlayıcı da yüksek vermiştir ya da düşük puan verdiği birey diğer puanlayıcılar tarafından da düşük puanlanmıştır.

Tablo 11 incelendiğinde çaprazlanmış desende yer alan öğrenci x soru ve soru x puanlayıcı ortak etkilerinin sırasıyla % 53,2 ve % 0 olduğu görülmektedir. Bu durum öğrencilerin sorudan soruya performansının büyük ölçüde değiştiğinin göstergesidir. Yani soruların güçlük düzeyleri öğrencilerin geçmiş yaşantıları vb. ye bağlı olarak farklılaşmaktadır. Soru-puanlayıcı ortak etkileşimine bağlı varyansın ise toplam hata varyansına katkısı olmadığı görülmektedir. Bu durum puanlayıcıların verdikleri puanların sorudan soruya farklılık göstermediğine işaret eder.

Artık varyans bileşenleri incelendiğinde, çaprazlanmış desen için toplam hata varyansının %17; yuvalanmış desen için ise % 62'sini açıkladığı görülmüştür. Olabildiğince düşük olması

istenen bu varyans bileşeni, yuvalanmış desende oldukça yüksek bir değerdedir. Bu durumun sebepleri arasında öğrenci x soru varyans bileşeninin yüksek olması yer alabilir. Aynı zamanda soru-puanlayıcı ya da ölçülemeyen tesadüfî hatalar da artık varyans bileşeninin yükselmesine sebep olan durumlardır.

Ö X S X P ve Ö X (S:P) desenlerine ait Genellenebilirlik düzeyleri (G ve Phi katsayıları), Tablo 12'de yer verilmiştir.

Tablo 12. Ö X S X P Ve Ö X (S:P) Desenlerine Ait G ve Phi Katsayıları

	Ö X S X P Deseni	Ö X (S:P) Deseni
G Katsayısı	0,87000	0,89097
Phi Katsayısı	0,86970	0,87640

Ö X S X P ve Ö X (S:P) desenlerine ait G ve Phi katsayıları Tablo 12'den incelendiğinde; çaprazlanmış desene ait G ve Phi katsayıları 0,87 bulunmuştur. Mutlak ölçmeler için hesaplanan Phi katsayısı daha fazla hata varyansı içerdiği için G katsayısından daha düşük bir değer alırken burada G ve Phi katsayılarının nerdeyse eşit olması puanlayıcı, soru ve soru-puanlayıcı varyans bileşenlerinin toplam hata varyansına katkısının 0 olmasından kaynaklanmaktadır.

Yuvalanmış desende ise G ve Phi katsayıları sırası ile 0,89 ve 0,88 bulunmuştur. Her iki desende de mutlak ve bağıl değerlendirmeler için kabul edilebilir güvenirlik parametreleri elde edildiği; bununla birlikte yuvalanmış desende az bir farkla daha yüksek G ve Phi katsayılarına ulaşıldığı görülmektedir. Yuvalanmış etkinin varlığına rağmen daha yüksek güvenirlik katsayıları elde edilmesi, bu desende kestirilen bağıl ve mutlak hata varyanslarının daha küçük olmasından kaynaklanmaktadır. Özellikle bağıl hata varyansı hesaplamasında dikkate alınan öğrenci yüzeyi ile diğer yüzeylerin etkileşimine bağlı hata kaynaklarının (öğrenci x soru, öğrenci-puanlayıcı, öğrenci x soru-puanlayıcı) çaprazlanmış desende toplam hata varyansına katkısı, yuvalanmış desene göre daha fazladır. Bu durum yuvalanmış desendeki G katsayısını çaprazlanmış desendeki G katsayısına göre daha yüksek kılmaktadır. Mutlak hata varyansı hesaplanmasında ise dikkate alınan hata kaynaklarının yüzdesi, yine çaprazlanmış desende küçük bir fark ile daha yüksektir. Bu durum Ö X (S:P) desenindeki Phi katsayısını Ö X S X P desenindeki Phi katsayısına göre yüksek kılmaktadır.

Dördüncü Alt Probleme Ait Bulgular ve Yorumlar

4. **Ölçeğin Ö X S X P ve Ö X (S:P) desenlerinde puanlayıcı ve soru sayılarının arttırılıp azaltılmasıyla yapılan Karar çalışmaları parametrelerinin değişimi nasıldır?**

   4.1. Ö X S X P ve Ö X (S:P) desenlerinde puanlayıcı ve soru sayılarının arttırılıp azaltılmasıyla yapılan Karar çalışmalarında mutlak ve bağıl hata varyanslarının değişimi nasıldır?

Tablo 13. Ö X S X P ve Ö X (S:P) Desenlerinde Puanlayıcı ve Soru Sayılarının Arttırılıp Azaltılmasıyla Yapılan Karar Çalışmalarından Elde Edilen Mutlak ve Bağıl Hata Varyansları

Ö X S X P				Ö X (S:P)			
n_p	n_s	Bağıl Hata Varyansı	Mutlak Hata Varyansı	n_p	n_s:P	Bağıl Hata Varyansı	Mutlak Hata Varyansı
1	15	0,206	0,208	2	5	0,197	0,227
1	18	0,184	0,185	2	6	0,164	0,189
1	21	0,168	0,170	2	7	0,141	0,162
1	24	0,156	0,158	2	8	0,123	0,142
1	27	0,147	0,149	2	9	0,110	0,126
1	30	0,140	0,141	2	10	0,099	0,114
2	15	0,136	0,137	3	5	0,131	0,152
2	18	0,134	0,135	**3**	**6**	**0,110**	**0,126**
2	21	0,120	0,121	3	7	0,094	0,108
2	24	0,110	0,110	3	8	0,082	0,095
2	27	0,102	0,102	3	9	0,073	0,084
2	30	0,095	0,096	3	10	0,066	0,076
3	15	0,136	0,137	4	5	0,099	0,114
3	**18**	**0,118**	**0,118**	4	6	0,082	0,095
3	21	0,104	0,105	4	7	0,070	0,081
3	24	0,094	0,095	4	8	0,061	0,071
3	27	0,087	0,087	4	9	0,055	0,063
3	30	0,080	0,081	4	10	0,049	0,057
4	15	0,128	0,128				
4	18	0,109	0,110				
4	21	0,096	0,097				
4	24	0,087	0,087				
4	27	0,079	0,079				
4	30	0,073	0,073				

n_p: puanlayıcı sayısı, n_s: soru sayısı

Tablo 13 incelendiğinde, bağıl hata varyanslarının yuvalanmış desende, çaprazlanmış desene oranla daha küçük olduğu görülmektedir. Mutlak hata varyanslarının ise birbirine yakın olduğu söylenebilir. Bununla birlikte hata varyanslarının değerlerinin çok yüksek olmadığı, elde edilen G ve Phi katsayılarının değerlerine dayandırılarak söylenebilir.

Ö X S X P deseni incelendiğinde, soru sayısının 18 puanlayıcı sayısının 3 olduğu araştırma senaryosu için mutlak ve bağıl hata varyanslarının 0,118 değerine sahip olduğu görülmektedir. Mutlak hata varyansı öğrenci dışındaki bütün varyans bileşenlerini içerdiği için bağıl hata varyansına oranla daha yüksek olması beklenmektedir. Fakat Ö X S X P deseninde puanlayıcı, soru ve soru x puanlayıcı varyans bileşenleri 0 bulunmuş, bu durumun bir sonucu olarak da bağıl ve mutlak hata varyansları eşit bulunmuştur.

Yuvalanmış desendeki bağıl ve mutlak hata varyansları soruların her 3 puanlayıcı içinde yuvalandığı araştırma durumu için sırasıyla 0,110 ve 0,126'dır.

Ö X S X P deseni için Tablo 13 incelenirse, soru sayısı sabit iken puanlayıcı sayısının azaltılmasının bağıl ve mutlak hata varyanslarını arttırdığı görülmüştür. Ancak bu artış kabul edilebilir düzeydedir. Puanlayıcı sayısının 1 olduğu durumda dahi bağıl ve mutlak hata varyansları çok yüksek olmayıp buna bağlı olarak kabul edilebilir güvenirlik parametreleri vermektedir. Puanlayıcı sayısının bir arttığı senaryoda ise bağıl ve mutlak hata varyanslarının sırası ile 0,009; 0,008 azaldığı görülmektedir. Ö X (S:P) deseni için de aynı durum söz konusudur. Puanlayıcı sayısı 1 arttırıldığında bağıl ve mutlak hata varyansı değerleri sırası ile 0,028 ve 0,031 miktar azalmıştır. Bu durum puanlayıcı sayısındaki artışın yuvalanmış desen için hata varyanslarını küçültmede daha etkili olduğunu göstermektedir.

Soru sayılarının değişimine bağlı olarak mutlak ve bağıl hata varyanslarının değişimi incelendiğinde, çaprazlanmış desen için soru sayısı 3 azaltıldığında sırasıyla bağıl ve mutlak hata varyansları 0,018 ve 0,019; yuvalanmış desen için ise 0,021 ve 0,026 artmıştır. Ancak bu durumda bile hata varyanslarının değerleri güvenirlik parametrelerini yetersiz kılmaya yetecek kadar yüksek değildir. Her iki desende de hata varyansları soru sayısının azalmasına bağlı olarak artmıştır. Soru sayısının 3 arttırıldığı senaryoda ise hata varyanslarının azaldığı görülmektedir.

Puanlayıcı sayısının 4 ve soru sayısının 24 ve üstü olduğu senaryolarda her iki desen içinde oldukça küçük hata varyansları elde edilmiş bu durum yüksek güvenirlik katsayılarının elde edilmesini sağlamıştır.

4.2.Ö X S X P ve Ö X (S:P) desenlerinde puanlayıcı ve soru sayısının arttırılıp azaltılmasıyla yapılan Karar çalışmalarında elde edilen G ve Phi katsayılarının değişimi nasıldır?

Tablo 14'te Ö X S X P ve Ö X (S:P) desenlerinde puanlayıcı ve soru sayısının arttırılıp azaltılmasıyla yapılan Karar çalışmalarından elde edilen G ve Phi katsayıları bulunmaktadır.

Tablo 14. Ö X S X P ve Ö X (S:P) Desenlerinde Puanlayıcı ve Soru Sayısının Arttırılıp Azaltılmasıyla Yapılan Karar Çalışmalarında Elde Edilen G ve Phi Katsayıları

Ö X S P Deseni				Ö X (S:P) Deseni			
n_p	n_s	G Katsayısı	Phi Katsayısı	n_p	n_s:P	G Katsayısı	Phi Katsayısı
2	15	0,8370	0,8360	2	5	0,8190	0,7970
2	18	0,8544	0,8539	2	6	0,8450	0,8250
2	21	0,8675	0,8670	2	7	0,8640	0,8460
2	24	0,8776	0,8771	2	8	0,8790	0,8630
2	27	0,8857	0,8851	2	9	0,8910	0,8760
2	30	0,8922	0,8917	2	10	0,9000	0,8870
3	15	0,8525	0,8521	3	5	0,8720	0,8550
3	**18**	**0,8700**	**0,8697**	**3**	**6**	**0,8910**	**0,8700**
3	21	0,8830	0,8827	3	7	0,9050	0,8920
3	24	0,8930	0,8927	3	8	0,9160	0,9040
3	27	0,9010	0,9006	3	9	0,9250	0,9140
3	30	0,9075	0,9071	3	10	0,9320	0,9220
4	15	0,8606	0,8604	4	5	0,9000	0,8870
4	18	0,8781	0,8779	4	6	0,9160	0,9040
4	21	0,8910	0,8908	4	7	0,9270	0,9170
4	24	0,9010	0,9007	4	8	0,9360	0,9260
4	27	0,9089	0,9086	4	9	0,9420	0,9340
4	30	0,9153	0,9150	4	10	0,9480	0,9400

$\boldsymbol{n_p}$: puanlayıcı sayısı, $\boldsymbol{n_s}$: soru sayısı

Ö X S X P deseni için Tablo 14 incelendiğinde, soru sayısının sabit puanlayıcı sayısının azaldığı durumda G ve Phi katsayılarının azaldığı; puanlayıcı sayısının arttığı durumda ise bu katsayıların arttığı görülmüştür. Aynı durum Ö X (S:P) deseni için de geçerlidir. Bununla birlikte G ve Phi katsayıları yuvalanmış desende çaprazlanmış desene göre daha yüksektir.

Çaprazlanmış desen için puanlayıcı sayısının 4, soru sayısının 24 ve üstü olduğu; puanlayıcı sayısının 3, soru sayısının 27 olduğu senaryolarda güvenirlik parametreleri 0,90'ın üzerindedir. Dolayısı ile bu senaryolarda yüksek güvenirlik elde edilmiştir.

Tablo 14 yuvalanmış desen için incelendiğinde, puanlayıcı sayısının 3, soru sayısının 21 ve üstü (n_s:P= 7, 8, 9, 10); puanlayıcı sayısının 4, soru sayısının 20 ve üstü olduğu senaryolarda (her bir puanlayıcıya 5 ve üstü soru düştüğü durumda) G ve Phi parametrelerinin 0,90'ın üzerinde olduğu görülmüştür. Ayrıca her iki desendeki bütün senaryolarda göreli ölçmeler için hesaplanan G katsayısı, mutlak ölçmeler için hesaplanan Phi katsayısından daha yüksek değerlere sahiptir. Bu durum bir önceki alt problemden de anlaşılacağı üzere mutlak hata varyans değerlerinin bağıl hata varyans değerlerinden daha

fazla varyans bileşenine sahip olması ve bundan kaynaklı olarak daha büyük değerler almasından kaynaklanmaktadır.

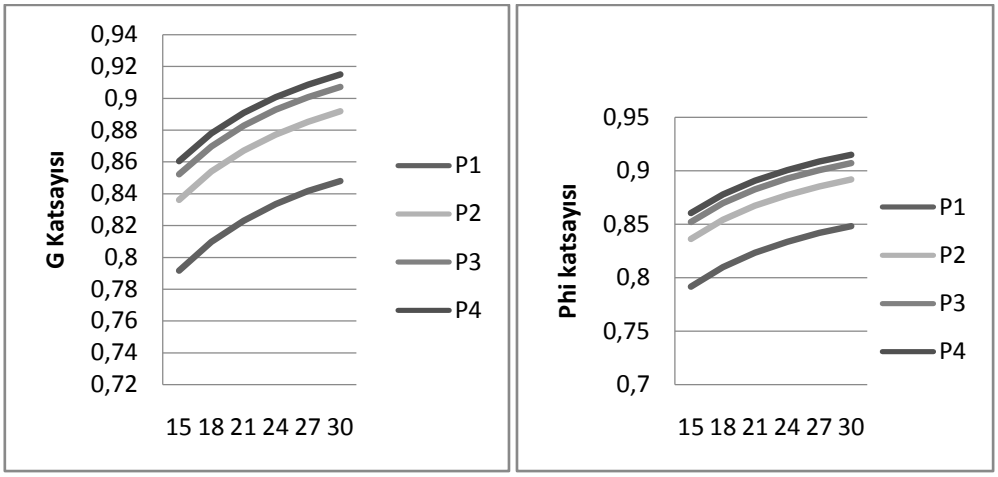
Beşinci Alt Probleme İlişkin Bulgular ve Yorumlar

5. Matematiksel muhakeme performansının belirlenmesinde kullanılan ölçek için her iki desende (Ö X S X P ve Ö X (S:P)) kabul edilebilir bir düzeyde genellenebilirlik katsayısı elde etmek için gerekli minimum soru ve puanlayıcı sayısı nedir?

5.1. Ö X S X P deseninde kabul edilebilir bir düzeyde genellenebilirlik katsayısı elde etmek için gerekli minimum soru ve puanlayıcı sayısı nedir?

Ö X S X P deseni için soru sayısının 18 ve puanlayıcı sayısının 3 olduğu araştırmada G ve Phi katsayıları sırasıyla 0,87 ile 0,8697 bulunmuştu. Cardinet ve diğerlerine göre G ve Phi katsayılarının en az 0,80 olması ölçek puanlarının güvenirliği için kabul edilebilir düzeydir. Bu durumda araştırma için tasarlanan senaryoda elde edilen güvenirlik parametreleri minimum güvenirlik düzeyinin üzerindedir.

Alanyazında yer alan bazı çalışmalarda ise G ve Phi katsayılarının 0,90 ve üzerinde olmasının yüksek güvenirliğe işaret ettiği yer almaktadır (Güler, 2012; Nalbantoğlu, 2009). Bu durumda araştırmacı soru ya da puanlayıcı yüzeylerinden hangisini arttıracağına karar vermelidir. Bu kararı vermesinde Şekil 4; hangi yüzeyin güvenirlik parametrelerini daha fazla arttırdığını koşul sayılarıyla birlikte vermesi açısından yararlı olacaktır. Bununla birlikte bazı durumlarda araştırmacı sadece bir yüzeyin sayısal niceliklerini değiştirebilme imkânına sahip olabilir. Dolayısı ile bahsedilen bütün durumlar için uygun olan senaryolar hakkında bilgi verilmesi en uygun kararı verebilmek açısından yararlı olacaktır.

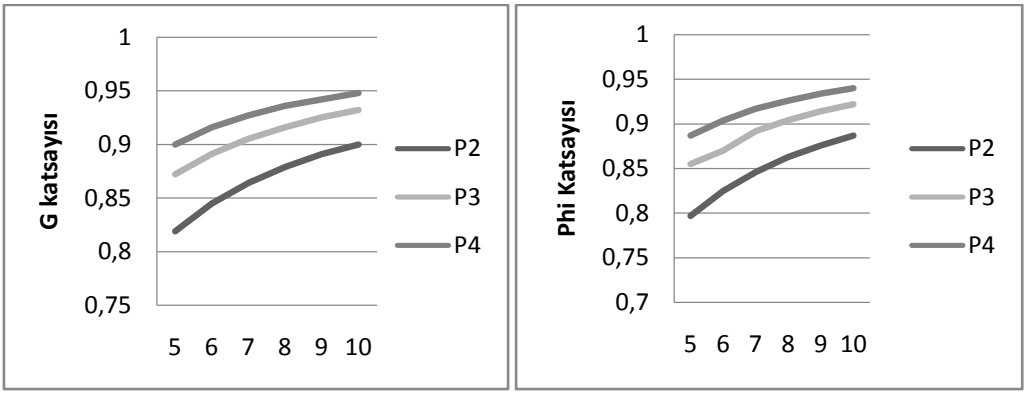


Şekil 3. Ö X S X P deseninde puanlayıcı ve soru sayılarının değişimine göre G ve Phi katsayılarının değişimi

Ö X S X P deseni için Şekil 3 incelendiğinde puanlayıcı sayısını 1 den 2 ye yükseltmenin G ve Phi katsayılarında önemli bir artışa sebep olduğu görülmektedir. Aynı düzeyde etki puanlayıcı sayısı 2'den 3'e; özellikle 3'ten 4'e çıkartıldığında görülmemektedir. Soru sayısındaki artış ise her aralıkta neredeyse aynı oranda G ve Phi katsayılarının artmasına sebep olmuştur. Bu durumda araştırma deseninde güvenirlik parametrelerini 0,90'a çıkarmak için puanlayıcı sayısının 3 olduğu senaryoda soru sayısını 27'ye yükseltmek makul görülmektedir.

5.2. Ö X (S:P) deseninde kabul edilebilir bir düzeyde genellenebilirlik katsayısı elde etmek için gerekli minimum soru ve puanlayıcı sayısı nedir?

Ö X (S:P) araştırma deseninde hesaplanan G ve Phi katsayıları sırasıyla 0,891; 0,87 bulunmuştur. Bu katsayıları 0,90'a yükseltmek için puanlayıcı ve soru sayılarının değişimine bağlı olarak değişen G ve Phi katsayılarının grafiğini incelemek faydalı olacaktır.



Şekil 4. Ö X (S:P) deseninde puanlayıcı ve soru sayılarının değişimine göre G ve Phi katsayılarının değişimi

Ö X (S:P) deseni için Şekil 4 incelendiğinde puanlayıcı sayısını 2'den 3'e yükseltmek G ve Phi katsayılarında etkili bir artışa sebep olmaktadır. Puanlayıcı sayısını 3'ten 4'e yükseltmek ise, aynı etkiyi oluşturmamakla birlikte güvenirlik parametrelerinde bir miktar artışı sağlamaktadır. Soru sayısındaki artışın ise 5-8 aralığında güvenirlik parametrelerine daha fazla katkı sağladığı görülmektedir. O halde Ö X (S:P) deseninde G ve Phi katsayılarının 0,90'a ulaşmasında makul görünen senaryolardan birisi puanlayıcı sayısı 3 iken soru sayısının 8'e yükseltildiği ($\boldsymbol{n_s}$:P=8, $\boldsymbol{n_s}$ =24) senaryodur. Böylece G ve Phi katsayıları sırası ile 0,916; 0,904'e yükselir. Soru yüzeyinin koşullarında bir artırmaya gidilememesi durumunda ise, puanlayıcı sayısının 4'e yükseltilmesi yine güvenirlik parametrelerinin 0,90 ve üzerinde olması için uygun bir senaryo olacaktır.

Altıncı Alt Probleme İlişkin Bulgular ve Yorumlar

6. **Klasik test kuramına göre; aşamalı puanlama anahtarı ile puanlanan matematiksel muhakeme performansı belirleme ölçeğinden elde edilen puanların Cronbach Alfa ve Tabakalanmış Alfa güvenirlik katsayıları nelerdir?**

Aşamalı puanlama anahtarı ile 3 puanlayıcı tarafından puanlanarak elde edilmiş puanların Klasik test kuramına göre güvenirlik analizi yapılırken;

i. Genellenebilirlik kuramında, öğrencilerin bütün sorulara cevap verdiği ve bütün puanlayıcılar tarafından puanlandığı Ö X S X P desenindeki G katsayısı ile karşılaştırılmak üzere, her bir puanlayıcı için Cronbach alfa katsayısı hesaplanmıştır.

ii. Öğrencilerin bütün sorulara cevap verdiği, bütün puanlayıcılar tarafından puanlandığı ve soruların puanlayıcılara yuvalandığı Ö X (S:P) desenindeki G katsayısı ile karşılaştırılmak üzere, her bir puanlayıcının ayrı sorulara vermiş oldukları puanlardan hesaplanan güvenirlik katsayılarını göz önüne alan, test bütünü için tek bir güvenirlik katsayısı oluşturan Tabakalanmış alfa katsayısı kullanılmıştır.

iii. Genellenebilirlik kuramındaki puanlayıcı değişkenlik kaynağına ait varyans bileşeni ile kıyaslanmak üzere puanlayıcılar arasındaki uyumun ölçüsü olan sınıf içi ilişki katsayısı hesaplanmıştır.

i. Tablo 15'te her 3 puanlayıcı için ayrı ayrı hesaplanmış Cronbach alfa katsayıları ve puanlar arasındaki ilişki derecelerini gösteren korelâsyon katsayıları yer almaktadır.

Tablo 15. Puanlayıcıların Puanları Arasındaki Korelâsyon Katsayıları ve Cronbach Alfa Değerleri

	1. Puanlayıcı	2. Puanlayıcı	3. Puanlayıcı
1. Puanlayıcı	-	0,930*	0,880*
2. Puanlayıcı		-	0,860*
Cronbach Alfa	0,903	0,904	0,896
*p<0,01			

Tablo 15'te her bir puanlayıcı için ayrı olarak hesaplanmış, Cronbach Alfa katsayıları bulunmaktadır. Hesaplanan değerler incelendiğinde, her üç puanlayıcı içinde Klasik test kuramı kapsamında yüksek düzeyde güvenirlik elde edildiği söylenebilir. Puanlayıcıların 18 soru için vermiş oldukları puanlar arasındaki ilişki katsayıları ise Tablo 15'te görüldüğü gibi 0,86 ve 0,93 arasında değişmektedir. Elde edilen korelâsyon katsayıları, puanların 0,01 düzeyde anlamlı ve beraber farklılaşma derecesinin yüksek olduğunu göstermektedir. En yüksek ilişkinin birinci ve ikinci puanlayıcı arasında olduğu görülmektedir.

ii. Tabloda ölçeğin bütününe ait tabakalanmış alfa katsayısını elde edebilmek için üç ayrı alt teste (6 soru) ait toplam, varyans ve Cronbach alfa değerleri sunulmuştur.

Tablo 16. Alt testlere ait varyans ve Cronbach Alfa değerleri

Puanlayıcıların Puanladığı Sorular	Alt testlere ait varyans ($\sigma_{x_j}^2$)	Alt testlere ait Cronbach α katsayıları
P1 (1, 2, 3, 4, 5, 6)	49,62	0,82
P2 (7, 8, 9, 10, 11, 12)	57,30	0,83
P3 (13, 14, 15, 16, 17, 18)	46	0,80
Toplam (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18)	308,48	

Tabakalanmış alfa katsayısı = 0, 91

Tablo 16'ya göre Tabakalanmış Alfa katsayısı 0,91 bulunmuştur. Bu değer toplam ölçek puanlarından elde edilen güvenirliği ifade etmektedir. Bu durumda her 3 puanlayıcının ölçeğin ayrı altı sorusunu puanlanması ile elde edilen toplam puanların güvenirliği oldukça yüksektir.

iii. Genellenebilirlik kuramındaki puanlayıcılara ait varyans bileşeni ile kıyaslanmak üzere puanlayıcılara arasındaki tutarlılık KTK' daki uyum indekslerinden biri olan sınıf içi ilişki katsayısı ile ölçülmüştür. Puanlayıcılar arasındaki uyumu test etmek için ikiden fazla puanlayıcı ve sürekli ölçek tipinin olduğu ölçme durumlarında hesaplanabilen sınıf içi ilişki katsayısı Tablo 17'de sunulmuştur.

Tablo 17. Farklı Puanlayıcıların Analitik Dereceli Puanlama Anahtarı ile Aynı Kişileri Puanlamaları Sonucu Elde Edilen Tutarlılık Katsayıları

ICC(2,1)	Tek Ölçüm	186	372	0,892	25,659*
	Ortalama Ölçüm	186	372	0,961	25,659*

p<0,01 (Sd: Serbestlik derecesi, ICC: Sınıf içi ilişki katsayısı)

Tablo 17'de görüldüğü gibi, 3 puanlayıcının analitik dereceli puanlama anahtarı ile verdikleri puanlar arasındaki uyum; mutlak anlaşma ve tutarlılık anlamında incelenmiştir. Hesaplanan değerlere göre, puanlar arasında manidar düzeyde yüksek bir ilişki olduğu görülmektedir (Tek ölçüm/mutlak anlaşma için hesaplanan ICC: 0,892; Ortalama ölçüm/tutarlılık için hesaplanan ICC: 0,961). Bu takdirde Klasik Test Kuramına göre puanlayıcıların puanlamaları arasında yüksek düzeyde tutarlılık olduğu söylenebilir.

Yedinci Alt Probleme İlişkin Bulgular ve Yorumlar

7. Matematiksel muhakeme performansının belirlenmesinde kullanılan ölçekten elde edilen puanların Genellenebilirlik ve Klasik Test Kuramına dayalı güvenirlik katsayıları arasında manidar farklılık var mıdır?

7.1. Ö X S X P deseninden elde edilen puanların güvenirlik katsayıları arasında manidar farklılık var mıdır?

Araştırmada analitik dereceli puanlama anahtarı ile 3 puanlayıcı tarafından puanlanarak elde edilen verilere, Klasik Test Kuramı ve Genellenebilirlik Kuramı'nın Ö X S X P desenine göre güvenirlik analizleri yapılmış ve elde edilen istatistikler Tablo 18'de sunulmuştur.

Tablo 18. Analitik Dereceli Puanlama Anahtarıyla Elde Edilen Puanların KTK ve G Kuramı (Ö X S X P deseni) Güvenirlik Analizi Sonuçları

Klasik Test Kuramı	Genellenebilirlik Kuramı Ö X S X P Deseni
1. Puanlayıcı İçin Cr α = 0,903 2. Puanlayıcı İçin Cr α =0,904 3. Puanlayıcı İçin Cr α=0,896 Cr α= 0,896-0,904	G=0,87 ϕ =0,8697 Öğrenci Yüzeyine Ait Varyans Bileşeni=0,79 Toplam Varyansı Açıklama Yüzdesi=27,5
Pearson Momentler Çarpımı Korelâsyon katsayıları: r_{12}=0,93 r_{13}=0,88 r_{23}=0,86 r= 0,86-0,93 Puanlayıcılar Arasındaki Uyum Katsayıları ICC= 0,89-0,96	Puanlayıcı Yüzeyine Ait Varyans Bileşeni=0,0012 Hata Varyansını Açıklama Yüzdesi=0

Tablo 18 incelendiğinde KTK'da her bir puanlayıcının öğrenci cevaplarına verdikleri puanlara ait ayrı ayrı hesaplanan Cronbach Alfa değerlerinin 0,896 ile 0,930 arasında değiştiği görülmektedir. G kuramında ise her üç puanlayıcının vermiş olduğu puanların aynı anda değerlendirilmesiyle elde edilen G ve Phi katsayıları 0,8700; 0,8697 bulunmuştur. KTK'ya göre elde edilen Cronbach Alfa katsayılarının göreceli olarak daha yüksek olduğu söylenebilir. Bağıl ölçmeler için kestirilen G ve Cronbach alfa katsayılarının aralarındaki bu farkın manidarlığını test etmek üzere Feldt'in (1969) geliştirdiği yöntem kullanılmıştır.

Tablo 19. Ö X S X P deseninden elde edilen G ve Cronbach Alfa Katsayılarının Karşılaştırılmasına Yönelik F testi Sonuçları

			N	W
P1	G katsayısı	0,870	187	1,34*
	Cronbach α	0,903	187	
P2	G katsayısı	0,870	187	1,35*
	Cronbach α	0,904	187	
P3	G katsayısı	0,870	187	1,25
	Cronbach α	0,896	187	

*p<0,05 (P1: birinci puanlayıcı; P2: ikinci puanlayıcı; P3: üçüncü puanlayıcı)

Tablo 19'da KTK'da her bir puanlayıcı için hesaplanmış Cronbach α değerleri ve G analizi ile kestirilen G katsayıları görülmektedir. Birinci ve ikinci puanlayıcının puanlarından elde edilen Cronbach alfa katsayıları G katsayısından anlamlı derecede farklı bulunurken (W_1, $W_2 > F_{187,187}$); üçüncü puanlayıcı için hesaplanan Cronbach alfa katsayısı için aynı durum ($W3 < F_{187,187}$) gözlenmemiştir. Bu durumda KTK'da birinci ve ikinci puanlayıcılar ile elde edilen ölçümlerin güvenirlik katsayıları, G kuramından daha yüksek bulunmuştur. Üçüncü puanlayıcının ölçümleri ile kestirilen güvenirliğin ise G kuramı ile eşdeğer olduğu görülmüştür.

G kuramındaki puanlayıcı varyans bileşenine karşılık, her üç puanlayıcının birbirinden bağımsız olarak vermiş oldukları puanlar arasındaki uyum, KTK da sınıf içi ilişki katsayısı ile incelenmiştir ve mutlak anlaşma indeksi 0,89; tutarlılık indeksi ise 0,96 bulunmuştur. Ayrıca her 3 puanlayıcının puanları arasındaki korelâsyon katsayılarının da 0,86-0,93 arasında değiştiği görülmüştür. Buna karşılık G kuramında puanlayıcı değişkenlik kaynağına ait kestirilen varyans bileşeni 0,0012 gibi oldukça küçük bir değer bulunmuştur ki, bu yüzeyin toplam hata varyansına katkısı bulunmamaktadır. Puanlayıcıya ait kestirilen varyans bileşeninin çok küçük bulunmasına rağmen KTK, Tablo 18'de görüldüğü gibi diğer hata bileşenlerine dair bir kestirimde bulunamamıştır. Bu sebeple alanyazındaki pek çok araştırmada olduğu gibi Genellenebilirlik kuramının hata varyans kaynakları hakkında daha detaylı bilgi verdiği söylenebilir (Güler, 2008; Deliceoğlu, 2009; Öztürk, 2011; Büyükkıdık, 2012).

7.2.Ö X (S:P) deseninden elde edilen puanların güvenirlik katsayıları arasında manidar farklılık var mıdır?

Araştırmada analitik dereceli puanlama anahtarı ile 3 puanlayıcı tarafından puanlanarak elde edilen verilere, Klasik Test Kuramı ve Genellenebilirlik Kuramı'nın Ö X (S:P) desenine göre güvenirlik analizleri yapılmış ve elde edilen istatistikler Tablo 20'de sunulmuştur.

Tablo 20. Analitik Dereceli Puanlama Anahtarıyla Elde Edilen Puanların KTK ve G Kuramı (Ö X (S:P) deseni) Güvenirlik Analizi Sonuçları

Klasik Test Kuramı	Genellenebilirlik Kuramı Ö X (S:P) Deseni
Tabakalanmış alfa katsayısı= 0,91	G=0,890 ϕ =0,876 Öğrenci Yüzeyine Ait Varyans Bileşeni=0,90 Toplam Varyansı Açıklama Yüzdesi=28,3
	Puanlayıcı yüzeyine ait varyans bileşeni=-0,03 Hata Varyansını Açıklama Yüzdesi=0

Tablo 20 incelendiğinde, her üç puanlayıcının 6 ayrı soruyu puanlamasıyla elde edilen puanların KTK' da tabakalanmış alfa güvenirliği 0,91 bulunmuşken; G Kuramında Ö X (S:P) deseni için G ve Phi katsayıları sırası ile 0,890; 0,876 bulunmuştur. Her iki kurama göre elde edilen güvenirlik indekslerinin birbirine yakın sonuçlar verdiği görülmekle birlikte, aralarındaki farklılıkların manidarlığını test etmek üzere Feldt'in (1969) yöntemi kullanılmıştır.

Tablo 21. Ö X (S:P) deseninden elde edilen G ve Cronbach alfa Katsayılarının Karşılaştırılmasına Yönelik F testi Sonuçları

		N	W
G katsayısı	0,89	187	1,22
Tabakalanmış alfa katsayısı	0,91	187	

*p<0,05

Tablo 21'de görüldüğü gibi, tabakalanmış alfa katsayısı ile yuvalanmış desende kestirilen G katsayısı arasında 0,05 düzeyinde manidar bir farklılık ($W<F_{186,186}$) bulunmamıştır. Bu

durumda her iki kuramdan elde edilen güvenirlik katsayılarının da birbirine denk olduğu söylenebilir.

Ayrıca G Kuramında puanlayıcı değişkenlik yüzeyine ait varyans bileşeni 0'a oldukça yakındır ve bu varyans bileşeninin toplam hata varyansına katkısı %0'dır.

BÖLÜM IV

SONUÇ, TARTIŞMA VE ÖNERİLER

Bu bölümde araştırmada sunulan bulgulara yönelik sırası ile sonuçlara yer verilmiştir. Sonuçlar tartışılarak, araştırmanın alana ve uygulamaya yönelik katkıları sunulmuş, araştırmacılara ve öğretmenlere yönelik öneriler ile bölüm sonlandırılmıştır.

Araştırmada yedinci sınıfta öğrenim gören 187 ortaokul öğrencisinin matematiksel muhakeme becerilerini ölçmek üzere uygulanan ölçeğin, 3 bağımsız puanlayıcı ile puanlanmasından elde edilen puanları güvenirliği KTK ve G kuramı ile incelenmiş ve her bir bulguya ait elde edilen sonuçlar aşağıda sırasıyla sunulmuştur:

Birinci Alt Probleme İlişkin Elde Edilen Sonuçlar ve Tartışma

Ö X S X P deseni ile yapılan G çalışması sonucunda çalışmanın objesi olan öğrenci değişkenlik kaynağına ait kestirilen varyans, en büyük ikinci varyans bileşeni bulunmuştur. Bu durum öğrencilerin matematiksel muhakeme performansları açısından birbirlerinden farklılaştığı sonucunu göstermektedir. Evren puanına katkı sağlayan öğrenci varyans bileşeninin hata varyansına katkısının olmaması; bu varyans bileşeninin değerinin olabildiğince yüksek kestirilmesini öngörür (Güler, 2012, s. 69). Dolayısı ile araştırmada ölçme objesi durumunda olan öğrencilere ait varyans payının toplam varyans içinde ikinci sırayı alması ölçme aracının öğrencileri iyi ayırt edebildiğini göstermiştir.

Kestirilen en büyük hata varyans bileşeni ise öğrenci x soru değişkenlik kaynağına ait çıkmıştır. Bu durum, çaprazlanmış desende öğrencilerin sorudan soruya performanslarının önemli ölçüde değiştiğini göstermiştir (Brennan, 2011). Bir başka deyişle soruların zorluk düzeyleri öğrenciler arasında farklılaşmaktadır.

Puanlayıcılara ait kestirilen varyans bileşeninin hata varyansına katkısının olmadığı bulunmuştur. Bu değişkenlik kaynağının hata varyansına katkısının hiç olmaması üç bağımsız puanlayıcının puanlamaları arasında yüksek derecede tutarlılık olduğunu

göstermiştir. Uygulamada puanlayıcılar arası mükemmel uyum zor görünse de alanyazında aşamalı puanlama anahtarı kullanılarak yapılmış puanlamalar arasında G kuramı kapsamında puanlayıcıya ait varyans bileşeninin 0 ya da 0'a oldukça yakın olduğu çalışmalar mevcuttur (Gelbal ve Güler, 2010, Büyükkıdık, 2012; Alkan, 2013).

Öğrenci x soru x puanlayıcı etkileşiminden kaynaklanan varyans bileşeni ise toplam hata varyansının % 16,7'sini açıklamıştır. Düşük çıkması istenen artık hata varyansının bu değeri, çalışmada tesadüfî yollarla oluşan ya da öğrenci x soru x puanlayıcı etkileşiminden kaynaklı tanımlanamayan hata bileşeninin olduğunu göstermiştir.

Shavelson ve Webb (1991), G kuramında kestirilen G ve Phi katsayılarının 0,80 ve üzerinde iken kabul edilebilir güvenirlik düzeyinde ve anlamlı olduğunu ifade etmiştir. Ö X S X P çapraz deseninde kestirilen G ve Phi katsayıları sırası ile 0,87 ve 0,8697'dir. Dolayısı ile bu desen için güvenirliğin sağlandığı görülmüştür. Puanlayıcı ve soru sayılarının artırılıp azaltılmasıyla oluşturulan senaryolarla yapılan Karar çalışmaları sonucunda;

Puanlayıcı sayısı sabit iken, soru sayısının 27'ye yükseltildiği senaryoda G ve Phi katsayılarının 0,90 üzerinde kestirildiği görülmüştür. Puanlayıcı sayısı, soru sayısı sabit iken, bir arttırıldığında ise G ve Phi katsayıları 0,88'e yükselmiştir. Puanlayıcı sayısının bire düşürüldüğü senaryoda hata varyansları artsa da güvenirlik katsayıları 0,80'in üzerinde dolayısı ile kabul edilebilir düzeyde kestirilmiştir. Genel olarak puanlayıcı ve soru sayılarının artışı ile hata varyanslarının azalmasına bağlı olarak bağıl ve mutlak kararlar için kestirilen G ve Phi katsayılarının arttığı; yüzeylerin koşul sayılarının azalması ile de bu katsayıların azaldığı görülmüştür.

Ayrıca çalışmada iki yüzeyin güvenirlik katsayılarına katkısını birebir kıyaslamak için soru ve puanlayıcı sayılarının sırası ile birer arttırılıp azaltıldığı senaryolar üretilmiş ve puanlayıcı yüzeyinin soru yüzeyine göre küçük bir farkla, G ve Phi katsayılarını arttırmaya yönelik daha fazla katkı sağladığı sonucuna ulaşılmıştır. Fakat G analizi sonucunda puanlayıcı ve soru değişkenlik kaynaklarına ait kestirilen varyansın 0 olması, bu farkın önemsenmeyecek düzeyde küçük olmasını göstermiştir.

İkinci Alt Probleme İlişkin Elde Edilen Sonuçlar ve Tartışma

Ö X (S:P) deseni ile yapılan G çalışması sonucunda, ölçme objesi durumundaki öğrenciler için kestirilen varyans bileşeni %28,3 olup ikinci büyük varyans bileşenidir. Yüksek

olması istenen bu varyans bileşeni, yuvalanmış desen için de, ölçeğin öğrencileri matematiksel muhakeme becerisi açısından ayırt edebilmekte başarılı olduğunu göstermiştir.

Puanlayıcı varyans bileşenin bu desen için de, toplam hata varyansına katkısının olmadığı görülmüştür. Puanlayıcıların puanlamaları arasındaki uyum mükemmeldir. Ö X S X P deseninden farklı olarak Ö X (S:P) deseninde sorulara ait varyans bileşeni, puanlayıcı yüzeyine yuvalanmış olarak kestirilmektedir. Yuvalanmış etkiyi gösteren bu varyans bileşeninin hata varyansına katkısı % 10 bulunmuştur.

Kestirilen en yüksek varyans bileşeni, öğrenci-soru-puanlayıcı ortak etkisine ait varyans bileşeni bulunmuştur. Bu durum çalışmada ölçülemeyen, tesadüfî ya da bilinmeyen hata kaynaklarının olabileceğini göstermektedir.

Ö X (S:P) deseninde kestirilen G ve Phi katsayıları sırası ile 0,89 ve 0,87'dir. Puanlayıcı ve soru sayılarının artırılıp azaltılmasıyla oluşturulan senaryolarla yapılan Karar çalışmaları sonucunda;

Puanlayıcı sayısı sabit iken, soru sayısının 24 olduğu durumda G ve Phi katsayılarının 0,90 üzerinde olduğu görülmüştür. Soru sayısının birer azaltıldığı durumda ise bağıl ve mutlak hata varyanslarının artışına bağlı olarak güvenirlik katsayıları küçülmüş olmakla birlikte 0,80'in üzerinde kabul edilebilir düzeydedir (Shavelson ve Webb, 1991).

Puanlayıcı sayısının bir arttırıldığı durumda soru sayısı 20 iken, güvenirlik katsayıları 0,90'ın üzerinde kestirilmiştir. Genel olarak puanlayıcı ve soru sayılarının artışı ile hata varyanslarının azalmasına bağlı olarak bağıl ve mutlak kararlar için kestirilen G ve Phi katsayılarının arttığını; yüzeylerin koşul sayılarının azalması ile de bu katsayıların azaldığı görülmüştür. Ayrıca bu desen için de, puanlayıcı yüzeyindeki artışın soru yüzeyindeki artışa göre küçük bir farkla daha etkili olduğu görülmüştür.

Üçüncü Alt Probleme İlişkin Elde Edilen Sonuçlar ve Tartışma

Ö X S X P ve Ö X (S:P) desenlerinden G çalışmasıyla kestirilen varyans ve toplam varyansı açıklama oranları karşılaştırıldığında;

Öğrenci ana etkisine ait varyans bileşeni iki desende de yüksek ve yuvalanmış desende küçük bir oranla fazla olmakla birlikte hemen hemen aynı bulunmuştur. Ölçme objesi durumunda olan bu değişkenlik kaynağının yüksek olması iki desen için de öğrencilerin

matematiksel muhakeme performansları açısından farklılaştığını gösterir. Dolayısı ile her iki desende de, matematiksel muhakeme performansının belirlenmesine yönelik kullanılan ölçek; öğrencileri ayırt edebilmede eşdeğerdir.

Puanlayıcı ana etkisine ait kestirilen varyans bileşeni her iki desende de 0'dır. Bu değer her iki desen için de puanlayıcılar arasında uyumsuzluğun olmadığını, tutarlılığın oldukça yüksek derecede olduğunu göstermiştir.

Ö X S X P deseninde sorulara ait varyans bileşeni 0 iken; Ö X (S:P) deseninde puanlayıcılardan bağımsız olarak kestirilemeyen bu yüzeyin hata varyansına katkısı % 9,5 bulunmuştur. Bu durum, çaprazlanmış desende soruların güçlük düzeyleri arasında bir farklılık bulunmadığına, yuvalanmış desende ise her bir puanlayıcının puanladığı sorulardan alınan puanların farklılık gösterdiğine işaret etmektedir (Güler, 2012).

Öğrenci x puanlayıcı ortak etkisine ait varyans bileşeninin, yuvalanmış desende toplam hata varyansına bir katkısı yok iken; çaprazlanmış desen de toplam hata varyansını açıklama oranı % 2,6 gibi küçük bir değer bulunmuştur. Dolayısı ile her iki desen için de puanlayıcıların verdikleri puanların bireyden bireye çok fazla farklılık göstemediği sonucuna ulaşılmıştır.

Ö X S X P deseni, yuvalanmış desenden farklı olarak, öğrenci x soru, soru x puanlayıcı varyans bileşenlerini kestirmeye de imkan vermiştir. Zira öğrenci x soru etkileşiminden kaynaklanan varyans bileşeni, toplam varyansı en yüksek oranda açıklayan hata kaynağı olarak bulunmuştur. Bu durum öğrencilerin sorudan soruya performanslarının büyük ölçüde değiştiğini göstermiştir (Brennan 2011; Güler, 2012). Soru x puanlayıcı varyans bileşeninin ise hata varyansına bir katkısı bulunmamıştır. Böylece puanlayıcıların sorudan soruya verdikleri puanlar arasında farklılık bulunmadığı sonucuna ulaşılmıştır.

Artık varyans bileşenine ait kestirilen değerler yuvalanmış desende çaprazlanmış desene göre daha yüksek bulunmuştur. Küçük olması istenen bu değerin çaprazlanmış desenden yüksek olmasının, öğrenci x soru etkileşiminin de artık varyansın içinde yer almasından kaynaklanabileceği düşünülmüştür.

Sonuç olarak, kesitirilen varyans bileşenleri birbirlerine paralel olmakla birlikte yuvalanmış desende kestirilen artık hata varyansı daha yüksek bulunmuştur. Ayrıca çaprazlanmış desenin, öğrenci x soru, soru x puanlayıcı varyans bileşenlerini ayrı olarak kestirmeye olanak sağladığı görülmüştür (Shavelson ve Webb,1991).

G çalışmaları sonucunda Ö X S X P ve Ö X (S:P) desenlerinde kestirilen G ve Phi katsayıları ise birbirine oldukça yakın bulunmuştur. Çaprazlanmış desende kestirilen G ve Phi katsayılarını eşit olması, puanlayıcı, soru ve soru x puanlayıcı değişkenlik yüzeylerinin hata varyansına katkısının olmamasından kaynaklanmıştır. Sonuç olarak, soruların puanlayıcılara yuvalandığı Ö X (S:P) deseninde güvenirlik katsayılarının bir miktar daha yüksek kestirildiği görülmüştür. Alanyazında Yılmaz; Gelbal (2011) ve Alkan (2013), Genellenebilirlik kuramında farklı desenleri kıyasladıkları çalışmalarında, yuvalanmış etkinin olmasına rağmen yuvalanmış desende daha yüksek güvenirlik indekslerine ulaşmışlardır.

Dördüncü Alt Probleme İlişkin Elde Edilen Sonuçlar ve Tartışma

Ö X S X P ve Ö X (S:P) desenlerinde puanlayıcı ve soru sayılarının arttırılıp azaltılmasıyla yapılan Karar çalışmalarında:

Soru sayısının sabit tutulduğu ve puanlayıcı sayısının arttırıldığı senaryolarda her iki desende de bağıl ve mutlak hata varyanslarının azaldığı; puanlayıcı sayısının azaltıldığı senaryolarda ise hata varyanslarının arttığı gözlenmiştir. Zira puanlayıcı sayısı hata varyansı hesabının paydasında yer aldığı için, artışa bağlı olarak doğrudan azalacaktır. Yuvalanmış desende, puanlayıcı sayısının değişimi, çaprazlanmış desene göre hata varyanslarında daha fazla etki yaratmıştır.

Puanlayıcı sayısının sabit tutulup soru sayısının arttırıldığı senaryolarda her iki desende de hata varyansları azalmış; soru sayısının azaltıldığı senaryolarda ise hata varyansları artmıştır.

Genel olarak bağıl hata varyansı kestirimlerinin yuvalanmış desende daha küçük çıkma eğiliminde olduğu buna bağlı olarak G katsayılarının daha yüksek değerler aldığı sonucuna varılmıştır. Mutlak hata varyansı değerleri her iki desen içinde hemen hemen birbirine yakın bulunmuştur. Ayrıca her iki desen için kestirilen G ve Phi katsayıları değerlerinin kabul edilebilir düzeyde olması, mutlak ve bağıl hata varyanslarının çok yüksek derecede olmadığını göstermiştir.

Beşinci Alt Probleme İlişkin Elde Edilen Sonuçlar ve Tartışma

G çalışmaları sonucu Ö X S X P ve Ö X (S:P) desenlerinde kestirilen G ve Phi katsayıları 0,87; 0,87 ve 0,89; 0,88'dir. Cardinet vd. (2010)'e göre G ve Phi katsayılarının en az 0,80 olması ölçek puanlarının güvenirliği için kabul edilebilir düzeydir. Alanyazında yer alan bazı çalışmalarda ise yüksek güvenirlik için bu katsayıların 0,90 ve üzerinde olması gerektiği yer almaktadır (Güler, 2012; Nalbantoğlu, 2009). Dolayısı ile çalışmada güvenirliği 0,90 ve üzerine çıkarmak amacı ile geçekleştirilen Karar çalışmaları sonucunda;

Ö X S X P deseni için en ideal senaryo puanlayıcı sayısı üç iken soru sayısının 27'ye yükseltilmesi ile elde edilmiştir.

Ö X (S:P) deseni için en makul senaryo ise puanlayıcı sayısı üç iken soru sayısının 24'e yükseltildiği senaryo olmuştur. Ayrıca soru yüzeyinin koşullarında bir arttırmaya gidilememesi durumunda, puanlayıcı sayısının dörde yükseltilmesi yine güvenirlik kestirimlerinin yüksek düzeyde olması için uygun bir senaryo olacaktır.

Sonuç olarak, yuvalanmış desende yüksek düzeyde güvenirlik katsayıları elde edebilmek için gereken senaryodaki yüzey sayılarının, daha ekonomik olduğu görülmüştür. Zira araştırma senaryosu için elde edilen G ve Phi katsayıları yuvalanmış desende daha yüksek olması bu sonucu destekler niteliktedir.

Altıncı Alt Probleme İlişkin Elde Edilen Sonuçlar ve Tartışma

Brennan (2011)'e göre G çalışmasındaki yüzeyin koşul sayısı K çalışmasındakine eşit olduğu zaman, herhangi bir metrikte G katsayısı Cronbach α katsayısına eşit olacaktır. Zira her iki güvenirlik katsayısı da bağıl kararlar için kestirilmekte ve alanyazında pek çok çalışmada karşılaştırılmaktadır (Hoyt ve Melby, 1999; Chafolues, Christ, Riley-Tillman ve Briesch, (2007); Brennan, 2011; Büyükkıdık, 2012). Klasik test kuramına göre her 3 puanlayıcı için ayrı ayrı hesaplanmış Cronbach alfa katsayıları ve puanlar arasındaki ilişki derecelerini gösteren korelâsyon katsayıları sonuçlarına göre;

Cronbach Alfa katsayıları incelendiğinde, her üç puanlayıcı içinde Klasik test kuramı kapsamında yüksek düzeyde güvenirlik elde edildiği görülmüştür.

Ayrıca yuvalanmış desende kestirilen G katsayısı ile kıyaslanmak üzere KTK'da hesaplanan Tabakalanmış alfa katsayısı; soruların puanlayıcılara yuvalanarak oluşturulduğu desenin KTK'ya göre güvenirliğinin yüksek olduğunu göstermiştir.

Analitik puanlama anahtarı ile farklı puanlayıcılar tarafından verilen puanlar arasındaki uyum; KTK'ya göre hesaplanan korelasyon katsayıları ve sınıf içi ilişki katsayısı ile yüksek düzeyde bulunmuştur.

Yedinci Alt Probleme İlişkin Elde Edilen Sonuçlar ve Tartışma

Matematiksel muhakeme performansını belirlenme ölçeğinden elde edilen puanların güvenirlik kestirimlerinin, Genellenebilirlik kuramındaki Ö X S X P deseni ve Klasik Test kuramından elde edilen sonuçları karşılaştırıldığında; birinci ve ikinci puanlayıcılara ait Cronbach α katsayılarının, G katsayısından anlamlı derecede ($p<0,05$) büyük olduğu görülmüştür. Alanyazında KTK ve GK'nın karşılaştırıldığı bazı araştırmalarda, GK'da daha yüksek ya da KTK ile eşdeğer güvenirlik kestirimi yapıldığı görülmüştür (Güler, 2008; Deliceoğlu, 2009; Büyükkıdık, 2012). Bununla birlikte KTK'da çok küçük farkla da olsa, daha yüksek güvenirlik kestirimlerinin yapıldığı çalışmalarda mevcuttur (Güler ve Gelbal, 2010; Yelboğa ve Tavşancıl, 2010). Çalışmaların genelinde güvenirlik katsayıları arasındaki farkın manidarlığı sınanmamış olmakla birlikte, çok küçük farklılıkların eşdeğer kabul edildiği görülmüştür. G katsayısının, Cronbach Alfa katsayısından anlamlı derecede küçük olması diğer hata varyansı kaynaklarının varlığını gösterebilmektedir. Zira KTK tek bir hata kaynağı ile ilgilenirken G kuramında birden çok hata kaynağı aynı anda kestirilir (Brennan, 2011). Üçüncü puanlayıcı için hesaplanan Cronbach α katsayısı ile G katsayısı arasında anlamlı bir farklılık görülmemiştir.

G kuramında puanlayıcı varyans bileşeninin hata varyansına katkısı yok iken; KTK' da puanlayıcılar arasındaki uyumu kestirebilmek için mutlak anlaşma ve tutarlılık anlamında sınıf içi ilişki katsayıları hesaplanmıştır. Puanlayıcılar arasındaki uyum, her iki durumda da yüksek bulunmuştur. Ayrıca her bir puanlayıcının verdiği puanlar arasındaki korelâsyon katsayılarıyla da puanlamalar arasında yüksek derecede ilişki bulunmuştur. KTK' da yapılan bu analizlere göre puan sıralamalarının her üç puanlayıcı için de paralel olduğu görülmüştür. O halde her iki kuram için de puanlayıcıların puanlamaları arasında tutarlılık yüksek olmakla birlikte alanyazında pek çok araştırmada olduğu gibi, GK'ya göre KTK'da puanlayıcıların puanlamaları arasındaki ilişki katsayıları, göreceli olarak biraz daha düşük sonuçlar vermiştir (Deliceoğlu, 2009; Güler ve Gelbal, 2012; Büyükkıdık, 2012).

Sonuç olarak, KTK'da birinci ve ikinci puanlayıcılar için kestirilen güvenirlik katsayıları GK'da kestirilen G katsayılarından anlamlı derecede yüksek olmakla birlikte, KTK'da tek

bir analizle bütün hata varyanslarını dikkate alan ortak bir güvenirlik katsayısı bulunmamaktadır. Zira G kuramında, çalışmadaki bütün değişkenlik kaynaklarının hata varyansına katkısı tek bir analizle hesaplanabilmişken, KTK'da yalnızca puanlayıcı değişkenlik kaynağı ile kıyaslanacak değerler için bile birçok analiz yapılmıştır. Ayrıca Ö X S X P deseninde G çalışması ile kestirilen soru, soru x puanlayıcı, öğrenci x soru x puanlayıcı, öğrenci x puanlayıcı gibi pek çok hata varyansı bileşeninin KTK'da karşılaştırılacak kestirimlerinin olmadığı görülmüştür.

Genellenebilirlik kuramındaki Ö X (S:P) deseni ve Klasik Test kuramından elde edilen sonuçlar karşılaştırıldığında; ölçeğin bütününden elde edilen Tabakalanmış Alfa katsayısı ile G katsayısı arasında manidar düzeyde ($p<0,05$) bir farklılığın olmadığı görülmüştür. Bu durumda her iki kuramdan elde edilen güvenirlik katsayılarının birbiri ile eşdeğer olduğu söylenebilir.

Ö X (S:P) deseninde puanlayıcı değişkenlik yüzeyinin toplam hata varyansına katkısının olmadığı görülmüştür. O halde puanlayıcıların puanlamaları arasında fark yoktur. KTK'nın, yuvalanmış desendeki puanlayıcılara ait varyans bileşenine karşılık bir kestirimi bulunmamaktadır. Bu durum G kuramının KTK'ya göre daha detaylı bilgi verdiğini desteklemektedir. Zira yuvalanmış desende kestirilen diğer hata varyansı bileşenlerinin de KTK'da bir karşılığı olmadığı görülmüştür. Dolayısı ile hata varyansı kaynaklarının birden fazla olduğu çalışmalarda, G kuramının kullanılmasının daha yararlı olacağı düşünülmektedir (Shawelson ve Webb, 1991; Brennan, 2011).

G kuramının ayrıca Karar çalışmaları ile ileriye dönük çalışmalar için uygun senaryolar hakkında bilgi vermesi, KTK'da mümkün olmamakta ve bu yönden de GK; KTK'ya üstünlük sağlamaktadır.

Öneriler

Araştırma Sonuçlarına Yönelik Öneriler

1. Öğrencilerin matematiksel muhakeme yeteneği gibi üst düzey zihinsel becerilerinin performansa dayalı değerlendirilmesinde, objektifliğin ve güvenilir ölçmenin sağlanabilmesi açısından birden fazla puanlayıcı (gözlemci) ve onlara rehber olacak puanlama anahtarları kullanılmalıdır.

2. Geniş çaplı sınavlar için her ne kadar birden fazla puanlayıcı kullanımı mümkün olsa da, küçük gruplar için uygulamada bu mümkün olmamaktadır. Bu durumlarda bütün puanlayıcıların bütün soruları puanlamaları daha fazla çaba ve vakit aldığından soruların ya da öğrencilerin puanlayıcılar ile yuvalandığı bir senaryo ile puanlama yapmak ekonomiklik sağlayacaktır.
3. Üst bilişsel becerilerin ölçümünde çoktan seçmeli testlerin kullanımı yerine öğrencilerin uzun süreli performansları (bir soruya verdiği cevabı yapılandırması, eleştirel düşünmesi vb.) gözlemlenmeli, ya da bilgiyi gerçek yaşam problemlerine aktarabilmeyi sağlayacak açık uçlu sorulardan oluşan ölçme araçları tercih edilmelidir.

İleride Yapılacak Araştırmalara Yönelik Öneriler

1. Çalışmada Genellenebilirlik ve Karar çalışmalarında ölçeğin güvenirliği faktörlerine göre ayrı ayrı incelenmemiştir, ileriki araştırmalarda ölçeğin her bir faktörü için ayrı G ve K çalışması yapılabilir ve elde edilen G katsayıları KTK' da her bir boyut için kestirilen Cronbach Alfa katsayısı ile karşılaştırılabilir.
2. Matematiksel muhakeme becerisi açısından sınıflar arasında bir faklılık olup olmadığı, sınıf değişkenlik kaynağı da eklenerek test edilebilir. Zira ölçeğin beşinci sınıf öğrencileri için geliştirilmiş ve çalışmada yedinci sınıf öğrencilerine uygulanmış olması sınıf değişkeninin incelenmesi gerektiğinin göstergesidir.
3. Genellenebilirlik kuramında yapılan G ve K çalışmalarının ardından çalışmada yüzeylerin G ve Phi katsayılarına etkisinin derecesini görmek üzere, G-yüzey analizi yapılabilir. Böylece, (örneğin) madde yüzeyi için hangi maddenin güvenirliği büyük ölçüde düşürdüğü görülebilir ve ileriki aşamada da G ve Phi katsayılarını yükseltmek için bu maddenin atılmasına karar verilebilir.
4. Ölçek maddelerinin, madde analizi sonucu hesaplanan güçlük düzeyleri ile G çalışması ile kestirilen madde değişkenlik kaynağına ait varyans bileşeni karşılaştırılabilinir.
5. Klasik Test kuramı ve Genellenebilirlik kuramı ile birlikte Madde Tepki kuramındaki güvenirlik indekslerinin de dâhil edildiği ve karşılaştırıldığı bir araştırma yapılabilir.

KAYNAKLAR

Acar, M.,& Anıl, D. (2009). Sınıf öğretmenlerinin performans değerlendirme sürecindeki değerlendirme yöntemlerini kullanabilme yeterlikleri, karşılaştıkları sorunlar ve çözüm önerileri. *TUBAV Bilim Dergisi*,*3*(2), 354-363.

Aktaş, M. (2013). *Aynı performans görevinin farklı sayıda puanlayıcılar tarafından üç farklı teknikle puanlanmasından elde edilen puanların güvenirliklerinin genellenebilirlik kuramına göre incelenmesi.* Yüksek Lisans Tezi, Mersin Üniversitesi Eğitim Bilimleri Enstitüsü, Mersin.

Alkan, M. (2013). *PISA 2009 okuma becerileri açık uçlu sorularının puanlanmasında genellenebilirlik kuramındaki farklı desenlerin karşılaştırılması.* Doktora Tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Altıparmak, K., & Öziş, T. (2005). Matematiksel ispat ve matematiksel muhakemenin gelişimi üzerine bir inceleme. *Ege Eğitim Dergisi*, *6*(1), 25-37.

Anıl, D., & Büyükkıdık, S. (2012). Genellenebilirlik kuramında dört facetli karışık desen kullanımı için örnek bir uygulama. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*,*3*(2*)*, 291-296.

Arterberry, B. J., Martens, M. P., Cadigan, J. M., & Smith, A. E. (2012). Assessing the dependability of drinking motives via generalizability theory. *Measurement and Evaluation in Counseling and Development*, *45*(4*)*, 292-302.

Ateş, C., Öztuna, D., & Genç, Y. (2009). The use of intraclass correlation coefficient (ICC) in medical research: review. *Turkiye Klinikleri Journal of Biostatistics*, *1*(2), 59.

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

Aybek, E. C., & Demirtaşlı, R. N. (2014). A comparison of psychometric properties of a generalability test which administered in paper-pencil and computer based form, *Elementary Education Online, 13*(4), 1400-1413.

Baki, A., Güven, B.,& Karataş, İ. (2002). *Klinik mülakat yöntemi ile problem çözme becerilerinin değerlendirilmesi.* V. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi'nde sunulmuş bildiri, ODTÜ, Ankara.

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

Brennan, R. L. (2001). *Generalizability Theory*. USA: Springer-Verlag New York Inc.

Brennan, R. L. (2011). *Using generalizability theory to address reliability issues for PARCC assessments:* A whitePaper.

Briesch, A. M., Chafolueas S. M., & Tillman T. C. (2010). Generalizability and dependability of behavior assessment methods to estimate academic engagement: a comparison of systematic direct observation and direct behavior rating. *School Psychology Review, 39*(3), 408-421.

Briesch, A. M., Chafolueas S. M., Tillman T. C., & Boice, C. H. (2010). Direct behavior rating (dbr): Generalizabilityanddependabilityacrossratersandobservations. *Educational and Psychological Measurement, 70*(5), 825-843.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: a practical guide to study design, implementation, and interpretation. *Journal of School Psychology, 52*,13–35.

Brown, J. D. (2005). Generalizability and decision studies. *Shicken: JALT Testing & Evaluation SIG Newsletter. 9* (1),12-16.

Büyükkıdık, S. (2012). *Problem çözme becerisinin değerlendirilmesinde puanlayıcılar arası güvenirliğin klasik test kuramı ve genellenebilirlik kuramına göre karşılaştırılması.* Yüksek Lisans Tezi,Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Büyükturan, E. B., & Demirtaşlı, N. Ç. (2013). Çoktan seçmeli testler ile yapılandırılmış gridlerin psikometrik özellikleri bakımından karşılaştırılması, *Journal of Faculty of Educational Sciences, 46*(1), 395-415.

Büyüköztürk, Ş. (2007). Performansa dayalı durum belirleme nedir? *İlköğretmen Dergisi, 8*, 28-32.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. London: Routledge.

Chafolues, S. M., Christ, T. J., Riley-Tillman, T. C., Briesch, A. M., & Chanese, A. M. (2007). Generalizability and dependability of direct behavior ratings to assess social behavior of preschoolers. *School Psychology Review, 36*(1), 63-79.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

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

Çakıcı, D. (2011). *Genellenebilirlik kuramı ve lojistik regresyona dayalı hesaplanan puanlayıcılar arası tutarlığın karşılaştırılması*. Yüksek Lisans Tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Çoban, H. (2010). *Öğretmen adaylarının matematiksel muhakeme becerileri ile biliş ötesi öğrenme stratejilerini kullanma düzeyleri arasındaki ilişki*. Yüksek Lisans Tezi, Gaziosmanpaşa Üniversitesi Sosyal Bilimler Enstitüsü, Tokat.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2010).*Sosyal Bilimler için Çok Değişkenli İstatistik.* Ankara: Pegem.

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

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for twotests. *Psychometrika, 34,* 363–373.

Fraenkel, J. R., & N. E. Wallen. (2003). *How to design and evaluate research in education.* New York: MacGraw-Hill 15 Kasım 2014 tarihinde http://doha.ac.mu/ebooks/Research%20Methods/DesigningAndEvaluatingResearchInEducation.pdf sayfasından erişilmiştir.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Psysical Education and Exercises Science, 5*(1), 13-14.

Guildford, J. P. (1956). *Fundamental statistics in psychology and education*. New York: McGraw-Hill.

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

Güler N., & Gelbal, S. (2010). Açık uçlu matematik sorularının güvenirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi. *Educational sciences: theory & practice*. *10*(2), 989-1019.

Güler, N., Taşdelen, G., Uyanık, K., & Gülden P. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem.

Güloğlu, B., & Aydın, G. (2001). Coopersmith özsaygı envanterinin faktör yapısı. *Eğitim ve Bilim, 26*, 66-71.

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

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability ıs not enough: teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41(2), 56–64.

Hoyt, W. T., & Melby, J. N. (1999). Dependability of measurement in counseling psychology: an introduction to generalizability theory, *The Counseling Psychologist*, 27(3), 325-352.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing, 17*, 123–139.

Kaya, G. (2011). *Genellenebilirlik kuramının doldurma kavram haritası değerlendirme çalışmasına uygulanması,* Yüksek Lisans Tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Kılıç, S. (2009). *Ölçümlerin Uyumluluğu ve Tıptaki Uygulamaları,* Yüksek Lisans tezi, Çukurova Üniversitesi Sağlık Bilimleri Enstitüsü, Adana.

Krulik, S., Rudnick, J. A. (1993). *Reasoning and problem solving. A handbook for elementary school teachers.* Mass: Allyn & Bacon.

Kutlu, Ö., Doğan, C. D., & Karakaya. İ. (2009). *Öğrenci başarısının belirlenmesi.* Ankara: Pegem.

Liu, P. H. (1996). Do teachers need to incorporate the history of mathematics in their teaching?. *The Mathematics Teacher, 96*(6), 416.

Moskal, B. M. (2000). Scoringrubrics: what, when, how? *Practical Assessment Research and Evaluation, 7*(3).

NAEP, (2002). Mathematics framework for the 2003 national assessment of educational progress. Washington, DC: National Assessment Governing Board.

Nalbantoğlu, F. (2009). *Performans ölçümlerinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması*. Yüksek Lisans Tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Özberk, E. H. (2012). *Genellenebilirlik kuramı karar çalışmalarında kullanılan farklı katsayıların karşılaştırılması.* Yüksek Lisans Tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Öztürk, M. E. (2011). *Voleybol becerileri gözlem formu ile elde edilen puanların genellenebilirlik ve klasik test kuramına göre karşılaştırılması*. Yüksek Lisans Tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Peresini, D., Webb, N. (1999). Developing mathematical reasoning in grades. Lee V. Stiff (Ed.), *Analyzing mathematical reasoning in students' responses across multiple performance assessment tasks,* National Council of Teachers of Mathematics, Reston,Virginia.

Pilten, P. (2008). *Üst biliş stratejileri öğretiminin ilköğretim beşinci sınıf öğrencilerinin matematiksel muhakeme becerilerine etkisi.* Doktora Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Roebroeck, M. E., Harlaar J., & Lankhorst G. J. (1993). The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. *PhysicalTherapy, 73*(6), 386-395.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin, 86*(2), 420.

Shavelson, J. R., Webb, N. M., & Rowley, G. (1989). Generalizability theory. *American Psychologist, 44*(6), 922-932.

Shavelson, R. J., & Webb, M. N. (1991). *Generalizability theory: A Prime*. SAGE Publication, Inc., California.

Steen, L.A. (1999) *Twenty questions about mathematical reasoning. Developing mathematical reasoning in grades K-12.* National Council of Teachers of Mathematics, Reston: Virginia.

Şahin, D. B., & Gülleroğlu, H. D. (2013). Likert tipi ölçeklere madde seçmede kullanılan farklı madde analizi teknikleri ile oluşturulan ölçeklerin psikometrik özelliklerinin incelenmesi, *Asian Journal of Instruction, 1*(2),18-28.

Tall, D. (1995). *Cognitive growth in elementary and advenced mathematical thinking*. Proceedings of the International Conference forthe Psychology of Mathematics Education, Brazil: I, 161-175.

Tan, Ş. (2009). Misuses of KR-20 and Cronbach's alpha reliability coefficients. *Education and Science, 34*(152).

Tan, Ş. (2012). *Öğretimde ölçme ve değerlendirme.* Ankara: Pegem.

Tavşancıl, E. (2002). *Tutumların ölçülmesi ve SPSS ile veri analizi*. Ankara: Nobel.

Tekindal, S. (2014). *Okullarda ölçme ve değerlendirme yöntemleri*.Ankara: Nobel.

Tezbaşaran, A. (1996). *Likert tipi ölçek geliştirme kılavuzu.* Ankara: Özyurt.

Tracy, L. & Gibson, B. A. (2005). *Development of an instrument to ases student attitudes toward educational process in an under graduate core curriculum.* Doktora Tezi, Univercity of Arkansas.

Tural, H. (2005). *İlköğretim matematik öğretiminde oyun ve etkinliklerle öğretimin erişi ve tutuma etkisi*. Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi Eğitim Bilimleri Enstitüsü, İzmir.

Turgut, M.F. (1997). *Eğitimde ölçme ve değerlendirme metotları*. Ankara: Gül.

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

Umay, A. (2003). Matematiksel muhakeme yeteneği. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi,24*, 234-243.

Volpe R. J.,&Briesch A. M. (2012). Generalizability and dependability of single-item and multiple item direct behavior rating scales for engagement and disruptive behavior. *School Psychology Review, 41*(3), 246-261.

Yelboğa, A. (2007). *Klasik test kuramı ve genellenebilirlik kuramına göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi.* Doktora Tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Yurdugül, H. (2010). Ölçme kuramı ve güvenirlik katsayıları. 31 Temmuz 2015 tarihinde

http://yunus.hacettepe.edu.tr/~yurdugul/3/indir/Guvenirlik.pdf

sayfasından erişilmiştir.

EKLER

Ek 1. Matematiksel Muhakeme Performansının Belirlenmesinde Kullanılan Ölçek

1-5 sorularda verilen problemleri çözünüz. Altında verilen boşluklara nasıl çözüldüğünüzü, kullandığınız çözüm yolunu neden seçtiğinizi ayrıntılı bir şekilde yazınız.

1. Bir öğrenci bir kitabın önce 7/9'unu, sonra kalanın 1/4' ünü okuyor. Geriye 25 sayfa kaldığına göre kitap kaç sayfadır?
    a. 60
    b. 75
    c. 90
    d. 150

……
……
……
……

2. Saatte 85 km hızla giden bir otomobil gideceği yolun 5/8'inin 2/5'ini gittikten sonra kalan yolu 3 saatte tamamlamıştır. Otomobilin aldığı bütün yol kaç kilometredir?
    a. 340
    b. 350
    c. 400
    d. 480

……
……
……
……
……

3. 40 kişilik bir otobüste çocuk yolcuların sayısı, büyüklerin 4 katıdır. Otobüste kaç tane çocuk yolcu vardır?

   a. 6

   b. 12

   c. 22

   d. 32

……
……
……
……

4. Bir kitapla bir dergi 34 TL ye alınmıştır. Kitabın fiyatı, derginin fiyatının 8 katından 2,5 TL fazladır. Kitabın fiyatı kaç TL'dir?

   a. 3,5

   b. 28

   c. 30,5

   d. 3,15

……
……
……
……

5. Bir bilim adamı yaptığı deneyde kullanmak için 1 litre çözeltiye ihtiyaç duymaktadır. Fakat elinde bulunan derecesiz büyük kabın içerisindeki çözeltiden 1 litre elde etmek için kullanabileceği 3 litre, 5 litre ve 7 litre büyüklüklerinde 3 adet deney tüpü bulunmaktadır. Bilim adamı bu tüpleri kullanarak 1 l çözeltiyi ne şekilde elde eder?

……
……
……
……

6-9 sorularda problemler ve bu problemlere ait çözüm yolları verilmiştir. Belirtilen çözüm yolunun doğru olup olmadığını düşününüz. Yanlış ise hatanın nerede yapıldığını yazınız.

6. Problem:

İki kardeşin 75 cevizi vardır. Büyük kardeş küçüğün 4 katından 10 eksik ceviz alıyor. Küçük kardeş kaç ceviz almıştır?

Çözüm Yolu:

Eksik olan miktar toplam miktardan çıkarılır ve birim sayısına bölünürse küçük kardeşin aldığı ceviz sayısı bulunur.

Küçük kardeş	:	1 birim
+ Büyük kardeş	:	+ 4 birim (-10)
75		5 birim (-10)

75 – 10 = 65

65 : 5 = 13 küçük kardeşin aldığı cevizdir.

a. Doğru

b. Yanlış (Çünkü)

……
……
……
……
……

7. Problem:

Bir maratonu birinci bitiren sporcu 3 saat 47 dakika 51 saniyede, sonuncu bitiren sporcu ise 5 saat 32 dakika 27 saniyede tamamlamıştır. Yarışı birinci bitiren sporcu, sonuncu bitiren sporcudan ne kadar önce bitirmiştir?

Çözüm Yolu:

Yarışı sonuncu bitiren sporcunun bitirme süresinden, yarışı birinci bitirenin süresi çıkartılır. Bunun için aşağıdaki işlemler yapılır.

5 saat, 32 dakika, 26 saniye = 4 saat, 92 dakika, 26 saniye

= 4 saat, 91 dakika, 126 saniye

4 sa, 91 dk, 106 sn

\- 3 sa, 47 dk, 51 sn

1 sa, 44 dk, 54 sn

a. Doğru

b. Yanlış (Çünkü)

……

……

……

……

8. Problem:

0,5 saat kaç dakikadır?

Çözüm Yolu:

0,5 sa = 1/5 sa ve 1 sa = 60 dk olduğundan,

1/5 x 60 dk = 12 dkdır.

a. Doğru

b. Yanlış (Çünkü)

……

9. Problem:

6 işçinin 12 günde yaptığı işi kaç işçi 8 günde bitirebilir?

Çözüm Yolu:

6 işçi ⟶ 12 günde bitirirken

? işçi ⟶ 8 günde bitirir

$$\frac{6 \times 8}{12} = 12$$

a. Doğru

b. Yanlış (Çünkü)

……

……

……

……

10-12 sorularda bir sınıfa sorulan sorular ve öğrencilerin en fazla verdikleri yanlış cevaplar görülmektedir. Sizce öğrenciler neden belirtilen yanlış cevabı vermiş olabilirler?

10. $\frac{12}{13} + \frac{7}{8} =$ işleminin sonucu nedir?

Bir sınıfta bulunan öğrencilerden yukarıdaki soruyu çözemeyenlerin çoğunluğunun vermiş oldukları yanlış cevap 19' dur. Sizce bu öğrenciler tarafından yapılmış olan hata ne olabilir?

……

……

……

……

11. Ahmet 2003 yılında 12 yaşındayken, babası, 1989 yılında 43 yaşındaydı. Ahmet doğduğu zaman babası kaç yaşındaydı?

Bir sınıfta bulunan öğrencilerden yukarıdaki soruyu çözemeyenlerin çoğunluğunun vermiş oldukları yanlış cevap 55'tir. Sizce bu öğrenciler tarafından yapılmış olan hata ne olabilir?

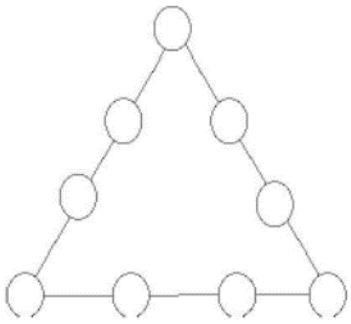
……
……
……
……

12. 4 katı 48 olan sayının $\frac{1}{3}$'ü kaçtır?

Bir sınıfta bulunan öğrencilerden yukarıdaki soruyu çözemeyenlerin çoğunluğunun vermiş oldukları yanlış cevap 36'dır. Sizce bu öğrenciler tarafından yapılmış olan hata ne olabilir?

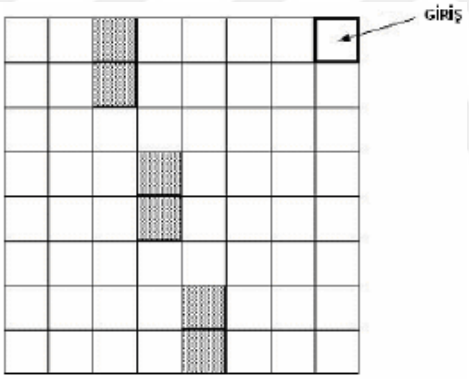
……
……
……
……

13. Aşağıda her bir kenarı üzerinde 4 halka olan bir üçgen verilmiştir. Sizden 1'den 9'a kadar olan rakamları bu halkalara yerleştirmeniz isteniyor. Üçgenin her bir kenarı üzerindeki 4 halkaya yazacağınız rakamların toplamının 20 olması gerektiğini ve 1 den 9'a kadar olan rakamları sadece bir kez kullanabileceğinizi unutmayın.



14. Aşağıdaki labirentte bir yürüyüş yapmanız isteniyor. Yalnız yürüyüş esnasında şu kurallar unutulmamalıdır:

- Yürüyüşe giriş karesinden başlanacak ve yine bu noktada bitirilecektir.
- Açık renkli kareler üzerinde yürünecek, her adımda sadece bir tane kareye basılacaktır.
- Labirentin üzerinde bulunan açık renkli karelerin hepsine bir kez basmak zorunludur. Açık renkli karelerden üzerine basılmayan kalmamalıdır.
- Bir kez üzerine basılan kareye tekrar basılmayacaktır.
- Koyu renkli karelere basılmayacaktır.
- Çapraz adım atmak yasaktır. Sadece yukarı-aşağı ya da sağa-sola adım atılabilir.



15. Aşağıdaki şekilde 4 parçaya ayrılmış durumda bir pasta görülmektedir. Bu pastayı size verilen bıçağı kullanarak ve verilen kurallara uyarak en fazla kaç parçaya bölebilirsiniz.

<u>Kurallar:</u>

Bıçağı üç kez kullanabilirsiniz.

Bıçağı kullanırken elinizi kaldıramazsınız.

Yalnızca düz kesimler yapabilirsiniz.

Pastayı eşit büyüklükte parçalara ayırmak zorunda değilsiniz.

16. 1'den 100'e kadar olan tek doğal sayıların toplamı (1+3+5+7+9+…+99) ile ilgili aşağıda verilen tabloyu inceleyiniz ve bir genellemede bulununuz.

Toplanan Eleman Sayısı (n)	Toplanan Elemanlar	TOPLAM
1	1	1
2	1+3	4
3	1+3+5	9
4	1+3+5+7	16
5	1+3+5+7+9	25
6	1+3+5+7+9+11	36
7	1+3+5+7+9+11+13	49
8	1+3+5+7+9+11+13+15	64
↓	↓	↓
50	1+3+5+7+9+11+13+15+...+99	2550

17. 1'den 100'e kadar olan çift doğal sayıların toplamı (2+4+6+8+10+…+100) ile ilgili aşağıda verilen tabloyu inceleyiniz ve bir genellemede bulununuz.

Toplanan Eleman Sayısı (n)	Toplanan Elemanlar	TOPLAM
1	2	2
2	2+4	6
3	2+4+6	12
4	2+4+6+8	20
5	2+4+6+8+10	30
6	2+4+6+8+10+12	42
7	2+4+6+8+10+12+14	56
8	2+4+6+8+10+12+14+16	72
↓	↓	↓
50	2+4+6+8+10+12+14+16+…+100	2550

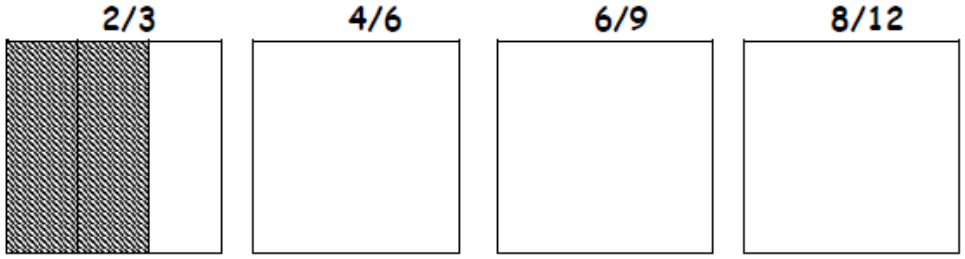
……

……

……

……

18. Verilen kesirlerin değerlerini altında bulunan modeller üzerinde gösteriniz. Verilen kesirlerle oluşturduğunuz modelleri birlikte değerlendiriniz. Kesirlerde denkliği tanımlayınız.



……..

……..

……

Ek 2. Analitik Puanlama Anahtarı

Sorunun İlgili Olduğu Ölçek Boyutu	Puan	Gözlenecek Öğrenci Davranışı
Çözüm Yolu/Sonucun Doğruluğuna Karar Verme	4	Öğrenci çözüm yolu/sonucun doğruluğunu karar vermede uygun kriterleri kullanır. Çözüm yolu/sonucun neden en doğru olduğunu tam olarak açıklar.
	3	Öğrenci çözüm yolu/sonucun doğruluğunu karar vermede doğru kriterleri kullanır. Çözüm yolu/sonucun neden en doğru olduğunu tam olarak açıklamaz.
	2	Öğrenci çözüm yolu/sonucun doğruluğunu karar vermede kullandığı kriter durumla ilgilidir ama en uygun olan değildir veya öğrenci verilen kriterler içerisinde en uygun seçeneği belirleyemez.
	1	Öğrenci karar vermede problem durumu ile ilgili olmayan kriter kullanır.
	0	Öğrenci herhangi bir yargıda bulunmaz.
Rutin Olmayan Problemleri Çözme	4	Öğrenci bir engelin veya zorluğun üstesinden gelmede en etkili çözümün yolunu seçer ve bunun olası çözüm yolları içerisinde neden en etkilisi olduğunu tam olarak açıklar. Öğrencinin vermiş olduğu cevap çözüm sürecini açık veya tam olarak gösterir niteliktedir.
	3	Öğrenci bir engelin veya zorluğun üstesinden gelmede en etkili çözümün yolunu seçer ve bunun olası çözüm yolları içerisinde neden en etkilisi olduğunu tam olarak açıklayamaz. Öğrencinin vermiş olduğu cevap çözüm sürecini gösterir niteliktedir.
	2	Öğrenci bir engelin veya zorluğun üstesinden gelmede doğru bir çözüm yolu seçer ama bu en etkili olan değildir. Öğrencinin vermiş olduğu cevap çözüm sürecini kısmen olsa da gösterir niteliktedir.
	1	Öğrencinin seçmiş olduğu çözüm yolu engelin veya zorluğun üstesinden gelebilecek nitelikte değildir. Öğrencinin vermiş olduğu cevap çözüm sürecini göstermez.
	0	Öğrenci herhangi bir yargıda bulunmaz.

Çözüme İlişkin Mantıklı Tartışmalar Geliştirme	4	Öğrenci düşündüklerini çok iyi ifade eden ayrıntılı tartışmalar geliştirir. Tartışmalarda herhangi bir mantık hatası bulunmamaktadır.
	3	Öğrenci düşündüklerini çok iyi ifade eden ancak ayrıntılı olmayan tartışmalar geliştirir. Tartışmalarda herhangi bir mantık hatası bulunmamaktadır.
	2	Öğrenci çözüme ilişkin tartışmalarda bulunmuştur ama bunlar düşüncelerini çok iyi ifade eder nitelikte ve ayrıntılı değildir. Tartışmalarda bazı mantıksal hatalar bulunmaktadır.
	1	Öğrencinin tartışmaları açık değildir ve çok fazladır. Mantıksal olarak geçersizdir.
	0	Öğrenci herhangi bir yargıda bulunmaz.
Genelleme Yapma	4	Öğrenci geçerli bir genelleme oluşturur ve oluşturduğu genellemenin mantığını açık bir şekilde ifade eder.
	3	Öğrenci geçerli bir genelleme oluşturur fakat oluşturduğu genellemenin mantığını açık bir şekilde tanımlayamaz.
	2	Öğrenci tanımlamış olduğu özellikle ilgili bir takım ilişkiler içeren genelleme oluşturur; özellikler genellemeyi tamamen desteklemeyebilir.
	1	Öğrenci genelleme oluşturamaz veya genelleme tanımlanan özellik tarafından desteklenmemektedir.
	0	Öğrenci herhangi bir yargıda bulunmaz.
Uygun Muhakemeyi Belirleme ve Kullanma	4	Öğrenci doğru cevap vermiş. Geliştirdiği muhakeme tam ve açık ve muhakemeyi doğru kullanmış.
	3	Öğrenci doğru cevap vermiş. Fakat geliştirdiği muhakeme tam ve açık değil.
	2	Öğrenci yanlış cevap vermiş. Fakat doğru muhakemeyi belirlemiş ve kullanma girişiminde bulunmuş fakat tamamlayamamış.
	1	Öğrenci yanlış cevap vermiş, geliştirdiği muhakeme kısmen doğru ve problemin bir bölümünde kullanabilmiş.
	0	Öğrenci herhangi bir yargıda bulunmaz.

Kaynak: Pilten, P. (2008)

Ek 3. Araştırma İzni