Int. J. Asst. Tools in Educ., Vol. 4, Issue 2, (2017) pp. 149-165

http://www.ijate.net

e-ISSN: 2148-7456 © IJATE

INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

"Review Article"

A Generalizability Analysis of the Reliability of Measurements: "An Example of Cell Division and Heredity Unit"

Gülşah BAŞOL[*,1] **Muammer YÜKSEL**[2]

[1]Gaziosmanpaşa Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Böl., Taşlıçiftlik Yerleşkesi, 60100 Tokat, Türkiye.
[2]Ayşe Temizel Ortaokulu, 45500 Soma/Manisa, Türkiye.

Abstract

The purpose of the study is to measure students' performance through different measurement tools and compare the findings through G Theory in order to identify the errors associated with the raters and items to improve the future applications. The sample consisted of 48 eighth graders in Kars. Two different types of exams (a multiple choice test and an essay) were applied and essays were graded by three raters. G and K analyses were performed on the results. According to the findings, the error rate was higher for the essays in comparison to multiple-choice test. The mean score was higher for the multiple-choice test, the variances were similar. There were no differences among the essay scores given by different raters. Findings of decision study indicated Student facet as the main source of the variation in the data for both types of the measurements.

Article Info

Received
03 February 2017

Revised
23 March 2017

Accepted
28 March 2017

Keywords
Measurement,
Generalizability Theory,
Reliability,

Bir Genellenebilirlik Analizi Çalışması: "Hücre Bölünmesi & Kalıtım Ünitesi"

Özet

Araştırmanın amacı, aynı konu alanı ile farklı ölçme araçlarından elde edilen puanların G Kuramı ile karşılaştırılmasıdır. Örneklem Kars'ta öğretim gören 48 sekizinci sınıf öğrencisinden oluşmuştur. İki farklı sınav türü (çoktan seçmeli test ve yazılı sınav) öğrencilere uygulanmış ve yazılı sınav üç puanlayıcı tarafından puanlanmış, sonuçlar üzerinde G ve K analizleri yapılmıştır. Sonuçlar çoktan seçmeli test ve yazılı sınav için karışan hata varyanslarının yazılı sınavda daha fazla olduğunu, çoktan seçmeli testin puan ortalamasının daha yüksek olduğunu, varyansların ise iki sınav türü için birbirine yakın olduğunu ortaya koymuştur. Yazılı sınav için puanlayıcılar arasında herhangi bir fark bulunmamıştır. Ayrıca, karar çalışmasından elde edilen sonuçlar verideki varyansın ana kaynağının Öğrenci faktörü olduğunu ortaya koymuştur.

Makale Bilgisi

Makale Gönderim
03 Şubat 2017

Makale Düzeltme:
23 Mart 2017

Makale Kabul
28 Mart 2017

Anahtar Kelimeler:
Ölçme,
Genellenebilirlik Kuramı,
Güvenirlik,

[*]Corresponding Authors' E-mails: gulsah.basol@gop.edu.tr yuksel.muammer.my@gmail.com

Bu çalışma, 1-3 Eylül 2016'da Antalya'da gerçekleştirilen V. Eğitimde Ölçme ve Değerlendirme Kongresi'nde bildiri olarak sunulmuştur.

2148-7456 /© 2017 DOI: 10.21449/ijate.303991

1. GİRİŞ

Son yıllarda ön plana çıkan yaşam boyu öğrenme anlayışı, bireylerin gelişmelerine büyük oranda katkı sağlamaktadır. Hayatımızın her safhasında yeni şeyler öğrenme ve kendimizi geliştirme imkanı sağlayan bu görüş, öğrenmelerin ve eğitimin önemini daha çok ortaya çıkarmaktadır. Eğitim süreci bünyesinde yer alan hazırbulunuşluk, güdülenme, öğretim, ölçme ve değerlendirme gibi kavramlar da artık hayatın içinde daha sık benimsenmektedir. Her biri birbirinin tamamlayıcısı ve önemli bir parçası olan içi içe geçmiş bu süreçleri bilmek eğitim ve öğretimin kalitesini arttırır.

Eğitim istendik davranış oluşturma veya istendik davranış değiştirme süreci olarak, toplumun süzgeçten geçirilmiş değerlerinin, ahlak standartlarının, bilgi ve beceri birikimlerinin yeni nesillere aktarılmasıdır (Senemoğlu, 2002). Eğitim süreci sonunda bireylerin belli konularda bilgi, beceri ve tutum kazanması beklenir. Bu istendik bilgi, beceri ve tutumların kazanılma düzeyinin; sürecin verimliliğini göstermesi ve dönüt sağlayarak süreci zenginleştirmesi beklenir. Eğitim sistemimizde bireylerin bu kazanımları başarı olarak nitelendirilmekte ve farklılaşan başarı düzeylerinin doğru olarak ölçülmesine çalışılmaktadır.

Etkinlikler sonunda beklenen kazanımların; bir kısmının oluştuğu, bir kısmının yeterli düzeyde oluşmadığı, istenmeyen kazanım şeklinde ortaya çıktığı veya planlandığı şekilde oluşmadığı görülmektedir. Bu durum eğitimde kontrol ihtiyacını doğurur (Turgut ve Baykul, 2010). Burada yer alan kontrol kavramı eğitim sürecinin ve ürünlerinin gözden geçirilmesi ve bir sonuca varılması anlamına gelmektedir. Kontrol süreci eğitimi hem planlı hale getirir hem de var olan eksikliklerin giderilmesine ve kalitenin arttırılmasına olanak sağlar.

Öğrencilerin başarılarının belirlenmesinde öncelikle ölçme ve sonrasında bunu da içine alan değerlendirme sürecine yer verilmelidir. Eğitim sürecindeki bireylerin eğitimden ne kadar yararlandıkları ya da öğrenilmesi beklenilen kazanımlara ne ölçüde ulaşıldığı sürekli merak konusudur. Çünkü hem eğitimin niteliği hem de bireyler hakkında verilecek kararlar için kazanımların ulaşılma düzeyleri saptanmalıdır. Burada da devreye ölçme ve değerlendirme süreci girer.

Kazanımla ifade edilen hedefleri gerçekleştirme yolunda öğretim etkinlikleri planlanır. Öğretimde izlenen yöntemi de dikkate alarak farklı ölçme araçları arasından, öğrenmenin gerçekleşip gerçekleşmediğini yoklamak için en uygun olanı seçilir. Değerlendirmenin amacına göre kullanılan ölçme araçları da çeşitlilik gösterir. Ölçme yönteminin hedeflenen kazanımlara uygun olması ölçme sonuçlarının geçerliği için önemlidir. Bu nedenle ölçülmek istenilen kazanımların niteliğine en uygun olabilecek ölçme aracının seçilmesine gerekli önem verilmelidir.

Kullanılacak ölçme aracına; öğrencinin hazırbulunuşluk düzeyi, sınavın yapılacağı ortam, zaman sınırlaması olup olmadığı ve uygulama koşulları gibi faktörler dikkate alınarak karar verilir. Farklı ölçme araçlarıyla elde edilen sonuçların benzer olup olmadığı araştırılmaya değer bir konudur. Bu sayede farklı kaynaklardan ulaşılan ölçme sonuçlarının güvenilir olup olmadığını anlamak mümkün olur. Eğitim sürecinin önemli bir parçası olan ölçme için bireylerin farklı ölçmeler neticesinde ortaya çıkan sonuçların birbirleri ile ilişkisinin nasıl olduğu bir merak konusudur. Bu soruların cevaplarının bulunması elbette performans, not ve başarı seviyesi olarak ilerleyen sürecin daha anlamlı şekilde açıklanmasını sağlayabilir.

Katılımcıların performansının ölçülmesinin amaçlandığı araştırmalarda araştırma sürecinde yer alan ve araştırmayı etkileyen ya da etkileyebilecek pek çok değişken kaynağı bulunmaktadır.

Bu değişken kaynaklarının etkilerinin olup olamadığı ya da etkilerinin ne ölçüde olduğunun ortaya konmasında farklı ölçme kuramlarından yararlanılmaktadır. Bu kuramlardan biri olan Genellenebilirlik Kuramı (G Kuramı); hata kaynaklarını aynı anda ele alması ve birbirleri ile ilişkilerine yer vermesi nedeni ile araştırmada değişken kaynaklarının birbirleri ile karşılaştırılmasına olanak vermektedir.

G Kuramı ölçme sonuçlarının güvenirliğinin belirlenmesini, güvenilir gözlemlerin tasarımını, araştırılmasını ve kavramsallaştırılmasını sağlayan istatistiksel bir kuramdır. G Kuramı, Klasik Test Kuramı (KTK)' nın bir uzantısıdır (Cronbach ve diğerleri, 1972; Brennan, 2001). G Kuramı, KTK' nın günümüzde hala popüler olan gerçek puan modelinin sınırlılıklarına olan cevap vermek amacıyla Cronbach ve arkadaşları (1963) tarafından ortaya atılmıştır. KTK, bir tek gerçek puana sahip her bir gözlem ya da test puanının paralel gözlemlerin bir grubuna ait tek bir güvenirlik katsayısı üretmesi fikri etrafında merkezlenir (Lord ve Novick, 1968; Baykul, 2000). G Kuramı ölçüm prosedürlerinin geliştirilmesine uygulanmış olmakla birlikte, özellikle eğitim araştırmaları içinde uygulaması sınırlı kalmıştır (Bottema-Beutel, Lloyd, Carter ve Asmus, 2014).

Shavelson ve Webb'e (1991) göre, G Kuramı dört farklı açıdan KTK' nın daha genişletilmiş bir halidir: 1. Genellenebilirlik Kuramı, çoklu varyans kaynaklarını tek bir analizde ele alır. 2. Her bir varyans kaynağının büyüklüğünün belirlenmesini sağlar. 3. Bireylerin performanslarına dayalı hem bağıl kararlar hem de mutlak kararlar alınmasına ilişkin iki farklı güvenirlik katsayısının (sırasıyla; G katsayısı ve Phi katsayısı) hesaplanmasına olanak sağlar. 4. Belirli bir amaca bağlı olarak, ölçme hatasının en aza indirgenebileceği ölçmelerin düzenlenmesine (Karar "K" çalışmaları) imkân tanır.

G Kuramı farklı hata kaynaklarının varyans analizi yoluyla ayrı ayrı ve bir arada rapor edilerek kestirmesini sağlar. Genellenebilirlik Kuramında yer alan çoklu hata kaynakları bir örnek üzerinden açıklanabilir. Bir başarı testinin iki ya da daha fazla puanlayıcı tarafından puanlandığı bir durumda, kestirilebilecek hata kaynağı ile aynı testin paralel formlarından elde edilen puanlara ilişkin kestirilen hata kaynağı aynı olmayacaktır. Klasik Test Kuramında bu hata kaynaklarını aynı anda kestirmek mümkün değildir (Güler, 2009).

G Kuramına göre değişkenlik kaynakları çapraz (crossed) ya da yuvalanmış (nested) şekilde olabilir (Rentz, 1987). Çaprazlanmış desende değişkenlik kaynağının koşulları başka bir değişkenlik kaynağının koşullarıyla örtüşmektedir (Brennan 2001). Çaprazlanmış desende değişkenlik kaynakları arasına 'x' işareti konulmaktadır. Araştırmada bir değişken kaynağı diğer değişken kaynağının tüm koşulları ile örtüşmüyor, sadece belli koşulları ile örtüşüyor ise bu çalışma desenine yuvalanmış desen denilmektedir. Yuvalanmış desende değişkenlik kaynakları arasına ' : ' işareti konulur.

G Kuramında güvenirliğin araştırılması iki aşamadan oluşmaktadır. Bunlardan ilki Genellenebilirlik çalışması (G-çalışması) ve ikincisi Karar çalışması (K-çalışması) şeklindedir (Kaya, 2011). G çalışması, ölçüm hatasını makul ve ekonomik olarak çok yönlü yalıtmak ve tahmin etmek, uygulama yapabilmek üzere tasarlanmıştır (Shavelson ve Webb, 2005). G çalışmasının amacı, ölçmenin birden çok kullanımını kestirmek ve bu sayede varyans kaynakları ile ilgili mümkün olan en çok bilgiye ulaşmaktır. G çalışması, mümkün olan en çok değişkenlik kaynağını içerecek biçimde tasarlanmalıdır. Bir başka deyişle G çalışması, kabul edilebilir gözlemlerin evrenini mümkün olan en geniş şekilde tanımlar (Shavelson ve Webb, 1991).

G-çalışması sürecinde, örneklemin evrene genellenebilmesi için, puanların değişkenliğinin tüm kaynakları (varyans bileşenleri) ve bunlar arasındaki etkileşimler aynı anda ANOVA yöntemi

kullanılarak kestirilmektedir. Kestirilen bu varyans bileşenleri bir sonraki aşama olan K-çalışmasında kullanılır (Kaya, 2011). G çalışması sonucunda elde edilen sonuçların K çalışmasında kullanımı söz konusudur ya da araştırmacı isterse devam etmeyip, çalışmasını G çalışması olarak sonlandırabilir.

K-çalışması, karar vermek üzere belirli bir amaç için veri toplanan çalışmadır ve yapılan bir K çalışmasında, incelenen bireyleri tanımlamak için veri toplanabilir (Kaya, 2011). Bir G çalışmasına karşılık, birden fazla K çalışması yapılabilir. K çalışması ile güvenirlik katsayısına benzeyen genellenebilirlik katsayısına (*G katsayısı*) ve güvenirlik indeksine (*Phi katsayısı*) ulaşılır. G katsayısı evren puanı varyansının kendisi ile bağıl puan varyansının toplamına oranıdır ve bağıl modellerde çalışılmaktadır (Çakıcı Eser, 2011).

G katsayısı KTK' daki güvenirlik katsayısına benzemektedir. G katsayısı, göreli karar modelinde gerçek varyansın, göreceli varyans ve gerçek varyansın toplamına bölünmesi ile bulunmaktadır. Öte yandan güvenilirlik endeksi ya da Phi (Φ) katsayısı mutlak karar modeli ile kullanılmaktadır. Phi katsayısı, gerçek varyansın, mutlak hata varyansı ve gerçek varyansın toplamına bölünmesi ile hesaplanır. Diğer bir deyişle, bu iki katsayı hatanın ne koşullarda kabul edileceğine göre farklılık göstermektedir (Alharby, 2006). Sonuç olarak, tek bir G-çalışmasından elde edilen aynı varyans kestirimlerine dayalı pek çok K-çalışması düzenlenebilir. K-çalışmasında kullanılan formül Spearman-Brown formülüne benzerdir (Mushquash ve O'Connor, 2006).

Cronbach ve arkadaşları tarafından 1963 yılında temelleri atılan G Kuramı ile ilgili çalışmalar yurt dışında aynı tarihleri takriben başlarken, ülkemizde 2004 yılından itibaren ve daha çok yüksek lisans ve doktora tezleri üzerinde yoğunluk göstermiştir. Bu yeni kuram; başlarda tezlerde yapılan araştırmalarla, günümüzde makalelerle ve üzerine yazılan bir kitap ile (Güler, Kaya Uyanık ve Taşdelen Teker, 2012) daha çok dikkat çekmeye başlamıştır.

Ülkemizde henüz yangınlaşmaya başlayan G Kuramı çalışmaları genellikle performansın ölçülme süreci, puanlayıcılar ve klasik ölçme araçları üzerinde yoğunlaşmıştır. Puanlayıcıların, bireylerin ve maddelerin etkileri araştırılırken farklı desenlerin incelendiği araştırmalar (Wang, 2005; Au, Prahardhi ve Shiell 2008; Lane ve Sabers, 1989; Nalbantoğlu Yılmaz ve Uzun Başusta, 2012; Nalbantoğlu , 2009) daha çok yoğunluk kazanmıştır.

Atılgan (2004); Güler (2008) ve Alkahtani (2012) G Kuramı ile yaptıkları çalışmalarında, KTK yanında Çok Değişkenli Rasch Modelini (ÇDRM) kullanmışlar; maddelerin zorluk düzeyleri ve puanlayıcıların puan verme eğilimleri hakkında bilgiye ulaşmaya çalışmışlardır. Kuramların ve modellerin karşılaştırılmasının yanında, bazı çalışmalarda Lojistik Regresyon Analizi kullanılması, farklı kesme puanları hesaplama yöntemlerinin karşılaştırılması, farklı ölçeklerin güvenirliklerinin araştırılması çalışmaları G Kuramı yardımıyla yapılmıştır.

Araştırmada aşağıda yer alan alt problemlere cevaplar aranmıştır:

1. Çoktan seçmeli test için G Kuramına göre kestirilen parametrelerin varyansları ve toplam varyansları açıklama yüzdeleri nedir?

2. Çoktan seçmeli test için yapılan K çalışması sonuçlarına göre G ve Phi katsayılarının değişimleri nasıldır?

3. Yazılı sınav için yapılan G Kuramına göre kestirilen parametrelerin varyansları ve toplam varyansları açıklama yüzdeleri nedir?

4. Yazılı sınav için yapılan K çalışmasına göre farklı senaryolara göre G ve Phi katsayılarının değişimi nasıldır?

2. YÖNTEM

Başol (2008)'e göre betimsel araştırmalar ne ve nasıl sorularına sistematik olarak cevap vererek, olay ve durumların detaylı olarak betimlenmesi amacıyla yapılır. Araştırma, G Kuramı ile mevcut sistemde öğretmenlerin aynı konu hakkında kullandıkları ölçme araçları arasında ilişkiyi belirleme çalışması olduğundan betimsel bir araştırma niteliği taşımaktadır.

2.1. Evren ve örneklem

Araştırmanın çalışma evrenini 2013-2014 Eğitim-Öğretim yılında Kars il merkezinde öğrenim gören 8. sınıf öğrencileri oluşturmaktadır. Araştırma örneklemini ise Kars il merkezinde yer alan bir ortaokulda öğrenim gören 48 öğrenci oluşturmaktadır. Araştırmada uygulama kolaylığından dolayı amaçlı örnekleme gitmiştir.

2.2. Veri toplama aracı

Araştırma için gerekli veriler, araştırmacılar tarafından hazırlanan yazılı sınav (essay) ve ölçme sürecinde daha önce kullanılmış olan sorular arasından seçilen çoktan seçmeli test sorularına verilen cevaplardan elde edilmiştir**.** Araştırma soruları için 8. sınıf fen bilimleri dersinde yer alan 'Hücre Bölünmesi ve Kalıtım' ünitesine ait 20 kazanım ele alınmış olup öğrenci seviyesi de düşünülerek çoktan seçmeli test için ilk olarak 40 madde seçilmiştir. Hazırlanan bu sınav öncelikle iki konu alanı uzmanı ve bir dil uzmanına danışılarak deneme formatı için hazır hale getirilmiştir. Deneme uygulaması Kars il merkezinde yer alan farklı üç okulda öğrenim gören 96 öğrenci üzerinde yapılmış ve büyük ölçüde eksik olduğu belirlenen altı katılımcının cevapları çıkarılmıştır. Geriye kalan 90 kişinin cevapları dikkate alınmış ve deneme uygulamasının yapıldığı 90 kişiden oluşan grup nihai uygulamaya dahil edilmemiştir.

Deneme uygulaması için test ve madde istatistikleri TAP.exe (Brooks ve Johanson, 2003) uygulaması kullanılarak elde edilmiştir. Konu alanı ve kazanımların ağırlıkları incelenmiş ve alan uzmanların görüşüde alınarak 40 madde hazırlanmış ancak konu alanını daha iyi temsil ettiği ve bazı kazanımlar için yazılan soru sayısının dağılımın farklı olduğu için madde güçlük katsayıları ve madde ayırt edicilik güçleri incelenerek ağırlılıklı olarak orta güçlük seviyesinde, madde ayırtdiciliği .40 üzerinde olan maddeler seçilerek her biri dört seçenekli 22 maddelik coktan seçmeli testi oluşturmuştur.

Yazılı sınav için iki konu alanı uzmanının görüşüne başvurularak kapsam geçerliliğinin sağlanması amacıyla sorular hazırlanmış ve bir dil uzmanına danışılarak uygulama formu hazır hale getirilmiştir. Soruların yanlış anlaşılmalara neden olmaması ve tarafsızlığa hizmet etmesi açısından, bir kız ve bir erkek öğrenciye önceden çözdürülmüştür. Sınavın uygulandığı bu iki öğrenci için uygulanan sınav sonrası öğrenci görüşleri ele alındığında cinsiyete göre yanlılığının olmadığı sonucuna ulaşılmıştır. Ayrıca bu iki öğrenci nihai uygulama grubu arasında yer almamıştır.

Çoktan seçmeli test ve yazılı sınav Kars il merkezinde yer alan bir ortaokulda öğrenim gören 48 katılımcıya birer hafta ile uygulanmış ve uygulamalar birinci araştırmacı tarafından bireysel olarak gözlemlenmiştir.

2.3. Verilerin analizi

Araştırmacılar tarafından geliştirilen ölçme araçlarından elde edilen verilerin analizinde TAP.exe (Brooks ve Johanson, 2003) , SPSS (16. Sürüm, SPSS Inc, Chicago, 2007) ve G Kuramı

analizleri için EduG software (EduG version 6.1-e, Quebec, Canada, 2012) paket programları kullanılmıştır.

İlk olarak belirlenen ölçme araçları ile gerekli uygulamalar yapılmış, çoktan seçmeli nihai test maddeleri ortak sonuçlar doğuracağından tek bir puanlayıcı tarafından, hazırlanmış olan yazılı sınav ise üç farklı puanlayıcı tarafından puanlanmıştır. Puanlayıcılara araştırmacı tarafından puanlama cetveli verilmiş ve puanlama için gerekli süre sağlanmıştır. Puanlayıcılar birbirlerini tanımamakta, farklı okullarda görev yapmakta ve farklı kıdem düzeylerinde bulunmaktadır.

2.4. Sınırlılıklar

Araştırma 2013-2014 eğitim-öğretim yılı Kars il merkezinde yer alan bir ortaokulda öğrenim gören 8. sınıf öğrencilerinden seçilen 48 kişi ve Fen ve Teknoloji dersi 8. sınıf ' *Hücre Bölünmesi ve Kalıtım'* ünitesi ile sınırlıdır.

3. BULGULAR

Bu bölümde araştırmanın alt problemleri için toplanan verilerden elde edilen bulgular, tablo ve açıklamalarıyla birlikte verilerek bunlara dayalı yorumlar yapılmıştır.

Performansın Ölçülmesinde Kullanılan Çoktan Seçmeli Teste Ait Özellikler: Çoktan seçmeli test için belirtke tablosuna göre oluşturulan 40 soruluk ön uygulama için betimsel istatistikler Tablo 1' de verilmiştir.

Tablo 1. Çoktan Seçmeli Testin Ön Uygulamasına Ait Betimsel İstatistikler

Öğrenci Sayısı (N)	90
Madde Sayısı (K)	40
Aritmetik Ortalama	50.16
Varyans (s^2)	468.85
Standart Sapma (s)	21.65
En Düşük Puan (Min.)	15.00
En Büyük Puan (Max.)	92.00
Ortalama Güçlük	.523
Ortalama Ayırt Edicilik	.544

Çoktan seçmeli testin ön uygulamasından elde edilen madde istatistiklerine göre hazırlanan yazılı sınav sorularının doğrultusunda 22 madde nihai uygulama için seçilmiştir. Çoktan seçmeli test için KR-20 güvenirlik değeri hesaplanmış ve bu katsayının .896 olduğu görülmüştür. KR-20 ile hesaplanan güvenirlik katsayısı testin kendi içinde tutarlılığının bir ölçüsü olup bu değerin yüksek çıkması testin güvenilir olduğu anlamına gelmektedir (Başol, 2016).

Öğrencilere ilk olarak uygulanan çoktan seçmeli test önceden belirlenen kazanımları temsil eden 22 test maddesi ile değerlendirilmiştir. Bunun için öncelikle öğrencilerin doğru cevapları hesaplanmış, 100 üzerinden puanlara dönüştürülmüştür. Çoktan seçmeli teste ait istatistikler Tablo 2' de verilmiştir.

Tablo 2. Çoktan Seçmeli Test İle Yapılan Nihai Uygulamaya Ait Betimsel İstatistikler

Soru Sayısı	n	Ortalama	Medyan	Mod	Mak.	Min.	Ranj	Çarpıklık	Basıklık
22	48	68.37	68.18	68.18	100	18.18	81.82	-.55	-.33

Uygulanan çoktan seçmeli testte her bir maddeden alınabilecek en düşük puan bir, testten alınabilecek en yüksek puan 22' dir. Puanlar 100 üzerinden değerlendirmeye alınmış ve istatistiksel işlemler bu puanlar üzerinden yapılmıştır. Dönüştürülen puanlara göre çoktan seçmeli testin ortalaması 68.37, medyanı 68.18, modu 68.18' dir. Bu durumda puanların normal dağılım gösterdiğine işaret etmektedir. Testten alınan en yüksek puan 22 sorunun hepsini doğru cevaplayan üç kişi için 100, testten alınan en düşük puan ise dört doğru ile 18.18 olarak hesaplanmıştır. Puanların ortalamasının 50'den yüksek olması öğrencilerin başarı seviyelerini %50 den yüksek olduğunun göstermektedir.

Alt Problem 1: 'Çoktan seçmeli test için G Kuramına göre kestirilen parametrelerin varyansları ve toplam varyansları açıklama yüzdeleri nedir?'

Çoktan seçmeli test için birey (b) ve madde (m) değişkenlerinin değişimlerini ve varyans kaynaklarının oranlarını belirlemek için tek değişkenli G (Genellenebilirlik) çalışması yapılmıştır.

Tablo 3. Tek Değişkenli G Çalışması Sonucunda Ölçmenin Kestirilen Varyansları ve Toplam Varyansı Açıklama Oranları

Varyans Kaynağı	Sd	Toplam Kareler	Kareler Ortalaması	Varyans	%
b	47	45.814	.975	.037	16.9
m	21	20.235	.964	.017	7.6
bm	87	162.311	.165	.164	75.5
Toplam					100

Tablo 3 incelendiğinde birey (b) ana etkisi için kestirilen varyans bileşeninin (.037) toplam varyansın %16.9' unu açıkladığı görülmektedir. Tek değişkenli modelle yapılan incelemede bireyler için kestirilen varyans bileşeni, toplam varyans içinde en yüksek ikinci paya sahip olan varyans bileşenidir. Genellenebilirlik çalışmalarında, birey ana etkisi evren puanı varyansı olarak değerlendirilir ve ölçülen özellik açısından bireyler arası farklılaşmayı ifade eder (Shavelson ve Webb, 1991; Brennan, 2001). Bireyler için kestirilen varyansın toplam varyans içindeki oranının daha fazla olması istenilen bir durumdur. Bu ölçme ile elde edilen boyutta bireyler arası farklılıkların ortaya çıkarılabildiğinin bir göstergesidir (Güler, 2008).

Madde (m) ana etkisi için tek değişkenli modelle yapılan G çalışmasında kestirilen varyans bileşeni (.017) toplam varyansın %7.6' sını açıklamaktadır. Madde ana etkisinin varyans bileşeni büyüklüğün, toplam varyans değişkeni büyüklüğünde üçüncü ve en az orana sahiptir.

Birey x madde ortak etkisi (.164) toplam varyansın %75.5' ini açıklamaktadır. Birey x madde ortak etkisi tek değişkenli modelle yapılan G çalışmasında elde edilen en büyük varyans

değeridir. Bu durum; bu ölçme için birey x madde ortak etkisinden kaynaklanan farklılığın büyük olduğunu, belli bireylerin bağıl durumlarının bir maddeden diğerine çok farklılaştığını göstermektedir. Ayrıca birey x madde varyans değerinin büyük olması birey ve madde ortak etkisi veya tesadüfü hataların büyük olabileceği anlamına gelebilir.

Alt Problem 2: 'Çoktan seçmeli test için yapılan K çalışması sonuçlarına göre G ve Phi katsayılarının değişimleri nasıldır?'

Performansın ölçülmesinde kullanılan çoktan seçmeli test için 22 madde ve madde sayısının arttırılıp azaltılması durumlarında G Kuramı çalışması ile yapılan K çalışması sonucu elde edilen G ve Φ katsayıları Tablo 4' de verilmiştir.

Tablo 4. Performansın Ölçülmesine İlişkin Yapılan K çalışması İle Madde Sayıları Senaryolarına Göre G ve Phi Katsayıları

Madde Sayısı	Φ	G
18	.801	.785
20	.818	.803
22	**.831**	**.817**
24	.843	.829
26	.854	.841

Tablo 4' te çoktan seçmeli testin madde sayılarının arttırılıp azaltılması durumlarına göre hesaplanan G ve Φ katsayıları verilmiştir. Tabloya göre, madde sayısının nihai testteki değerine göre yapılan analiz sonuçlarına göre; Φ katsayısı .831 ve G katsayısı .817 olarak kestirilmiştir.

Tablo 4 incelendiğinde, madde sayısının azaltılması durumlarında Φ katsayısı ve G katsayılarının azaldığı, madde sayısının arttırıldığı durumlarda Φ katsayısı ve G katsayılarının arttığı gözlemlenmiştir. Ayrıca, 20 madde için elde edilen değerlerin KTK'da Cronbach α değerine karşılık gelmekte ve madde sayısını azaltıp-arttırmanın sonucunda elde edilen güvenirliğin yine KTK'da kestirilebilmekte; G katsayısının avantajı sadece mutlak değerlendirmeler için kullanılabilecek bir güvenirlik değerinin elde edilmesine imkan tanınmasıdır.

Performansın Ölçülmesinde Kullanılan Yazılı Sınava Ait Özellikler: Performansın ölçülmesine yönelik uygulanan yazılı sınav 11 maddeden oluşmaktadır. Uygulanan sınav üç farklı puanlayıcı tarafından puanlanmış ve puanlayıcılar üzerinden elde edilen verilerle işlemler gerçekleştirilmiştir. Yazılı sınava yönelik puanlayıcılardan elde edilen puanlara ait betimsel istatistikler Tablo 5' te verilmiştir.

Tablo 5 incelendiğinde, 48 öğrencinin 11 madde üzerinden aldıkları puanlara ilişkin en yüksek ortalama birinci puanlayıcıya aittir ve 56.187 şeklindedir. En düşük ortalama ise 34.708 ile üçüncü puanlayıcıya aittir. İkinci puanlayıcı 45.479 ile puanlayıcı ait ortalama değeri ise bu iki değer arasında yer almaktadır. Birinci puanlayıcıya ilişkin ortanca değer aritmetik ortalamadan yüksektir ve puanların hafif sola çarpık bir dağılım gösterdiği söylenebilir. İkinci ve üçüncü puanlayıcıya ilişkin ortanca değerlerinin aritmetik ortalamadan küçük olması ise puanların hafif sağa çarpık bir dağılım gösterdiğini ortaya koymaktadır. Bu durum çarpıklık katsayılarının birinci puanlayıcıya ait puan değerleri için hafif negatif, ikinci ve üçüncü puanlayıcılara ait puan değerleri

içinse hafif pozitif çıkmasıyla da görülmektedir. Puanlayıcıların verdikleri puan değerlerine ait Cronbach alfa (α) güvenirlik katsayıları birbirine yakın ve yüksek değerlerdir. Puanlayıcıların verrmiş oldukları puan değerlerinin ortalamalarının birbirlerinden farklı olmasının; öğrencilerin sorulara verdikleri yanıtlara açıklık derecelerine göre ya bütüncül ya da ayrıntılı olarak puanlama yapmış olmalarının neden olduğu düşünülmektedir.

Tablo 5. Performansın Ölçülmesinde Yapılan Yazılı Sınav İçin Üç Puanlayıcı Ait Betimsel İstatistikler (N=48)

İstatistikler	1. Puanlayıcı	2. Puanlayıcı	3. Puanlayıcı
Ortalama	56.187	45.479	34.708
Medyan	56.5	44.5	29.5
Mod	34	43	30
Std. Sapma	2.362	2.358	2.411
Varyans	557.943	556.170	581.360
Çarpıklık	-.184	.248	.609
Basıklık	-.970	-.990	-.769
Minimum	8	8	2
Maksimum	96	83	84
α güvenirliği	.850	.870	.870

Alt Problem 3: 'Yazılı sınav için yapılan G Kuramına göre kestirilen parametrelerin varyansları ve toplam varyansları açıklama yüzdeleri nedir?'

Matematik performansının ölçülmesine yönelik hazırlanan 11 maddelik yazılı ölçme aracının G çalışması ile elde edilen varyanslarını ve varyans yüzdelerini hesaplamak için tümüyle çaprazlanmış b x m x p modeli uygulanmıştır. Ölçmenin uygulandığı 48 öğrenci, 11 madde ve üç puanlayıcıdan oluşan verilerde tek değişkenli modelle yapılan G çalışması için; kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri b, m ve p ana etkileri ile bm, bp, mp, ve bmp ortak etkileri Tablo 6' da verilmiştir.

Tablo 6. Tek Değişkenli G Çalışması Sonucunda Ölçmenin Kestirilen Varyansları ve Toplam Varyansı Açıklama Oranları

Varyans Kaynağı	sd	Toplam Kareler	Kareler Ortalaması	Varyans	%
b	47	5193.657	110.503	3.082	16.0
m	10	44.326	4.433	-.027	.0
p	2	127.971	63.986	.130	.7
bm	470	8652.705	18.410	1.153	6.0
bp	94	502.514	5.346	-.873	.0
mp	20	96.807	4.840	-.211	.0
bmp	940	14052.708	14.950	14.950	77.4
Toplam					100

Tablo 6'ya göre, birey (b) ana etkisi için kestirilen varyans bileşenini (3.08173) toplam varyansın %16' sını açıklamaktadır. Tek değişkenli modelle bireyler için kestirilen varyans bileşeni, toplam varyans içinde en yüksek ikinci sırada paya sahiptir.

Madde (m) ana etkisi için kestirilen varyans bileşeni tek değişkenli modelle yapılan G çalışmasına göre kestirilen varyans bileşeni eksi değer aldığı için (-.02686) toplam varyansı açıklama yüzdesi içinde (%0) bir etkiye sahip olmadığı görülmüştür. Varyansın *sıfir* alınmasının nedeni G Kuramı çalışmalarında varyans değerinin negatif çıkması durumlarında uygulanan dört farklı yöntemden biri olmasıdır (Brennan, 2001). Shavelson ve Webb (1981)'e göre negatif varyanslar örnekleme hatalarından ya da yanlış model seçiminden kaynaklanmış olabilir.

Shavelson ve Webb (2005) negatif varyans söz konusu olduğunda dört çözüm önerisi olduğunu belirtmiştir: Cronbach, Gleser, Nanda ve Rajaratnam (1972) negatif varyans değerinin yerine sıfır yazmayı önermişler, ikinci öneri olarak (Brennan, 2001) negatif varyansların sıfır alınmasını ancak beklenen ortalama kareler eşitliğinde negatif varyansların olduğu gibi kullanılmasını, üçüncü öneri ise (Shavelson ve Webb) Bayesian metot kullanılarak tahmin edilen varyans için en küçük değerin sıfır olarak değiştirilmesini, son olarak Searle (1987) maksimum olabilirlik modeli kullanılarak negatif varyansların önüne geçilmesini önermiştir (Akt. Shavelson ve Webb, 2005). Mevcut çalışmada Cronbach, Gleser, Nanda ve Rajaratnam (1972)'ın önerisi dikkate alınarak negatif varyanslar 0 alınmıştır.

Buna göre, puanlayıcı ana etkisinin G çalışması ile kestirilen varyans bileşeni (.13021) toplam varyansın %0.7' ini açıklayarak toplam varyans içinde dördüncü sırada yer almaktadır. Puanlayıcı etkisinin tek değişkenli modelle yapılan G çalışması ile kestirilen varyans oranının düşük olması, puanlayıcıların tüm bireyler için yaptıkları puanlamalar arasında bir fark bulunmadığını, puanlamalar arasında da bir tutarlılığın olduğunu göstermektedir.

Birey x madde (bm) ortak etkisi (1.15344) toplam varyansın %6' sını açıklamaktadır. Birey x madde ortak etkisi tek değişkenli modelle kestirilen en yüksek üçüncü değere sahip varyans değeridir. Bu da birey x madde ortak etkisinden kaynaklanan farklılığın büyük olduğunu, belli bireylerin bağıl durumlarının bir maddeden diğer maddeye çok farklılaştığını göstermektedir (Güler, 2008).

Birey x puanlayıcı (bp) ortak etkisi (-.87307) toplam varyansın %0' ını açıklamaktadır. Madde x puanlayıcı (mp) ortak etkisi (-.21061) 0' ın altında değer aldığı için toplam varyans içerisinde açıklama yüzdesi %0 dır.

Madde x puanlayıcı etkisinin tek değişkenli modele göre madde x puanlayıcı ortak etkisinden kaynaklanan bir farklılığın olmadığı yorumu yapılabilir.

Birey x madde x puanlayıcı (artık) ortak etkisi varyans bileşenide (14.94969) toplam varyansın %77.4' ünü açıklamaktadır. Bu oran varyans değerleri arasından en büyük değerdir. Birey x madde x puanlayıcı (artık) varyansın büyük olması; birey, madde ve puanlayıcı ortak etkisi veya tesadüfi hataların büyük olabileceğinin bir göstergesi olabilir.

Alt Problem 4: 'Yazılı sınav için yapılan K çalışmasına göre farklı senaryolara göre G ve Phi katsayılarının değişimi nasıldır?'

Uygulanan yazılı sınava ait veriler üzerinden madde sayısı ve puanlayıcı sayılarının arttırılıp azaltılması durumlarına göre G Kuramı kullanılarak K çalışması yapılmıştır. Yapılan K çalışmasına ait G ve Φ katsayılarının değişimi Tablo 7' de verilmiştir.

Tablo 7. Performansın Ölçülmesine İlişkin Yapılan K çalışması ile Madde ve Puanlayıcı Sayıları Senaryolarına Göre Phi ve G Katsayıları

Madde Sayıları	Puanlayıcı Sayıları					
	2		3		4	
	G	Φ	G	Φ	G	Φ
9	.763	.751	.819	.810	.850	.843
11	.797	.784	**.847**	**.837**	.874	.866
13	.823	.809	.867	.857	.891	.883

Tablo 7'ye göre tek değişkenli modelle yapılan ölçme sonuçlarına göre 11 madde ve üç puanlayıcıya göre kestirilen G katsayısı .847 ve Φ katsayısı da .837 olarak kestirilmiştir. Kestirilen katsayı değerlerine bakılarak G katsayısının Φ katsayısından daha yüksek olduğu görülmektedir. Gerek bağıl değerlendirme durumlarında kullanılan G katsayısı ve gerek mutlak değerlendirme durumlarında kullanılan Φ katsayılarının madde sayılarının ve puanlayıcı sayılarının artması durumunda yükseldiği ortaya çıkmıştır. Tüm madde ve puanlayıcı senaryolarında G katsayıları, Φ katsayılarından yüksek değerde çıkmıştır. Madde sayısının aynı kalması durumunda puanlayıcı sayısının artması senaryolarında ortaya çıkan G ve Φ katsayıları; puanlayıcı sayılarının aynı kalması durumunda madde sayısının arttırılması ile kestirilen G ve Φ katsayılarına göre daha yüksek değerlerde ortaya çıkmıştır.

4. TARTIŞMA

Araştırma bulgularına göre, bireylerin çoktan seçmeli testten aldıkları puanlar ile yazılı sınavdan aldıkları puanların dağılımlarının paralellik gösterdiği gözlenmiştir. Çoktan seçmeli testten alınan puanların daha yüksek olduğu ortaya çıkan bulgular arasındadır. Ranj değerlerinin değişimine baktığımızda yazılı sınav için her bir puanlayıcının vermiş olduğu puanlar ile çoktan seçmeli teste ait ranj değerinin birbirleri ile çok yakın olduğu görülmektedir.

Yazılı sınav ve çoktan seçmeli test için gerek ortanca gerekse standart sapma değerlerinin ortalama ekseninde değişimleri için belirlenen başarı puanlarının çoktan seçmeli test için dağılımları ile paralellik gösterdiği görülmüştür. Ancak bu araştırmada başarı puanları açısından çoktan seçmeli testten alınan puanların daha yüksek olduğu ortaya çıkmıştır.

Çetin (2009) yapmış olduğu araştırmasında; performans görevi, yazılı sınav ve çoktan seçmeli test arasındaki ilişkiyi farklı değişkenlerle incelemiştir. Çetin, araştırma sonuçlarına göre başarı puanlarının çoktan seçmeli test için daha yüksek olduğu sonucuna ve üç sınav arasında ilişkinin orta düzeyde olduğu sonucuna ulaşmıştır. Ancak ikili ilişkilere bakıldığında çoktan seçmeli test ile yazılı sınav arasındaki ilişkinin daha ileri düzeyde olduğu gözlenmiştir. Diğer ikili karşılaştırmalara göre, yapılmış olan bu araştırmada çoktan seçmeli test ve yazılı sınav arasında ilişki yüksek bulunmuş; uygulama amacına göre sınavların uygulanmasında araştırmacının istediği özelliklere göre her iki sınavında kullanılabilirliği sonucuna varılmıştır. Yazılı sınavda soru sayısının az olması gibi dezavantajlarının yanında puanlayılar arası tutarlılığın sağlanması halinde çoktan seçmeli teste yakın sonuçlar verdiği ortaya çıkmıştır.

Eser (2011) sınav türleri konusunda öğrenci tercihlerini çalışmış olduğu betimsel tarama modelindeki araştırmasında, öğrenciler, başarı puanları daha yüksek olduğu için çoktan seçmeli testleri, yazılı sınavlara göre daha çok tercih ettiklerini belirtmişlerdir. Araştırma sonuçlarına göre

en az tercih edilen sınav türü yazılı sınav türü olarak belirtilmiştir. Yapılan bu çalışmada ise tercih türleri araştırılmamış ancak çoktan seçmeli test puanlarının dağılımlarının yazılı sınav türünden elde edilen puan dağılımlarına göre daha yüksek olduğu sonucuna ulaşılmıştır. Öğrencilerin çoktan seçmeli testlere daha çok aşina olması bu sonuçta etkili olmuş olabilir.

Öğrencilerin bilgilerini kullanarak bir ürün ortaya çıkarılmasını isteyen yazılı sınavların öğrencilerde kaygı ve korkuya neden olduğu ve bu nedenle öğrencilerin başarılarının düşük olduğu farklı araştırmalarda ortaya konulmuştur. Ömür (2002) çalışmasında, öğrencilerin cevap üretmek yerine verilen cevaplar arasından birini seçmeyi daha çok tercih ettiklerini belirtmiştir. Ayrıca başarının yazılı sınavlarda çoktan seçmeli testlere göre daha yüksek olduğu sonucu bu çalışmadaortaya çıkan bir diğer bulgudur.

Bunun aksine bazı çalışmalar da yazılı sınavda ortaya konulan performansın çoktan seçmeli testlere göre daha yüksek olduğu sonucuna ulaşılmıştır. Önder (2008) matematik başarısının ölçülmesi ve sınav kaygı düzeyi üzerine yapmış olduğu çalışmasında; yazılı sınava hazırlanan öğrencilerin başarılarının daha yüksek olduğunu belirtmiştir. Ayrıca çalışmada, hangi tür sorularla sınavlara hazırlanırsa hazırlansınlar, öğrencilerin yazılı sınavlarda daha başarılı oldukları sonucu elde edilmiş; yazılı sorularla sınava hazırlanan öğrencilerin yanı sıra, çoktan seçmeli test sorularla sınava hazırlanan öğrencilerin de yazılı sınavlardan daha iyi bir performans gösterdiği bulunmuştur. Oysa, bu araştırmanın bulgularından birisi öğrencilerin performans puanlarının, çoktan seçmeli sınav için yazılı sınava göre daha yüksek olduğudur. Alan yazın incelendiğinde farklı sınav türlerinin karşılaştırıldığı ve üzerinde G Kuramı çalışması yapılan araştırmalara rastlanmamıştır. Daha çok performansın belirlenmesinde puanlayıcıların birbirleri ile tutarlılığının incelendiği ve farklı desenlere göre karşılaştırılmaların yapıldığı araştırmalar mevcuttur.

Yapılan analizlere göre; G Kuramına göre puanlayıcılara ait puanlayıcı değişkenliğinin etkisinin düşük olduğu ortaya çıkmıştır. Ortaya çıkan bu sonuçların benzer başka çalışmalarda da ortaya çıktığı görülmüştür. Güler (2008) farklı kuramlara göre karşılaştırma yaptığı çalışmasında; matematik başarısını belirlemede uygulanan klasik sınav verileri üzerinden KTK, G Kuramı ve ÇDRM çalışmaları yapmıştır. Elde edilen bulgulara göre G Kuramı çalışması sonuçlarına göre puanlayıcılar arasında tutarlılığı yüksek bulunmuştur. Nalbantoğlu (2009) puanlayıcıların birlikte ve dönüşümlü olarak puanlamalarında sonuçlar arasında paralellik olduğu ve puanlamaların birbirleri ile tutarlı olduğu sonucuna ulaşılmıştır.

LLabre 1978'deki çalışmasında farklı modlar ve farklı yazma becerilerini değerlendiren puanlayıcıların vermiş oldukları puanların aradaki zaman ve farklı ortamlara rağmen tutarlı sonuçlar verdiğini belirtmiştir. Puanlayıcı sayısının artması halinde güvenirlik değerinin yükseldiği sonucu araştırmadan çıkan sonuçlardandır.

Çoktan seçmeli ve yazılı sınavlarda madde sayısının artması sonucu güvenirlik değerinin arttığı bulgularda gözlenmiştir. Puanlayıcı ve madde sayısının artması araştırmanın güvenirliği açısından önemli bir özelliktir. Ancak uygulama, maliyet ve zaman gibi etkenlerden dolayı araştırmalarda hangisinin tercih edilebileceği hakkında bir noktaya varılmak istendiğinde bulgular dahilinde çoktan seçmeli test için madde sayısının arttırılmasının; yazılı sınav için puanlayıcının sayısının arttırılmasının güvenirlik değerlerini daha çok yükselttiği görülmektedir.

Her iki sınav türü içinde güvenirlik çalışması yapılmış ve güvenirlik indeksleri olarak KTK için α ve G Kuramı için G katsayısı hesaplanmıştır. Araştırma için hesaplanan bu değerlere göre α ve G katsayıları oldukça yüksek ve birbirlerine yakın bulunmuştur. Wang (2005) benzer bir çalışmada farklı güvenirlik indekslerini hesaplamış ve karşılaştırmıştır. Çalışmanın sonucunda α ve G katsayısının birbirine yakın olduğu sonucuna ulaşılmıştır.

5. KAYNAKLAR

Aiken, L.R. (1991). *Psychological testing and assessment* (7. baskı). Boston: Allyn and Bacon.

Alharby, E.R. (2006). *A comparison between two scoring methods, holistic vs. Analytic using two measurement models, the Generalizability Theory and the many facet Rasch measurement within the context of performance assessment*. Unpublished doctoral dissertation. The Pennsylvania State University Faculty of Education, Pennsylvania.

Alkahtani, S.F. (2012). *Oral performace scoring using generalizability Theory and many-facet Rasch measurement: a comparison study*. Unpublished doctoral dissertation. The Pennsylvania State University, Pennsylvania.

Atılgan, H. (2004). *Genellenebilirlik Kuramı ve çok değişkenlik kaynaklı rasch modelinin karşılaştırılmasına ilişkin bir araştırma*. Yayınlanmamış doktora tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Au, F., Prahardhi, S., & Shiell, A. (2008). Reliability of two instruments for critical assessment of economic evaluations. *Value in Health, 11*, 435- 439.

Başol, G. (2008). Bilimsel araştırma süreci ve yöntem. İçinde Kılıç, O. & Cinoğlu M. (Ed.), *Bilimsel araştırma yöntemleri*, Bölüm 5, İstanbul: Lisans Yayıncılık.

Başol, G. (2016). *Eğitimde ölçme ve değerlendirme*. Genişletilmiş 4. Baskı, Ankara: Pegem Yayıncılık.

Bottema-Beutel, K., Lloyd, B., Carter, E.W. & Asmus, J.M. (2014). Generalizability and decision studies to inform observational and experimental research in classroom settings. *American Journal on Intellectual and Developmental Disabilitie, 119*(6), 589–605.

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

Brooks, G.P. & Johanson, G.A. (2003). Test Analysis Program. *Applied Psychological Measurement*, 27, 305-306.

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

Burns, K.J. (1998). Beyond classical reliability: Using Generalizability Theory to assess dependability. *Research in Nursing and Healty, 21*, 83-90.

Burton, E.B. (1998). *An investigation of the school-level generalizability of performance assessment results*. Unpublished doctoral dissertation. Rutgers University, New Jersey.

Çakıcı Eser, D. (2011). *Genellenebilirlik Kuramı ve lojistik regresyona dayalı hesaplanan puanlayıcılar arası tutarlığın karşılaştırılması*. Yayınlanmamış yüksek lisans tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

EduG 2012 software. EduG version 6.1-e, Generalizability Study. Société Suisse pour la Recherche en Education, Groupe de travail Edumétrie – Qualité de l'évaluation en éducation; software prepared by Maurice Dalois and Léo Laroche, Educan Inc., Longueuil, Quebec, Canada.

Gao, X. & Brennan, R.L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education, 14*, 191-203.

Güler, N. (2008). *Klasik Test Kuramı, Genellenebilirlik Kuramı ve Rasch modeli üzerine bir araştırma*. Yayınlanmamış doktora tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Güler, N. (2009). Genellenebilirlik Kuramı ve SPSS ile GENOVA programlarıyla hesaplanan G ve K çalışmalarına ilişkin sonuçların karşılaştırılması. *Eğitim ve Bilim, 34*, 154.

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

Kaya, G. (2011). *Genellenebilirlik Kuramının doldurma kavram haritası değerlendirme çalışmasına uygulanması*. Yayınlanmamış yüksek lisans tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Moon, S.Y. (1995). *Performance assessment: Measurement ıssues of generalizability, dependability of scoring and relative information on student performance.* Unpublished doctoral dissertation. The Florida State University, Tallahassee.

Mushquash, C. & O'Connor, B. P. (2006). SPSS and SAS programs for Generalizability Theory analysis. *Behavior Research Methods*. *38*(3), 542-547.

Nalbantoğlu, F. (2009). *Performans ölçümlerinde Genellenebilirlik Kuramıyla farklı desenlerin karşılaştırılması*. Yayınlanmamış yüksek lisans tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

O' Neill & O' Neill (2015). Improving QST reliability-moreraters, tests, oroccasions? A multivariate generalizability study. *The Journal of Pain, 16*(5), 454-462.

Rentz, J.O. (1987). Generalizability Theory: a comprehensive method for assessing and improving the dependability of marketing measures. *Journal of Marketing Research, 24*(1), 19-28.

Shavelson, R.J. & Webb, N.M. (1981). Generalizability Theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133–166.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory*: *A primer*. Newbury Park, CA: Sage.

Shavelson, R.J. & Webb, N.M. (2005). *Generalizability Theory*. Web: http://web.stanford.edu/dept/SUSE/SEAL/Reports_Papers/methods_papers/G%20Theory%20AERA.pdf adresinden alınmıştır.

Senemoğlu, N. (2002). *Gelişim öğrenme ve öğretim: kuramdan uygulamaya*. Ankara: Gazi Kitapevi.

SPSS Inc. Released 2007. SPSS for Windows, Version 16.0. Chicago, SPSS Inc.

Solano-Flores, G. & Li, M. (2006). The use of Generalizability (G) Theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice, 25*(1), 13-22.

Suen, H.K., & Lei, P.W. (2007). Classical versus Generalizability Theory of measurement. *Educational Measurement, 4*, 1-13.

Tekindal, S. (Ed.) (2011). *Eğitimde ölçme ve değerlendirme*, Ankara: PegemA Yayıncılık.

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

Wang, Z. (2005). *Estimating reliability under a Generalizability Theory model for writing scores in C-base*. Yayınlanmamış doktora tezi. University of Missouri, Columbia.

Yelboğa, A. (2007). *Klasik Test Kuramı ve Genellenebilirlik Kuramına göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi*. Yayınlanmamış doktora tezi. Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Summary

Introduction

Education is the process of transferring past values, moral standards, knowledge and skills to new generations of society as a process of creating the desired behaviors (Senemoğlu, 2002). Through the education, individuals are expected to acquire knowledge, skills and attitudes in certain topics. It is always a matter of curiosity how much the individuals in the training process have benefited from the training or to what extent the learning objectives are achieved. Through the evaluation, it is expected to define the level of knowledge, skills and attainment of attitudes. Therefore, assessment results can be considered as the identificators of the efficiency of the teaching process and expected to enrich the teaching methods that are currently in use in order to get more effectiveness and satisfaction.

Taking into consideration the method used in teaching, among the different measuring instruments, the most suitable one is selected to check whether the learning goals has been achieved or not. The measuring instruments used according to the purpose of the evaluation also vary.

The instrument to be used is decided according to the factors such as the student's readiness, the environment for the examination, time limit and exam conditions. It is worth investigating whether the results obtained with different measuring tools are similar. In this way, it could be possible to know whether the measurement results obtained from different sources are reliable.

In the research studies aimed at measuring the performance of the participants, the researchers used statistical studies using the Classical Test Theory (CTT). CTT is preferred more often because of its ease of use and familiarity. However, with CTT, it is not possible fully understand the inconsistencies in the scores. On the other hand, Generalizability Theory is particularly well suited to distinguish the sources of inconsistencies in observed scores. The Generalizability Theory (G Theory) allows comparison of research results with the reason that it handles the sources of errors at the same time and places them in relation to each other.

The Generalizability (G) Theory is a statistical theory that enables the determination of the reliability of measurement results, the design, the investigation and the conceptualization of reliable observations. Generalizability Theory is an extension of the Classical Test Theory (Cronbach et al., 1972, Brennan, 2001).

According to Shavelson and Webb (1991), Generalizability Theory is an extension of Classical Test Theory from four different perspectives: 1. Generalizability Theory deals with multiple variance sources in a single analysis. 2. It defines each variance source. 3. It allows calculating two different reliability coefficients (G coefficient and phi coefficient) for making relative decisions based on both individual performances as well as absolute decisions about individual performances. 4. Depending on a specific purpose, it is possible to arrange measures (Decision "K" studies) that can reduce the measurement error to the greatest extent possible.

Reliability search in G Theory is conducted in two steps; first is the Generalizability study (G-study) and the second is the Decision study (K-study) (Kaya, 2011). Among these, G study is designed to provide a reasonable and economically versatile isolation and estimation of the measurement error (Shavelson & Webb, 2005). In the G-study process, all sources of variability (variance components) and interactions between them are estimated using the ANOVA method to

generalize the sample to the equation. These predicted variance components are used in the next phase of the K-study (Kaya, 2011).

With the K study, the generalization coefficient (G coefficient) and the reliability index (Phi coefficient), which are similar to the reliability coefficient, are reached. In the G model, the Phi (Φ) coefficient is used with the absolute decision model.

Methodology

The purpose of the study is to investigate the consistency of the achievement scores obtained from different measurement instruments on the same content. It is also aimed to determine the error amounts of the measurement results obtained in the same individuals, in different situations with different measurement tools separately and for each variable and their combinations with each other. The current study was carried out based on the Theory of Generalizability because different measurement tools and multiple error sources were considered.

The population is the 8th grade students in Kars province during 2013-2014 school years. The study sample, selected through purposeful sampling, composed of 48 students, attending Atatürk Middle School in Kars city center.

The items selected for the multiple-choice test were at different item difficulty levels, the KR 20 reliability value was found to be .90. In order to ensure the validity of the written exam, the questions were prepared and consulted to an area expert and a language expert and the application form was prepared. The items were given to a female and male student beforehand to ensure that there was no misunderstanding or confusion, also to make sure they serve neutrality, both the items of the multiple choice test and the questions in the essay.

Results and Discussion

According to the measurement results with the univariate model, the G coefficient is estimated as .817 and the Φ coefficient as .831 according to the 22 items included in the multiple choice exam. G coefficient is estimated with respect to three scorers and 11 items are estimated as .847 and Φ coefficient is estimated as .837.

According to the analysis of the variance components for the multiple choice exam; the variance component predicted for the individual has the second highest share, the main effect of the substance has the third and least proportion of the total variance, and the common effect of the individual and the substance has the greatest variance value.

According to analysis of variance components for the written exam; as for the indiviudal component, the variance component was found to be the highest in the total variance and the total variance did not have an effect in the percentage of the explanatory value (0%), as the variance component predicted for the G run with the univariate model predicted for the substance main effect was negative.

The low variance ratio predicted by the G study with the univariate model showed that there was no difference between the scorers for all individuals, therefore there was a consistency between the scorers. Individual x item x scorer (error) source had the largest variance proportion in the common effect variance component.

According to the research findings, it was observed that the distribution of the grades of the individuals in the written exams and the grades they got from the multiple choice test were parallel.

For the written test and multiple choice tests, the median and the standard deviations were found to be consistent with the distributions for the multiple-choice exam.

The results of the analyzes showed that, in terms of CTT, the high correlation between the scorers indicated low scorer effect and according to G Theory the low scorer effect meant high consistency among the scorers. The increase in the number of items of multiple-choice and written exam lead to more reliable scores. The number of items and scorers are important in terms of the reliability of the research. Reliability studies were carried out in both types of tests and Cronbach's alpha (α) for Classical Test Theory and Generalizability for G Theorem were calculated as reliability indices. According to the results, Cronbach's alpha (α) and G coefficients are very high and close to each other.