KLASİK TEST KURAMI VE GENELLENEBİLİRLİK KURAMINDAN PUANLAYICILAR ARASI TUTARLILIĞIN FARKLI YÖNTEMLERE GÖRE KARŞILAŞTIRILMASI

THE COMPARISON OF INTERRATER RELIABILITY BY USING ESTIMATING TECNIQUES IN CLASSICAL TEST THEORY AND GENERALIZABILITY THEORY

BULUT YILDIZTEKİN

Hacettepe Üniversitesi

Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin

Eğitim Bilimleri Anabilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalı için

Öngördüğü

Yüksek Lisans Tezi

Olarak Hazırlanmıştır.

2014

Eğitim Bilimleri Enstitüsü Müdürlüğü'ne

Bulut YILDIZTEKİN'in hazırladığı "Klasik Test Kuramı ve Genellenebilirlik Kuramından Puanlayıcılar Arası Tutarlılığın Farklı Yöntemlere Göre Karşılaştırılması" başlıklı bu çalışma jürimiz tarafından Eğitim Bilimleri **Anabilim Dalı**, Ölçme ve Değerlendirme **Bilim Dalı**'nda **Yüksek Lisans** olarak kabul edilmiştir.

Başkan	Prof. Dr. Selahattin GELBAL
Üye (Danışman)	Doç. Dr. Duygu ANIL
Üye	Doç. Dr. Şeref TAN
Üye	Doç. Dr. Burcu ATAR
Üye	Yrd. Doç. Dr. Derya ÇOBANOĞLU AKTAN

ONAY

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri üyeleri tarafından 29.05.2014 tarihinde uygun görülmüş ve Enstitü Yönetim Kurulunca 05/06/2014 tarihinde kabul edilmiştir.

Prof. Dr. Bekir AKMAN
Eğitim Bilimleri Enstitüsü Müdürü

KLASİK TEST KURAMI VE GENELLENEBİLİRLİK KURAMINA GÖRE PUANLAYICILAR ARASI TUTARLIĞIN KARŞILAŞTIRILMASI

Bulut YILDIZTEKİN

ÖZ

Bu araştırmada 7. Sınıf matematik öğrencilerine uygulanan ve problem çözme becerisini ölçen açık uçlu sorular, analitik ve bütünsel dereceli iki ayrı puanlama anahtarı kullanılarak 5 farklı matematik öğretmeni tarafından puanlanmıştır. Elde edilen puanların Klasik test kuramı (KTK) ve Genellenebilirlik kuramına (GK) göre güvenirlik kestirimleri yapılmış ve puanlayıcılar arası tutarlık dereceleri belirlenmeye çalışılmıştır. İki kuramdan farklı tekniklerle belirlenen güvenirlik ve tutarlık düzeylerinde farklılaşma olup olmadığı ve kullanılan tekniklerden hangisinin daha fazla bilgi sunduğu belirlenmeye çalışılmıştır.

Araştırma için Ankara ilindeki bir ortaokulun 7. sınıfında öğrenim gören 84 öğrenci seçilmiştir. Bu öğrencilere 2013-2014 Eğitim-Öğretim yılı bahar döneminde, tam ve rasyonel sayılarda problem çözme becerisini ölçen 6 adet açık uçlu sorudan oluşan bir test uygulanmıştır. Elde edilen cevaplar alanında uzman 5 matematik öğretmeni tarafından, analitik ve bütünsel dereceli puanlama anahtarları kullanılarak (ADPA-BDPA) 20-25 gün arayla puanlanmıştır. Araştırmanın verileri elde edildikten sonra Klasik test kuramından, Pearson momentler çarpımı korelasyon katsayısı (PMÇKK), Spearman sıra farkları korelasyon katsayısı (SSFKK), Cronbach Alpha, Kappa ve Krippendorf Alpha katsayıları ile Genellenebilirlik kuramından b x m x p çapraz deseninde değişkenlik kaynakları ve yüzdeleri belirlenerek güvenirlik analizleri yapılmıştır.

Araştırma sonucuna göre, KTK ve GK ‘na göre elde edilen güvenirlik katsayıları birbirine parelel ve oldukça yüksektir. Ancak Kappa istatistiği orta düzeyde uyumu işaret etmektedir. Yine aynı sonuçlarda genellenebilirlik kuramında oluşturulan birey, madde ve puanlayıcı deseninden (b x m x p) elde edilen sonuçlarda da puanlayıcıların iki farklı dereceli puanlama anahtarı kullanarak verdikleri puanlar arasında değişkenliğe etki etmedikleri görülmüştür. Ayrıca puanlayıcılar arası tutarlık düzeyinin yüksek olduğu ve analitik dereceli puanlama anahtarı ile elde

edilen puanların tutarlığının bütünsel dereceli puanlama anahtarı ile elde edilen puanların tutarlığından göreceli olarak daha yüksek olduğu belirlenmiştir..

Anahtar sözcükler: Problem çözme becerisi, klasik test kuramı, genellenebilirlik kuramı,kappa tekniği , krippendorf alfa, güvenirlik, puanlayıcılar arası tutarlık, analitik ve bütünsel dereceli puanlama anahtarları.

Danışman: Doç. Dr. Duygu ANIL, Hacettepe Üniversitesi, Eğitim BilimleriAnabilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalı

THE COMPARISON OF INTERRATER RELIABILITY BY USING ESTIMATING TECHNIQUES IN CLASSICAL TEST THEORY AND GENERALIZABILITY THEORY

Bulut YILDIZTEKİN

ABSTRACT

In this research, analytical and holistic rubrics, which evaluate the problem solving ability of seventh grade students,were used to be scored by five different mathematics teachers. The reliability analysis of obtained scores from raters have been made with respect to Classical test theory and Generalizability theory and interrater agreement level has been examined. Whether there is any difference between reliability coefficents and interrater agreement level obtained from different techniques from two theories and which tecniques used to gather more information was determined.

In this research the convenient sample is used. The sample consists of 84 seventh grade students in Ankara. The data collection instrument includes 6 open-ended questions which measures the problem solving ability about integers and rational numbers in 2013-2014 spring semester. Student's answers were scored with respect to analytical and holistic rubrics by five mathematics teachers in Ankara following twenty- twenty five days. For data analysis, pearson product-moment correlation coefficient (PPMCC), spearman's rank correlation coeeficient (SRCC), cronbach alpha, kappa statistics and krippendorf alpha coefficient in the classical test theory and crossed design b x m x p which examines sources of variation and percentages in the generalizability theory are used to determine reliability and interrater agreement level.

Consequently, it was found that the obtained reliability coefficients with respect to classical test theory and generalizability theory are parallel and relatively high. However, the kappa statistics states middle level agreement. Furthermore, the crossed design (b x m x p) also states that raters does not have effect on variation between scores obtained from them according to two rubrics. Finally, inter rater agreement level between teachers is high and the consistency of the scores obtained from analytical rubric is relatively higher than the consistency of the scores obtained from holistic rubric.

Keywords: Problem solving ability, classical test theory, generalizability theory, kappa statistics, krippendorf alpha coefficients, interrater agreement level, analytical and holistic rubrics.

Advisor: Doç. Dr. Duygu ANIL, Hacettepe University, Department of Educational sciences Division ofEducational Measurement and Assessment

ETİK BEYANNAMESİ

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- Tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- Görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- Başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- Atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- Kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- Bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı beyan ederim.

Bulut YILDIZTEKİN

vii

TEŞEKKÜR

Tez çalışmam boyunca değerli katkı ve yorumlarıyla beni yönlendiren danışmanım Doç. Dr. Duygu ANIL' a teşekkürlerimi sunarım.

Ölçme ve Değerlendirme alanına ilk adım attığım günden itibaren yardım ve desteğini esirgemeyen çok kıymetli hocam Prof. Dr. Selahattin Gelbal' a teşekkürü borç bilirim.

Eleştiri ve önerileriyle çalışmama katkıda bulunan çok değerli jüri üyelerim Doç. Dr. Şeref TAN' a, Doç. Dr. Burcu ATAR' a ve Yrd. Doç. Dr. Derya ÇOBANOĞLU AKTAN' a şükranlarımı sunarım.

Tez çalışmam boyunca bilgi ve birikimini benimle paylaşan Ar. Gör. Levent YAKAR' a, her zaman yanımda hissettiğim Ar. Gör. Sinan YAVUZ' a, görüşleri ve dostluğu ile yanımda olan Ar. Gör. Haydar KARAMAN' a, tüm oda ve mesai arkadaşlarıma (Hepsinin yeri ayrı ☺) ayrı ayrı teşekkürü borç bilirim.

Çalışmam sürecinde geçirdiği rahatsızlıkla tüm ailemizi üzen, ancak kısa sürede iyileşmesiyle bizi de hayata bağlayan kız kardeşime, hep yanımda, yüreğimde olan kardeşlerime, emektar anne ve babama teşekkür ederim.

Tez döneminden önce, tez süresince ve tez döneminden sonra da desteğini yüreğimde hissettiğim, sevgisi, sabrı ve güveniyle hep yanımda olan çok kıymetli eşime ve henüz doğmamış olan fakat varlığıyla beni heyecanlandıran tatlılar tatlısı kızıma teşekkür ederim.

Yüksek lisans eğitimim süresince maddi katkılarıyla çalışmanın ilerlemesine yardımcı olan TÜBİTAK' a teşekkürlerimi sunarım.

Çalışmaya özveri ve ciddiyetle destek veren çok değerli öğrenci, öğretmen ve uzman arkadaşlarıma, ismini sayamadığım dostlarıma ve emeği geçen herkese teşekkür ederim.

İÇİNDEKİLER

ÇİZELGELER DİZİNİ

ŞEKİLLER DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

SPSS: Statistical Package for Social Sciences

KTK: Klasik Test Kuramı

G-: Genellenebilirlik

K Çalışması: Karar Çalışması

GK: Genellenebilirlik Kuramı

ADPA: Analitik Dereceli Puanlama Anahtarı

BDPA: Bütünsel Dereceli Puanlama Anahtarı

PMÇKK: Pearson Momentler Çarpımı Korelasyon Katsayısı

SSFKK: Spearman Sıra Farkları Korelasyon Katsayısı

1. GİRİŞ

Bu bölümde problem durumu, araştırmanın amacı ve önemi üzerinde durulmuştur. Ayrıca problem cümlesi, alt problemler, sayıltılar, sınırlılıklar ve araştırmanın kuramsal temeli hakkında bilgiler sunulmuştur.

Geçmişten günümüze ilerleyen teknoloji ve bilgi çağının getirdiği yenilikler ile ihtiyaç duyulan kişi profilinde de bazı farklılıklar ortaya çıkmaktadır. Elde edilmesi planlanan temel bilgi ve becerinin yanında, analitik, eleştirel ve yaratıcı düşünme becerileri ile problem çözme becerilerine duyulan ihtiyaç artmıştır.

Üst düzey becerilerden olan problem çözme becerilerini ölçmek ve ölçme işlemini gerçekleştirirken elde edilen puanların tutarlığı hakkında sağlıklı yorumlar yapmak her zaman mümkün değildir. Çünkü, üst düzey becerilerin ölçülmesinde ölçüm sonuçlarının bütünsellik ve tutarlık teşkil etmesi oldukça güçtür.

Üst düzey becerileri ölçmek için kullanılan çoktan seçmeli testler, objektifliği ve hızlı puanlamayı olağan kılsa da, her zaman yeterli olamayabilmektedir. Çünkü çoktan seçmeli sorular süreç odaklı değerlendirme yapmak yerine, sonuç odaklı değerlendirme yapılmasını sağlar. Ayrıca çoktan seçmeli testler ile ortaya çıkan şans başarısı faktörü de bu yöntemin sınırlılıklarından biridir. Örneğin; temel eğitimden ortaöğretime geçiş sınavlarında (TEOG) öğrenci kazanımı elde etmemiş olsa bile doğru cevaba ulaşma ihtimali %25' dir.

Çoktan seçmeli sınavların bu sınırlılığını aşmak için sıklıkla açık uçlu sorular kullanılmaktadır. Açık uçlu sorular şans başarısı faktörünü ortadan kaldırsa da nesnel ve güvenilir değerlendirmeler yapılabilmesi, her zaman mümkün olmamaktadır. Yapılan değerlendirmelerin ne kadar güvenilir olduğu ile puanlayıcıdan puanlayıcıya ne kadar tutarlık gösterdiği araştırmanın çıkış noktasını oluşturmuştur.

Bu araştırmada, problem çözme becerisini ölçen çoktan seçmeli soruların alternatifi olarak kullanılan açık uçlu soruların, güvenirliğinin ve puanlayıcılar arası tutarlığının belirlenmesine yönelik çalışma yürütülmüştür. Güvenirliğin ve tutarlığın belirlenmesi için klasik test kuramı ve genellenebilirlik kuramında farklı yöntemlerden yararlanılmıştır.

1.1. Problem Durumu

Mevcut eğitim sisteminde öğrencilerin başarıların ölçülmesi için birçok yöntem kullanılmaktadır. Örneğin yazılı sınavlar, sözlü sınavlar, kısa cevaplı testler, sınıflandırma gerektiren testler, çoktan seçmeli testler, projeler, açık uçlu sorular ve gözlem bunlardan bazılarıdır. Bu yöntemler kullanılarak öğrencilerin başarısı en doğru ve en objektif biçimde belirlenmeye çalışılmaktadır. Hatalardan arınık (güvenilir) ve amacına uygun (geçerli) yöntemler ile öğrencilerin başarısı artırılması ile ölçme ve değerlendirme işlevinin gerçekleştirilmesi hedeflenmektedir.

Kullanılan yöntemlerden biri olan çoktan seçmeli sorular, şans başarısı içermektedir. Şans başarısı; gerekli bilgi sahibi olmadan soruyu doğru yanıtlama ihtimali olarak tanımlanabilir. Örneğin; dört seçenekli bir soruda öğrencinin soruyu okumadan soruyu %25 ihtimalle doğru cevaplama şansı bulunmaktadır. Yapılan araştırmalara göre, şans başarısı geçerliği ve güvenirliği olumsuz etkilemektedir (Turgut & Baykul,2012).Fakat puanlama kolaylığı açısından birçok öğretmenin ilk tercihi çoktan seçmeli testler olmaktadır. Puanlayıcıların görüşlerine göre objektif bir test olması ve uygulama kolaylığı açısından cazip olsa da sonuç odaklı değerlendirme yapması, hazırlama zorluğu gibi sebeplerle tercih edilmediği durumlar da olmaktadır.

Çoktan seçmeli testlerin dışında sıklıkla kullanılan diğer bir ölçme aracı da açık uçlu sorulardan oluşan yazılı sınavlardır. Çoktan seçmeli sorular ile karşılaştırıldığında, açık uçlu soruların puanlanmasıdaha fazla zaman alır. Bu nedenle kullanışlılığı düşüktür. Ancak hazırlaması çoktan seçmeli sınavlara göre daha kolay olduğundan sıklıkla tercih edilmektedir.Puanlamanın objektifliğinden bahsetmek de oldukça güçtür. Aynı yazılıyı değerlendiren farklı puanlayıcılar, aynı öğrencinin aynı soruya verdiği cevabına farklı puanlar verebilirler. Bu güçlük, hem her cevabın dikkatle okunması zorunluluğundan, hem de cevapların kesinlikle doğru veya yanlış olarak sınıflanamayışından doğar. Ayrıca, puanlayıcı kanısı gerektiğinden, puanlama işlemine çeşitli hatalar karışabilmektedir (Turgut & Baykul, 2012-145). Yapılan hataların en aza indirilmesi için Kan (2001) değerlendirmenin belirli bir ölçekle yapılmasını önermektedir.Bu ölçekler performansa dayalı durum belirleme başlığı altında anlatılmıştır.

Araştırmada problem çözme becerisini ölçen açık uçlu sorulardan yararlanılmıştır. Öğrencilerin uygulamaya verdikleri cevaplar, iki farklı zamanda, farklı dereceli puanlama anahtarları ile 5 farklı puanlayıcı tarafından puanlanmıştır. Elde edilen puanlar KTK ve GK ile değerlendirilmiş, güvenirlik düzeyleri ve puanlayıcılar arası tutarlık belirlenmeye çalışılmıştır.

Aşağıda problem ve problem çözme becerileri hakkında tanım ve bilgiler yer almaktadır.

1.1.1. Problem ve Problem Çözme Becerileri

Alan yazında problem farklı biçimlerde tanımlanmıştır. Problem insanların karşı karşıya kaldığı, çözüm gerektiren ve çözüm yolu anında bilinemeyen olgulardır (Posameinter & Krulik, 1998).Başka bir tanımda ise problem; kişide çözme arzusu uyandıran ve çözüm süreci hazırda olmayan fakat kişinin bilgi ve deneyimlerini kullanarak çözebileceği durumlara denir (Olkun & Toluk, 2001). Sözlükteki problemin karşılığı ise "teoremler ve kurallar yardımıyla çözülmesi istenen soru, mesele" olarak karşımıza çıkmaktadır (Türk Dil Kurumu, 2005).Ayrıca problem denildiğinde aklımıza yalnız matematik alanındaki problemler gelmez. Cüceloğlu (1999) problemi, hayatın her alanında bireyin hedefe ulaşmasını engelleyen ketler olarak ifade etmiştir.John Dewey, problemi insan zihnini karıştıran, ona meydan okuyan ve inancı belirsizleştiren her şey olarak tanımlamaktadır. Bu tanımla problem, belirsizliklerin ortadan kaldırılması demek olur.

Problem çözme becerisi ise insanoğlunun neslinin varlığını sürdürülmesi için gerekli olan en hayati yetenek olarak yorumlanabilir. İnsan hayatında ne zaman problemlerle karşılaşılacağı önceden belli değildir. Bununla baş edebilmek için eğitim sistemi kendi kendine güçlüklerin üstesinden gelebilecek bireyler ve toplumlar yetiştirmeyi hedef edinmiştir. Bu yüzden bireyin problem çözme becerisini kazanması büyük önem arz etmektedir. Problem çözme becerilerinin değerlendirilmesi için performansa dayalı durum belirleme teknikleri kullanılır. Aşağıda performansa dayalı durum belirleme ve puanlama yöntemleri hakkında bilgi verilmiştir.

1.1.2. Performansa Dayalı Durum Belirleme

Performansa dayalı durum belirleme, gerçek yaşam öğeleri ve şartlarını içeren görevlerle, bireylerin sahip olduğu, bilgi, kavram ve becerileri değerlendirme yöntemi olarak tanımlanmıştır (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Ayrıca Mehrens'e göre (1992) bu belirleme yöntemi, bireylerin, gerçek yaşam koşullarında karmaşık görevleri yaparken edindikleri temel bilgileri kullanma becerisini ölçmeye çalışır.

Performansa dayalı durum belirleme iki kategoride incelenebilir.Bu kategorilerden birincisi performans görevi ve ikincisi dereceli puanlama anahtarlarıdır (Kutlu, Yıldırım ve Bilican,2009). Performans görevleri öğrencilerin kendi anlayışlarını ortaya koydukları, kendi çözüm yöntemlerini ürettikleri ve sonunda kendilerinden izler taşıyan yanıtlar ortaya koydukları ölçme araçlarıdır. Çoktan seçmeli ya da kısa yanıtlı bir testte seçeneklerden farklı yanıtlara yer yoktur. Fakat etkili bir öğretimde amaç bilgiyi yapılandırmadır ve bu da performans görevleri ile gerçekleşebilir (Marzano, Pickering ve Mctighe, 1993).

Performansa dayalı durumların daha etkin ve daha güvenilir şekilde değerlendirilmesi için doğru araçlara ihtiyaç duyulmaktadır. Bu aşamada dereceli puanlama anahtarlarının yaygın olarak kullanıldığı söylenebilir. Dereceli puanlama anahtarlarını ortaya konulan performansın en düşük seviyesi ile en yüksek seviyesi arasında gözlenebilen ve önceden belirlenmiş ölçütler bütünü olarak tanımlanmaktadır (Callison, 2000).

1.1.3. Performansa Dayalı Durum Belirlemede Puanlama Yöntemleri

Performansa dayalı durum belirleme için kullanılan dereceli puanlama anahtarları ile ilgili ayrıntılı bilgi devam eden başlıklarda sunulmaktadır. Ayrıca dereceli puanlama anahtarları ve çeşitleri ile kullanımı hakkında bilgilerde bulunmaktadır.

1.1.3.1 Dereceli Puanlama Anahtarları

Puanlama anahtarları, kazanımı veya içeriği tanımlarken bu içeriklere ve doğruluk derecesine göre uygun görülen ölçütleri içeren puanlama araçlarıdır. Anahtarlar ölçme etkinliğinin tamamının ya da kısımlarının nasıl puanlanacağını ana hatlarıyla belirtir. Dereceli puanlama anahtarları ile öğrenciler hangi ölçütlere göre değerlendirileceklerini, kendilerinden tam manasıyla ne beklendiğini ve yaşadıkları

süreç ve ürünlerinin hangi puana denk geldiğini öğrenirler. Böylelikle düzey ve hedef belirleme ile ilgili de mesafe almış olurlar (Doğan, Karakaya ve Kutlu, 2010).

Dereceli puanlama anahtarları geliştirilirken, sırasıyla, anahtarın amacı belirlenir, puanlanacak öğeler belirlenir ve yeterlik düzeylerine karar verilir (Mertler,2001). Diğer bir dereceli puanlama anahtarı hazırlama aşamaları sınıflandırmasında ise Popham (1997);

- Değerlendirme ölçütlerini belirleme
- Ölçüt tanımlamalarını yapma
- Puanlama stratejisini belirleme adımlarına vurgu yapmıştır.

Bu sınıflandırmalara ek olarak Mertler (2001) ise daha ayrıntılı bir listeyi alan yazına kazandırmıştır.

- Verilen performans görevi ile ölçülecek hedeflerin belirlenmesi
- Performansta öğrenciden istenen gözlenebilir bilgi ve becerilerinbelirlenmesi,
- İkinci adımda belirlenen her bilgi veya beceri için ortalamanın üstü,ortalama ve ortalamanın altı olabilecek durumların belirlenmesi,
- Her bilgi veya beceri için gözlemlenebilir en yüksek, orta ve en düşükperformansların bir araya getirilerek, ayrı ayrı grupların oluşturulması.
- Her bir seviye için öğrencilerin gözlemlenmesi ve dereceli puanlamaanahtarının doldurulması,
- Dereceli puanlama anahtarının gözden geçirilmesi ve dereceli puanlamaanahtarı üzerinde gerek varsa değişikliklerin yapılması (Akt: Hızarcıoğlu,2013).

Dereceli puanlama anahtarı hazırlarken dikkat edilmesi gerekenler:

- Performans değerlendirmesinde gözlenmesi ve karşılaşılması muhtemel öğrenci davranışları belirlenmelidir.
- Değerlendirilecek anahtar davranışların boyutları belirlenmelidir.
- Değerlendirilecek davranışlara somut örnekler gösterilmelidir.
- Ne tür bir derecelendirme ölçeği kullanılacağına kara verilmelidir.

- Performans belirlemek için mükemmelliğin standartları ve ölçütleri belirlenmelidir.
- Değerlendirmenin kim tarafından yapılacağına karar verilmelidir.

Dereceli Puanlama Anahtarlarının Avantajları:

- Öğrenciler kendilerinden ne beklendiğini bilebilir.
- Ölçütlerin açık ve net oluşu ile kullanışlıdır.
- Kullanışlı oluşu zaman tasarrufu sağlar.
- Eğer etkin kullanılabilirse öğrencilerin eksik oldukları kavramları kavramaları kolaylaşır.
- Değerlendirmenin sınırları belirlendiğinden tarafsızlık da artırılmış olur.
- Öğrenciler gelişim süreçlerini izleyebilirler.
- Öğretmen ve öğrenci arasındaki pozitif iletişim artar.
- Dereceli puanlama anahtarları sayesinde öğrenci performans değerlendirmesi sürecinde ve sonucunda kendisinden hangi performansı göstermesi konusunda bilgi sahibi olur. Öğrenci ve velide değerlendirme ölçütü hakkında bilgi sahibidir. Kısacası öğrenciden beklenenler rubriklerle birlikte öğrenci ve velilere duyurulmaktadır (Sezer, 2006).

Dereceli Puanlama Anahtarlarının Sınırlılıkları:

- Dereceli puanlama anahtarları birçok hazırlık süreci ve ölçüt belirleme gerektirdiğinden hazırlamak zaman alabilir,
- Ölçütler iyi belirlenmezse puanlayıcılar arası tutarlık ve dolayısıyla güvenirlik düşebilir,

Öznel ölçütlerle puanlamada vakit kaybı yaşanabilir (Büyükkıdık, 2012)

1.1.3.1.1 Bütünsel Dereceli Puanlama Anahtarı

Bütünsel dereceli puanlama anahtarı adından da anlaşılacağı üzere, puanlamanın bir bütün olarak yapıldığı durumlarda kullanılır. Bir diğer adıyla genel izlenimle puanlama da denmektedir (Mertler, 2001).

Öğrencilerin gösterdikleri performans, hazırladıkları proje veya yazılı yoklamada yaptıkları işlemler, “mükemmel”, “iyi”, “geliştirilmeli”şeklinde, bu ifadelere denk

gelen sayılarla (0,1,2 gibi) ifade edilir. Böylece öğrencinin puanı belirlenmiş olur. Puanlamanın kısa sürmesi, ölçülen bazı davranışın boyutlara ayrılmasının zor olması ve cevapların bütün olarak değerlendirilmesinin gerektiği durumlarda sıklıkla tercih edilmektedir (Gündüz, Sefer, 2006).

Bütünsel dereceli puanlama anahtarları süreçten çok ürünle ilgilenir. Bu nedenle bütüncül dereceli puanlama anahtarlarında puanlama ölçütleri bir problemin çözümünün önemli bölümleriyle ilgili olması gerekir.

Aşağıda 6 düzeyli bütünsel dereceli puanlama anahtarının yapısını gösteren genel bir çizelge verilmiştir(Mertler, 2001)

Çizelge 1.1: Bütünsel Dereceli Puanlama Anahtarı Yapısı

Performans Düzeyi	Performans Tanımları
5 (Mükemmel)	Problemin tamamının anlaşıldığını gösterir. Cevaplar ya da ürün performans görevine ilişkin bütün gerekleri içermektedir.
4 (Başarılı)	Problemin önemli ölçüde anlaşıldığını gösterir. Cevaplar ya da ürün performans görevine ilişkin gerekleri içermektedir.
3 (Gelişmekte)	Problemin kısmen anlaşıldığını gösterir. Cevaplar ya da ürün performans görevine ilişkin çoğu gereği karşılamaktadır.
2 (Başlangıç)	Problemin çok az anlaşıldığını gösterir. Cevaplar ya da ürün performans görevine ilişkin çoğu gereği karşılamaktan yoksundur.
1 (Başarısız)	Problemin tamamıyla anlaşılmadığını gösterir.
0 (Yetersiz)	Cevap ya da çözüm yok/ performans görevi yapılmamış.

Kaynak:Mertler, C. A. (2001). *Designing scoring rubrics for your clasroom*. Practical Assessment,Research andEvaluation

1.1.3.1.2 Analitik Dereceli Puanlama Anahtarı

Analitik dereceli puanlama anahtarında performansın süreci veya ürünün, çözümün parçaları ayrı ayrı puanlanır, ardından bu parça puanlar toplanarak toplam puan elde edilir. Analitik dereceli puanlama anahtarı ürün kadar sürece de değer verir.

Analitik dereceli puanlama anahtarının daha analitik olması için kategori ve ölçütlerin sayısını artırmak gerekir. Kriterleri artırmak bu puanlama anahtarını hazırlamayı ve puanlamayı zorlaştırır. Yani bütünsel puanlama anahtarı ile yapılan puanlamalar daha kolaydır. Ancak kategori ve ölçütler artıkça öğretmen ve

öğrenciler çok daha güvenilir dönütler almış olurlar. Örneğin; yazılı anlatım becerisinin ölçüldüğü kompozisyon yazma sınavlarında, öğretmen kompozisyonu farklı başlıklar altında inceler; sayfa düzeni, dili etkin kullanma, yazım ve noktalama kurallarına uyma, içerik ve akıcılık gibi. Böylesine kategorize edilmiş ölçekler ile öğrenci daha güvenilir, yani daha hatasız bir puan almış olur.

1.2. Araştırmanın Amacı ve Önemi:

Tezin genel amacı puanlayıcılar arası tutarlık derecelerini, Klasik test kuramı ve genellenebilirlik kuramı kullanılarak karşılaştırmak ve bilgi edinmek olmuştur. Hangi güvenirlik belirleme kullanmak gerektiği ile problem çözme becerisini ölçen açık uçlu soruların puanlanmasında hangi dereceli puan ölçeğinin kullanmanın daha uygun olduğu belirlenmeye çalışılmıştır.Puanlayıcıların bütünsel ve analitik dereceli puanlama anahtarlarını kullanma sıklıkları ve bu puanlama anahtarlarına yönelik görüşleri incelmiştir. Ayrıca öğrencilerin problem çözme becerilerinin de değerlendirilmesi sağlanarak bu konudaki farkındalığın arttırılması hedeflenmiştir.

Daha önce de belirtildiği gibi öğrencilere problem çözme becerilerini kazandırmak gelişen ve sürekli değişen dünya için kaçınılmaz bir zorunluluk olmuştur. Öğrencileri hayata hazırlamak ve gerçek hayatta karşılaşacakları, içine düşecekleri problemlere ve zorluklara çözüm üretecek pozisyona getirmek için daha sıkı çalışmak gerekliliği doğmuştur. Ancak bu öğretilerin daha nesnel bakış açılarıyla, daha hatasız değerlendirilmeleri için tutarlık vazgeçilmez bir nitelik olarak karşımıza çıkmaktadır.

Ayrıca değişen sınav sistemleriyle değeri artırılan okul başarılarını ve ulusal rekabeti göz önünde bulunduracak olursak, öğrencilerin becerilerinin en adil ve en tutarlı şekilde değerlendirme gerekliliğini anlamış oluruz. Bundan yola çıkarak okullarda sıklıkla kullanılan açık uçlu yazılı sorularının değerlendirilmesi ve puanlayıcılar arası tutarlık dereceleri saptamak ve bu tutarlık derecelerinin artması için gerekli dönütleri vermek bu araştırmanın önemine katkıda bulunmaktadır. Genel olarak bu araştırmanın amacı , öğrenci becerilerinin en adil ve en tutarlı şekilde puanlanması ve değerlendirilmesi için geri bildirim sağlamaktır.

1.3. Problem Cümlesi:

İlköğretim matematik öğretmenlerinin, öğrencilerin problem çözme becerilerini ölçen açık uçlu soruları, analitik ve bütünsel dereceli puanlama anahtarı ile

puanlamalarında puanlayıcılar arası tutarlık KTK ve G-Kuramına göre nasıl değişir?

1.3.1. Alt Problemler:

1.Puanlayıcıların analitik ve bütünsel dereceli puanlama anahtarlarından elde ettikleri puanlar,Klasik test kuramındaki puanlayıcılar arası tutarlık belirleme tekniklerine göre nasıl değişmektedir?

a) Analitik ve bütünsel dereceli puanlama anahtarları kullanılarak elde edilen puanların tutarlığı Pearson Momentler Çarpımı Korelasyon Katsayısı ve Spearman Sıra Farkları Korelasyon Katsayısına göre nasıl değişmektedir?

b) Analitik ve bütünsel dereceli puanlama anahtarları ile elde edilen puanların tutarlığı Kappa Tekniğine göre nasıl değişmektedir?

c) Analitik ve bütünsel dereceli puanlama anahtarları ile elde edilen puanların tutarlığı Krippendorf Alfa Tekniğine göre nasıl değişmektedir?

2.Puanlayıcıların analitik ve bütünsel dereceli puanlama anahtarları ile farklı zamanda elde ettikleri öğrenci puanlarının tutarlık dereceleri Genellenebilirlik kuramına göre birey (b), madde (m) ve puanlayıcı (p) değişkenin çapraz tasarlandığı b x m x p deseninde nasıl değişmektedir?

a) Analitik dereceli puanlama anahtarı ile elde edilen puanların oluşturduğu b x m x p deseni G çalışması sonucunda kestirilen varyans bileşenlerini ve toplam varyansı açıklama yüzdelerini nasıl etkilemektedir?

b) Bütünsel dereceli puanlama anahtarı ile elde edilen puanların oluşturduğu b x m x p deseni G çalışması sonucunda kestirilen varyans bileşenlerini ve toplam varyansı açıklama yüzdelerini nasıl etkilemektedir?

3. Klasik Test Kuramına ve Genellenebilirlik Kuramına göre elde edilen güvenirlik katsayıları tutarlık göstermekte midir?

a) Analitik dereceli puanlama anahtarıyla elde edilen puanların güvenirlik Katsayılarıtutarlı mıdır?

b) Bütünsel dereceli puanlama anahtarıyla elde edilen puanların güvenirlik Katsayıları tutarlı mıdır?

1.4. Sayıltılar:

1. Uzmanlar geliştirilen test ve puanlama anahtarları ile ilgili ankete ve görüşmelere ciddiyetle katılmışlardır.

2. Öğretmen ve öğrenciler gönüllülük esası ve benzer şartlar altında çalışmaya dahil olmuşlar ve tüm çalışmalarını içtenlik ve ciddiyetle sürdürmüşlerdir.

3. Puanlayıcıların yaptıkları puanlamalar birbirinden bağımsızdır.

4. Ölçme araçlarından kaynaklanan yapısal özelliklerin tüm puanlamalar için sabit olduğu varsayılmaktadır.

1.5. Sınırlılıklar:

1. Araştırma 2013–2014 öğretim yılı Ankara ilinde görev yapan 5 öğretmen ve 84 öğrenci ile sınırlıdır.

2. Araştırma açık uçlu sorulardan oluşan 6 soruluk bir ölçme aracı ile sınırlıdır.

3. Araştırma KTK ve GK'dan farklı tutarlık belirleme yöntemleri ile sınırlıdır.

1.6. Araştırmanın Kuramsal Temelleri

Bu alanda araştırmanın kuramsal temellerini oluşturan klasik test kuramı ve genellenebilirlik kuramı üzerinde durulmuştur.

1.6.1 Puanlayıcılar Arası Güvenirliğini Kestirme Yöntemleri

Bir ölçme sonucuna hata karışmasına, ölçme yapılan ortam, ölçmeye tabi tutulan birey, hazırlanan ölçek ve puanlayıcı gibi birçok etken neden olabilir. Bu

araştırmada puanlayıcıların ölçme üzerindeki etkileri ve ölçeğin güvenirliğini ne derece etkiledikleri belirlenmeye çalışılmıştır. Puanlama, ölçmenin en kritik ve en önemli adımlarından biri olarak görülmektedir. Puanlama hatalarının ölçek güvenirliğini, ölçekten, ölçmeciden ve ölçme ortamından kaynaklanan sebeplere göre daha fazla etkilediği öngörülmektedir.

Puanlayıcılar arası güvenirlik iki ya da daha fazla puanlayıcı arasındaki uyum ve tutarlığın derecesidir (Crocker & Algina, 1986). Puanlayıcı güvenirliği, yargıcı güvenirliği, gözlemci güvenirliği ve derecelendiriciler arası güvenirlik gibi çeşitli şekillerde adlandırılan puanlayıcılar arası güvenirlik; iki ya da daha fazla puanlayıcı arasındaki belirli bir ölçüm ile ilgili uzlaşmanın veya tutarlığın derecesidir. Puanlayıcılar arası güvenirliği hesaplamanın en kolay yolu korelasyon hesaplamaktır (Jay Cohen & Swerdlik, 2010). Puanlayıcılar arasındaki uyumu belirlemek için sıklıkla Pearson Momentler Çarpımı Korelasyon katsayısı kullanılır. Bu korelasyon tekniğinin kullanılabilmesi için puanların sürekli değişken olması ve en az eşit aralıklı ölçekle gösterilebiliyor olması gerekmektedir. Fakat elde edilen değerler puanlayıcıların verdikleri puanların değişkenliği gösterip, güvenirliği açıklamakta yetersiz kalmaktadır. Bu yüzden Burry-Stock ve diğerleri (1996) korelasyon katsayısının puanların birlikte değişkenliğini açıkladığını fakat puanlayıcılar arasındaki uyumu göstermekten uzak olduğunu ifade etmişlerdir. Korelasyon tekniğinin ve benzer şekilde kullanılan diğer yöntemlerden paralel formlar (eşdeğer formlar), test tekrar test yöntemi, eşdeğer yarılar gibi yöntemlerin sınırlılıkları nedeniyle farklı yöntem ve tekniklere ihtiyaç duyulmaktadır. Bunları Cohen'in Kappası, ağırlıklandırılmış Kappa, Kendall'ın W'si, Krippendorff'un Alfası, Scott'un Pisi, Holsti'n güvenirliği, Lin'in konkordans korelasyon katsayısı, Cochran'ın Q testi, Lojistik regresyon, Loglineer analiz vb. olarak sıralayabiliriz. Ayrıca puanlayıcılar arası tutarlığın belirlenmesinde, verilerin sürekli değişken niteliğinde olduğu durumlarda varyans analizi de kullanılmaktadır. Varyans analizine dayalı teknik olarak sınıf içi korelasyon katsayısı hesaplanmaktır (Bıkmaz, 2011). Genellenebilirlik Kuramı da klasik test kuramına paralel bir şekilde güvenirlik katsayısına benzer olan genellenebilirlik katsayısını verir ve puanlayıcılar arasındaki tutarlığın ve güvenirliğin kestirilmesi için kullanılır. (Güler, Kaya Uyanık, & Taşdelen Teker, 2012). Bu yöntemler arasından Klasik test kuramının öğelerinden, Cronbach Alfa, PMÇKK, SSFKK, Kappa tekniği,

KrippendorffAlfa tekniği ile Genellenebilirlik kuramı kullanılarak puanlayıcılar arası tutarlık ve güvenirlik dereceleri belirlenecektir.

1.6.2. Klasik Test Kuramı

Klasik test kuramında amaç gözlenen özelliklerin gerçek değerini bulmaktır. Ancak ölçmeye karışan çeşitli hata kaynakları gerçek puanların ölçme kanalıyla elde edilmesine engeller. Klasik test kuramına göre gözlenen puanlardan yola çıkarak gerçek puanları elde etmeye çalışır. Bu nedenle Klasik test kuramına **gerçek puan teorisi** de denir (Baykul, 2010). Klasik test teorisine en büyük katkının Sperman tarafından verilmeye başlandığı (1907–1913) ifade edilmektedir. Bu alana katkı getiren diğer bilim adamlarını ise sırasıyla, Linquist (1954), Ebelhart, Hull, Rush (Gulliksen, 1950, s.2, Akt: Baykul, 2010:107), Lazarsfeld (Lord & Novick, 1968), Lord ve Novick olarak sayılabilir (Akt: Baykul, 2010, s.107).

Teoremde bahsi geçen gerçek, gözlenen ve hata puanları arasındaki ilişki aşağıdaki denklemle ifade edilebilir. (Sperman, 1904; Croker &Algina, 1986: 107, Baykul, 2010: 113,126,Cohen & Swerdlik, 2013: 139). Bu denklem kuramın temel denklemi olarak kabul edilir.

$$X_{Gözlenen}=X_{Gerçek} + X_{Hata}\text{…….. (1)}$$

Klasik test kuramında güvenirlik katsayısı gerçek puan varyansının gözlenen puan varyansına oranı olarak ifade edilir .

$$\sigma_{gözlenen}^{2} = \sigma_{gerçek}^{2} + \sigma_{hata}^{2} \quad \text{.....(2)}$$

$$\rho = \frac{\sigma_{gerçek}^{2}}{\sigma_{gözlenen}^{2}} \quad \text{.....(3)}$$

Klasik test kuramının varsayımları (temel özellikleri) şöyle, ifade edilmektedir.

- Hata puanlarının beklenen değeri sıfırdır.
- Gerçek puanlarla hata puanları arasındaki korelasyon sıfırdır.
- Farklı ölçümlere karışan hata puanlarının korelasyonu sıfırdır.

1.6.2.1 Kappa Tekniği

Kappa istatistiği (κ) ilk olarak 1960'da Cohen tarafından önerilmiştir (Fleiss, 1971;Fleiss ve Cohen, 1973; Landis ve Koch, 1977; Brennen ve Prediger, 1981; Hunt, 1986; Banerjee ve diğerleri, 1999; Akt: Bıkmaz, 2011). Bu nedenle Cohen'in Kappası olarak da bilinir. Kappa istatistiği en sık kullanılan puanlayıcılar arası güvenirlik belirleme yöntemlerinden birisidir. Sınıflama düzeyinde puanlama yapan iki puanlayıcının uyum derecesini özetleyen betimsel bir istatistiktir.Kappa Tekniğinin uygulanması için bazı varsayımların karşılanması gerekmektedir.. Brennen ve Prediger'e göre (1981) bu varsayımlar;

- Kategorize edilen nesne ve bireyler bağımsızdır
- Puanlayıcı puanlamaları bağımsızdır
- Puanlamada kullanılan kategoriler birbirinden bağımsızdır.

1.6.2.1.1 İki Puanlayıcı ve İki Kategori İçin Kappa İstatistiği Hesaplama

Kappa istatistiğinin (κ) hesaplanması için puanlayıcıların verdikleri puanlardan oluşan çapraz tablolar kullanılır. Çapraz tablolar ölçülen özelliğe verilen puanların frekans değerlerini içerir.

Çizelge: 1.2: Kappa İstatistiği Çapraz Çizelge Örneği

	Puanlayıcı 1			
Puanlayıcı 2		1	2	Satır Toplam
	1	a	b	g_1
	2	c	d	g_2
	Sütun Toplam	f_1	f_2	n

Kaynak: Sim, J. and Wright, C. C. (2005) *"The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements"* in Physical Therapy. Cilt. 85, say. 257-268

Çizelge 1.2' yi incelediğimizde;

- "a" ile gösterilen hücre iki puanlayıcının da ortak olarak kategori 1'e koydukları değer,
- "d" ile gösterilen hücre iki puanlayıcının da ortak olarak kategori 2'ye koydukları değer,
- "f_1" ile gösterilen hücre 1. puanlayıcının kategori 1'de puan verdiği toplam birey sayısı,
- "f_2" ile gösterilen hücre 2. puanlayıcının kategori 2'de puan verdiği toplam birey sayısı,
- "g_1" ve "g_2" ile gösterilen hücreler ise iki kategoride puan verilen toplam değeri ifade etmektedir.

Çizelge 1.2' nin ana köşegen üzerindeki "a" ve "d" değerleri puanlayıcılar arasındaki uyumu, "b" ve "c" değerleri ise puanlayıcılar arasındaki uyumsuzluğu ifade etmektedir (Crawforth, 2001). Kappa istatistiği (κ) hesaplanırken bu "a" ve "d" değerlerinden yola çıkılarak uyum oranını gösteren formül aşağıda gösterildiği gibidir.

Çizelge 1.2' nin ana köşegen üzerindeki "a" ve "d" değerleri puanlayıcılar arasındaki uyumu, "b" ve "c" değerleri ise puanlayıcılar arasındaki uyumsuzluğu ifade etmektedir (Crawforth, 2001). Kappa istatistiği (κ) hesaplanırken bu "a" ve "d" değerlerinden yola çıkılarak uyum oranını gösteren formül aşağıda gösterildiği gibidir.

$$P_0 = \frac{a+d}{n} \quad \text{....(4)}$$

Fakat uyum oranının belirlenmesinde kullanılan bu hesaplamaya karışacak tesadüfi değişkenleri ve şans başarısını ortadan kaldırmak için "tesadüfi uyumluluk" değeri de bulunarak Kappa (κ) istatistiğine son hali verilmiş ve formülize edilmiştir(Sim ve Wright, 2005, Akt: Bıkmaz, 2011).

$$P_c = \frac{\left(\frac{f_1 * g_1}{n}\right) + \left(\frac{f_2 * g_2}{n}\right)}{n} \quad \text{....(5)}$$

$$\kappa = \frac{gözlenen\ uyumluluk - tesadüfi\ uyumluluk}{1 - tesadüfi\ uyumluluk}$$

$$\kappa = \frac{P_0 - P_c}{1 - P_c} \quad \text{....(6)}$$

1.6.2.1.2 İki Puanlayıcı ve İkiden Fazla Kategori İçin Kappa İstatistiği Hesaplama

Birim ve kategori sayısı arttığında Cohen'in Kappası aşağıda verilen tablo değerlerinden yola çıkılarak verilen formüller ile hesaplanmaktadır.

Çizelge: 1.3: Kappa İstatistiği İki Puanlayıcı İkiden Fazla Kategori ÇizelgeÖrneği

Puanlayıcı 1						
Puanlayıcı 2		1	2	…	k	Satır toplam
	1	P_{11}	P_{12}	…	P_{1k}	P_1
	2	P_{21}	P_{22}	…	P_{2k}	P_2
	…	....	…	…	…	…
	k	P_{k1}	P_{k2}	…	P_{kk}	P_k
	Sütun Toplam	P_1	P_2	…	P_k	P_{ij}=1

Gözlenen uyum ya da diğer bir adıyla isabet oranı iki kategorili çapraz tabloda olduğu gibi ana köşegen üzerindeki değerlerin toplanmasıyla bulunur (Fleiss ve Cohen, 1973).

$$P_0 = \sum_{i=1}^{k} P_{ii} \quad \text{….(7)}$$

Ortaya çıkan şans oranı satır ve sütün toplamları hesaplanarak bulunur.

$$P_c = \sum_{i=1}^{k} P_i P_i \quad \text{….(8)}$$

Yine gözlenen uyumdan beklenen uyumun çıkarılmasıyla iki puanlayıcı ve k kategorili kappa istatistiği hesaplanmış olur.

$$\kappa = \frac{P_0 - P_c}{1 - P_c} \quad \text{....(9)}$$

1.6.2.1.3 İkiden Fazla Puanlayıcı ve İkiden Fazla Kategori İçin Kappa İstatistiği Hesaplama

Cohen'in İki puanlayıcı*İki kategorili ve İki puanlayıcı* "k" kategorili hesapları bu durumda yetersiz kalmaktadır. Fleiss (1971) tarafından ikiden çok puanlayıcı arasındaki uyumu hesaplamak adına Cohen'in kappa istatistiğinin genellenmiş hali olan bir kappa istatistiği önerilmiştir. Buna göre n tane puanlayıcı tarafından "i" birey ve "j" kategorideki değerlendirmelerin genel oranı $\mathbf{P_j}$ ile gösterimiştir.

$$P_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}$$

….(10)

P_i j tane kategori içinbelirlenen tüm değerlerin oranı, P_i'lerin ortalaması ise uyumun toplam derecesini ifade etmektedir.

$$P_i = \frac{1}{n(n-1)} \sum_{i=1}^{k} n_{ij} \left(n_{ij} - 1\right)$$

$$= \frac{1}{n\,(n-1)} \left(\sum_{j=1}^{k} n_{ij}^2 - n \right)$$

....(11)

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i$$

$$= \frac{1}{N\,n\,(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - N\,n \right)$$

….(12)

Şansla beklenen uyumun oranı da hesaplamaya karışmış olabilir. Bu nedenle bu uyumun da hesaplanması gerekmektedir. Şansla beklenen uyum **"$\mathbf{P_C}$ "**ile gösterilir.

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2$$ ….(13)

Son olarak bulunan değerler kullanılarak Fleiss'in Kappası bulunur.

$$\kappa = \frac{\bar{P} - \bar{P}_c}{1 - \bar{P}_c}$$ ……(14)

1.6.2.1.4 Kappa İstatistiğinin Yorumlanması

Fleiss'e göre (1971) Kappa değerleri -1 ve +1 arasında değişen değerler alabilirler. Bu değerlerden pozitif olanları puanlayıcılar arasında şansla beklenen değerden daha yüksek bir uyum olduğu ve negatif değerler ise puanlayıcılar arasında şansla beklenenden daha düşük bir uyum olduğu ifade eder. Landis ve Koch (1977) aşağıdaki tabloyu önermiş ve birçok kesim tarafından kullanılabilirliği kabul edilmiştir.

Çizelge: 1.4 Kappa İstatistiği Değerleri ve Uyum Yorumları

Kappa değerleri	Uyum düzeyi
0.00-0.20	Önemsiz
0.21-0.40	Düşük düzeyde
0.41-0.60	Orta düzeyde
0.61-0.80	Yüksek düzeyde
0.81-1.00	Çok yüksek düzeyde

Kaynak: Landis, J. R. ve Koch, G. G. (1977) "The measurement of observer agreement for categorical data" ,*Biometrics*. Cilt. 33, say. 159-174

1.6.2.1.5 Kappa İstatistiğinin Tercih Sebepleri

- Hesaplamak ve yorumlamak kolaydır.
- Tesadüfi uyum da düşünülerek hesap yapıldığından güvenirliği yüksektir.

- Kategori sayısı ve puanlayıcı sayısından etkilenmemektedir.(Silcocks, 1983).

1.6.2.1.6 Kappa İstatistiğinin Sınırlılıkları

- Her ne kadar kabul görse de yorum için önerilen aralıklar kesin değildir.
- Kategorik veriler dışında uygulamak mümkün değildir.
- Grubun büyüklüğü ve istatiksel dağılım modelinden etkilenmektedir (Landis ve Koch, 1977)

1.6.2.2 KrippendorffAlfa Tekniği

Puanlayıcılar arasındaki uyumun belirlenmesi için Krippendorf tarafından geliştirilen ve önerilen yöntemdir (Krippendorff, 2007). KrippendorffAlfa istatistiği Kappa İstatistiği, Benini'nin Betası ile benzerlikler gösterse de aslında oldukça farklıdır (Krippendorff, 2007). Bu istatistiğin hesaplanma adımları aşağıda formülleri ile ifade edilmiştir. Öncelikle 1. adımda gözlenen uyumsuzluğu temsil eden D_o değeri bulunur. Ardından ikinci adımda beklenen uyumsuzluğu simgeleyen D_e bulunur. Son olarak Alfa değerini bulmak için 3. Adımda uyumsuzluk değerinin beklenen uyumsuzluk değerine oranı bulunarak 1'den çıkartılır.

$$D_o = \frac{1}{n}\sum_c\sum_k o_{ck\,\text{metric}}\,\delta^2_{ck} \quad \text{..... (15)}$$

$$D_e = \frac{1}{n(n-1)}\sum_c n_c \sum_k n_{k\,\text{metric}}\,\delta^2_{ck} \quad \text{....(16)}$$

$$\alpha = 1 - \frac{D_o}{D_e} \quad \text{.... (17)}$$

Krippendorff 'a göre (2007)Alfa'nın 1 olması beklenen uyumun mükemmel olmasına, 0 olması ise tam uyumsuzluğu simgelemektedir.

1.6.2.2.1 Krippendorff Alfasının İki Puanlayıcı ve İki Kategori ile Hesaplanması

Öncelikle iki puanlayıcı arasındaki uyumsuzlukları tespit etmek gerekmektedir. Ardından bu uyumsuzluk 1'den çıkarılmalıdır. Böylelikle uyum ölçüsü elde edilmiş olur. Uyumsuzlukları gözlemlemek için puanlamalarla matris oluşturulur. Oluşturulan matriste puan çiftleri oluşturulur. İki puanlayıcının iki kategori için verdikleri puanlar aşağıdaki gibi toplanır.

$$O_{12} = X_{12} + X_{21} \quad \text{.....(18)}$$

1	2	3	4	5	6	7	8	9	10
0	1	0	0	0	0	0	0	1	0
1	1	1	0	0	1	0	0	0	0

.....(19)

Gözlenen uyumsuzluk değerleri için iki puanlayıcının birbirinden farklı verdikleri puanlar hesaplanır. Örneğin; 4 durumda (birinci, üçüncü, altıncı ve dokuzuncu durumlar) farklı düşündükleri görülmektedir $O_{01}=O_{10}=4$. Ardından uyumlu oldukları durumlar $O_{11}=2$ ve $O_{00}=10$ olarak belirlenir. Buna göre gözlenen uyum matrisi;

	0	1	
0	o_{00}	o_{01}	n_0
1	o_{10}	o_{11}	n_1
	n_0	n_1	$n=2N$

	0	1	
0	10	4	14
1	4	2	6
	14	6	20

.....(20)

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - (n-1)\frac{o_{01}}{n_0 \cdot n_1}$$

$$\alpha = 1 - (20-1)\frac{4}{14 \cdot 6} = 0.095$$

.....(21)

Yapılan hesaplamalar sonucu KrippendorffAlfasının 0.095 olduğu bulunmuştur. Sıfıra yakın olduğu için oldukça düşük uyum olduğu ifade edilebilir.

1.6.2.2.2 Krippendorff Alfa İstatistiğinin Yorumlanması

Alfa değeri 0 ile +1 arasında reel değerler almaktadır. Uyumun 1 çıkması mükemmelliğini, uyumun 0 'a yaklaşması ise uyumun zayıfladığını ifade eder. Daha ayrıntılı bir sınıflama da Alfanın 0,80 den yüksek olması yüksek düzeyde uyuma, 0,67 ile 0,80 arasında olması orta düzeyde uyuma, 0,67 den düşük olması ise zayıf düzeyde uyuma işaret eder(Krippendorff, 2007).

1.6.2.2.3 Krippendorff Alfa İstatistiğinin Tercih Sebepleri

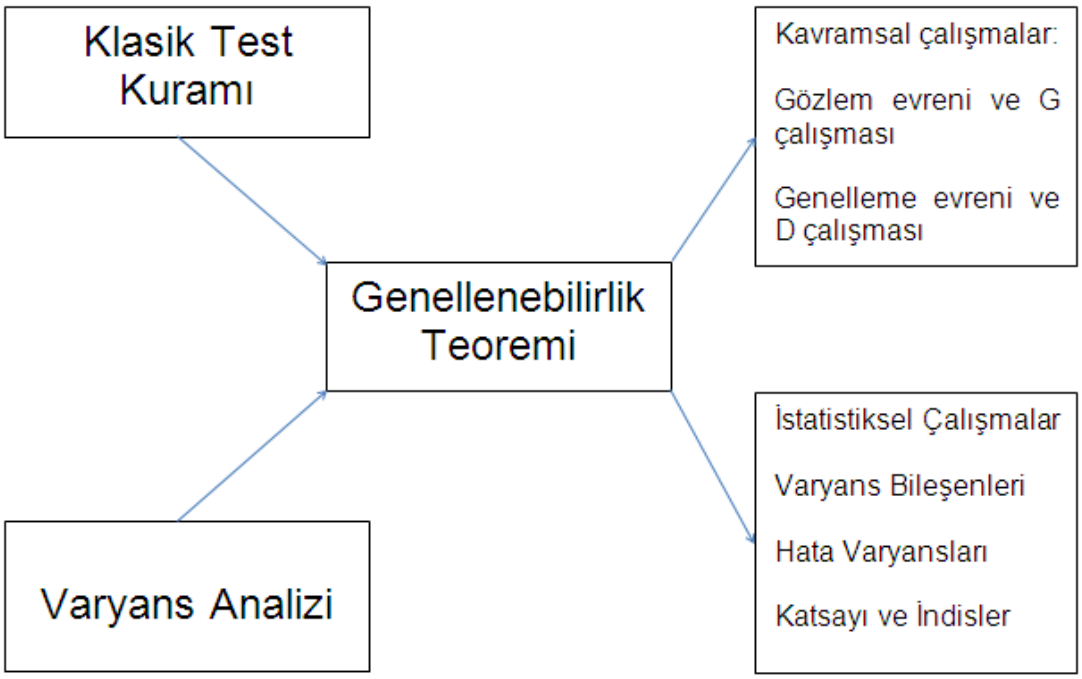
- Tüm ölçek türlerinde kullanılabilir.
- Şansla oluşan yüzdeyi hesaba katarak devre dışı bıraktığından daha güvenilir uyum derecesi verir.
- Eksik veri olması halinde de kullanılabilmektedir.

1.6.2.2.4 Krippendorff Alfa İstatistiğinin Sınırlılıkları

- Paket bilgisayar programlarıyla hesaplanmadığından kullanılması zor ve yaygınlaşamamıştır.
- Grubun çok büyük olmasından ya da çok küçük olmasından etkilenmektedir.

1.6.3. Genellenebilirlik Kuramı

Genellenebilirlik kuramının temelleri ilk olarak 1963–1965 arasında Cronbach, Rajaratman ve Gleser tarafından yapılan çalışmalar sonucu atılmıştır. 1983 yılında Brennan 'ın "Elements of Generalizability Theory" adlı kitabıyla geliştirilen kuram Shavelson ve Webb'in 1991 yılında yayımladıkları "Generalizability Theory: A primer" kitaplarıyla daha kolay anlaşılmaya ve çalışmalarda kullanılmaya başlanmıştır. 2001 yılında Brennan tarafından yayımlanan "Generalizability Theory" kitabı ve sonrasında geliştirdiği bilgisayar programlarıyla daha sık kullanılmaya başlanmıştır. Brennan 'a göre (2001) genellenebilirlik kuramı varyans bileşenleri ve onların analizi ile ilgilenmektedir. Brennan genellenebilirlik kuramının kökenini ve kavramsal çerçevesini aşağıdaki şekil ile açıklamaktadır.



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

Şekil 1. 1 : Genellenebilirlik Kuramının Kökeni ve Kavramsal Çerçevesi

1.6.3.1. Genellenebilirlik Çalışması ve Karar Çalışması

Genellenebilirlik kuramını ve genellenebilirlik çalışmalarını kavrayabilmek için öncelikle G kuramı ve KTK arasındaki farkları kavrayabilmek gerekir. Brennan (2001), Güler ve diğerleri (2012) bu farkları şu şekilde ifade etmektedirler. Genellenebilirlik kuramını Klasik test kuramına göre üstün olmasını farklı hata kaynaklarını aynı anda belirlemesine borçludur. KTK' da madde sayısının güvenirliğe olan etkisi kestirilebilirken, G kuramı ile güvenilir sonuçlar elde etmek için ihtiyaç duyulan durum, test formu ve uygulayıcı gibi var olan tüm hata kaynakları sayısının ne olması gerektiği belirlenebilir. KTK ile güvenirlik hakkında tek bir yoruma ulaşılırken, G kuramı puanların güvenirliğini farklı açılardan değerlendirmeyi sağlar.

Genellenebilirlik çalışması yani G çalışması için araştırmacının amacı ölçme yaptığı örneklemi, ölçmenin evrenine genellemektir diğer bir deyişle kabul edilebilir gözlemlerin evreniyle birleşen varyans bileşenlerinin hesabını elde etmektir (Brennan, 2001, s. 8) . Karar çalışmasında ise araştırmacı belirli bir amaç üzerinde karar vermek için G çalışmasının sonuçlarından ve değerlendirmelerinden

yararlanır (Akt: Güler, Kaya Uyanık, & Taşdelen Teker, 2012, s. 5-6). K çalışması; iyi belirlenmiş ölçme süreçleriyle, kararlar alabilmek adına, varyans bileşenlerinin kestirimleri, kullanılması ve yorumlanması olarak düşünülebilir (Brennan, 2001)

1.6.3.2. Çaprazlanmış Desen Yuvalanmış Desen

Bu deseni bu araştırmanın yönteminden bahsederek örneklendirebiliriz. Örneğin yaptığımız çalışmada öğrencilerin açık uçlu sorularda problem çözme becerilerini ölçen başarı testinde soruları cevaplayan her bir öğrenci (b), her bir soru (m) ve bu soruları puanlayan her bir puanlayıcı (p) ile oluşturulan desen "b x m x p" şeklinde çaprazlanmış desen olarak ifade edebiliriz. Eğer farklı öğrencilere (b) farklı maddeler (m) sorulup, farklı puanlayıcılar (p) tarafından puanlansalardı bu şekilde oluşturulan desen de yuvalanmış desen adını alırdı. Bu araştırmada da kullanılan b x m x p çaprazlanmış deseni iki yüzeyli ölçmelere örnek olarak verilebilir. Bu değişimin olası kaynakları madde ve puanlayıcı iken, madde ve puanlayıcıların sayısındaki değişim de bu değişim kaynaklarının koşulu olarak ifade edilebilir . İki yüzeyli ölçmelerde değişkenlik kaynakları aşağıdaki çizelgede gösterilmiştir (Brennan,2001).

Çizelge 1.5: İki Yüzeyli Ölçmelerde Değişkenlik Kaynakları

Değişkenlik Kaynağı	Değişkenlik Türü	Varyans Sembolü
Birey (b)	Evren Puanı (Ölçme Objesi)	σ^2_b
Madde (m)	Bir maddenin diğerine değişen birey davranışlarından kaynaklı bütün bireyler üzerindeki sabit etkisi	σ^2_m
Puanlayıcı (p)	Puanlayıcıların sıkılığının neden olduğu bütün bireyler üzerindeki sabit etki	σ^2_p
b x m	Bir maddenin diğerine bireyin cevaplarındaki farklılık	σ^2_{bm}
b x p	Bireyin puanlanmasında puanlayıcılar arasındaki tutarsızlık	σ^2_{bp}
m x p	Bir maddeden diğerine puanlayıcı sıkılığı arasındaki farkın neden olduğu sabit etki	σ^2_{mp}
b x m x p, e	Artık varyans	$\sigma^2_{bmp,e}$

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

Brennan (2001)' a göre G kuramında bağıl değerlendirmeler için genellenebilirlik katsayısı hesaplanır. Bu nedenle öncelikle bağıl hata varyansının belirlenme şartı vardır. Buna göre bağıl değerlendirme üç çeşit hata kaynağından etkilenir. Bunları birey-madde ortak etkisi, birey-puanlayıcı ortak etkisi ve kalan varyansı olarak sıralanabilir.

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

Bağıl hata varyansı (22) (Brennan,2001)

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

Genellenebilirlik katsayısı (23) (Brennan,2001)

Genellenebilirlik kuramında mutlak değerlendirmeler için Phi katsayısı Φ kullanılır. Fakat bu hesaplamalar için öncelikle mutlak hata terimlerinin belirlenmesi şartı aranmaktadır. Mutlak değerlendirmelerin etkilendiği varyans bileşenleri sırasıyla maddelerden, puanlayıcılar, madde-puanlayıcı etkileşimi, birey-madde etkileşimi, birey-puanlayıcı etkileşimi ve kalan varyans olarak açıklanır (Brennan, 2001; Atılgan, 2004).

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

Mutlak hata varyansı (24) (Brennan,2001)

$$\Phi = \frac{\sigma_{b}^{2}}{\sigma_{b}^{2} + \sigma_{\Delta}^{2}}$$

Phi katsayısı (25) (Brennan,2001)

2.İLGİLİ ARAŞTIRMALAR

Bu bölümde araştırmanın konusu ile benzerlik gösteren, yurtiçi ve yurtdışında yürütülmüş akademik çalışmalar yer almaktadır.

2.1. Yurtiçinde Yapılan İlgili Araştırmalar

Atılgan (2004), "Genellenebilirlik Kuramı ve Çok Değişken Kaynaklı RaschModelinin Karşılaştırılması" adlı çalışmasında müzik öğretmenliği özel yetenek seçme sınavı verileri ile birden çok görev için bireylerin gözlenmesi ve puanlanması durumunda genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch modeli ile kestirilen istatistikleri karşılaştırmıştır. Bu çalışmalara 2002 yılında 499 öğrenci, 3 puanlayıcı katılmış ve araştırma 19 madde ile tamamlanmıştır. 2003 yılındaki 28 maddelik çalışmaya ise 689 öğrenci, 4 puanlayıcı katılmışlardır. Araştırmasonuçlarına göre, G kuramı ve çok değişkenlik kaynaklı Rasch yaklaşımlarıyla eldeedilen sonuçların kısmen tutarlı sonuçlar verdiği görülmüştür. Genellenebilirlikkuramının tek değişkenli ve çok değişkenli modellerin aynı ölçme durumu için altboyutlardan oluşan testlerde farklı sonuçlar ürettiği belirlenmiştir. Ayrıca G ve phi katsayılarınınmodelden etkilendiği ve alternatif karar çalışmalarıyla farklı puanlayıcı sayılarısenaryoları karşısında elde edilen G ve phi katsayılarının gerçek durumdakestirilenlerden farklı olduğu sonucuna varılmıştır.

Gündüz Sefer (2006), " Matematik Dersinde Problem Çözme Becerilerinin Dereceli Puanlama Anahtarı Kullanılarak Değerlendirilmesi" adlı çalışmasında 5. Sınıf öğrencilerinden biri deney diğeri kontrol grubu olarak belirlediği iki grup ve toplamda 42 öğrenci ile çalışmıştır. Bu yarı deneme modellerinden eşitlenmemiş kontrol gruplu araştırma deseni ile deney grubu öğrencilerinin dereceli puanlama anahtarı ile problem çözme becerilerinin diğer gruba göre daha fazla artış gösterip göstermediği belirlenmeye çalışılmıştır. T-testi ile yapılan analizler sonucunda deney grubunun problem çözme becerisi kontrol grubuna göre daha fazla artış gösterse de iki grubun ortalamaları arasında 0,05'lik manidarlık düzeyinde anlamlı farklılık ortaya çıkmamıştır. Ayrıca deney grubuna uygulanan anket çalışması

sonrasında öğrencilerin problem çözme becerilerinin dereceli puanlama anahtarı ile değerlendirilmesi konusunda olumlu görüşe sahip oldukları ve problem çözme aşamaları ile ilgili farkındalıklarının arttığı belirlenmiştir.

Erdem ise 2007 yılında yayınlanan çalışmasında dokuzuncu sınıf matematik öğrencilerinin genel izlenimle ve performansa dayalı olarak elde ettikleri sözlü puanların geçerliğini incelemiştir. Geçerlik çalışmasında başarı testi kullanan Erdem, bu başarı testini bütünsel ve analitik dereceli puanlama anahtarları ile puanlamıştır. 88 öğrencinin katıldığı çalışmada analitik puanlama ile verilen sözlü puanlarının geçerliğinin daha yüksek olduğu bulunmuştur. Ayrıca genel izlenimle verilen sözlü puanları ve bütünsel dereceli puanlama anahtarı kullanılarak verilen sözlü puanları arasında ve analitik dereceli puanlama anahtarı kullanılarak verilen sözlü puanları ile bütünsel dereceli puanlama anahtarı kullanılarak verilen sözlü puanları arasında manidar bir fark gözlenmemiştir.

Yelboğa (2007), ”Klasik Test Kuramı ve Genellenebilirlik Kuramına Göre Güvenirliğin Bir İş Performansı Ölçeği Üzerinde İncelenmesi” başlıklı çalışmasında klasik test kuramı ve genellenebilirlik kuramından elde edilen güvenirlik katsayılarını karşılaştırmıştır. Çalışmasını 2005 ve 2006 yıllarında Ankara’da faaliyet gösteren bir hizmet sektörünün 176 çalışanına, 32 soruluk bir ölçekle uygulamıştır. 3 puanlayıcı tarafından puanlanan ölçekler neticesinde İş Performansı ölçeğinin klasik test kuramı ve genellenebilirlik kuramının çok değişkenli modeline göre elde edilen katsayılarının uyumlu sonuçlar verdiğini ve araştırmada kullanılan iş performansı ölçeğinin geçerliği ve güvenirliği yüksek bir ölçme aracı olarak değerlendirilebileceği ifade edilmiştir.

Güler 2008 yılında “Klasik Test Kuramı, Genellenebilirlik Kuramı ve Rasch Modeli Üzerine Bir Araştırma” başlıklı tez araştırmasında, TIMMS 1999 açık uçlu matematik sorularından 24’ünü 203 öğrenciye uygulamış ve ardından dört puanlayıcı yanıtları, bütünsel dereceli puanlama anahtarına göre puanlamışlardır.Elde edilen puanların güvenirliği KTK’de Cronbach Alfa iç tutarlık katsayısı 0.92, Kendall’ın uyum istatistiği 0.52 olarak bulunurken, puanlayıcılar arası ilişki katsayıları 0.90 ve 0.97 arasında değişkenlik göstermiştir.G kuramında b x g x p deseninde ise; G katsayısı 0.92, phi katsayısı 0.90, ek olarak puanlayıcı değişkenlik kaynağının toplam varyansı açıklama yüzdesi 2,1 gibi düşük bir değer bulunmuştur. Çok değişkenlik kaynaklı Rasch ölçme modeline göre ise

puanlayıcılar arası güvenirlik 0.99 bulunmuştur. Araştırma neticesinde; Matematik başarısını ölçen ölçeklerden elde edilen puanların güvenilir olduğu ve farklı puanlayıcılar arasında uyumlu puanlamalar gerçekleştiği saptanmıştır.

Kasap (2008) " Dereceli Puanlama Anahtarı ve Puanlama Anahtarından Elde Edilen Puanların Karşılaştırılması" başlıklı çalışmasında puanlama anahtarı ile dereceli puanlama anahtarının karşılaştırılmasını amaçlamıştır. Öncelikle bu araştırmada lise 1. Sınıfta öğrenim gören 195 öğrenciye uygulanan 1. Matematik yazılıları öğretmenlere puanlama anahtarı ve analitik dereceli puanlama anahtarları kullanılarak puanlamaları sağlanmıştır. Yazılıdan bir hafta sonra soruların doğru cevapları öğrencilere çözülmüş ve dereceli puanlama anahtarları ile ilgili öğrencilere kısa bir ders verilmiştir. Ardından yazılı kâğıtları öğrencilere verilerek öğrencilerin kendilerini ve akranlarını değerlendirmeleri istenmiştir. Puanlama anahtarı ve dereceli puanlama anahtarı ile elde edilen puanların tutarlığı incelenmiştir. Ayrıca 56 öğrenciye dereceli puanlama anahtarı ile ilgili anket uygulanmıştır. Sonuçlara göre; öğretmenlerin dereceli puanlama anahtarı kullanarak verdikleri puanların ortalaması daha yüksek ve iki puanlama anahtarı ile elde edilen puanların tutarlığı yüksektir. Ayrıca öğrencilerin yaptıkları puanlamaların ortalamaları öğretmenlerinkinden daha yüksek olsa da puanlamalar arasındaki uyum yok denecek kadar azdır. Bununla birlikte, yapılan anketin sonucuna göre öğrenciler dereceli puanlama anahtarları sorunun her basamağını değerlendirdiğinden, bu puanlama anahtarına oldukça olumlu bakmaktadırlar.

Ömür(2009)'ün "Dereceli Puanlama Anahtarıyla, Genel İzlenimle Ve İkili karşılaştırmalar Yöntemiyle Yapılan Değerlendirmelerin Karşılaştırılması" başlıklı tezinde üç farklı değerlendirme biçiminin psikometrik nitelikleri ele alınmıştır. Kompozisyonların değerlendirilmesi 194 Türkçe öğretmeni ile dereceli puanlama anahtarının güvenirlik çalışması ise 21 Türkçe öğretmeni ile yapılmıştır. Puanlayıcılar kompozisyonları önce ikili karşılaştırma, sonra genel izlenime dayalı ve en son da dereceli puanlama anahtarı ile değerlendirmişlerdir. Kompozisyonlar dereceli puanlama anahtarı ile 1–2 hafta Aralıkla 2 defa puanlanmış, iki değerlendirme arasında pozitif yönde manidar korelasyon belirlenmiştir. Genel izlenim ve dereceli puanlama anahtarı ile verilen puanlar arasında farklılık olduğu ve genel izlenim ve dereceli puanlama anahtarı ile puanlanmada puanlayıcı güvenirliğine bakılmış, puanlayıcıların tutarlı puanlama yaptığı belirlenmiştir.

Deliceoğlu (2009), “Futbol Yetilerine İlişkin Dereceleme Ölçeğinin Genellenebilirlikve Klasik Test Kuramına Dayalı Güvenirliklerinin Karşılaştırılması” adlıçalışmasında Ankara ilindeki profesyonel spor kulüplerinfutbol takımlarının alt yapılarındaki toplam 72 futbolcunun teknik yetilerinin tespitedilmesinde kullanılan Futbol Yetilerine İlişkin Dereceleme Ölçeğini kullanmıştır. Bu ölçeğin 4 puanlayıcı ile puanlanmasından elde edilen puanlar, Klasik Test Kuramı (KTK) ve Genellenebilirlik Kuramı (G) çerçevesinde belirlenen güvenirlik katsayılarına göre kıyaslanmıştır. Ölçeğe ilişkin G katsayısı 0,82 olup, alt boyutlarailişkin Cronbach Alfa katsayıları 0,71 ve 0,90 arasında değişmektedir. Phikatsayısı 0,77, Kendall W Güvenirlik katsayısı ise boyutlar için 0,72 ile 0,83arasında değişmektedir.

Parlak (2010) “Öğrenci Performansının Belirlenmesinde Puanlama Anahtarı Ve Dereceli Puanlama Anahtarının Karşılaştırılması” başlıklı çalışmasında 70 öğrenci ve 6 öğretmen ile çalışmıştır. Öğretmenlerin puanlamaları arasında yüksek bir tutarlık olduğu gözlenmiştir. Bununla birlikte puanlama anahtarı ile verilen puanlamaların ortalamasının, dereceli puanlama anahtarı ile verilen puanların ortalamasından düşük olduğu belirlenmiştir. Dereceli puanlama anahtarları hakkında öğretmenlerden alınan görüşler doğrultusunda; dereceli puanlama anahtarları ile puanlamanın daha nesnel ve öğrenci başarısı üzerindeki etkisi dahayüksek olduğu , hazırlamanın daha zor ve daha fazla zaman aldığı belirlenmiştir.

Şanlı (2010) “Bilimsel Süreç Becerilerinin Ölçülmesinde Bütünsel ve AnalitikPuanlama Anahtarlarının Güvenirliklerinin Karşılaştırılması” adlı çalışmasında bilimsel süreç becerilerininölçülmesinde iki puanlayıcının bütünsel ve analitik puanlama anahtarları kullanarakvermiş olduğu puanlarda puanlayıcılar arası güvenirlikleri Sperman Brown SıraFarkları Katsayısı ile hesaplamıştır. Analizler sonucunda analitik dereceli puanlama anahtarı ile elde edilen sonuçların tutarlığı, bütünsel dereceli puanlama anahtarı ile elde edilen sonuçların tutarlığından daha yüksek çıkmıştır.

Bıkmaz (2011), “Üst Düzey Zihinsel Özelliklerin Ölçülmesinde Puanlayıcılar ArasıGüvenirlik Belirleme Tekniklerinin Karşılaştırılması” adlı çalışmasında Fen ve Teknoloji dersi gören 50 öğrenciye verdiği performans görevini 10 farklı puanlayıcının iki farklı puanlama anahtarı ile puanlaması neticesinde gerçekleştirmiştir. Puanlayıcılar arası güvenirliği Kappa İstatistiği ve

Krippendorff'un Alfa tekniği ile belirleyen Bıkmaz, araştırma sonucunda Kappa ve KrippendorffAlfa tekniklerinin birbirine yakın sonuçlar verdiğini görmüştür. Bununla birlikte, Log-lineer analiz tekniğinin ise değişkenler arasındaki etkileşimleri ve uyumsuzluk kaynağını gösterdiği ve daha geniş bilgi sağladığını belirlemiştir.

Öztürk (2011), "Voleybol Becerileri Gözlem Formu İle Elde Edilen PuanlarınGenellenebilirlik ve Klasik Test Kuramına Göre Karşılaştırılması" adlı çalışmasında,voleybolbecerilerine ilişkin dereceleme ölçeğinden elde edilen ölçmelerin; klasik test kuramıve genellenebilirlik kuramı ile elde edilen güvenirlik katsayılarını ve belirlemiş ve karşılaştırmıştır. Genellenebilirlik kuramında bx g x p deseni sonuçları incelendiğinde en yüksek varyans bileşenin g x p ortaketkisi için kestirilen varyans bileşeni olduğu görülmüştür. Güvenirlik katsayıları iç ölçütlere göre incelendiğindeG, phi katsayısı, Kendall uyuşum katsayıları veCronbach Alfa katsayılarının beklenenden düşük olduğu belirlenmiştir. Analizler neticesinde klasik test kuramında ve genellenebilirlik kuramının çokdeğişkenli modelle elde edilen güvenirlik katsayılarınınbirbirleriyleuyumlu olduğu görülmüştür. Bu araştırma sonucunda, Voleybol becerilerine ilişkin dereceleme ölçeğiningüvenilir bir ölçek olarak değerlendirilemeyeceği ifade edilmiştir.

Güler (2011) " Rastgele Veriler Üzerinde Genellenebilirlik Kuramı ve Klasik Test Kuramı'na Göre Güvenirliğin Karşılaştırılması" başlıklı çalışmasında 125 öğrencinin 18 maddeye verdiği cevapları 4 farklı puanlayıcı puanlamıştır. Değişkenlik kaynağının maddeler olduğu (b x m) çapraz desen için hesaplanan G katsayısı ile Cronbach Alfa değerleri her puanlayıcı için ayrı ayrı hesaplanmış ve çok düşük değerler elde edilmiştir. Değişkenlik kaynağının maddeler ve puanlayıcılar olduğu tümüyle çapraz desen (b x m x p) için Genellenebilirlik Kuramına dayalı G- Katsayısı ve Phi katsayısı sırasıyla 0.457 ve 0. 456 olarak bulunmuştur.

Büyükkıdık (2012) "Problem Çözme Becerisinin Değerlendirilmesinde Puanlayıcılar Arası Güvenirliğin Klasik Test Kuramı Ve Genellenebilirlik Kuramına Göre Karşılaştırılması" başlıklı çalışmasında matematik problemi çözme becerilerinin değerlendirilmesine yönelik olarak hazırlanan iki adet performans görevi ile araştırmasını tamamlamıştır. Bu performans görevlerini analitik ve bütünsel dereceli puanlama anahtarlarıyla, dört puanlayıcı tarafından puanlanmasından elde edilen puanlara klasik test kuramı ve genellenebilirlik

kuramı ile puanlayıcılar arası güvenirlik sınanması yapılmıştır. Araştırmanın sonucunda genellenebilirlik kuramından elde edilen katsayıların, klasik test kuramından elde edilen katsayılara göre göreceli olarak daha yüksek olduğu, G kuramı ile elde edilen bilgilerin daha detaylı olduğu ve her iki kuramdan elde edilen güvenirlik katsayılarının yüksek olduğu tespit edilmiştir. Analitik dereceli ve bütünsel dereceli puanlama anahtarları ile elde edilen güvenirlik katsayıları yüksek olsa da, analitik dereceli puanlama anahtarından elde edilen güvenirlik katsayısının göreceli olarak daha yüksek olduğu görülmüştür.

Özmen Hızarcıoğlu (2013) "Problem Çözme Sürecinde Dereceli Puanlama Anahtarı Kullanımında Puanlayıcı Uyumunun İncelenmesi" başlıklı araştırmasında 6. Sınıf seviyesinde 4 adet açık uçlu problem ölçeğini kullanmıştır. Bu ölçek 27 öğrenciye uygulanmış, problem yanıtları 15 farklı matematik öğretmeni tarafından puanlanmıştır. Öğretmenler problemleri öncelikle dereceli puanlama anahtarı kullanmadan, iki hafta ardından dereceli puanlama anahtarı kullanarak puanlamışlardır. Ayrıca öğrenciler kendi yanıtlarını dereceli puanlama anahtarı kullanarak puanlamışlardır. Öğretmenlerin kendi puanlamaları arasındaki uyumu belirlemek için Kendall'ın W testi kullanılmıştır. Öğretmenlerin dereceli puanlama anahtarı kullanarak yaptıkları puanlamaların tutarlığı yüksek, kullanmadan gerçekleştirdikleri puanlamaların tutarlığı ise düşük çıkmıştır. Ayrıca öğrencilerin öz değerlendirme çerçevesinde kendilerine verdikleri puanlar arasında yüksek uyum gözlenmemiştir. Son olarak öğretmenlerin dereceli puanlama anahtarları hakkındaki görüşleri neticesinde, dereceli puanlama kullanmayı bildikleri fakat hazırlama, uygulama ve çeşitleri hakkında yetersiz oldukları ifade edilmiştir.

Yapılan çalışmalar incelendiğinde, bütünsel ve analitik dereceli puanlama anahtarının karşılaştırıldığı çalışmalar ile bütünsel ve analitik dereceli puanlama anahtarları ile elde edilen ölçümlerin Klasik test kuramı ve Genellenebilirlik kuramları ile güvenirlik saptamaları yapılmıştır. Çalışmalarda klasik ve Genellenebilirlik kuramlarını bir arada kullanan ve kıyaslayan çok az araştırmaya rastlanmıştır. Alanyazın incelendiğinde açık uçlu matematik testinin klasik test kuramına göre puanlayıcılar arasındaki güvenirlik ve tutarlığın belirlenmesi ile ilgili yapılmış ve bu çalışma ile birebir örtüşen bir çalışmanın bulunmadığı tespit edilmiştir.

2.2. Yurtdışında Yapılan İlgili Araştırmalar

Lane ve Sabers (1989) yazılı sınavlarda genellenebilirlik kuramını kullanarak puanlama güvenirliğini belirlemeye çalışmışlardır. Bu amaçla farklı öğrenim kademelerinden 15 öğrenci verilen bir konuda kompozisyon yazmıştır. Bu kompozisyonlar belirlenen 4 kategori altında 8 farklı puanlayıcı tarafından puanlanmıştır. Elde edilen veriler üzerinde "öğrenci x kategori x puanlayıcı" çaprazlanmış desenine göre genellenebilirlik çalışması yapılmış, bağıl ve mutlak hata varyansları analiz edilmiştir. Araştırma neticesinde puanlayıcı sayısı arttıkça standart sapmanın azaldığı, grup değerlendirmeleri için bir puanlayıcının uygun olduğu ve bireysel değerlendirmeler için puanlayıcı sayısının arttırılması gerektiği tespit edilmiştir.

Stuhlmann ve diğerleri (1999) genellenebilirlik çalışmasıile öğretmenlerin puanlama eğitimini almalarının, dereceli puanlama anahtarıile yapılan puanlamalarda güvenirliğe olan etkisini araştırmışlardır. Bu araştırmada 40 okulöncesi ve birinci sınıf öğrencisi 20 yazma görevini yerine getirmiş, 23'ü puanlamaeğitimi almış, 17'si puanlama eğitimi almamış toplam 40 puanlayıcı öğrencileripuanlamıştır. Çalışmadan elde edilen bulgular neticesinde öğretmenlerin puanlama eğitimialmalarının puanlama güvenirliğini arttırmadığı ortaya çıkmıştır.

Swartz ve diğerleri (1999) "Analitik ve Bütünsel Puanlama Yöntemleriyle Elde Edilen Yazma Puanlarının Güvenirliğinin Genellenebilirlik Kuramı Kullanılarak Hesaplanması" adlı araştırmalarında farklı özellikleri olan 251 öğrencinin yazdığı hikâye havuzundan yirmisini rasgele seçmiş ve bunları analitik ve bütünsel dereceli puanlama anahtarlarıyla puanlayarak G ve K çalışmaları yapmışlardır. Çalışma neticesinde; iyi eğitim almış puanlayıcılar önemli bir değişkenlik kaynağı olmazken, bireysel farklılıklar ve birey ve puanlayıcı etkileşiminin değişkenliğe kaynaklık ettikleri görülmüştür. Ayrıca çalışma bulgularından bir tanesi de puanlayıcı sayısı ile kestirilen güvenirliğin pozitif ilişkili olduğudur.

Goodrich'in 2001 yılında yaptığı çalışmada dereceli puanlama anahtarının öğrencilerin yazılı kompozisyon becerileri üzerindeki etkileri araştırılmıştır. 3 er kompozisyon yazan gruplardan birine dereceli puanlama anahtarları ile öğretici

dönütler sunulurken, kontrol grubuna verilmemiştir. Araştırmanın sonucuna göre dönüt verilen grubun becerilerini geliştirmelerinin pozitif doğrultuda olduğu belirlenmiştir. Ayrıca öğretmenlerin dereceli puanlama anahtarlarıyla değerlendirmeyi daha nesnel bulduklarına dair ortak görüşleri olduğu belirtilmiştir.

Rae ve Hyland (2001) "Bir Adam Çiz Testi İçin Koppitz'in Puanlama SisteminiKlasik Test Kuramı ve Genellenebilirlik Kuramıyla Analizi" başlıklı araştırmalarında 85 çocuğun (8–9 yaş) çizimlerinde dört puanlayıcının ve iki durumun güvenirliğe etkisiniklasik test kuramı ve genellenebilirlik kuramı ile belirlemeye çalışmışlardır. Araştırma neticesinde Genellenebilirlikkatsayısının 0.47 olduğunu, Klasik test kuramında puanlayıcıgüvenirliğini G kuramına göre yüksek, test tekrar test güvenirliğini ise düşükbulmuşlardır. Sonuç itibariyle; çocukların birkaç durumda buteste tabii tutulmalarının beklenen düzeyde yüksek güvenirlikleri sağlayacağı belirtilmiştir.

Hafner (2003) yayımladığıaraştırmasında dereceli puanlama anahtarının geçerliliğini incelemiştir. Araştırmasını biyoloji bölümünde öğrenim gören 107 kişilik üçüncü sınıf öğrencileri üzerinde yürütmüştür. Biyoloji dersinde, her yılın sonunda öğrenciler sunum yapmış, yapılan sunumlar öğretmen ve öğrenciler tarafından dereceli puanlama anahtarı ile puanlanmıştır. Çalışmada elde edilen bulgular neticesinde öğretmeninverdiği puanlar ile öğrencilerin birbirlerine verdikleri puanların uyumlu olduğu anlaşılmıştır.

Smith ve Kulikowich (2004) "Bir Kompleks Problem Çözme BecerisiDeğerlendirmesi Kullanılarak Genellenebilirlik Kuramı ve Çok Değişkenli RaschÖlçümlerinin Kullanılması" başlıklı çalışmalarına, 4. Sınıfta öğrenim gören 44 öğrenci ve iki puanlayıcı katılmıştır. İki farklı şekilde puanlamanın yapıldığı çalışmada Rasch modeli ve G kuramı ile güvenirlik test edilmiştir. Araştırma neticesinde, çok değişkenli Rasch ölçümlerinin ve G kuramının değişkenlik kaynaklarının göreceli büyüklüklerinin değişimlerinde benzer, her bir ölçüm tekniğinin değişkenlik kaynaklarını ise farklı şekilde değerlendirdikleri görülmüştür.

Lei ve diğerleri (2007, Akt: Güler, 2008) güvenirlik kestirim çalışmalarında Genellenebilirlik kuramının sınırlılıklarını araştırmışlardır. Bu çalışmada gözleme dayalı ölçmeler ile elde edilen verilere tekrarlı ölçümler yapılarak ve küçük örneklemler kullanılarak ulaşıldığının altı çizilmiş ve bu verilerin güvenirlik

katsayılarını belirlemenin zor olduğu ifade edilmiştir. Özellikle tek bir bireyin yer aldığı gözlem çalışmalarında, Genellenebilirlik kuramının çok iyi açıklandığı şartlar altında kullanılabileceği vurgulanmıştır. Bunun için iki sayısal örnek kullanan araştırmacılar, ilk örnekte, tek bir denek üzerinde yapılan ölçmelerde, ölçme objesini birey değil, ölçme durumları, değişkenlik kaynağını ise gözlem yapan puanlayıcılar olduğu üzerinde durulmuştur. Çalışmanın ikinci örneğinde ise tek bir denek yerine çok küçük örneklemlerin seçildiği durumlarda da ölçme objesinin ölçme durumları olması gerektiği vurgulanmış ve 4 birey 10 farklı ölçme durumu üzerinde Genellenebilirlik çalışması yapılmıştır. Bu örnekte de ilk örneğe paralel olarak puanlayıcı sayısının artırılmasının Genellenebilirlik ve Güvenirlik katsayılarında artış sağladığı görülmüştür.

Christ ve diğerleri (2010) "Doğrudan Davranış Puanlama (DBR): Puanlayıcılar ve Gözlemciler Arasında Bağımlılık ve Genellenebilirlik" başlıklı çalışmalarında altı okul öncesi çocuğunun çözülemeyen bir lego ile uğraşmalarını içeren üçer dakikalık videoları, davranışları doğrudan gözlemleyen 125 puanlayıcı tarafından puanlanarak çaprazlanmış desenlerle Genellenebilirlik çalışmaları yapılmıştır. 6'lı, 10'lu ve 14'lü metrik ölçeklerle yapılan çalışmada 10'lu metrik ölçekte elde edilen puanların daha etkili ve duyarlı sonuçlar verdiğini bulmuşlardır.

Yurtdışında yapılan çalışmalar incelendiğinde açık uçlu matematik testinin klasik test kuramına ve genellenebilirlik kuramına göre puanlayıcılar arasındaki güvenirlik ve tutarlığın belirlenmesi ile ilgili yapılmış ve bu çalışma ile birebir örtüşen bir çalışmanın bulunmadığı tespit edilmiştir.

Tüm bu çalışmalar incelendiğinde klasik test kuramı ve genellenebilirlik kuramından puanlayıcılar arası tutarlığın ayrı yarı ya da birlikte farklı yöntemlere göre karşılaştırıldığı ancak içerik ve yöntem açısından farklılıklar taşıdığı düşünülmektedir.

3. YÖNTEM

Bu bölümde araştırmanın yöntemi, araştırma yapılan çalışma grubu, veri toplama araçları, veri toplama aşamaları, verilerin analizi ve araştırmanın iç ve dış geçerliği ile ilgili bilgilere yer verilmiştir.

3.1. Araştırmanın Türü

Çeşitli öğrenci gruplarının başarılarını belirlemek, öğretmenlerin, yöneticilerin ya da danışmanların davranışlarını tanımlamak için yapılan çalışmalara betimsel araştırmalar denir (Büyüköztürk ve diğerleri, 2012). Betimsel çalışmalardaki amaç var olan durumu belirlemek ve tanımlamak olarak düşünülebilir. Bu araştırmada da KTK ve GK yöntemlerini kıyaslamak, sınırlılıklarını ve kullanım sebeplerini belirlemek ile birlikte iki farklı dereceli puanlama anahtarıyla puanlama yapan öğretmenler arasındaki tutarlık ölçüsü belirlenmeye çalışılmıştır.Araştırma var olan durumu belirlemek ve gerekli saptamalar yapılması için oluşturulduğundan bu yönüyle betimsel bir araştırmadır.

3.2. Çalışma Grubu

Çalışmanın yapılması için hazırlanan açık uçlu sorular Pursaklar İlçesinde örneklem olarak seçilen bir ortaokulda 7. Sınıfında eğitim ve öğretim gören 84 öğrenciye uygulanmıştır. Elde edilen açık uçlu cevap kâğıtları yine Ankara'nın çeşitli okul ve özel eğitim kurumlarında çalışan 5 öğretmen tarafından iki farklı puanlama anahtarıyla puanlanmıştır. İlk puanlama analitik dereceli puanlama anahtarı kullanılarak yapılmıştır. İkinci puanlama ise aradan geçen 20-25 günün sonunda başlamış ve bütünsel dereceli puanlama kullanılarak elde edilmiştir. Öğretmenlerin iki farklı puanlamaları arasında 20-25 gün bırakılması bir önceki puanlamalarını hatırlamamaları ve ikinci puanlamanın geçerliğini arttırmak için alınmış bir önlemdir. Öğrenci katılımcıların demografik bilgileri aşağıdaki tabloda gösterilmiştir.

Çizelge 3.1: Öğrenci betimsel bilgileri ve frekansları

Cinsiyet	Frekans	Yüzde (%)
Kız	45	53,6
Erkek	39	46,4
Toplam	84	100

Katılımın diğer öğesi olan puanlayıcı öğretmenler ise 1- 10 yıl arası görev yapmakta olan alan uzmanı öğretmenlerden oluşturulmuştur.

3.3. Veri Toplama Araçları

Bu çalışmada veri toplama araçları olarak 7. Sınıf öğrencilerin tam sayılar ve rasyonel sayılarda problem çözme becerisini ölçen 6 soruluk açık uçlu yazılı formu, bu formun değerlendirilmesi için puanlayıcılar tarafından iki farklı puanlamada ihtiyaç duyulan analitik dereceli ve bütünsel dereceli puanlama anahtarları kullanılmıştır. Ayrıca form oluşturma sürecinde uzman kanılarını ve sorular hakkındaki eleştirilerini öğrenmek için anket ve çalışmaöncesinde çalışmaya katılan öğrenciler içinveli onayı ve öğretmenler için öğretmen gönüllü katılım formları hazırlanmıştır.

3.3.1. Açık Uçlu Sorulardan Oluşan Yazılı Formu

Öğrencilerin tam ve rasyonel sayılardaki problem çözme becerilerini ölçen bu uygulamanın oluşturulması için, farklı yayınevlerine ait 7. sınıf matematik kaynak kitapları taranmış, soruların seçilmesi, düzenlenmesi ve son halinin verilmesi için tüm süreçlerde uzman görüşlerinden yararlanılmıştır. Sorular oluşturulmadan Bloom taksonomisine göre belirtke tablosu hazırlanmıştır.

Elde edilen yazılının geçerliği ve güvenirliği alınan uzman görüşlerinden sonra belirlenmiş, son hali verilmiş ve uygulanmıştır. Bu aşamada 8 uzmanın görüşleri alınmıştır. Uzmanların görüşlerini almak için araştırmacının oluşturduğu mini bir anketten de faydalanılmıştır. Bu uzmanların görüşleri ve anketten elde edilen genel fikirler doğrultusunda oluşturulan forma son hali verilmiştir. Yazılıda uygulama süresinin sınırlılığı ve katılımcılardan daha etkin katılım almak adına soru sayısı sınırlı tutulmuştur. Ayrıca öğretmenlerin puanlama yaparken yaşayacakları sıkıntıları da en aza indirgemek için böyle bir yol tercih edilmiştir.

Açık uçlu sorulardan oluşan yazılı formu, Ek-1'de, uzman görüşlerini almak için oluşturulan küçük anket ise Ek-2'de yer almaktadır.

3.3.2. Analitik Dereceli ve Bütünsel Dereceli Puanlama Anahtarları

Analitik dereceli ve bütünsel dereceli puanlama anahtarları oluştururken ölçülen kazanımlar ve elde edilmek istenen süreçler belirlenmiştir. Öğrencilerin problem çözme sürecinde attıkları her olası adım ile beklentiler karşılaştırılmıştır. Bu karşılaştırma neticesinde ortak karar olarak beklenen davranışların önem ve beklenti düzeyine göre puanlaması yapılmıştır. Beklentilerin tümünü karşılayan eksiksiz/kusursuz davranışlardan, beklentiyi karşılamanın çok uzağında kalmış, eksik ve problemin çözüm sürecine katkıda bulunmayacak hatalı adımlara kadar kademeli olarak belirlenmiş ve derecelendirmeleri yapılmıştır.

Bu anahtarların oluşturulmasında puanlama yapan öğretmenlerden ikisi ile ölçme değerlendirme alanında çalışan 1 kişinin fikirlerinden yola çıkılmıştır. Oluşturulan formların değerlendirilmesi için yine matematik alanında ve ölçme alanında uzman olan iki kişinin görüşleri alınmıştır. Dönütlerden sonra yapılan değişiklikler ile dereceli puanlama anahtarları kullanıma hazır hale getirilmiştir. Oluşturulan ve kullanılan anahtarlar Ek-6 ve Ek-7'de yer almaktadır.

3.3.3. Veli Onay Formu ve Gönüllü Katılım Formu

Öncesinde de bahsedildiği gibi çalışma tamamen gönüllülük ilkesine dayanmaktadır. Çalışmanın etik ve bilimsel değerler çerçevesinde kalması ve elde edilen neticelerin genellenebilir ve savunulabilir olması için çalışmaya katılan öğrencilerin velilerin izni "Veli Onay Formu" ile alınmıştır. Ayrıca öğrenci cevaplarını puanlayacak olan öğretmenlerden "Gönüllü Katılım Formu" istenmiştir. Bu formlar çalışmanın gönüllülük esasına dayandığını, verilerin çalışmadan sonra imha edileceğini, gizliliğin korunacağını, katılımcı kimliklerinin üçüncü kişiler ile paylaşılmayacağını ve çalışmanın istenildiği zaman bırakılabileceğini garanti etmektedir. Bu formlar da Ek-4 ve Ek-5'de yer almaktadır.

3.4. Verilerin Analizi

Araştırmanın amacı doğrultusunda, alt problemlere çözüm bulabilmek için analitik ve bütünsel dereceli puanlama anahtarlarından elde edilen puanlar klasik test kuramından ve genellenebilirlik kuramından farklı tutarlık belirleme tekniklerine göre incelenmiştir. Kullanılan yöntemler aşağıda başlıklar halinde verilmiştir.

3.4.1. Pearson Momentler Çarpımı ve Spearman Sıra Farkları Korelasyon Katsayısı

İlk olarak puanlayıcıların yaptıkları farklı puanlamalarda puanlar arasındaki ilişkinin yönünü ve derecesini belirlemek için Pearson momentler çarpımı ve Spearman sıra farkları korelasyon katsayıları hesaplanmıştır. Böylelikle puanlayıcıların farklı zamanlarda bütünsel dereceli ve analitik dereceli puanlama anahtarı kullanarak yaptıkları puanlamalar arasındaki ilişki belirlenmeye çalışılmıştır. Bu hesaplamaların yapılabilmesi için SPSS 15.0 for Windows paket programından yararlanılmıştır. Bu istatistik birinci alt problemin a şıkkının cevaplanması için kullanılmıştır.

3.4.2. Kappa İstatistiği

Kappa istatistiği iki puanlayıcı arasındaki uyumu belirlemek için kullanılır. SPSS 15.0 for Windows paket programında yer alan Kappa istatistiği iki puanlayıcı arasındaki uyumun belirlenmesine olanak tanır. Çalışmada analitik ve bütünsel dereceli puanlama anahtarlarını kullanan 5 puanlayıcı arasındaki uyumu belirlemek için “SPSS syntax (mkappasc. sps)” komut dosyası kullanılmıştır.Böylelikle analitik ve bütünsel dereceli puanlama anahtarları ile yapılan puanlamaların tutarlığı kestirilerek birincialt problemin b şıkkına cevap aranmıştır.

3.4.3. KrippendorffAlfa Tekniği

Puanlamalar sonucu elde edilen toplam puanlar ve her bir alt kategoriden elde edilen puanlar için ayrı ayrı hesaplanmıştır. Hesaplanması için SPSS 15.0 for Windows paket programı ve SPSS syntax (kAlfa. sps) kullanılmıştır. Böylelikle ilk alt problemin c şıkkı cevaplanmıştır.

3.4.4. Genellenebilirlik Kuramı

Birey (b), madde (m) ve puanlayıcı (p) değişkenin çapraz tasarlandığı b x m x p deseninin genellenebilirlik kuramı sonuçlarını belirlemek ve kestirilen varyans bileşenlerini ve toplam varyansı açıklama yüzdelerinin bütünsel dereceli ve analitik dereceli puanlama anahtarlarından nasıl etkilendiği gözlemlemek için EduG 6.1 programından yararlanılmıştır. Ayrıca bu program ile KTK ve G-kuramı ile elde edilen güvenirlik katsayıları arasında uyum derecelerinin de belirlenmesi hedeflenmiştir. Böylelikle 2. alt probleme cevap aranmıştır.

Tüm analizlerin ardından elde edilen bulgular karşılaştırılarak 3. Alt problem olan Klasik Test Kuramına ve Genellenebilirlik Kuramına göre elde edilen güvenirlik katsayıları tutarlı mıdır sorusuna cevap aranmıştır.

3.5. Araştırmanın Geçerliği

3.5.1.Araştırmanın İç Geçerliği

Öncelikle iç geçerliği sağlamak için, deneklerin seçiminde yansız davranılmıştır. Ayrıca çalışmanın yürütülmesi için gönüllülük ilkesi esas alınmıştır. Bununla birlikte çalışmanın yapıldığı grup genişliği çalışmanın devam ettirilmesi için yeterlidir.Uygulamanın geliştirilmesi için 7. sınıf matematik kaynak kitapları taranmış, soruların seçilmesi, düzenlenmesi ve son halinin verilmesi için soruları oluşturma sürecinde matematik eğitimi alanı ve eğitim bilimleri ölçme ve değerlendirme alanı uzmanlarının görüşleri yapılandırılmış ve yapılandırılmamış formlar yoluyla alınmıştır. Ardından sorulara son hali verilerek uygulamaya hazır hale getirilmiştir. Analitik dereceli ve bütünsel dereceli puanlama anahtarları da yine alanında yetkin kişilerin yardımlarıyla oluşturulmuştur.Puanlayıcıların iki farklı puanlama anahtarı ile yaptıkları puanlamalar arasında zamanın puanlama üzerindeki etkisini en aza indirmek için iki puanlama arasında 20-25 günlük süre bırakılmıştır.

3.5.2 Araştırmanın Dış Geçerliği

Araştırmanın dış geçerliğini sağlamak için dış geçerliği olumsuz etkileyen faktörlerin rolü en aza indirgenmeye çalışılmıştır. Dış geçerliği etkileyen faktörler şöyle sıralanabilir (Campbell ve Stanley, 1963, ss. 175–176, Akt: Karasar, 2012).

- Ölçme – bağımsız değişken etkileşimi
- Yanlı seçim- bağımsız değişken etkileşimi
- Deneme tepkisi
- Bağımsız değişkenlerin etkisi

Araştırmaya katılan gönüllülerin seçiminde yansızlık ilkesi gözetilmiştir. Ayrıca puanlayıcıların genel puanlama alışkanlıklarının etkisinin azaltmak için, puanlama öncesinde bir eğitime tabi tutulmuşlardır. Atılan tüm adımlar ve alınan önlemler ile çalışmanın iç geçerliği ve dış geçerliği ile iç-dış geçerlik dengesinin sağlandığı varsayılmaktadır.

4. BULGULAR VE TARTIŞMA

Bu bölümde, her bir alt problemeilişkin araştırma bulguları ve bu bulgularla ilgili yorumlar yer almaktadır. Ayrıca alt problemler için elde edilen bulgular ve yorumlara geçmeden önce, uygulamaya ait betimsel istatistiklere de yer verilmiştir.

4.1. Betimsel İstatistikler

Puanlayıcıların iki farklı dereceli puanlama anahtarı kullanarak verdikleri puanlara ait betimsel istatistikler Çizelge 4.1 ve Çizelge 4.2 'de verilmiştir.

Çizelge 4.1: ADPA'na Göre Yapılan Puanlamalara Ait Betimsel İstatistikler

Puanlayıcı	Minimum	Maksimum	Ortalama	Std. Sapma	Varyans	Çarpıklık	Basıklık
1. puanlayıcı	6	30	21,38	5,657	31,998	-,550	-,367
2. puanlayıcı	8	30	21,27	5,478	30,008	-,558	-,574
3.puanlayıcı	5	30	20,76	6,261	39,196	-,540	-,580
4.puanlayıcı	6	29	20,36	5,658	32,015	-,446	-,741
5.puanlayıcı	8	28	20,08	5,130	26,318	-,514	-,611

Çizelge 4.1 incelendiğinde analitik dereceli puanlama anahtarına göre yapılan puanlamalar neticesinde en yüksek ortalama 1. puanlayıcıya ait olup (21,4), en düşük ortalama 5. puanlayıcıya aittir (20,1). Bağımlı gruplar t-testi yapıldığında bu puanlar arasında 0.01 düzeyinde manidar bir fark bulunmuştur (t=6,178, sd=83, p=0,002). Bununla birlikte tüm puanlayıcıların yaptıkları değerlendirmeler ile elde edilen puanların çarpıklık ve basıklık katsayıları -1 ile +1 arasında değişmektedir. Çarpıklık katsayıları tüm puanlayıcılar için negatif değerler almıştır. Negatif değerler dağılımın sola çarpık olduğunu ifade etmektedir. Basıklık değerleri de çarpıklık değerleri gibi negatif değerler ile gösterilmiştir. Negatif basıklık değerleri dağılımın normalden daha basık olduğunu göstermektedir. Ayrıca çizilen sütun grafiği ve Q-Q yerleşim eğrisi ile puanların normal dağılıma sahip olduğu söylenebilir.

Çizelge 4.2: BDPA'na Göre Yapılan Puanlamalara Ait Betimsel İstatistikler

Puanlayıcı	Minimum	Maksimum	Ortalama	Std. Sapma	Varyans	Çarpıklık	Basıklık
1. puanlayıcı	5	29	19,44	5,535	30,635	-,408	-,574
2. puanlayıcı	10	28	20,17	5,232	27,369	-,409	-,890
3.puanlayıcı	6	29	20,20	6,058	36,693	-,593	-,546
4.puanlayıcı	8	29	19,52	5,465	29,867	-,294	-,715
5.puanlayıcı	8	29	19,56	4,988	24,876	-,424	-,586

Çizelge 4.2 incelendiğinde bütünsel dereceli puanlama anahtarına göre yapılan puanlamalar neticesinde en yüksek ortalama 3. puanlayıcıya ait olup (20,2), en düşük ortalama 1. puanlayıcıya aittir (19,4). Bağımlı gruplar t-testi yapıldığında bu puanlar arasında 0.01 düzeyinde manidar bir fark bulunmuştur (t=-3,653, sd=83, p=0,003).Bununla birlikte tüm puanlayıcıların yaptıkları değerlendirmeler ile elde edilen puanların çarpıklık ve basıklık katsayıları -1 ile +1 arasında değişmektedir. Ayrıca çizilen sütun grafiği ve Q-Q yerleşim eğrisi ile puanların normal dağılıma sahip olduğu söylenebilir.

Çizelge 4.3'de de öğrencilerin iki farklı puanlama anahtarlarından elde edilen puanlarının ortalamalarına ait betimsel istatistiklere yer verilmiştir.

Çizelge 4.3: ADPA ve BDPA'na Göre Yapılan Puanlamaların Ortalamalarına Ait Betimsel İstatistikler

	Analitik Dereceli Puanlama Anahtarı	Bütünsel Dereceli Puanlama Anahtarı
Ortalama	20,7714	19,7786
Ortanca	21,7000	20,700
Varyans	30,860	28,256
Std. Sapma	5,55517	5,3156
Minumum	7,00	8,20
Maksimum	29,00	28,60
Çarpıklık	-,530	-,455
Basıklık	-,619	-,728

ADPA ve BDPA'na göre yapılan puanların ortalamalarına ait betimsel istatistikler, her bir puanlayıcı için hesaplanan betimsel istatistiklerle benzerlik göstermektedir. Puanlayıcıların yaptıkları puanlamalarla elde edilen puanların her bir öğrenciye göre ortalaması alınmış ve soruların güvenirliği Cronbach Alfa kullanılarak hesaplanmıştır. Buna göre Cronbach Alfa değeri α=0,86 olarak bulunmuştur. Bu da puanlayıcıların dereceli puanlama anahtarlarıyla ne kadar tutarlı puanlar

verdiğini göstermektedir. Baykul'a (2010) göre bu değer yüksek iç-tutarlığı işaret etmektedir.Puanlayıcıların, farklı dereceli puanlama anahtarıkullanarak verdikleri puanların Cronbach Alfadeğerleri de aşağıdaki Çizelge 4.4 de gösterilmiştir.

Çizelge 4.4: ADPA ve BDPA'na Göre Yapılan Puanlamalara Ait Cronbach Alfa (α) Değerleri

Anahtar/puanlayıcı	ADPA	BDPA
1. puanlayıcı	0,8776**	0,8792**
2. puanlayıcı	0,8490**	0,8553**
3. puanlayıcı	0,8923**	0,8205**
4. puanlayıcı	0,8620**	0,7979**
5. puanlayıcı	0,7894**	0,8299**

** p<0,01 olduğunu göstermektedir.

Çizelge 4.4'den anlaşılacağı üzere her bir puanlayıcı için hesaplanan iç tutarlık değerleri α =0,79 ile α =0,89 arasında değişim göstermektedir. Cronbach Alfa değerlerinin bu aralıkta olması sınavın iç tutarlığının yüksek olduğunu ifade eder (Cronbach, 1951; Novick &Lewis, 1967; Kaiser &Michael, 1975; Akt: Cohen & Swerdlik, 2013).

Puanlayıcılara ait betimsel istatistikler incelendikten sonra alt problemlere geçilmiştir.

4.2. Birinci Alt Probleme İlişkin Bulgular ve Yorumları

Bu bölümde birinci alt problem olan *"*Puanlayıcıların analitik ve bütünsel dereceli puanlama anahtarlarından elde ettikleri puanlar, Klasik test kuramındaki puanlayıcılar arası tutarlık belirleme tekniklerine göre nasıl değişmektedir?*"* sorusuna yanıt bulmak için birinci alt problemin çözümüne yönelik alt başlıklardan yararlanılmıştır. Elde edilen bulgular 1.a, 1.b ve 1.c altbaşlıklarında sunulmuş ve tartışılmıştır.

1.a*"Analitik ve bütünsel dereceli puanlama anahtarları kullanılarak elde edilen puanların tutarlığı Pearson Momentler Çarpımı Korelasyon Katsayısı ve Spearman Sıra Farkları Korelasyon Katsayısına göre nasıl değişmektedir?"*

Analitik ve bütünsel dereceli puanlama anahtarları kullanılarak belirlenen öğrenci puanları arasındaki ilişki katsayıları PMÇKK ve SSFKK ile belirlenmiş, PMÇKK değerleri, Çizelge 4.5 ve Çizelge 4.6' da, SSFKK değerleri, Çizelge 4.7ve Çizelge 4.8 'da gösterilmiştir.

Çizelge 4.5: ADPA' na Göre Yapılan Puanlamalar Arasındaki İlişki (PMÇKK)

ADPA	1. puanlayıcı	2. puanlayıcı	3. puanlayıcı	4. puanlayıcı	5. puanlayıcı
1. puanlayıcı	1	,967(**)	,987(**)	,963(**)	,941(**)
2. puanlayıcı	,967(**)	1	,976(**)	,958(**)	,973(**)
3. puanlayıcı	,987(**)	,976(**)	1	,974(**)	,954(**)
4. puanlayıcı	,963(**)	,958(**)	,974(**)	1	,939(**)
5. puanlayıcı	,941(**)	,973(**)	,954(**)	,939(**)	1

(**) p<0,01 olduğunu göstermektedir.

Çizelge 4.6: BDPA' na Göre Yapılan Puanlamalar Arasındaki İlişki (PMÇKK)

BDPA	1. puanlayıcı	2. puanlayıcı	3. puanlayıcı	4. puanlayıcı	5. puanlayıcı
1. puanlayıcı	1	,924(**)	,950(**)	,940(**)	,976(**)
2. puanlayıcı	,924(**)	1	,918(**)	,918(**)	,922(**)
3. puanlayıcı	,950(**)	,918(**)	1	,955(**)	,933(**)
4. puanlayıcı	,940(**)	,918(**)	,955(**)	1	,929(**)
5. puanlayıcı	,976(**)	,922(**)	,933(**)	,929(**)	1

(**) p<0,01 olduğunu göstermektedir.

Pearson momentler çarpımı korelasyon katsayısı ile elde edilen puanlayıcılar arası ilişki değerleri incelendiğinde, tüm puanlayıcıların analitik ve bütünsel dereceli puanlama anahtarları kullanarak yaptıkları puanlamalar arasında anlamlı, yüksek düzeyde ve pozitif yönlü ilişki olduğu sonucuna varılmaktadır (p<0,01). PMÇKK 0,918 ve 0,987 arasında değerler almaktadır. En düşük ilişki düzeyinin yüksek düzeyde olduğu göz önünde bulundurularak, puanlayıcıların arasındaki tutarlık düzeylerinin yüksek düzeyde olduğu savunulabilir (Kan, 2001; Howell, 2002).

Çizelge 4.7: ADPA' na Göre Yapılan Puanlamalar Arasındaki İlişki Değerlerin (SSFKK)

ADPA	1. puanlayıcı	2. puanlayıcı	3. puanlayıcı	4. puanlayıcı	5. puanlayıcı
1. puanlayıcı	1	,954(**)	,988(**)	,959(**)	,929(**)
2. puanlayıcı	,954(**)	1	,964(**)	,941(**)	,968(**)
3. puanlayıcı	,988(**)	,964(**)	1	,969(**)	,941(**)
4. puanlayıcı	,959(**)	,941(**)	,969(**)	1	,921(**)
5. puanlayıcı	,929(**)	,968(**)	,941(**)	,921(**)	1

(**) p<0,01 olduğunu göstermektedir.

Çizelge 4.8: BDPA' na Göre Yapılan Puanlamalar Arasındaki İlişki Değerleri (SSFKK)

BDPA	1. puanlayıcı	2. puanlayıcı	3. puanlayıcı	4. puanlayıcı	5. puanlayıcı
1. puanlayıcı	1	,921(**)	,938(**)	,930(**)	,975(**)
2. puanlayıcı	,921(**)	1	,909(**)	,905(**)	,919(**)
3. puanlayıcı	,938(**)	,909(**)	1	,953(**)	,925(**)
4. puanlayıcı	,930(**)	,905(**)	,953(**)	1	,924(**)
5. puanlayıcı	,975(**)	,919(**)	,925(**)	,924(**)	1

(**) p<0,01 olduğunu göstermektedir.

Spearman Sıra Farkları korelasyon katsayısı ile elde edilen puanlayıcılar arası ilişki değerleri incelendiğinde, tüm puanlayıcıların analitik ve bütünsel dereceli puanlama anahtarları kullanarak yaptıkları puanlamalar arasında anlamlı, yüksek düzeyde ve pozitif yönlü ilişki olduğu sonucuna varılmaktadır (p<0,01). SSFKK 0,900 ve 0,988 arasında değişen değerler almaktadır. Puanlayıcılar arasındaki ilişki katsayılarına göre , puanlayıcılar arasında yüksek düzeyde tutarlığın var olduğu söylenebilir (Cohen, 1988; Howell, 2002; Baykul, 2010). Aşağıdaki Çizelge 4.9' de aynı puanlayıcının iki farklı dereceli puanlama anahtarı ile verdiği puanlar arası ilişki değerleri verilmiştir.

Çizelge 4.9: Puanlayıcıların İki Farklı Puanlama Anahtarı İle Yaptıkları Puanlamaların İlişkileri

Puanlayıcılar	P (rs)
1.	0.962**
2.	0.933**
3.	0.967**
4.	0.915**
5.	0.924**

** p<0,01 olduğunu göstermektedir.

Çizelge 4.9 de görüldüğü gibi tüm ilişkiler anlamlıdır (p<0,01). Puanlayıcıların farklı zamanlarda yaptıkları puanlamaları arasında 0,915 ve 0,967 arasında değişen pozitif yönlü yüksek ilişki katsayıları hesaplanmıştır. Bu Spearman Rho değerleri ilişkinin pozitif yönlü ve çok kuvvetli olduğunu ifade eder. İlişki katsayılarının pozitif yönlü ve yüksey düzeyde ilişki göstermesi puanlayıcılara arası yüksek tutarlığa işaret etmektedir (Cohen, 1988; Howell, 2002; Baykul, 2010). Bu yüksek düzeydeki tutarlığın dereceli puanlama anahtarlarının yapısından kaynaklandığı söylenebilir. Yapıların birbirine yakın olması, puanların da birbirine yakın olmasını sağlamış ve puanlayıcıların iki farklı dereceli puanlama anahtarını kullanarak verdikleri puanların ilişki katsayıları pozitif yönlü ve yüksek çıkmıştır.

1.b*"Analitik ve bütünsel dereceli puanlama anahtarları ile elde edilen puanların tutarlığı Kappa Tekniğine göre nasıl değişmektedir?"*

Analitik dereceli ve bütünsel dereceli puanlama anahtarları kullanılarak elde edilen puanlara ait Kappa istatistiği değerleri ve puanlayıcıların uyum yüzdeleri Çizelge 4.10 ve Çizelge 4.11' de verilmiştir.

Çizelge 4.10: ADPA İle Elde Edilen Puanların Kappa İstatistikleri Ve Uyum Yüzdeleri

Madde	Kappa Değeri (κ)	Std. Hata	Z değeri	Uyum yüzdesi/100	P değeri
1. madde	0,5039	0,0240	21,0258	0,6399	0,000**
2. madde	0,5190	0,0235	22,0404	0,6446	0,000**
3. madde	0,5320	0,0204	26,0394	0,6387	0,000**
4. madde	0,5353	0,0205	26,1686	0,6417	0,000**
5. madde	0,5518	0,0205	26,9071	0,6536	0,000**
6. madde	0,5164	0,0277	18,6252	0,6667	0,000**

** p<0,01 olduğunu göstermektedir.

Çizelge 4.11: ADPA ve BDPA İle Elde Edilen Puanlara Kappa İstatistikleri Ve Uyum Yüzdeleri

Madde	Kappa Değeri (κ)	Std. Hata	Z değeri	Uyum yüzdesi/100	P değeri
1. madde	0,5020	0,0217	21,0044	0,6225	0,000**
2. madde	0.5075	0,0205	22,3252	0,6300	0,000**
3. madde	0.5144	0,0214	25,2110	0,6260	0,000**
4. madde	0.5227	0,0265	24,1556	0,6338	0,000**
5. madde	0.5337	0,0244	26,3504	0,6348	0,000**
6. madde	0,5014	0,0283	19,8554	0,6127	0,000**

** p<0,01 olduğunu göstermektedir

Çizelge 4.10 ve Çizelge 4.11 incelendiğinde, iki puanlama anahtarı ile elde edilen puanlarla tüm maddeler için hesaplanan Kappa (κ) değerleri anlamlıdır (p<0,01). Kappa değerleri 0,50 ve 0,55 arasında, uyum yüzdeleri de %61,3 ve %66,7 arasında değerler almaktadır. En düşük uyum BDPA ile puanlanan 6. Madde için κ=0,5014 olarak bulunmuştur. En yüksek uyum ise ADPA ile puanlanan 5. Madde için κ=0,5518 olarak hesaplanmıştır. Bu değerler orta dereceli uyuma işaret etmektedir (Landis ve Koch,1977). Ayrıca hesaplanan uyum yüzdelerinde de en düşük uyum yine BDPA ile elde edilen 6. madde için görülürken (%61,3), en yüksek uyum ADPA ile puanlanan 6. madde için (%66,7) bulunmuştur. En yüksek ve en düşük yüzdelerin aynı soruya aittir. Bufarkın temel sebebinin puanlama anahtarlarının yapısından kaynaklandığı düşünülmektedir. Kappa değerleri maddeden maddeye değişkenlik gösterse de birbirinden çok uzak ve aykırı

değerler almamıştır. Değerlerin birbirine yakın olması ve orta düzeyde uyuma işaret etmesi puanlayıcılar arası tutarlığın da orta düzeyde olduğunu ifade eder (Landis ve Koch,1977). Orta düzeyde gözlenen uyumun dereceli puanlama anahtarlarının ve ölçme aracının yapısından kaynaklandığı söylenebilir.

1.c*"Analitik dereceli puanlama anahtarları ile elde edilen puanların tutarlığı Krippendorff Alfa Tekniğine göre nasıl değişmektedir?"*

Analitik dereceli puanlama anahtarından elde edilen puanlara göre hesaplanan Krippendorff Alfa değerleri aşağıdaki Çizelge 4.12 da verilmiştir.

Çizelge 4.12: ADPA İle Elde Edilen Puanlara ait Krippendorff Alfa değerleri

Madde	Krippendorff Alfa değeri
1. madde	0,809**
2. madde	0,841**
3. madde	0,866**
4. madde	0,919**
5. madde	0,881**
6. madde	0,715**

** p<0,01 olduğunu göstermektedir.

ADPA'na göre hesaplanan Alfa (α) değerleri 0,715 ile 0,919 arasında değişkenlik göstermektedir. En düşük uyum yine kappa istatistiğinde olduğu gibi 6. Maddeye ait iken (α=0,72), en yüksek uyum dördüncü maddede karşımıza çıkmaktadır (α=0,92). 6. Madde için orta düzeyde uyum belirlenirken, diğer maddeler için yüksek düzeyde uyumlar elde edilmiştir. Krippendorff'a göre (2007) tüm maddelerin yüksek düzeyde güvenirliğe sahip olduğu ve puanlayıcılar arası tutarlığın yüksek olduğu savunulabilir.

Tüm bulgulara göre, analitik dereceli puanlama anahtarından elde edilen puanların, bütünsel dereceli puanlama anahtarından elde edilen puanlara göre KTK'da göreceli olarak daha yüksek güvenirlikler verdiği ve bu durumun bir çok araştırmada gözlendiği belirlenmiştir (Follman & Anderson, 1967; Bauer, 1981; Jonsson & Svingby, 2007; Ömür, 2009; Atmaz, 2009; Parlak, 2010; Bıkmaz, 2011; Büyükkıdık, 2012).KTK içinde seçilen tekniklerden PMÇKK ve SSFKK ile elde edilen sonuçların birbirine çok yakın olduğu ve çok yüksek düzeyde tutarlığa işaret

ettiği görülmektedir. Kappa ve Krippendorf teknikleri kullanılarak elde edilen bulgular incelendiğinde ise Kappa tekniği kullanılarak elde edilen tutarlık

derecelerinin daha düşük olduğu görülmektedir.Bu durum da Bıkmaz 'ın araştırması (2011) ile örtüşmektedir. Üst düzey becerilerin ölçülmesinde güvenirliğin ve tutarlık derecelerinin belirlenmesindehangi tekniğin kullanılacağı,ölçüm sonuçlarının hangi amaç doğrultusunda kullanılacağına, kullanılan ölçeğin türüne bağlı olarak değişkenlik gösterebilir.

4.3. İkinci Alt Probleme İlişkin Bulgular ve Yorumları

Puanlayıcıların analitik ve bütünsel dereceli puanlama anahtarları ile elde ettikleri öğrenci puanlarının tutarlık dereceleri Genellenebilirlik kuramına göre birey (b), madde (m) ve puanlayıcı (p) değişkenin çapraz tasarlandığı b x m x p deseninde nasıl değişmektedir? sorusuna yanıt bulmak için ikinci alt problemin çözümüne yönelik alt başlıklardan yararlanılmıştır. Elde edilen bulgular 2.a ve 2.b altbaşlıklarında sunulmuş ve tartışılmıştır.

2.a "*Analitik dereceli puanlama anahtarı ile elde edilen puanların oluşturduğu b x m x p deseni G çalışması sonucunda kestirilen varyans bileşenlerini ve toplam varyansı açıklama yüzdelerini nasıl etkilemektedir?*"

ADPA'na göre elde edilen puanlar ile G kuramında b x m x p deseni ile genellenebilirlik çalışması yapılmıştır. Her bir değişkenlik kaynağına ait kestirilen varyans bileşenleri ve toplam varyans açıklama yüzdeleri Çizelge 4.13 da verilmiştir.

Çizelge 4.13: ADPA İle elde edilen puanların oluşturduğu b x m x p desenine ait G çalışması sonucunda kestirilen varyans bileşenlerini ve toplam varyansı açıklama yüzdeleri

Varyans Kaynağı	Sd	Kareler Toplamı (S.S)	Kareler Ortalaması (M.S)	Varyans	%
b	83	2134,48	25,72	0,75828	46,7
m	5	214,61	42,92	0,00724	5,8
p	4	17,77	4,44	0,09379	0,4
bp	332	72,43	0,22	0,00229	0,1
bm	415	1226,06	2,95	0,54998	33,9
mp	20	15,61	0,78	0,00686	0,4
bmp,e	1660	339,39	0,21	0,20445	12,6
Toplam	2519	4020,34			100

b:birey, m:madde, p:puanlayıcı

Oluşturulan desene ait G çalışması incelendiğinde, birey grubunun toplam varyansın %46,7' sini açıkladığı görülmüştür. Bu haliyle bireyler (öğrenciler) tüm değişkenlikteki en fazla etkiye sahiptirler. Kestirilen varyans ve toplam varyans yüzdeleri incelendiğinde birey x puanlayıcı ortak etkileşiminin %0,1 ile en az etkiye sahip olduğu gözlenmiştir. Ancak puan ve madde x puanlayıcı etkileşimlerin etkisi de birey x puanlayıcı ortak etkileşimin etkisinden çok daha farklı düzeyde değildir (%0,4).

Birey (b) bireylerin problem çözme becerileri farklılık gösterdiğinden tüm değişkenliğin % 46,7' sini açıklamaktadır. Bireylerin varyansı σ^2_b =0,75828 ve açıklanan varyansın yarısına yakınına kaynaklık eden bu değişim bireylerin bireysel farklılıklarından ortaya çıkmaktadır. Madde (m) uygulanan 6 maddeyi temsil etmektedir. Öğrencilere uygulanan maddelerin varyansı σ^2_m =0,00724 olarak bulunmuş ve bu değişkenlik toplam varyansın %5,8 ini açıklamaktadır. Bu bilgiden hareketle soruların benzer güçlükte olduğu düşünülebilir. Puanlayıcı (p) puanlama yapan 5 öğretmeni temsil etmektedir. Puanlayıcıların varyansı σ^2_p =0,09379 olarak bulunmuştur. Bu varyans toplam varyansın %0,4 üne tekabül etmektedir. Bu düşük varyans yüzdesi puanlayıcıların puanların farklılaşmasında neredeyse hiç rol oynamadığını göstermektedir.

Önceki analiz neticelerinde de puanlayıcılar arası yüksek tutarlık olduğu defalarca görülmüştür. Birey, puanlayıcı etkileşimi (bp) ölçülmek istenmeyen puanlayıcı

etkisinin, ölçülmek istenen birey etkisi ile etkileşimi ile belirten değişkenlik kaynağıdır ve σ^2_{bp} =0,00229 olarak bulunmuştur. Bu da toplam varyansın %0,1 lik kısmını oluşturmaktadır. Buradan yola çıkarak istenmeyen puanlayıcı etkisinin ortaya çıkmadığı ve bireylerin puanlarının bir puanlayıcıdan diğer puanlayıcıya değişmediği söylenebilir.

Birey, madde etkileşimi (bm) ölçülmek istenmeyen madde etkisinin, ölçülmek istenen birey etkisi ile etkileşimini ifade etmektedir. σ^2_{bm} =0,54998 ve toplam varyansı açıklama yüzdesi %33,9 dur. Birey etkisine karışan madde etkisinin yüzde olarak yüksek olduğu ancak yüzde ne kadar yüksek olsa da çok büyük varyans farkını oluşturmadığı söylenebilir. Yine de oluşan bu farklılığın maddelerin aynı düzeyde olmamasından kaynaklandığı ve bireylerin farklı düzeyde maddeler için aynı problem çözme becerisini gösteremedikleri söylenebilir. Ayrıca madde sayısının 6 ile sınırlı kalması da bu farklılaşmanın kaynağı olabilir. Madde x puanlayıcı etkileşiminin (mp) varyansı σ^2_{mp} =0,00686 ve toplam varyansa etkisi %0,4 dür. Ölçülmek istenmeyen madde ve puanlayıcı etkisinin çok düşük kalması puanlayıcıların bir maddeden diğerine puanlama tutarlıklarında bir değişiklik olmadığını anlamamıza yardımcı olur.

Birey, madde ve puanlayıcı b x m x p deseninde açıklanamayan ve diğer bir deyişle ölçme hatalarından kaynaklanan artık varyans oranı toplam değişkenliğin %12,6'sını oluşturmaktadır. Birey, birey madde etkileşiminden sonra, dikkat çeken en büyük etki açıklanamayan varyansa aittir. $\sigma^2_{bmp,e}$ =0,20445 varyans oranı yine de çok yüksek değildir.

ADPA yapılan puanlamalardan elde edilen verilerin G-kuramına göre değerlendirilmesi ve kestirilen varyans bileşenleri ile toplam varyansı açıklama yüzdeleri açıklandığı gibidir. Genellenebilirlik katsayısı 0,88 olarak bulunmuştur. Bu değer yüksek güvenirliğe işaret etmektedir (Shavelson &Webb, 1991; Brennan 2001,2003; Atılgan, 2004).

2.b *"Bütünsel dereceli puanlama anahtarı ile elde edilen puanların oluşturduğu b x m x p deseni G çalışması sonucunda kestirilen varyans bileşenlerini ve toplam varyansı açıklama yüzdelerini nasıl etkilemektedir?"*

BDPA'na göre elde edilen puanlar ile G kuramında b x m x p deseni ile genellenebilirlik çalışması yapılmıştır. Her bir değişkenlik kaynağına ait kestirilen

varyans bileşenleri ve toplam varyans açıklama yüzdeleri Çizelge 4.14 da verilmiştir.

Çizelge 4.14: BDPA İle elde edilen puanların oluşturduğu b x m x p desenine ait G çalışması sonucunda kestirilen varyans bileşenlerini ve toplam varyansı açıklama yüzdeleri

Varyans Kaynağı	Sd	Kareler Toplamı (S.S)	Kareler Ortalaması (M.S)	Varyans	%
b	83	1954,40	23,55	0,69183	43,3
m	5	210,71	42,14	0,09139	5,7
p	4	7,80	1,95	0,00121	0,1
bp	332	112,86	0,34	0,00558	0,3
bm	415	1144,85	2,76	0,49044	30,7
mp	20	26,16	1,31	0,01192	0,7
bmp,e	1660	508,78	0,30	0,30649	19,2
Toplam	2519	3965,57			100

b:birey, m:madde, p:puanlayıcı

Oluşturulan desene ait G çalışması incelendiğinde, bireylerin toplam varyansın %43,3'ünü açıkladığı görülmüştür. Bu haliyle bireyler (öğrenciler) tüm değişkenlikteki en fazla etkiye sahiptirler. Kestirilen varyans ve toplam varyans yüzdeleri incelendiğinde puanlayıcıların %0,1 ile en az etkiye sahip oldukları gözlenmiştir. Ancak birey x puanlayıcı (bp), madde x puanlayıcı (mp) etkileşimlerinin etkisi de puanlayıcı etkisinden çok daha farklı düzeyde değildir (%0,3–0,7).

Birey (b) bireylerin problem çözme becerileri farklılık gösterdiğinden tüm değişkenliğin % 43,3 sini açıklamaktadır. Bireylerin varyansı σ^2_b =0,69183 ve bu değişkenliğin yine aynı bireylerin farklı beceri düzeylerine sahip olmalarından kaynaklandığı düşünülmektedir. Bu değerler ADPA ile elde edilen puanlarla çalışılan desene göre daha düşüktür. Madde (m) uygulanan 6 maddeyi temsil etmektedir. Öğrencilere uygulanan maddelerin varyansı σ^2_m =0,09139 olarak bulunmuş ve bu değişkenlik toplam varyansın %5,7'sini açıklamaktadır. Bu bilgiden hareketle soruların benzer güçlükte olduğu düşünülebilir.

Puanlayıcı (p) puanlama yapan 5 öğretmeni temsil etmektedir. Puanlayıcıların varyansı σ^2_p =0,00121 olarak bulunmuştur. Bu varyans toplam varyansın %0,1'ini oluşturmaktadır. Bu düşük varyans yüzdesi puanlayıcıların puanların farklılaşmasında neredeyse hiç rol oynamadığını göstermektedir. Birey, puanlayıcı etkileşimi (bp) ölçülmek istenmeyen puanlayıcı etkisinin, ölçülmek istenen birey etkisi ile etkileşimi ile belirten değişkenlik kaynağıdır veσ^2_{bp} =0,00558 olarak bulunmuştur. Bu da toplam varyansın %0,3 lik kısmını oluşturmaktadır. Böylelikle istenmeyen puanlayıcı etkisinin ortaya çıkmadığı ve bireylerin puanlarının bir puanlayıcıdan diğer puanlayıcıya değişmediği söylenebilir. Birey puanlayıcı etkileşiminin ADPA kullanılarak elde edilen puanlarla yapılan G çalışmasına göre varyans ve yüzde olarak fazla olsa da ikisi de %0,0 a çok yakındır ve aradaki fark önemsenmeyebilir.

Birey, madde etkileşimi (bm) ölçülmek istenmeyen madde etkisinin, ölçülmek istenen birey etkisi ile etkileşimini ifade etmektedir. σ^2_{bm} =0,49044 ve toplam varyansı açıklama yüzdesi %30,7 dur. Oluşan bu farklılığın maddelerin aynı düzeyde olmamasından kaynaklandığı ve bireylerin farklı düzeyde maddeler için aynı problem çözme becerisine sahip olmadıkları söylenebilir. Ayrıca madde sayısının 6 olmasının da bu farklılaşmayı sağlamış olabilir. Bir önceki değerlerden varyans ve yüzde olarak daha düşük olduğu görülse de iki çalışmada da birey değişkeninden sonra en fazla değişkenlik kaynağını oluşturmaya devam etmektedir. Madde puanlayıcı etkileşiminin (mp) varyansı σ^2_{mp} =0,01192 ve toplam varyansa etkisi %0,7'dir. Ölçülmek istenmeyen madde ve puanlayıcı etkisinin çok düşük kalması puanlayıcıların bir maddeden diğerine puanlama tutarlıklarında bir değişiklik olmadığını belirtir. Birey, madde ve puanlayıcı b x m x p deseninde açıklanamayan ve diğer bir deyişle ölçme hatalarından kaynaklanan artık varyans oranı toplam değişkenliğin %19,2'sini oluşturmaktadır. Birey, birey madde etkileşiminden sonra, dikkat çeken en büyük etki açıklanamayan varyansa aittir $\sigma^2_{bmp,e}$ =0,30649. BDPA ile yapılan puanlamalardan elde edilen verilerin G-kuramına göre değerlendirilmesi ve kestirilen varyans bileşenleri ile toplam varyansı açıklama yüzdeleri açıklandığı gibidir.Genellenebilirlik katsayısı 0,86 olarak bulunmuştur. Bu değer bir önceki tabloda 0,88 olarak belirlenmiştir. Bu değer yüksek güvenirliğe işaret etmektedir (Shavelson &Webb, 1991; Brennan 2001,2003; Atılgan, 2004). Genellenebilirlik katsayılarının birbirine çok yakın

olması farklı dereceli puanlama anahtarlarının değişkenliği oluşturan unsurları etkilemediği şeklinde yorumlanabilir.

4.4. Üçüncü Alt Probleme İlişkin Bulgular ve Yorumları

Klasik Test Kuramına ve Genellenebilirlik Kuramına göre elde edilen güvenirlik katsayıları tutarlık göstermekte midir? sorusuna yanıt bulmak için üçüncü alt problemin çözümüne yönelik alt başlıklardan yararlanılmıştır. Elde edilen Bulgular 3.a ve 3.b altbaşlıklarında sunulmuş ve tartışılmıştır.

3.a *"Analitik dereceli puanlama anahtarıyla elde edilen puanların güvenirlik katsayıları tutarlı mıdır?*

Analitik dereceli puanlama anahtarı ile elde edilen puanların KTK ve G kuramlarından farklı tutarlık belirleme teknikleri neticesinde ortaya çıkan bulgular Çizelge 4.15 de sunulmuş ve yorumlanmıştır.

Çizelge 4.15: ADPA İle Elde Edilen Puanların KTK Ve G-Kuramına Göre İncelenmesi

Kuram-ADPA	Cronbach Alfa (α)	PMÇKK (r)	SSFKK	Kappa (κ)	Krippendorff Alfa (α)
KTK	0,789-0,892	0,939-0,987	0,921-0,988	0,504-0,552 %63-%67	0,715-0,919
G-Kuramı (desen)	**Varyans**	**Varyans Açıklama Yüzdesi**	**G Katsayısı**		
b x m x p	σ^2_b =0,75828 σ^2_p =0,09379	b= %46,7 p= %0,4	G=0,88		

b:birey, m:madde, p:puanlayıcı

Çizelge 4.15 incelendiğinde Analitik dereceli puanlama anahtarı kullanılarak elde edilen puanların ilişkileri, tutarlık katsayıları, uyum oranları paralellik göstermektedir. Ayrıntılı olarak incelemek gerekirse Cronbach Alfa değerleri 0,789 ile 0,892 arasında değişmektedir. Ayrıca puanlayıcıların puanları arasındaki ilişki katsayıları da PMÇKK ile 0,939-0,987 aralığında, SSFKK ile 0,921-0,988 aralığında değişkenlik göstermiştir. Yine KTK'da Kappa değeri 0,504-0,552 arasında katsayı değeri ve %63-37 arasında değişen uyum yüzdesi ile orta

düzeyde tutarlığın göstergesidir. Bu kappa değeri şansla beklenen uyumdan daha yüksek orta derecede bir uyum olduğunu söyler. Krippendorff Alfa değeri ise 0,715–0,919 arasında değişen katsayılar ile yüksek ve çok yüksek düzeyde tutarlığın dayanağı olmuştur. KTK 'dan elde edilen tüm katsayılar yüksek düzeyde uyum ve ilişki gösterse de, göreceli olarak Kappa uyumu diğer istatistiklerden düşük, PMÇKK ve SSFKK yine göreceli olarak tüm istatistiklerden daha yüksek değerler almıştır. G-kuramına göre b x m x p çapraz deseninde sınanan puanlarda, en büyük değişkenlik kaynağının birey olduğu bulunmuştur (b= %46,7, σ^2_b =0,75828). Ölçülmek istenmeyen puanlayıcı, madde, puanlayıcı madde etkileşimi de çok küçük varyanslara sahip olup, değişime etki etmemişlerdir. Puanlayıcı etkisi (σ^2_p =0,09379, p= %0,4)olarak hesaplanmıştır. Ayrıca her bir ölçümden elde edilen ve tüm maddeleri kapsayan G katsayısı; bağıl hata kaynaklarına göre 0,88 ve mutlak hata kaynaklarına göre 0,87 olarak bulunmuştur.

3.b *"Bütünsel dereceli puanlama anahtarıyla elde edilen puanların güvenirlik katsayıları tutarlı mıdır?"*

Bütünsel dereceli puanlama anahtarı ile elde edilen puanların KTK ve G kuramlarından farklı tutarlık belirleme teknikleri neticesinde ortaya çıkan bulgular Çizelge 4.16 de sunulmuş ve yorumlanmıştır.

Çizelge 4.16: BDPA İle Elde Edilen Puanların KTK ve G-Kuramına Göre İncelenmesi

Kuram	**Cronbach Alfa (α)**	**PMÇKK (r)**	**SSFKK**	**Kappa (κ)**
KTK	0,798–0,879	0,918–0,976	0,905–0,975	0,501–0,534 %61-%64
G-Kuramı (desen)	**Varyans**	**Varyans Açıklama Yüzdesi**	**Gen. Katsayısı**	
b x m x p	σ^2_b =0,69183 σ^2_p =0,00121	b= %43,3 p= %0,1	G=0,86	

b:birey, m:madde, p:puanlayıcı

Çizelge 4.16 incelendiğinde Analitik dereceli puanlama anahtarı kullanılarak elde edilen puanların ilişkileri, tutarlık katsayıları, uyum oranları paralellik

göstermektedir. Ayrıntılı olarak incelemek gerekirse Cronbach Alfa değerleri 0,798 ile 0,879 arasında değişmektedir. Bu değerler Cronbach Alfa değerleri ile yakınlık göstermektedir. Ayrıca puanlayıcıların puanları arasındaki ilişki katsayıları da PMÇKK ile 0,918–0,976 aralığında, SSFKK ile 0,905–0,975 aralığında değişkenlik göstermiştir. Yine KTK'da Kappa değeri 0,501–0,534 arasında katsayı değeri ve %61–34 arasında değişen uyum yüzdesi ile orta düzeyde tutarlığın göstergesidir ve şansla beklenen uyumdan daha yüksek orta derecede bir uyum olduğunu söyler fakat ADPA'na göre göreceli olarak daha düşüktür. KTK'dan elde edilen tüm katsayılar yüksek düzeyde uyum ve ilişki gösterse de, göreceli olarak Kappa uyumu diğer istatistiklerden düşük, PMÇKK ve SSFKK yine göreceli olarak tüm istatistiklerden daha yüksek değerler almıştır. G-kuramına göre b x m x p çapraz deseninde sınanan puanlarda, en büyük değişkenlik kaynağının birey olduğu bulunmuştur (b= %43,3, σ^2_b =0,69183). Ölçülmek istenmeyen puanlayıcı, madde, puanlayıcı madde etkileşimi de çok küçük değişkenliğe sahip olup, varyansa etki etmemişlerdir. Puanlayıcı etkisi (σ^2_p =0,00121, p= %0,1)olarak hesaplanmıştır. Tüm maddeleri kapsayan G katsayısı; bağıl hata kaynaklarına göre 0,88 ve mutlak hata kaynaklarına göre 0,86 olarak bulunmuştur.

5. SONUÇ ve ÖNERİLER

Bu bölümde araştırmanın bulgu ve yorumlarına dayalı olarak ulaşılan sonuçların özetine ve araştırma ve uygulamaya dönük önerilerden oluşmaktadır.

5.1. Bulgulardan Elde Edilen Sonuçlar

5.1.1. Birinci Alt Problemin Sonuçları

1. ADPA ve BDPA kullanılarak yapılan puanlamaların Cronbach Alfa, Pearson momentler çarpımı ve Spearman sıra farkları korelasyon katsayısı ile puanlar arası ilişki incelendiğinde, çok yüksek değerler ile karşılaşılmıştır. Bu ilişki değerlerinin yüksek olması, puanlayıcıların hem kendi aralarında (farklı puanlama anahtarlarına göre) hem de birbirileri aralarında yüksek uyum olduğu çıkarımına varmamıza yardımcı olmuştur.
2. Tüm ilişkilerin manidar olduğu ve analitik dereceli puanlama anahtarı ile elde edilen puanların ilişkisinin göreceli olarak bütünsel dereceli puanlama anahtarlarından elde edilen puanların ilişkisinden yüksek olduğu ortaya çıkmıştır. Yapılan birçok çalışmada KTK'ya göre elde edilen tutarlık ve güvenirlik değerlerinin analitik dereceli puanlama anahtarı lehine çıktığını düşünürsek bu çalışmada da bu durum değişiklik göstermemiştir. (Follman & Anderson, 1967; Bauer, 1981; Jonsson & Svingby, 2007; Akt: Büyükkıdık, 2012).
3. KTK ile yapılan analizlerde; Cohen'in Kappası, Krippendorff 'un Alfa'sı ilişki katsayıları gibi çok yüksek değerlere sahip olmasalar da kendi içlerinde Kappa için orta düzeyde, Krippendorff için yüksek düzeyde uyum ve neticesinde orta ve yüksek düzeyde tutarlığın var olduğu ortaya çıkmaktadır.
4. Kappa ve Krippendorff değerlerinin birbirine yakınlık gösterdiği ancak maddeden maddeye farklılaşmalar içerdiği anlaşılmıştır. Bu değişimin en önemli nedenlerinden birisinin farklı maddelere farklı tepki gösteren bireyler olduğu düşünülmektedir. Tutarlıkların değişmesinde analitik ve bütünsel dereceli puanlama anahtarlarının yanında, puanlayıcılar da potansiyel kaynak gibi gözükse de asıl farkın bireylerden kaynaklandığı düşünülmektedir ve nitekim G-çalışmasında bu sonuç gözlenmiştir.

5. KTK ile elde edilen bulgular neticesinde iki farklı puanlama anahtarından ADPA'nın daha nesnel puanlamalara olanak tanıdığı, Cronbach Alfa değerlerine göre de maddelerin iç tutarlığının oldukça yüksek olduğu anlaşılmıştır.
6. PMÇKK ve SSFKK ile puanlayıcılar arası manidar ilişkinin, pozitif ve yüksek düzeyde olduğu söylenebilir. İki farklı anahtar için hesaplanan Kappa değerleri ile analitik dereceli puanlama anahtarı ile elde edilen puanlarla hesaplanan Krippendorff Alfa değerlerinin de puanlayıcılar arası yüksek tutarlığa ve test içi orta-yüksek güvenirliğe işaret ettiğinden söz edilebilir. Ancak bunu genellemenin ne derece doğru olduğu tartışılır. Bu testin kesin olarak güvenilir olduğunu savunmak hatalı bir düşünce olabilir. Nitekim bir test bir grup için güvenilir bulgular verirken bir diğer grup için güvenilir bulgulardan uzak kalabilir. Grubu oluşturan öğrencilerin bireysel farklılıkları, grubun homojen ya da heterojen yapısı, diğer dış ve iç faktörler testlerin güvenirliğini pozitif ya da negatif yönde etkileyebilir. Bu yüzden bu test güvenilirdir yerine bu testi puanlayanların tutarlı puanlamalar yaptıkları sonucuna varmak daha doğrudur.
7. Tutarlı puanlamalar ise yapılandırılmış ve puanlayıcıya rehberlik edebilen dereceli puanlama anahtarlarından kaynaklanmaktadır. Bu vesileyle de yapılandırılmış puanlama anahtarlarının öğrencilerin daha nesnel, daha geçerli ve nihayetinde daha güvenilir (daha az hatalı) sonuçlar almalarına yardımcı olduğu düşünülmektedir.

5.1.2. İkinci Alt Problemin Sonuçları

1. ADPA ve BDPA ile elde edilen puanlarla G çalışması yapılmış; kestirilen varyans bileşenlerine ait varyanslar ve toplam değişkenliği açıklama yüzdeleri bulunmuştur. ADPA ve BDPA ile elde edilen puanlara ait G çalışmasına ait bulgular incelendiğinde en yüksek değişkenlik kaynağının bireyler olduğu sonucuna varılmıştır. Bu bulgu Anıl ve Büyükkıdık' ın (2012) bulguları ile bezerlik göstermektedir.
2. Bireylerin değişkenliği arttırmasını bireylerin farklı düzeyde problem çözme becerisine bağlayabiliriz. Nitekim her bireyin farklı beceri ve öğrenme düzeyleri, uygulama ve pratiğe dönük artı ve eksileri, farklı ön öğrenmeleri bu değişkenlikte pay sahibi olabilir.
3. Birey madde etkileşiminin (bm) yüksek çıkması maddelerin bireylerden bireylere farklı anlaşıldığını ve maddelerin bireylerin gösterdiği beceri düzeylerinde

dolaylı da olsa pay sahibi olduğu sonucuna varmamızı sağlamıştır. Ayrıca açık uçlu başarı testindeki maddelerin yapısal özelliklerinin farklılık göstermesi de bu farklılığın kaynağı olarak düşünülebilir.

4. Ölçülmek istenmeyen puanlayıcı, madde, birey puanlayıcı etkileşimi, madde puanlayıcı etkileşimi, göreceli olarak düşük varyans değerlerine sahiptir ve dolayısıyla toplam varyansı açıklama yüzdeleri de küçük kalmıştır.
5. Çaprazlanmış b x m x p deseni ile yapılan genellenebilirlik çalışmasında elde edilen G katsayıları da ADPA için 0,88, BDPA için 0,86 bulunmuştur. Bu değerler mutlak ve bağıl hata kaynaklarına göre çok minik değişkenlik gösterseler de oldukça yüksek bir güvenirlik derecesini işaret eder. Brennan'a göre (2000) birey madde etkileşimi (bm) yüksek olması güvenirliği olumsuz etkilese de madde sayısının artması güvenirliği yükseltir. Buradan hareketle madde sayısını yüksek tutmanın güvenirliği etkilediği düşünülmektedir.

5.1.3. Üçüncü Alt Problemin Sonuçları

1. Problem çözme becerisini ölçmeye dönük açık uçlu soruların ADPA ve BDPA ile puanlamasından elde edilen puanlarla yapılan KTK ve G çalışması sonuçları puanlayıcılar arası tutarlığın ve maddelerin iç tutarlıklarının yüksek olduğunu ortaya koymuştur. Ancak KTK ve G çalışması sonuçlarını kıyasladığımızda G çalışmasının, değişmelerin hangi değişkenlerden kaynaklandığını ortaya koyması ile daha ayrıntılı bilgi verdiği görülmüştür. Aslında bu KTK ve G çalışmasını birlikte içeren önceki çalışmaların bu yönüyle de benzerlik göstermektedir (Yelboğa, 2007; Güler, 2008; Deliceoğlu, 2009; Öztürk, 2011; Büyükkıdık, 2012).
2. KTK'dan ve G çalışmasından elde edilen sonuçlar incelendiğinde güvenirlik katsayılarının birbirine yakın olduğu bulunsa da göreceli olarak kappa istatistiği daha düşük, G katsayısı da KTK'daki tekniklerle kıyaslayınca daha yüksek değerler vermiştir. Genellenebilirlik çalışmasında puanlayıcı değişkenin toplam değişkenliği açıklama yüzdesi çok düşük çıkarken (%0,1–0,4), KTK'da da bu sonucu desteklercesine; PMÇKK ve SSFKK ile yapılan analizler ile puanlayıcılar arasında anlamlı ($p<0,01$), pozitif yönlü, yüksek ilişki ve neticesinde yüksek tutarlık gözlenmiştir.

5.2. Öneriler

Bu bölümde araştırmanın konusu ile ilgili araştırmaya ve uygulamaya dönük öneriler bulunmaktadır.

5.2.1. Araştırmaya Dönük Öneriler

1. Belirlenen grubun düzey özelliklerine uygun problemler hazırlanmalıdır. Problem çözme becerisini ölçen problemlerin etkili bir şekilde çözülebilmesi için ön öğrenmelerin kazanılmış olması gerekir. Bu sebeple uygulama oluşturulurken öğrencilerin ön öğrenme düzeylerine dikkat edilmelidir.
2. Hazırlanan uygulamanın öğrenciler tarafından ne kadar sürede tamamlanacağı önceden hesaplanmalıdır. Çünkü çok soru sorarak güvenirliği arttırmak hedeflenirken, erişilemeyen soruların artabileceği göz ardı edilmemelidir.
3. Hazırlanan problemlere en uygun dereceli puanlama anahtarları oluşturulmalıdır.
4. Hazırlanan uygulama, dereceli puanlama anahtarları ve diğer formların geçerlik ve güvenirlik çalışmaları yapılmalı ve daha geçerli ve güvenilir çalışmalar ortaya çıkmalıdır.
5. Uygulama düzeninin kurulması, problem çözme becerisini ölçen soruların ve dereceli puanlama anahtarlarının hazırlanması olabildiğince erken yapılmalıdır. Çünkü gerçek verilerle çalışmak daha fazla zaman alır ve araştırmanın aksamasına ve yavaşlamasına sebep olabilir.
6. Uygulama yaparken öğrencilerin gönüllü olmasına dikkat edilmeli ve gerçek bir sınav ortamı oluşturulmalıdır. Öğretmenler sınav ciddiyetini sağlamak için bu tür çalışmaların neticelerini okul başarı notlarına yansıtma vaadiyle öğrencileri motive etmektedirler. Aslında pek uygun bir yol gibi gözükmese de bu seviyedeki öğrencilerin ciddiyetini arttırdığı belirlenmiştir.
7. Puanlamayı yapacak öğretmenlerin gerçek manada gönüllü olması gerekir. Unutulmamalıdır ki; katılımcılar arasında en çok zamanı puanlama yapan öğretmenler harcar. Öğretmenler puanlama yapmadan önce dereceli puanlama anahtarlarını nasıl kullanacaklarına dair eğitim almalıdırlar. Bu eğitime mümkünse tüm öğretmenlerin birlikte katılması size zaman kazandırır.

8. Bu eğitimde problem çözme becerisinin farklı yol ve yöntemler ile sergilenebileceği ve doğru sonuca ulaşmak için tek bir çözüm yolu olmadığı hususuna vurgu yapılmalıdır. Ayrıca eğitimde, puanlamanın sonuçtan ibaret olmayıp, problem çözme sürecini de kapsadığı hatırlatılmalıdır.

5.2.2. Uygulamaya Dönük Öneriler

1. Problem çözme becerisini ölçen farklı bir uygulama ile katılımcı ve puanlayıcı sayısını artırarak aynı teknikleri karşılaştıran farklı araştırmalar yapılabilir.
2. Bu çalışmada kullanılan tekniklere ek olarak farklı KTK tutarlık belirleme teknikleri ile teknikleri kıyaslayan, farklılıklarını ve farklılıklarının nedenleri inceleyen araştırma yapılabilir.
3. Farklı türde ölçümlerde KTK tutarlık belirleme teknikleri, G çalışması ve Madde tepki kuramı kullanılarak Puanlayıcılar arası tutarlık dereceleri bulunabilir ve karşılaştırma yapılabilir.
4. KTK, G çalışması ve MTK ile yapılan çalışmalarda değişen dereceli puanlama anahtarları etkisi ile değişen puanlayıcı sayısı ve katılımcı sayısı ile araştırmalar tekrarlanıp sonuçlar karşılaştırılabilir.
5. Genellenebilirlik kuramında bu çalışmada kullanılan b x m x p üçlü çaprazlanmış desenin yerine, dereceli puanlama anahtarlarının, bireylerin sosyo ekonomik düzeyinin, okul türlerinin (devlet-özel) de değişkenliğe dâhil olduğu farklı çaprazlanmış desenler ile güvenirlik çalışmaları yapılabilir.
6. Genellenebilirlik kuramı çerçevesinde karar çalışması yapılarak istenilen güvenirlik düzeyine ulaşmak için gerekli puanlayıcı sayısı belirlenip, KTK'dan elde edilen güvenirlik katsayıları ile karşılaştırılabilir.

KAYNAKÇA

American Educational Research Association, American Psychological Association,& National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Analitik ve Holistik Dereceli Puanlama Anahtarı[Çevrim-içi: http://alternatifolcme.com/performans/pdf/Analitik.pdf, Erişim tarihi*:* 19 Şubat 2014.

Anıl, D., Büyükkıdık, S. (2012). Genellenebilirlik Kuramında Dört Facetli Karışık Desen Kullanımı İçin Örnek Bir Uygulama[1] *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Kış 2012, 3(2)*, 291-296.

Atılgan, H. (2004). *Genellenebilirlik Kuramı ve Çok Değişkenlik Kaynaklı Rasch Modelinin Karşılaştırılmasına İlişkin Bir Araştırma.*Yayımlanmamış Doktora

Tezi. Ankara: Hacettepe Üniversitesi.

Banerjee, M., Capozzoli M., Mcweeney, L., Sinha, D. (1999). Beyond kappa: A rewiew of interrater agreement measures. *The Canadian Journal of Statistics, 27(1)*, 3-23.

Bauer, B. A. (1981). *A Study Of The Reliabilities And Cost-Efficiencies Of Three Methods Of Assessment For Writing Ability*. (ERIC Document Reproduction Service No. ED 216357).

Baykul, Y. (2010). *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması.* Ankara: Pegem Akademi.

Bıkmaz, Ö. (2011). *Puanlayıcılar Arası Güvenirlik Belirleme Teknikleri Üzerine Karşılaştırmalı Bir Araştırma.* Yayımlanmış Yüksek Lisans Tezi. Ankara: Hacettepe Üniversitesi.

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

Brennan, R. L. (2003). Coefficients and Indices in Generalizability Theory,*CASMA Research Report Number 1* , 11,13

Brennen, R. L. & Prediger, D. J. (1981). Coefficient kappa: Some Uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*(1981), 687-699.

Burry-Stock, J. A., Shaw, D. G., Laurie, C., & Chissom, B. S. (1996). Rater agreement indexies for performance assessment. *Educational and Psychological Measurement, 56*(2), 251-262.

Büyükkıdık, S. (2012), *Problem Çözme Becerisinin Değerlendirilmesinde Puanlayıcılar Arası Güvenirliğin Klasik Test Kuramı Ve Genellenebilirlik Kuramına Göre Karşılaştırılması,* Yayımlanmış Yüksek Lisans Tezi. Ankara: Hacettepe Üniversitesi

Büyüköztürk, Ş. (2012). Sosyal Bilimler İçin Veri Analizi El Kitabı. Ankara: Pegem Akademi.

Callison, D. (2000). Rubrics. *School Library Media Activities Monthly*, 17 (2).

Campbell, D. T., & J.C., S. (1963). *"Experimental and Quasi- Experimental Designs for Research on Teaching" Handbook of Research on Teaching. (G. N.L., D.)* Rand McNall

Christ, T. J., Tillman C., Chafouleas, S. M., & Boice C. H. (2010). Direct Behavior Rating (DBR): Generalizability and Dependability Across Raters and Observations. *Educational and Psychological Measurement*. *70*(5), 825–43.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Crawforth, K. (2001). *Measuring the interrater reliability of a data collection instrument developed to evaluate anesthetic outcomes.* Doctoral Dissertation.Available from Proquest Dissertations and Theses database. (UMI No. 3037063)

Crocker, L. M. & Algina, L. (1986). *Introduction to Classical and Modern Test Theory.* New York: Holt, Rinehart and Winston.

Cüceloğlu, D. (1999). İnsan ve Davranışı . İstanbul: Remzi Kitabevi.

Deliceoğlu, G. (2009). *Futbol Yetilerine İlişkin Dereceleme Ölçeğinin Genellenebilirlik Ve Klasik Test Kuramına Dayalı Güvenirliklerinin Karşılaştırılması.*Yayımlanmamış Doktora Tezi. Ankara: Ankara Üniversitesi

Doğan, D. C., Karakaya İ. & Kutlu, Ö. (2010).*Ögrenci basarısının belirlenmesi,Performans ve portfolyaya dayalı durum belirleme. Ölçme ve değerlendirme uygulamaları* (3. Baskı). Ankara: Pegem Yayıncılık.

Erdem, S. A. (2007). *İlköğretim 9. sınıf matematik dersinde öğrenci performansına dayalı verilen sözlü puanların geçerliliğinin incelenmesi.*Yayımlanmamış Yüksek Lisans Tezi, Ankara: Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*. *76(5),* 378-382.

Fleiss, J. L. ve Cohen, J. (1973). The equivalece of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613-619.

Follman, J. C. & Anderson, J. A. (1967). An Investigation of Reliability of Five Procedures for Grading English Themes. *Research in the Teaching of English*, 1, 190-200.

Freelon, D. (2013). ReCal for Ordinal, Interval, and Ratio-Level Data (testing). [Çevrim-içi: http://dfreelon.org/recal/recal-oir.php/, Erişim Tarihi: 28 Nisan 2014.

Goodrich, A. H. (2001). The effects of instructional rubrics on learning to write. *Current Issues ID Education*, 4(4). [Çevrim-içi: http://cie.asu.edu/volume4/number4, Erişim tarihi: 6 Şubat 2014.

Goodrich, A. H. (2005). Teaching with rubrics: The good, the bad and the ugly. *College Teaching, 53,* 27-30.

Gulliksen, H. (2013). Theory of Mental Tests. New York: Taylor and Frances Group.

Güler, N. (2008). *Klasik Test Kuramı Genellenebilirlik Kuramı ve Rasch Modeli Üzerine Bir Araştırma.* Yayımlanmamış Doktora Tezi. Ankara: Hacettepe Üniversitesi.

Güler, N. & Gelbal S. (2010). Açık Uçlu Matematik Sorularının Güvenirliğinin Klasik Test Kuramı Ve Genellenebilirlik Kuramına Göre İncelenmesi. *Kuram Ve Uygulamada Eğitim Bilimleri Dergisi(10) 2* 989–1019

Güler, N., Kaya Uyanık, G., & Taşdelen Teker, G. (2012). *Genellenebilirlik Kuramı.* Ankara: Pegem Akademi.

Gündüz Sefer, D. (2006). *Matematik Dersinde Problem Çözme Becerilerinin Dereceli Puanlama Anahtarı Kullanılarak Değerlendirilmesi*, Yayımlanmış Yüksek Lisans Tezi. Ankara: Hacettepe Üniversitesi

Hafner, J., & Hafner M. (2003). Quantitative analysis of the rubrics as an assessment tool : An emprical study of student peer- group rating. *İnternational Journal of Science Education.*

Howell, D.C.(2002). *Statistical Methods for Psychology*. (Fifth edition). Thomson Learning Academic Research center, USA.

Hunt, R. J. (1986). Percent agreement, pearson‟s correletion, and kappa as measures of ınter-examiner reliability. *Journal of Dental Research, 65(2),* 128-130.

Jay Cohen, R., & Swerdlik, M. E. (2010). *Psychological Testing and Assesssment* (Cilt 7). New York: McGraw-Hill Companies.

Jonsson, A. & Svingby, G. (2007). The Use Of Scoring Rubrics: Reliability, Validity And Educational Consequences. *Educational Research Review. 2*, 130, 144.

Kan, A. (2001), *Yazılı Yoklamaların Puanlanmasında Puanlama Cetveli ve Yanıt Anahtarı Kullanımının Puanlamaya Etkisi.*Yayımlanmamış Yüksek Lisans Tezi, Ankara: Hacettepe Üniversitesi

Kan, A. (2007). Performans degerlendirme sürecine katkıları açısından yeni program anlayısı içerisinde kullanılabilecek bir degerlendirme yaklasımı: Dereceli puanlama anahtarı puanlama yönergeleri. *Kuram ve Uygulamada Egitim Bilimleri, 7(1),* 129-152.

Karasar, N. (2012). Bilimsel araştırma Yöntemi (Cilt 24). Ankara: Nobel yayınları.

Kasap, Y. (2008). *Dereceli puanlama anahtarı ve puanlama anahtarından elde Edilen puanların karşılaştırılması*. Yayımlanmamış Yüksek Lisans Tezi, Ankara:

Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü.

King, J. E.(2004). "mkappasc. Sps" for Interrater reliability. [Çevrim-içi: http://www.ccitonline.org/jking/homepage/ Erişim tarihi: 24 Nisan 2014

Krippendorff, K. (2004). *Reliability In Content Analysis Some Common Misconceptions and Recommendations*.

[Çevrim-içi:http://faculty.washington.edu/jwilker/559/PAP/krippendorf%20-%20reliability.pdf. Erişim tarihi: 13 Mart 2014.

Krippendorff, K. (2007). Krippendorff's Alfa.

[Çevrim-içi: http://www.asc.upenn.edu/usr/krippendorff/

Erişim Tarihi: 14 Aralık 2013.

Kutlu, Ö.,Yıldırım, K. Ve Bilican, S. (2009). Öğretmenlerin Dereceli Puanlama Anahtarlarına İlişkin Tutum Ölçeği Geliştirme Çalışması. *Yüzüncü Yıl Üniversitesi, Eğitim Fakültesi Dergisi.* Aralık 2009. Cilt:VI, Sayı:II, 76-88

Landis, J, R. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Lane, S., Sabers, D. (1989). Use of Generalizability Theory for Estimating the Dependability of a Scoring System for Sample Essays. *Applied Measurement In Education, 2(3),*195-205.

Lei, P., Smith, M. & Suen, H. K. (2007). The Use of Generalizability Theory To Estimate Data Realibility in Single Subject Observational Research. *Psychology in Schools.* (44), (5), 433-439

Lord, F. M. & Novick, R. M. (1968) Statistical Theories of Mental Test Scores. California : Addison-Wesley Publishing Company.

Marzano, R. J., Pickering, D., & McTighe, J. (1993). Assessing student outcomes. *Alexandria: Association for Supervision and Curriculum Development*, *1(5*), 37-43.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes.*Educational Measurement. 11*, (4) 3-9.

Mertler, C. A. (2001). Designing scoring rubrics for your clasroom. Practical assessment, Research and Evaluation, 7 (25). http://PAREonline.net/getvn.asp?v=7&n=25, Erişim tarihi: 17 Aralık 2013.

Nitko, J. A. (2000). Educational assessment of student(2. Baskı). Englewood Cliffs,NJ: Merrill.

Olkun, S., & Toluk, Z. (2001). *İlköğretimde Matematik Öğretimi: 1-5 sınıflar.* Ankara: Artım.

Ömür S. (2009). *Dereceli Puanlama Anahtarıyla, Genel İzlenimle Ve İkili Karşılaştırmalar Yöntemiyle Yapılan Değerlendirmelerin Karşılaştırılması,* Yayımlanmış Yüksek Lisans Tezi. Mersin: Mersin Üniversitesi Sosyal Bilimler Enstitüsü.

Özdamar, K. (2004). Paket programlar ve istatistiksel veri analizi*.* Eskişehir: Kaan Kitabevi.

Özmen Hızarcıoğlu, B. (2013). *Problem Çözme Sürecinde Dereceli Puanlama Anahtarı Kullanımında Puanlayıcı Uyumunun İncelenmesi.*Yayımlanmış Yüksek Lisans Tezi, Bolu: Abant İzzet Baysal Üniversitesi.

Öztürk, M. E. (2011). *Voleybol Becerileri Gözlem Formu İle Elde Edilen PuanlarınGenellenebilirlik ve Klasik Test Kuramı'na Göre Karşılaştırılması.* Yayımlanmamış Doktora Tezi. Ankara: Hacettepe Üniversitesi.

Parlak, B. (2010). *Öğrenci Performansının Belirlenmesinde Puanlama Anahtarı Ve Dereceli Puanlama Anahtarının Karşılaştırılması.*Yayımlanmamış Yüksek Lisans Tezi, Ankara: Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü.

Popham, W. J. (1997). What's stil wrong- and what's stil right with rubric. EducationalLeadership, 55, 72-75.

Posameinter , A. S., & Krulik, S. (1998). *Problem Solving Strategies for Efficient and Elegant Solutions. A Research for the Mathematics Teacher .* California: Corvin Press.

Rae, G. & Hyland, P. (2001). Generalisability and Classical Test Theory Analysesof Koppitz's Scoring System for Human Figure Drawings. *British Journal of Educational Psychological, 71*, 369-382.

Sezer, S. (2006) Öğrencinin Akademik Başarısının Belirlenmesinde Tamamlayıcı Değerlendirme Aracı Olarak Rubrik kullanımı Üzerinde Bir Araştırma, *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi,* Sayı:18

Shavelson, J. &Webb, N. M. (1991). *Generalizability Theory: A Primer*. Sage Publications.

Silcocks, P. (1983). Measuring repetability and validity of historical diagnosis- a brief rewiew with some practical examples. *Journal of Clinical Pathology, 36*, 1269-1275.

Sim, J. and Wright, C. C. (2005) "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements" in *Physical Therapy. Cilt. 85*, say. 257—268

Smith, V. E. & Kulikowıch, J. E. (2004). An Application Of Generalizability Theory And Many-Facet Rasch Measurement Using A Complex Problem-Solving Skills Assessment. *Educational and Psychological Measurement, 64*, 617-639.

Stuhlmann, J., Danıel; C. Dellınger, A., Denny, R. K., Powers, T. (1999) A Generalizability Study Of The Effects Of Training On Teachers' Abilities To Rate Children's Writing Using A Rubric. *Journal Of Reading Pschology, 20*, 107-127.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., de Kruif, R. E. L.,Reed, M., Brown, T. T., Levine, M. D., & White, K. P. (1999). Using Generalizability Theory to Estimate the Reliability of Writing Scores Derived from Holistic and Analytical Scoring Methods. *Educational and Psychological Measurement, 59*: 492. DOI: 10.1177/00131649921970008

Şanlı, E. (2010*). Bilimsel Süreç Becerilerinin Ölçülmesinde Bütünsel ve Analitik Puanlama Anahtarlarının Güvenirliklerinin Karşılaştırılması.* YayımlanmışYüksek Lisans Tezi. Ankara: Ankara Üniversitesi.

Turgut, M. F., & Baykul, Y. (2012). *Eğitimde Ölçme ve Değerlendirme.* Ankara: Pegem Akademi.

Türk Dil Kurumu. (2005). Türkçe Sözlük. Ankara: Akşam Sanat Okulu Matbaası.

Yelboğa, A. (2007). *Klasik Test Kuramı ve Genellenebilirlik Kuramına Göre Güvenirliğin Bir İş Performansı Ölçeği Üzerinde İncelenmesi,*Yayımlanmış Doktora Tezi, Ankara: Ankara Üniversitesi.

EKLER DİZİNİ

Ek–1:Uygulama Soruları

UYGULAMA SORULARI

1. Bir otel her dolu oda için günlük 100 TL kar, her boş oda için günlük 30 TL zarar etmektedir. 110 odalık bu otelin ilk gün 53, ikinci gün 41 ve üçüncü gün 48 odası dolmuştur. **Bu üç günün sonunda otel kaç TL kar elde etmiştir?**

2. Bir dağcı deniz seviyesinden başlayarak birinci gün 1750 metre, ikinci gün birinci günden 600 metre daha az ve üçüncü gün ikinci güne göre 100 metre daha fazla tırmanmıştır ve sonrasında iniş yapmaya başlamıştır. İlk gün 1000 metre ve sonraki üç gün eşit mesafede iniş yaparak başladığı yere dönen bu dağcı **son gün kaç metre inmiştir?** (Deniz seviyesinin yüksekliği sıfır metre olarak kabul edilmektedir.)

3. $2\frac{3}{5} : \left[\frac{4+\frac{1}{2}}{1\frac{4}{5}} + \frac{3}{4} \right]$ **İşleminin sonucunu bulunuz.**

4. $2+\dfrac{16}{5-\dfrac{9}{2+\dfrac{1}{x}}}=10$ **işlemindeki** x **değerini bulunuz?**

5. Bir kilogram portakalın fiyatı bir kilogram muzun fiyatının $\frac{2}{3}$'ü kadardır. 2 kilogram portakal ve 3 kg muz alan kişi manava 13 TL ödüyor. **Buna göre 3 kilogram portakal ve 2 kilogram muz alan kişi manava kaç TL ödemelidir?**

6. Bir memur maaşının $\frac{1}{3}$'ünü ev kirası, $\frac{1}{4}$ 'ünü mutfak masrafları için kullanıyor. Kira ve mutfak masrafları için kullandığı miktarın $\frac{2}{7}$'si kadar da faturalarını ödemek için ayırıyor. Geriye kalan maaşının yarısı ile taksitlerini ödediğine göre **bu memurun geriye tüm maaşının kaçta kaçı kalmıştır?**

Ek–2: Uzman Görüşü Anketi

Değerli katılımcı ,

Bu anket formunda, size sunulan 7. Sınıf öğrencileri için hazırlanmış, tam ve rasyonel sayılarda problem çözme becerisini ölçen açık uçlu soruları, aşağıdaki ölçütlere göre değerlendirmeniz beklenmektedir. Çalışmaya içtenlikle katılacağınıza inanıyor, desteğiniz için teşekkür ediyorum.

Ar. Gör. Bulut YILDIZTEKİN

		Hiç katılmıyorum	Katılıyorum	Kararsızım	Katılıyorum	Tamamen katılıyorum
1	Problem öğrenci seviyesine uygundur					
2	Problem kazanıma uygun olarak belirlenmiştir					
3	Problemde verilen bilgiler açık ve anlaşılırdır					
4	Problemde sorulan sorular açık ve anlaşılırdır					
5	Problem bütünsel dereceli puanlama anahtarı ile puanlanabilir					
6	Problem analitik dereceli puanlama anahtarı ile puanlanabilir					

Ek–3:Etik Komisyonu Onay Belgesi

T.C.
HACETTEPE ÜNİVERSİTESİ
Genel Sekreterlik

Yazı İşleri Müdürlüğü

Sayı : 88600825 /GÜ– 1167

27 Mart 2014

Konu :

EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

İlgi: 03.03.2014 tarih ve 446 sayılı yazınız

Enstitünüz Eğitim Bilimleri Anabilim Dalı Eğitimde Ölçme ve Değerlendirme Bilim Dalı yüksek lisans programı öğrencilerinden **Bulut YILDIZTEKİN**'in hazırladığı **"Klasik Test Kuramı ve Genellenebilirlik Kuramından Puanlayıcılar Arası Tutarlılığın Farklı Yöntemlere Göre Karşılaştırılması"** başlıklı tez çalışması Üniversitemiz Senatosu Etik Komisyonunun **04.03.2014** tarihinde yapmış olduğu toplantıda incelenmiş olup, etik açıdan uygun bulunmuştur.

Bilgilerinizi ve gereğini saygılarımla rica ederim.

Prof. Dr. Ü. Şebnem HARPUT
Rektör a.
Rektör Yardımcısı

Ek: Tutanak

Hacettepe Üniversitesi Genel Sekreterlik, Yazı İşleri Müdürlüğü, 06100 Sıhhiye-Ankara
Telefon: 0 (312) 305 1008 - 1039 • Faks: 0 (312) 310 5552
E-posta: yazimd@hacettepe.edu.tr

Ayrıntılı Bilgi için:

Ek–4: Öğretmen Gönüllü katılım Formu

HACETTEPE ÜNİVERSİTESİ
KLASİK TEST KURAMI VE GENELLENEBİLİRLİK KURAMINDAN PUANLAYICILAR ARASI TUTARLIĞIN FARKLI YÖNTEMLERE GÖRE KARŞILAŞTIRILMASI
*****GÖNÜLLÜ KATILIM FORMU*****

Bu araştırmanın amacı, puanlayıcılar arası tutarlık derecelerini, Klasik test kuramı ve genellenebilirlik kuramı kullanılarak karşılaştırmak ve bilgi edinmek olacaktır. Bu araştırmanın evrenini tüm matematik öğretmenleri oluştururken, örneklem tesadüfi olarak seçilen 5 öğretmenden oluşmaktadır. ,
Araştırmanın verileri öğrencilere uygulanan açık uçlu başarı testinin öğretmenler tarafından puanlanmasıyla elde edilecektir.
Tüm istatiksel hesaplamalar, bilgisayar ortamında SPSS programı ve diğer yardımcı hesaplama programları kullanılarak elde edilecektir.

BİLGİ:

Açık uçlu soruların puanlanması iki kez gerçekleşecektir. Birinci puanlama bütünsel dereceli puanlama anahtarı ile ikinci puanlama ise analitik dereceli puanlama anahtarı ile gerçekleşecektir. Yapacağınız puanlamalar yalnızca bilimsel araştırma için araştırmacı tarafından kullanılacaktır. Bilimsel nitelikteki bu araştırmanın herhangi bir idari veya siyasi yönü yoktur. Puanlama verilerinin üzerine isminizi veya kimliğinizi belirten bir şey yazmayınız. İçtenlikle doldurduğunuz takdirde doğru verilere ulaşılmasında büyük katkılarınız olacaktır.

RİSKLER:

Bu çalışmanın katılımcılar için maddi-manevi herhangi bir risk unsuru oluşturmaması öngörülmektedir.

YARARLAR:

Bu çalışma ile bütünsel ve analitik dereceli puanlama anahtarlarından hangisinin kullanılmasının daha uygun olduğunun cevabı aranacaktır. Ayrıca bu iki puanlama anahtarının karşılaştırılması ve hangisinin puanlayıcı tutarlığını artırttığı tartışılacaktır. Tüm bunlara ek olarak, puanlayıcılar arası güvenirlik belirleme yöntemleri, klasik test kuramı için, Kappa Krippendorf Alfa teknikleri ve Genellenebilirlik yöntemlerinin incelenmesi ve karşılaştırılması sağlanacaktır.

GİZLİLİK:

Bu çalışmanın bilgileri kesinlikle gizli tutulacaktır. Elde edilen bilgiler sadece araştırmacılar tarafından görülecek ve değerlendirilecektir. Yazılı metinde kesinlikle isminiz geçmeyecek, bilgiler anonim olarak iletilecektir. Puanlama sonuçları çalışma bitiminde imha edilecektir.

İLETİŞİM:

Çalışma ile ilgili bir sorunuz olduğu takdirde araştırmacılar ile istediğiniz zaman iletişime geçebilirsiniz.

Doç. Dr. Duygu ANIL Eğitim Bilimleri Bölümü Beytepe, Ankara.

Tel: 312–2978550/106

E-mail: duygu.anil73@gmail.com

Yük. Lisans Öğrencisi: Ar. Gör. Bulut Yıldıztekin

Eğitim Bilimleri Bölümü Beytepe, Ankara.

Tel: 312–2978564

E-mail: bulutyildiztekin@hotmail.com

Eğer bu formda sözü edilen kurallar dışında davranıldığı veya katılımcı olarak gizliliğiniz tehlikeye atıldığı takdirde Hacettepe Üniversitesi Senato Etik Komisyonu ile iletişime geçebilirsiniz.

Hacettepe Üniversitesi Senato Etik Komisyonu

Hacettepe Üniversitesi Rektörlük 1. Kat 06800, Ankara

Tel: 312–2976062

KATILIM:

Bu çalışmaya katılımınız tamamen **gönüllülük** ilkesine dayanmaktadır. İstediğiniz takdirde katılmayı herhangi bir yaptırıma maruz kalmaksızın reddedebilirsiniz. Katıldığınız takdirde herhangi bir aşamasında çalışmayı yarıda bırakabilirsiniz. Eğer çalışmayı veri toplama işlemi bitiminden önce bıraktıysanız isteğiniz halinde verileriniz size teslim edilecek veya imha edilecektir. Puanlamalarınız tamamen gizli kalacak ve cevaplarınızdan elde edilen sonuçlar kesinlikle sizi işaret etmeyecektir.

ONAY:

Bu belgeyi okudum ve kopyasını aldım. Gerekli gördüğüm bütün cevapları almış durumdayım ve bu çalışmaya katılmayı kabul ediyorum.

Öğretmenin İmzası:……

Tarih:……

Gönüllü Formu Tarihi:……

Ek–5: Veli Onay Formu

HACETTEPE ÜNİVERSİTESİ
KLASİK TEST KURAMI VE GENELLENEBİLİRLİK KURAMINDAN PUANLAYICILAR ARASI TUTARLIĞIN FARKLI YÖNTEMLERE GÖRE KARŞILAŞTIRILMASI
*****VELİ ONAY FORMU*****

Bu araştırmanın amacı, puanlayıcılar arası tutarlık derecelerini, Klasik test kuramı ve genellenebilirlik kuramı kullanılarak karşılaştırmak ve bilgi edinmek olacaktır. 7. Sınıf öğrencilerinin tam ve rasyonel sayılarda problem çözme becerilerini ölçen 6 soruluk açık uçlu sınav uygulanması gerekmektedir. Bu uygulamada çocuğunuzun kimliğini belirtmesi gerekmemektedir. Tüm istatiksel hesaplamalar, bilgisayar ortamında SPSS programı ve diğer yardımcı hesaplama programları kullanılarak elde edilecektir.

BİLGİ:

Bilimsel nitelikteki bu araştırmanın herhangi bir idari veya siyasi yönü yoktur. Sonuçlar çocuğunuzun akademik başarısına herhangi bir etki içermemektedir. Başarı testinin üzerine çocuğunuzun ismini veya kimliğini içeren bilgiler yazılmayacaktır. Açık uçlu sorular içtenlikle cevaplandığı takdirde doğru verilere ulaşılmasında çok önemli katkılar sağlayacaktır.

RİSKLER:

Bu çalışmanın katılımcılar için maddi-manevi herhangi bir risk unsuru oluşturmaması öngörülmektedir.

YARARLAR:

Bu çalışma ile bütünsel ve analitik dereceli puanlama anahtarlarından hangisinin kullanılmasının daha uygun olduğunun cevabı aranacaktır. Ayrıca bu iki puanlama anahtarının karşılaştırılması ve hangisinin puanlayıcı tutarlığını artırttığı tartışılacaktır. Tüm bunlara ek olarak, puanlayıcılar arası güvenirlik belirleme yöntemleri, klasik test kuramı için, Kappa, Krippendorf Alfa teknikleri ve Genellenebilirlik yöntemlerinin incelenmesi ve karşılaştırılması sağlanacaktır.

GİZLİLİK:

Bu çalışmanın bilgileri kesinlikle gizli tutulacaktır. Elde edilen bilgiler sadece araştırmacılar tarafından görülecek ve değerlendirilecektir. Yazılı metinde kesinlikle isminiz geçmeyecek, bilgiler anonim olarak iletilecektir. Cevap kağıtları çalışma bitiminde imha edilecektir.

İLETİŞİM:

Sınav soruları, cevap kâğıtları veya çalışma ile ilgili bir sorunuz olduğu takdirde araştırmacılar ile istediğiniz zaman iletişime geçebilirsiniz.

Doç. Dr. Duygu ANIL Eğitim Bilimleri Bölümü Beytepe, Ankara.

Tel: 312–2978550/106

E-mail: duygu.anil73@gmail.com

Yük. Lisans Öğrencisi: Ar. Gör. Bulut Yıldıztekin

Eğitim Bilimleri Bölümü Beytepe, Ankara.

Tel: 312–2978564

E-mail: bulutyildiztekin@hotmail.com

Eğer bu formda sözü edilen kurallar dışında davranıldığı veya katılımcı olarak gizliliğiniz tehlikeye atıldığı takdirde Hacettepe Üniversitesi Senato Etik Komisyonu ile iletişime geçebilirsiniz.

Hacettepe Üniversitesi Senato Etik Komisyonu

Hacettepe Üniversitesi Rektörlük 1. Kat 06800, Ankara

Tel: 312–2976062

KATILIM:

Çocuğunuzun bu çalışmaya katılması tamamen **gönüllülük** ilkesine dayanmaktadır. İstediğiniz takdirde katılmayı herhangi bir yaptırıma maruz kalmaksızın reddedebilirsiniz. Çocuğunuzun çalışmaya katılmasına onay verseniz bile fikriniz değişirse çalışmayı durdurma hakkına sahipsiniz. Eğer çalışmayı veri toplama işlemi bitiminden önce bıraktıysanız isteğiniz halinde cevaplama kâğıdınız ve kopyaları size teslim edilecek veya imha edilecektir. Cevap kâğıdında verilen cevaplar ve problem çözümleri 5 öğretmen tarafından puanlanacak ve sonuçları tamamen **gizli** kalacaktır.

ONAY:

Bu belgeyi okudum ve kopyasını aldım. Gerekli gördüğüm bütün cevapları almış durumdayım ve çocuğumun bu çalışmaya katılmasını velisi olarak kabul ediyorum.

Velinin imzası: ……

Tarih: … ……

Veli onay formu tarihi :……

Ek–6: Analitik Dereceli Puanlama Anahtarı

ANALİTİK DERECELİ PUANLAMA ANAHTARI

Seviye → / Ölçütler / Bölümler ↓	Beceri yok (0)	Düşük beceri (1)	Orta beceri (2)	Yüksek beceri (3)	Puan
Problemi anlama ve kavrama	*Cevap veya açıklama yok	*Problem çok az anlaşılmış *Verilenler ve istenenler birbirine karıştırılmış *Problem ile ilgisi olmayan ve anlamsız notlar mevcut *Verilen ve istenenler kısmen anlaşılmış	*Problem kısmen anlaşılmış *Ölçütlerden bazıları anlaşılmış fakat eksik anlama mevcut. * Verilen ve istenenlerden birbirine karıştırılan bilgiler var	*Problem tamamen anlanmış gözüküyor *Verilen ve istenenler tamamen not edilmiş. *Birbirine karıştırılan kavram yok.	
Çözüm yolu geliştirme veya var olan yolu belirleme	*Cevap veya açıklama yok	*Belirlenen çözüm yolu problemin çözümüyle kısmen alakalı *Doğru olmayan çözüm yolları tercih edilmiş	*Belirlenen yöntem eksiklikler içeriyor. *Doğru çözüme götürecek plan belirlenmiş fakat eksikler mevcut	*Çözüm yolu tamamen doğru *Belirlenen çözüm yolu ile doğru cevaba ulaşılabilir	
Çözüm aşaması	*Cevap veya açıklama yok	*Çözüm var ancak çok fazla hata içeriyor *Çözüm kısmen doğru ve istenen doğru cevapların bir kısmını içeriyor	*Çözüm yolu doğru ancak, doğru cevaba ulaşırken son adımda işlem hatası mevcut	*Çözüm yolu ve uygulaması tamamen doğru ve doğru cevap(lara) ulaşılmış.	

Ek–7: Bütünsel Dereceli Puanlama Anahtarı

BÜTÜNSEL DERECELİ PUANLAMA ANAHTARI

Puan	Ölçüt	Öğrenci Puanı ve Dönütler
Beceri yok **(0)**	*Problem hiç anlaşılmamış veya hiçbir çözüm girişimi yok *Çözüme götürecek yollar belirlenmemiş *Problemdeki veriler aynen kopya edilmiş fakat uygulama yok *İşlemsiz, sadece sonuç var ve hiç açıklama yok	
Düşük beceri **(1)**	*Problem çok az anlaşılmış *Problem anlaşılmış ancak uygulama basamağı çok fazla işlem hatası içeriyor *Problem kısmen anlaşılmış, doğru olmayan çözüm yolu seçilmiş ve uygulama çözüme götürmüyor	
Geliştirilebilir beceri **(2)**	*Problem anlaşılmış ancak çözüm yolu veya uygulaması yok *Problem kısmen anlaşılmış ancak uygulamada işlem hataları var *Çözüme götüren yol seçilmiş ancak işlemler tamamlanmamış	
Orta düzey beceri **(3)**	*Problem anlaşılmış, çözüm yolu doğru olarak seçilmiş ancak uygulama aşamasında hatalar var	
Orta-Üstü beceri **(4)**	*Problem çok iyi anlaşılmış, doğru çözüm yolu seçilmiş ancak son aşamada işlem hatası var *Problem çok iyi anlaşılmış, herhangi kavram yanılgısı yok, farklı çözüm yolları ile sonuca ulaşmış fakat açıklamalar eksik	
Yüksek düzey beceri **(5)**	*Problem doğru anlaşılmış, çözüm yol(ları) doğru olarak belirlenmiş eksiksiz olarak hatasız bir uygulama ile cevap verilmiş	

EK-8: Kappa İstatistiği İçin Kullanılan Makro

```
*****************************************************************
* MACRO NAME: MKAPPASC.SPS *
* *
* README FILE: MKAPPASC.RM *
* *
* SPSS REQUIREMENTS: Release 4.0 or above *
* Advanced Statistics Module *
* *
* AUTHOR: David Nichols (nichols@spss.com) *
* *
* LAST UPDATED: 04/08/97 *
*****************************************************************
data list list /rater1 rater2 rater3 rater4 rater5 .
begin data

*veriler buraya giriliyor/

end data .
preserve.
set printback=off mprint=off.
save outfile=C:\Documents and Settings\user\Desktop\TEZ\tez\kaparr.sav'.

define cohensk (vars=!charend('/')).
set mxloops=100000.
count ms__=!vars (missing).
select if ms__=0.

*This section sets up a matrix(x) based on the raw data file, a matrix(y)
with rows equal to the number of items and
*columns equal to the number of categories, then determines for y the
number of responses per category for each
*case or item.

matrix.
get x /var=!vars.
compute c=mmax(x).
compute y=make(nrow(x),c,0).
loop i=1 to nrow(x).
loop j=1 to ncol(x).
loop k=1 to c.
do if x(i,j)=k.
compute y(i,k)=y(i,k)+1.
end if.
end loop.
end loop.
end loop.

*This section computes the basic information and kappa and its related
statistics.

compute pe=msum((csum(y)/msum(y))&**2).
compute k=ncol(x).
compute n=nrow(y) .
compute r=ncol(y) .
compute pa=mssq(y)/(nrow(y)*k*(k-1))-(1/(k-1)).
compute ck=(pa-pe)/(1-pe).
compute num=2*(pe-(2*k-3)*(pe**2)+2*(k-2)*msum((csum(y)/msum(y))&**3)).
compute den=nrow(y)*k*(k-1)*((1-pe)**2).
```

```
compute ase=sqrt(num/den).
compute z=ck/ase.
compute sig=1-chicdf(z**2,1) .
compute ckul=ck+1.96*ase .
compute ckll=ck-1.96*ase .

*This section computes the alternate standard error and related
statistics based on Fliess' corrected formula.

compute nm=sqrt(n*k*(k-1)) .
compute vectora=csum(y)/msum(csum(y)) .
compute vectorb=1-csum(y)/msum(csum(y)) .
compute vectorc=1-2*(csum(y)/msum(csum(y))) .
compute vectord=vectora&*vectorb.
compute vectore=vectora&*vectorb&*vectorc .
compute e=msum(vectord) .
compute f =msum(vectore) .
compute fse=(sqrt(2)/(e*nm))*(sqrt(e**2-f)) .
compute fsez=ck/fse .
compute fsesig=1-chicdf(fsez**2,1) .
compute fseu=ck+1.96*fse .
compute fsel=ck-1.96*fse .

* This section computes the kappas for the individual categories.  Each
statistic, e.g., k or standard error, is computed
* as a vector.  The vectors are then assembled into a matrix of all six
statistics.  As part of this process, -99999.0 is
* imputed as the missing data value.

compute matz=k-y .
compute mata=y&*matz .
compute vectorf=csum(mata)+(.0001) .
compute vectorg=vectord*(n*k*(k-1))+(.0001) .
compute vectorh=1-(vectorf&/vectorg) .
compute vectori=(1+(2*(k-1)*(csum(y)/msum(csum(y)))))&**2 .
compute vectorj=(2*(k-1))*vectord .
compute vectork=(n*k*(k-1)**2)*vectord+(.0001) .
compute vectorse=sqrt((vectori+vectorj)&/vectork) .
compute vectorz=vectorh&/(vectorse+.0001).
compute vectorp=1-cdfnorm(vectorz) .
compute vectorll=vectorh-(1.96*vectorse) .
compute vectorul=vectorh+(1.96*vectorse) .
loop i=1 to ncol(vectorh) .
do if (vectorh(i)=0.00) .
compute vectorh(i)=-99999) .
end if .
end loop .
loop i=1 to ncol(vectorh) .
do if (vectorh(i)=-99999) .
compute vectorse(i)=-99999 .
end if .
end loop .
loop i=1 to ncol(vectorh) .
do if (vectorh(i)=-99999) .
compute vectorz(i)=-99999 .
end if .
end loop .
loop i=1 to ncol(vectorh) .
do if (vectorh(i)=-99999) .
compute vectorp(i)=-99999 .
end if .
end loop .
```

```
loop i=1 to ncol(vectorh) .
do if (vectorh(i)=-99999) .
compute vectorul(i)=-99999.
end if .
end loop .
loop i=1 to ncol(vectorh) .
do if (vectorh(i)=-99999) .
compute vectorll(i)=-99999 .
end if .
end loop .
compute ikstat={vectorh;vectorse;vectorz;vectorp;vectorll;vectorul} .
save ikstat /outfile='ikstat1.sav' .

*This section saves the data and prepares the reporting formats.

save {k,n,r,pa,ck,ase,z,sig,ckll,ckul,fse,fsez,fsesig,fseu,fsel}
/outfile='k__tmp2.sav'
     /variables=k,n,r,pa,ck,ase,z,sig,ckll,ckul,fse,fsez,fsesig,fseu,fsel
end matrix .
get file='k__tmp2.sav'.
formats k (f8.0) /n (f8.0) /r (f8.0) /pa (f8.4) /ck (f8.4) /ase (f8.4) /z
(f8.4) /sig (f8.4) /ckul (f8.4) /ckll (f8.4)
 /fse(f8.4) /fsez (f8.4) /fsesig (f8.4) /fseu (f8.4) /fsel (f8.4) .
variable labels k 'Number of Raters' /n 'Number of Items' /r 'Number of
Categories' /pa 'Percent of Rater Agreement'
 /ck 'Kappa' /ase 'Standard Error' /z 'z'/sig 'p' /ckul 'Upper 95%
Confidence Limit' /ckll 'Lower 95% Confidence Limit'
 /fse 'Fleiss SE' /fsez 'z' /fsesig 'p' /fseu 'Upper 95% Confidence
Limit' /fsel 'Lower 95% Confidence Limit' .
report format=list automatic align(center)
  /variables=k,n,r,pa
  /title "Basic Information" .
report format=list automatic align(center)
  /variables=ck ase z sig ckll ckul
  /title "Cohens Kappa".
report format=list automatic align(center)
  /variables=fse fsez fsesig fsel fseu
  /title "Cohens Kappa -- Fleiss Adjusted Standard Error" .
get file='ikstat1.sav' .
flip .
delete variable case_lbl .
compute n1=$casenum .
formats n1 (f8.0) /var001 (f8.4) /var002 (f8.4)  /var003 (f8.4)  /var004
(f8.4)  /var005 (f8.4)  /var006 (f8.4)  .
variable labels n1 'Coding Category' /var001 'k' /var002 'Standard Error'
/var003 'z' /var004 'p'
  /var005 'Lower 95% Confidence Limit' /var006 'Upper 95% Confidence
Limit' .
save outfile='ikstat.sav' .
report format=list automatic align(center)
  /variables=n1 var001 var002 var003 var004 var005 var006
  /Title "Individual Category Statistics" .
!enddefine.
restore.
COHENSK VARS = rater1 to rater5 .
```

EK-9: Krippendorff Alfa İstatistiği İçin Kullanılan Makro

```
/* This macro computes Krippendorff's Alpha reliability estimate for judgments */.
/* made at any level of measurement, any number of judges, with or */.
/* without missing data. The macro assumes the data file is set up */.
/* in a SPSS data file with judges as the variables and the units being */.
/* judged in the rows. The entries in the data matrix should be */.
/* the coding (quantified or numerically coded for nominal judgments) given */.
/* to the unit in that row by the judge in that column. Once the macro is */.
/* activated (by running the command set below), the syntax is */.
/* */.
/* KALFA judges = judgelist/level = a/detail = b/boot = z.
/* */.
/* where 'judgelist' is a list of variable names holding the names of the */.
/* judges, 'a' is the level of measurement (1 = nominal, 2 = ordinal, */.
/* 3 = interval, 4 = ratio), 'b' is set to 1 if you desire SPSS to print */.
/* the coincidence and delta matrices, and 'z' is the number of bootstrap */.
/* samples desired for inference; z must be at least 1000 and is truncated to the */.
/* lowest 1000 entered (for exampole, 2300 is truncated to 2000) */.
/* The '/level' and '/detail' and '/boot' subcommands are */.
/* optional and default to 1,0, and 0, respectively, if omitted */.
/* */.
/* Missing data should be represented with a 'period' character */.
/* Units that are not coded by at least one judge are excluded from */.
/* the analysis */.
/* */.
/* This macro is version 3.0, updated on February 5, 2011 */.
/* */.
/* */.
/* Written by Andrew F. Hayes */.
/* School of Communication */.
/* The Ohio State University */.
/* hayes.338@osu.edu */.
/* http://www.afhayes.com */.
DEFINE kAlfa (judges = !charend ('/')/level = !charend('/') !default(1)/detail = !charend('/')
!default(0)/boot = !charend('/') !default(0)).
PRESERVE.
SET MXLOOP = 900000000.
SET LENGTH = NONE.
SET SEED = RANDOM.
SET PRINTBACK = OFF.
MATRIX.
get dat/variables = !judges/file = */names = vn/missing = -9999999.
compute btn = !boot.
do if (!boot > 0).
compute btn = trunc(!boot/1000)*1000.
end if.
do if (!boot > 0 and btn = 0).
print/title = "Number of bootstraps must be at least 1000.".
end if.
compute btprob = 0.
/* FIRST WE CREATE THE DATA FILE EXCLUDING OBJECTS WITH ONLY ONE
JUDGMENT */.
```

```
/* THAT DATA FILE IS HELD IN DAT AND DAT3 */.
compute rw = 1.
loop i = 1 to nrow(dat).
compute good = 0.
loop j = 1 to ncol(dat).
do if (dat(i,j) <> -9999999).
compute good = good + 1.
end if.
end loop.
do if (good > 1).
compute dat(rw,:) = dat(i,:).
compute rw = rw+1.
end if.
end loop.
compute dat = dat(1:(rw-1),:).
compute nj = ncol(dat).
compute nobj = nrow(dat).
compute dat3 = dat.
/* NOW WE CREATE A SINGLE COLUMN OF DATA TO FIGURE OUT HOW MANY
*/.
/* UNIQUE JUDGMENTS ARE MADE, AND WE SORT IT */.
compute m = reshape(t(dat),(nobj*nj),1).
compute allm = nobj*nj.
compute j = 0.
loop i = 1 to nrow(m).
do if m(i,1) <> -9999999.
compute j = j + 1.
compute m(j,:)=m(i,:).
end if.
end loop.
compute m = m(1:j,1).
compute mss = nrow(m).
compute mss = allm-mss.
compute mtmp = m.
compute mtmp(GRADE(m)) = m.
compute m = mtmp.
compute m2 = make(nrow(m),1,m(1,1)).
compute yass = csum((m = m2))/nrow(m).
do if (yass <> 1).
compute des = design(m).
compute uniq = ncol(des).
compute coinc = make(uniq,uniq,0).
compute delta = coinc.
compute map = make(uniq,1,0).
loop i = 1 to nrow(m).
loop j = 1 to uniq.
do if (des(i,j) = 1).
compute map(j,1) = m(i,1).
end if.
end loop.
end loop.
loop i = 1 to nobj.
loop j = 1 to nj.
```

```
do if dat(i,j) <> -9999999.
loop k = 1 to uniq.
do if dat(i,j) = map(k,1).
compute dat(i,j) = k.
BREAK.
end if.
end loop.
end if.
end loop.
end loop.
compute datms = (dat <> -9999999).
compute mu = rsum(datms).
compute nprs = csum(mu&*(mu-1))*.5.
compute btalp = make((btn+1),1,-999).
/* THIS CONSTRUCTS THE COINCIDENCE MATRIX FROM THE MATRIX DATA */.
loop k = 1 to nobj.
compute temp = make(uniq, uniq, 0).
loop i = 1 to nj.
loop j = 1 to nj.
do if (dat(k,i) <> -9999999 AND dat(k,j) <> -9999999 AND i <> j).
compute temp(dat(k,i),dat(k,j)) = temp(dat(k,i),dat(k,j)) + (1/(mu(k,1)-1)).
end if.
end loop.
end loop.
compute coinc = coinc + temp.
end loop.
compute q = reshape(coinc, (nrow(coinc)*ncol(coinc)), 1).
compute q = csum(q > 0).
compute nc = rsum(coinc).
compute n = csum(nc).
compute coinct = coinc.
compute dmat = diag(coinc).
compute nzero = csum(dmat > 0).
compute bootm = nprs.
compute nx = (dmat/n)&**bootm.
compute nx=rnd(btn*csum(nx)).
compute numone = 0.
/* THIS CONSTRUCTS THE EXPECTED MATRIX */.
compute expect = coinc.
loop i = 1 to uniq.
loop j = 1 to uniq.
do if (i = j).
compute expect(i,j)=nc(i,1)*(nc(j,1)-1)/(n-1).
else if (i <> j).
compute expect(i,j)=nc(i,1)*nc(j,1)/(n-1).
end if.
end loop.
end loop.
compute tst = 100*q.
compute tst = {tst; (nj-1)*n}.
compute bootm2 = cmin(tst).
loop z = 1 to (btn + 1).
/* HERE IS WHERE WE START DOING THE BOOTSTRAPPING */.
```

```
do if (z > 1).
compute rand = uniform(bootm2,1).
compute numsum = 0.
loop i = 1 to bootm2.
loop j = 2 to indx+1.
do if (rand(i,1) <= pmat(j,1)).
do if (rand(i,1) >= pmat(j-1,1)).
compute numsum = numsum + pmat(j,2).
end if.
end if.
end loop.
end loop.
compute Alfa = 1 - (numsum*(1/(expdis*bootm2))).
do if (Alfa< -1).
compute alpha = -1.
end if.
do if (alpha = 1 and nzero = 1).
compute alpha
end if.
do if (alpha = 1 and nzero > 1).
compute numone = numone + 1.
end if.
compute btalp(z,1) = alpha.
end if.
do if (z = 1).
do if (!level = 2).
compute delta = make(uniq,uniq,0).
loop i = 1 to uniq.
loop j = i to uniq.
do if (i <> j).
compute delta(i,j) = (csum(nc(i:j,1))-(nc(i,1)/2)-(nc(j,1)/2))**2.
compute delta(j,i) = delta(i,j).
end if.
end loop.
end loop.
compute v = {"Ordinal"}.
do if (z = 1).
compute deltat = delta.
end if.
end if.
do if (!level = 1).
compute delta = 1-ident(uniq).
compute v = {"Nominal"}.
compute deltat = delta.
end if.
do if (!level = 3).
loop i = 1 to uniq.
loop j = i to uniq.
do if (i <> j).
compute delta(i,j) = (map(i,1)-map(j,1))**2.
compute delta(j,i) = delta(i,j).
end if.
end loop.
```

```
end loop.
compute v = {"Interval"}.
compute deltat = delta.
end if.
do if (!level = 4).
loop i = 1 to uniq.
loop j = i to uniq.
do if (i <> j).
compute delta(i,j) = ((map(i,1)-map(j,1))/(map(i,1)+map(j,1)))**2.
compute delta(j,i) = delta(i,j).
end if.
end loop.
end loop.
compute v = {"Ratio"}.
compute deltat = delta.
end if.
compute num = csum(rsum(delta&*coinc)).
compute den = csum(rsum(delta&*expect)).
do if (den > 0).
compute alp = 1-(num/den).
compute btalp(1,1)=alp.
compute expdis=csum(rsum((expect&*delta)))/n.
end if.
/* NOW WE COMPUTE THE FUNCTION FOR BOOTSTRAPPING */.
compute pcoinc = 2*(coinc/n)-(mdiag(diag(coinc))/n).
compute temp = mdiag(diag(coinc))/n.
compute pmat = make((uniq+((uniq*(uniq-1))/2)),2,0).
compute psum = 0.
compute ct = 1.
loop i = 1 to uniq.
loop j = i to uniq.
compute psum = psum+pcoinc(j,i).
compute pmat(ct,1)=psum.
compute pmat(ct,2)=delta(j,i).
compute ct=ct+1.
end loop.
end loop.
compute indx = nrow(pmat).
compute t3 = {0,0}.
compute pmat = {t3;pmat}.
end if.
end loop.
compute alpfirst = btalp(1,1).
/* NOW WE CALCULATE CI AND P(Q) FROM BOOTSTRAPPING */.
do if (btn > 0).
compute btalp=btalp(2:nrow(btalp),1).
/* FIRST WE CORRECT DISTRIBUTION OF NEED BE */.
do if (nx > 0 and nzero > 1).
compute chk1 = 0.
compute chk2 = 0.
loop i = 1 to nrow(btalp).
do if (nx >= numone and btalp(i,1) = 1 and chk1 < numone).
compute btalp(i,1) = 0.
```

```
compute chk1 = chk1 + 1.
end if.
do if (nx < numone and btalp(i,1) = 1 and chk2 < nx).
compute btalp(i,1) = 0.
compute chk2 = chk2 + 1.
end if.
end loop.
end if.
/* NOW WE SORT THE BOOTSTRAP ESTIMATES */.
compute btalptmp = btalp.
compute btalptmp(GRADE(btalp)) = btalp.
compute btalp = btalptmp.
compute btalp = btalp(1:nrow(btalp),1).
compute mn = csum(btalp)/btn.
compute low95 = trunc(.025*btn).
compute high95 = trunc(.975*btn)+1.
compute low95 = btalp(low95,1).
compute high95 = btalp(high95,1).
compute median = btalp(0.50*btn).
compute q = {.9, 0; .8, 0; .7, 0; 0.67, 0; .6, 0; .5, 0}.
loop i = 1 to 6.
compute qcomp = (btalp < q(i,1)).
compute qcomp = csum(qcomp)/btn.
compute q(i,2)=qcomp.
end loop.
end if.
do if (btalp(1,1) = -999).
compute btprob = 1.
end if.
print/title = "Krippendorff's Alpha Reliability Estimate".
do if (btn = 0 or btprob = 1).
compute res = {alpfirst, nobj, nj, nprs}.
compute lab = {"Alpha", "Units", "Obsrvrs", "Pairs"}.
end if.
do if (btn > 0 and btprob = 0).
compute res = {alpfirst, low95, high95, nobj, nj, nprs}.
compute lab = {"Alpha", "LL95%CI", "UL95%CI", "Units", "Observrs", "Pairs"}.
end if.
print res/title = " "/rnames = v/cnames = lab/format = F10.4.
do if (btn > 0 and btprob = 0).
print q/title = "Probability (q) of failure to achieve an alpha of at least alphamin:"/clabels =
"alphamin" "q"/format = F10.4.
save btalp/outfile = alpboot.sav/variables = alpha.
print btn/title = "Number of bootstrap samples:".
end if.
print vn/title = "Judges used in these computations:"/format = a8.
do if (!detail = 1).
print/title = “KRİP”
```

```
print coinct/title = "Observed Coincidence Matrix"/format = F9.2.
print expect/title = "Expected Coincidence Matrix"/format = F9.2.
print deltat/title = "Delta Matrix"/format F9.2.
compute tmap = t(map).
print tmap/title "Rows and columns correspond to following unit values"/format = F9.2.
end if.
else.
print/title = "ERROR: Input Reliability Data Matrix Exhibits No Variation.".
end if.
do if (btprob = 1).
print/title = "A problem was encountered when bootstrapping, so these results are not printed".
end if.
print/title = "Examine output for SPSS errors and do not interpret if any are found".
END MATRIX.
RESTORE.
!ENDDEFINE.
```

ÖZGEÇMİŞ

Adı Soyadı	Bulut Yıldıztekin
Doğum Yeri	Ankara
Doğum Yılı	1989
Medeni Hali	Evli

Eğitim ve Akademik Durumu

Lise	Eskişehir Yunus Emre Anadolu Öğretmen Lisesi	2003–2007
Lisans	ODTÜ-İlköğretim Matematik Öğretmenliği	2007–2012
Yandal	ODTÜ-İstatistik	2010–2012
Yabancı Dil	İngilizce	İyi düzeyde
	Almanca	Başlangıç düzeyi
İş Deneyimi	Hacettepe Üniversitesi Eğitim Bilimleri Bölümü-Araştırma Görevlisi	2012–