International
Journal of Human Sciences
ISSN:2458-9489

Volume 16 Issue 1 Year: 2019

A generalizability Analysis: The reliability of measurements: "Let's circuit electric"[1]

Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim"

Gülşah Başol[2]
Banu Şevran[3]

Abstract

The study's purpose was to compare the results of G&K studies of some measurement instruments, used in the current study and also check their reliability. These are a multiple-choice test, a structured grid, and a performance task applied for measuring students' performance. It is based on Generalizability Theory (GT). The population was the 6th grade students in Antalya, a city in the Mediterranian area in Turkey. During the 2014-2015 school years, forty students (23 girls and 17 boys) from two 6th grade classes randomly selected from two middle schools. The first author was the instructor and it took a month to teach "Let's Circuit Electricity" unit in a regular classroom environment. For the study's purpose, a multiple-choice test, a structured grid, and a performance task, developed by the researchers, were utilized to measure students' achievement in each. The structured grids and performance tasks were graded by three teachers. According to the findings, KR-20 reliability coefficient of the multiple-choice test was .68; while G coefficient .68 and Phi (Φ) coefficients was .64. The Cronbach's Alpha coefficients for structured grid and G coefficient were .96 and

Özet

Araştırmanın amacı; Genellenebilirlik Kuramını kullanarak öğrenci performansının ölçülmesi amacıyla kullanılan çoktan seçmeli test, yapılandırılmış grid ve performans görevinin Genellenebilirlik (G) ve Karar (K) çalışmalarının yapılarak, sonuçların karşılaştırılmasıdır. Araştırma evrenimiz 2014-2015 eğitim-öğretim yılında Antalya'da öğrenim görmekte olan ortaokul altıncı sınıf öğrencilerini kapsamaktadır. Araştırma çalışma grubu ise Antalya'da iki farklı okuldan rastgele seçilen birer altıncı sınıf şubesinde öğrenim gören 23'ü kız 17'si erkek toplam 40 öğrenciden oluşmaktadır. Elektriği İletelim ünitesi üç haftada araştırmanın birinci yazarı tarafından normal sınıf ortamında işlenmiş; veri toplama aracı olarak araştırmacılar tarafından geliştirilen çoktan seçmeli test, yapılandırılmış grid ve bir performans görevi çalışma grubundaki öğrencilere uygulanmıştır. Bunların değerlendirilmesinde üç puanlayıcının görüşüne başvurulmuştur. Sonuçlara göre çoktan seçmeli testin, G katsayısı .68, Phi (Φ) katsayısı ise .64 bulunmuştur. Yapılandırılmış gridin G katsayısı .91, Phi (Φ) katsayısı ise .83 bulunmuştur. Performans görevi sonuçlarına bakıldığında, G katsayısı .89, Phi (Φ) katsayısının ise .86 şeklinde olduğu görülmüştür. Karar

[1]Çalışma, ikinci yazarın aynı başlıklı Dr. GülşahBaşol'un danışmanı olduğu YL tez çalışmasından üretilmiştir. 1-3 Eylül 2016'da V. Eğitimde Ölçme ve Değerlendirme Kongresinde (EÖD) sunuldu.

[2]Prof. Dr., Tokat Gazisomanpaşa Üniversitesi, gulsahbasol@gmail.com
[3]M.A.,, banusevranbasat@gmail.com

Submitted: 2018-06-18 Published: 2019-03-31

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

.91, respectively whereas Phi (Φ) coefficient was as .83. Cronbach's Alpha coefficient for performance task was found as .89; G coefficient as .89, and Phi (Φ) coefficient as .86. According to the results of decision studies, variation of G and Phi coefficient for all three measurements were in line. We also found that three graders' results were consistent with each other for the structured grid and performance tasks' grading results.

Keywords: Measurement; evaluation; measurement instruments; generalizability theory; reliability.

(Extended English summary is at the end of this document)

çalışmaları sonucuna göre, üç araç için de elde edilen G ve Phi katsayılarının değişiminin paralellik göstermektedir. Ek olarak yapılandırılmış grid ve performans görevinin değerlendirme sonuçlarının üç puanlayıcı için paralellik gösterdiği de bulgularımız arasındadır.

Anahtar Kelimeler: Ölçme; değerlendirme; ölçme araçları; genellenebilirlik kuramı; güvenirlik.

Giriş

Öğrenmenin nasıl gerçekleştiğine dair ortaya atılan teoriler, eğitim alanında farklı öğretim modellerinin oluşturulmasına öncülük etmiştir. Eğitimde yapılan her değişiklik sistemin diğer parçalarını da etkilemektedir. Bu durum ölçme ve değerlendirmede bazı anlayışların değişmesini tetiklemiştir. Artık ölçme ve değerlendirmede sonuç odaklı bir anlayışın yerine sürecin ölçülmesi önem kazanmış; yazma temelli ödevlerin yerini gerçek hayatla ilişkilendirilmiş performans görevleri; üstü kapalı bir anlayışın yerini açık yönergeler, sonul değerlendirmelerin yerini süreç değerlendirme, aralıklı değerlendirmeler yerine ise sürekli değerlendirmeler tercih edilir olmuştur (McMillian, 1997). Ölçme araçlarının bundan elli yıl öncesine göre çok fazla çeşitlilik gösterdiğini de söylemek mümkündür.

Öğrenme bilişsel alandan ibaret değildir, bu nedenledir ki öğrenmeyi ele alırken bilişsel, duyuşsal ve devinişsel becerilerin üçü de göz önünde bulundurulmalıdır. Öğretmen öğrencinin ortaya koyduğu davranışlardan diğer bir deyişle durumu kapasitesi hakkında fikir yürütmeye çalışır. Ayrıca bu sayede eksik ve yanlış öğrenmeleri ortaya koymak mümkün olacaktır. Öğretimde tüm bu işaretler performans olarak tanımlanır (Oosterhof, 1999). Öğrencinin performansı kapasitesinin bir yansımasıdır diyebiliriz. Verilen kararların doğru ve yerinde olması ise geçerli ve güvenilir ölçümlere dayanır.

Öğrenci performanslarının ölçülmesinde ve değerlendirilmesinde pek çok faktör etkili olur. Bunlar ölçme aracı, öğrenci, puanlayıcı, zamanlama, motivasyon, sınav ortamı vb. gibi faktörlerdir. Sağlıklı bir ölçme değerlendirme yapılabilmesi için oluşabilecek tüm hata kaynaklarının en aza düşürülmesi gerekmektedir.

Eğitimde geleneksel ölçme değerlendirme araçlarının dışında tamamlayıcı ölçme değerlendirme araçları da sıklıkla kullanılmaya başlanmıştır. Farklı ölçme araçları öğrenci kapasitesini ortaya çıkarmada ne kadar etkili olmaktadır? Her ölçme aracı güvenilir sonuçlar vermekte midir? Hangi ölçme aracı daha güvenilir sonuç vermektedir?

Klasik Test Teorisi kapsamında ölçme araçlarının güvenilirliğinin test edilmesi önerileri 1960'li yıllar kadar eskiye dayanır (Lord ve Novick, 1968). İlerleyen yıllarla birlikte ölçümlerin güvenirliği söz konusu olduğunda Klasik Test Kuramı dışında Genellenebilirlik Kuramı (G Kuramı) da kullanılmaya başlanmıştır. Brennan (2001) Genellenebilirlik Kuramını

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

(GK)öğrencilerin performans ölçümlerinde güvenirliğin değerlendirilmesi, gözlemlerin araştırılması ve kavramlaştırılmasını sağlayan bir istatistiksel kuram olarak tanımlamıştır.

Kuramın temelleri, Cronbach, Rajaratman ve Gleser tarafından 1963 ve 1965 tarihlerinde yayınlanan makalelerde derinlemesine ele alınmıştır. Varyans analizinin güvenirlik çalışmalarında kullanılması bu çalışmaların öncesine dayanıyor olsa da Cronbach ve arkadaşlarının bu konudaki katkısı önemlidir. Burt (1936), 1941 yılında Jackson ve Ferguson (1941), Hoyt (1941)çalışmalarında varyans analizinin güvenirliğin kestirilmesinde kullanılması konusunu tartışmışlardır. İlerleyen yıllarda bu çabalara Alexender (1947), Ebel (1951), Finlayson (1951), Loveland (1952) ve Burt (1955) katıda bulunmuştur. 1972'de Cronbach, Gleser, Nanda ve Rajaratman; 1983'te Brennan'ın katkılarıyla Genellenebilirlik Kuramı genişletilmiştir (Güler, Kaya Uyanık ve Taşdelen Teker, 2012). Sonuç olarak Genellenebilirlik (G) kuramının temeli varyans analizine (ANOVA) dayanmaktadır. Varyans analizi temel olarak toplam varyansın desendeki varyans bileşenlerine bölünmesidir. Bu sayede bir ölçme sonucunu farklı varyans kaynaklarına ayırarak gözlenen puanların evren puanlarına (gerçek puanlarına) genellenebilmesi mümkün olur (Brennan, 2001).

Genellenebilirlik Kuramı; varyans analizi yöntemiyle farklı hata kaynaklarını bir arada kullanarak onların büyüklüklerinin tahmin edilmesini sağlar. Tek bir analizle bir ölçmedeki birden fazla hata kaynağının ele alınabilmesi, Genellenebilirlik Kuramının avantajıdır (Shavelson ve Webb,1991). Genellenebilirlik Kuramı sayesinde çeşitli test senaryoları denenerek genellenebilirlik katsayısını nasıl etkilediğini ortaya koymak da mümkündür (Brown, 2005). Böylece, farklı test senaryolarının etkisi değerlendirilerek ekonomik kayba neden olacak bir girişimde bulunmadan genellenebilirliği en yüksek olan senaryo bulunabilir (Asmus, Bottema-Beutel, Carter ve Lloyd, 2014).

Ölçmedeki hata kaynaklarının hem ayrı ayrı, hem de birbirleriyle etkileşimlerinin ortaya çıkardıkları sonuçlar vardır (Thompson, 2003). Klasik Test Kuramının aksine Genellenebilirlik Kuramı bu durumu göz önünde bulundurmaktadır. G kuramı sayesinde hata kaynakları ayrı ayrı ortaya konulduğu gibi, bunların etkileşimini de belirlemek mümkündür. Ayrıca Genellenebilirlik Kuramı ile elde edilen güvenirlik puanları araştırmacının farklı yorumlara ulaşmasını sağlar. Örneğin bir sınavda puanlara bakarak bireylerin birbirlerine göre durumları hakkında yorum yapabiliriz. Buna Klasik Test Kuramında bağıl değerlendirme denir ve yapılabilen tek yorumdur. Genellenebilirlik Kuramı mutlak değerlendirme hakkında da bilgi vermektedir (Shavelson ve Webb, 1991). Bir diğer ifadeyle Genellenebilirlik Kuramı ile hem bağıl hem de mutlak değerlendirme yapılabilmektedir.

Genellenebilirlik Kuramında güvenirlik kavramı, araştırmacının genellemek istediği basit veya karmaşık her evrene uygulanabilir. Genellenebilirlik Kuramında, değişkenlik kaynağı (facet) puanlayıcı, madde ya da zaman gibi hata kaynaklarına verilen addır. Değişkenlik kaynağı, ölçme hatasının olası kaynağı olarak tanımlanabilir. Değişkenlik kaynağıyla ilişkili bir varyans istenilen türden bir varyans olmayıp bu varyansın olabildiğince küçük tutulması istenir (Alharby, 2006). Tek değişkenlik kaynaklı evren en basit evreni, değişkenlik sayısının artması ise karmaşık evrenleri oluşturmaktadır (ShavelsonveWebb, 1991).

Genellenebilirlik Kuramının çerçevesi iki analiz çalışmasını birleştirir. İlk durum Genellenebilirlik (G) çalışması olarak adlandırılır. Bu analizde kabul edilebilir gözlemlerin evrene genellenebilmesi yapılır. İkinci durum Karar (K) çalışması olarak adlandırılır. G çalışması hata varyansının büyüklüğünü hesaplar. K çalışması ise en etkili biçimde, en güvenilir ölçümleri elde etmek ve en iyi ölçme desenine karar vermek için bilgileri kullanır (Eason, 1989: 9).

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

Çaprazlanmış ve Yuvalanmış Desenler

Çaprazlanmış (crossed) desen; bir faktörün bütün koşullarının diğer değişkenlik kaynağının bütün koşulları ile gözlemlenmesiyle oluşan desendir. Yuvalanmış (nested) desen ise bir değişkenlik kaynağının sadece bazı koşulları diğer değişkenlik kaynağının bazı koşullarınca gözlemleniyorsa oluşur (Shavelson ve Webb, 1991).

Çaprazlanmış desenlerde değişkenlik kaynaklarının arasına "x" işareti konulur. Yuvalanmış desenlerde değişkenlik kaynakları arasına ":" konulur (Shavelson ve Webb, 1991; Brennan, 2001). Genellenebilirlik Kuramında tamamen çaprazlanmış desenler veya hem çaprazlanmış hem de yuvalanmış faktörlerin karışımı uygulanabilmektedir.

Tesadüfi ve Sabit Değişkenlik Kaynakları

Bir değişkenlik kaynağında yer alan tüm durumlar, o değişkenlik kaynağında yer alabilecek olası tüm diğer durumlar ile değiştirilebilir olma özelliğine sahipse, bu değişkenlik kaynağı tesadüfi olarak tanımlanır (Kieffer, 1998). Bir örneklemin tesadüfi olarak adlandırılabilmesi için örneklem boyutunun evren boyutundan oldukça küçük olması gerekmektedir.

Tesadüfi durumlar sonucunda oluşan değişkenlik kaynaklarına bağlı çalışmalar, ilgili değişkenlik kaynağına ilişkin elde edilen bütün sonuçların evrene genellenebilmesini mümkün kılar. Araştırmacı çalışmasında oluşan değişkenlik kaynağına bağlı sadece belli başlı durumlarla ilgileniyorsa ve bunun dışındaki durumlara genelleme yapmak gibi bir amacı söz konusu değilse ilgili değişkenlik kaynağı sabit olarak tanımlanır (Crocker ve Algina, 1986). Ölçmenin genellenebilirliği değişkenlik kaynağının tesadüfi veya sabit olmasından etkilenmektedir (Shavelson ve Webb, 1991; 12). G ve Phi katsayısının hesaplanacağı bir karar (K) çalışmasında değişkenlere ait koşullar tesadüfi (random) ya da sabit (fixed) olarak belirlenebilir (Rentz, 1987).

G (Genellenebilirlik) ve K (Karar) Çalışmaları

Genellenebilirlik Kuramında güvenirliğin analizi iki adımda gerçekleşir: Birincisi; Genellenebilirlik (G) çalışması ve ikincisi Karar (K) çalışması olarak adlandırılmaktadır. Ölçmenin birden çok kullanımını tahmin etmek ve böylece varyans kaynakları ile ilgili mümkün olan en çok bilgiyi sağlamak G çalışmasının amacıdır. G çalışması, mümkün olduğunca çok değişkenlik kaynağını içerecek biçimde tasarlanmalıdır. Yani G çalışması, kabul edilebilir gözlemlerin evrenini en geniş şekilde tanımlamalıdır (Shavelson ve Webb, 1991). K çalışması ise sosyal bilimlerdeki ölçmelerin belli amaçları doğrultusunda G çalışmasından elde edilen bilgi kullanılarak en iyi sonuçlara ulaşacak desenler oluşturmak için kullanır (Shavelson ve Webb, 1991).

G çalışması yeterli genellemeyi yapabilecek K çalışmasının planlanmasına katkı sağlar. Bu sebeple G çalışması, K çalışmasında kullanılabilecek bütün desenleri kapsayacak şekilde tasarlanmalıdır. Çalışmanın desenine bakılarak çalışmayı G çalışması veya K çalışması olarak sınıflamak mümkün değildir (Crocker ve Algina, 1986; 158). Bir tek G çalışmasına dayanarak pek çok K çalışması yapılabilir.

Göreceli ve Mutlak Model

Genellenebilirlik Kuramında, ölçmenin ne derece genellenebilir olduğunu anlamak için eldeki verilerle nasıl karar çalışması yapıldığına bakılabilir. Sosyal bilimlerdeki ölçmeler genel olarak iki amaçla yapılır: (1) bireyleri sıralamak ve (2) bireyin bilgi, beceri veya tutumunu mutlak bir düzeyde belirlemek. Eğer öğrencinin bir sınavdan geçmesi veya kalması sınava giren diğer öğrencilerin o sınavdaki performanslarına bağlı ise araştırmacının kullanacağı model göreceli modeldir; öğrencinin sınavdan geçmesi veya kalması diğer öğrencilerin o sınavdaki performanslarına bağlı değilse; yani, mutlak bir değerlendirme anlayışı benimsenmişse, mutlak model kullanılmalıdır.

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

Genellenebilirlik Katsayısı

İdeal ölçme sonucu, bireyin, tüm kabul edilebilir gözlemleri üzerinden elde edilecek olan puanların ortalamasıdır ve bu değere "evren puanı" adı verilir (Shavelson, WebbveRowley, 1989). G kuramı sosyal bilimlerdeki ölçmelerde hataya sebep olan varyans kaynaklarına odaklanmasına rağmen, genellenebilirlik (G) katsayısı adında bir güvenirlik katsayısı da sağlamaktadır. G katsayısı, bireyin gözlenen puanından o bireyin evren puanının ne derece doğru genellenebileceğinin bir göstergesidir. Klasik Test Kuramındaki güvenirlik katsayısı gibi, genellenebilirlik katsayısı bireylerin puanlarındaki çeşitliliğin oranını yansıtır (Güler, 2008).

G katsayısının tanımı ölçmenin nasıl kullanılacağına bağlıdır. Hata varyansı göreceli ve mutlak modellerde farklılık gösterdiği için, G katsayısının büyüklüğü de kullanılan modele bağlı olarak değişir (Shavelson ve Webb, 1991; 14). Genellenebilirlik katsayısı, gözlenen puan varyansının evren puan varyansına oranlanması olarak tanımlanabilir (Crocker ve Algina, 1986). Gözlenen puan varyansı K çalışmasının desenine bağlıdır. Bu sebeple K çalışmasının farklı desenleri farklı katsayı sonuçları verecektir. Evren puanı ve evren puan varyansı genelleme evrenine bağlıdır. Bu nedenle, aynı K çalışması için farklı genelleme evrenlerini kullanan iki araştırmacı farklı Genellenebilirlik katsayıları elde eder (Crocker ve Algina, 1986). G kuramında ölçme hataları, göreceli kararlar için ölçme hataları ve mutlak kararlar için ölçme hataları olarak iki başlıkta değerlendirilebilir. G kuramıyla hesaplanan Genellenebilirlik (G) katsayısını göreceli ve mutlak değerlendirmeler için ayrı ayrı hesaplamak mümkündür (Shavelson ve Webb, 1991; Brennan, 2001). G Kuramında, göreceli modeller için G-Katsayısı, mutlak modeller için ise Phi katsayısı (güvenirlik indeksi) elde edilir. Değişkenlik kaynağında yapılacak değişiklikler, G Kuramında elde edilecek güvenirlik katsayılarını da etkilemektedir.

Phi (güvenirlik) Katsayısı

Phi katsayısı güvenirlik katsayısı olarak adlandırılır. Bu katsayısı mutlak değerlendirmeler için uygundur ve hata varyansı olarak mutlak hata varyansı kullanılır. Phi katsayısı evren puanının varyansının, evren puanının varyansı ve mutlak hata varyansının toplamına oranıdır.

G katsayısı Klasik Test Kuramında güvenirlik katsayısı ile benzerdir. G katsayısının hesaplanmasında göreli karar modelinde gerçek varyans, göreceli varyansla gerçek varyans toplamına oranlanarak bulunur.

Phi (Φ) katsayısı ise mutlak karar modelinin kapsamında kullanılır. Phi katsayısının elde edilmesinde evren puan varyansı, mutlak hata varyansıyla evren puan varyansının toplamına oranlanarak hesaplanır. Sonuç olarak, bu katsayılar (G ve Phi katsayısı) hatanın hnagi koşullarda kabul edileceğine göre değişiklik gösterir(Alharby, 2006). G çalışması sonucunda elde edilen varyans kestirimleri kullanılarak pek çok K çalışması düzenlemek mümkündür.

Brennan (2001), G ve Phi katsayılarının yeterlik ölçütlerinin isteğe bağlı olarak değiştiğini ancak araştırmacıların G ve Phi katsayılarının .80'den büyük olması durumunda "yüksek" olarak değerlendirilebileceğini ifade etmektedir. Shavelson ve Webb (1991) de .80 ve üzeri G ve Phi katsayılarının anlamlı olduğunu belirtmişlerdir.

Literatürde Genellenebilirlik Kuramıyla ilgili çalışmalar yer almaktadır. Eroğlu (2010) araştırmasında yapılandırılmış grid ve kavram haritasının geçerlik ve güvenirliklerini Genellenebilirlik Kuramına göre incelemiştir. Sonuçlara göre yapılandırılmış gridin güvenirlik katsayısı kavram haritasına göre daha yüksek bulunmuştur.

Nalbantoğlu (2011) araştırmasında, puanlayıcıların birlikte ve dönüşümlü olarak puanlama yaptıkları durumlarda sonuçlar arasında paralellik olduğunu göstermişlerdir. Yani puanlamaların birbiriyle tutarlı olduğu sonucuna ulaşılmıştır. Yapılan çalışmalar ve analizler göstermiştir ki, G kuramına göre puanlayıcılar arasında bulunan yüksek korelasyon puanlayıcılar arasında yüksek uyuma işaret etmektedir.

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

Büyüktura ve Demirtaşlı (2012)araştırmasında, yapılandırılmış grid maddelerinin çoktan seçmeli maddelere göre daha kolay olduğunu tespit etmiştir. İki testten elde edilen puanların ölçüt geçerliklerine bakılmış fakat aralarında anlamlı bir farka ulaşılamamıştır. Sonuçlara göre yapılandırılmış grid, çoktan seçmeli teste göre daha güvenilir bulunmuştur. Çoktan seçmeli test puanlarının varyansı yapılandırılmış grid testi puanlarından daha yüksek bulunmuştur. Aynı şekilde yapılandırılmış grid testi puanlarının aritmetik ortalamasının da çoktan seçmeli test puanlarının aritmetik ortalamasından anlamlı derecede yüksek olduğu görülmüştür.

Doğan (2012) kavram haritası ve yapılandırılmış grid tekniğinin çoktan seçmeli testlerle karşılaştırılması çalışmasında kavram haritası ölçeğinin test-tekrar test güvenirliğini .99 olarak bulmuştur. Yapılandırılmış grid ölçeğinin Cronbach Alpha güvenirliği .85; çoktan seçmeli testin KR-20 güvenirliği .82 şeklinde bulunmuştur. Araştırmada, kavram haritası, yapılandırılmış grid ve çoktan seçmeli testin öğrencilerin karne puanlarının SBS düzeltilmiş puanlarını yordama düzeyini kestirebilmek için basit doğrusal regresyon analizi kullanılmıştır. İncelenen değerlendirme teknikleri ile karne puanları, SBS puanlarını daha iyi yordamaktadır. Karne puanlarının yordanmasında en etkili ölçme aracının çoktan seçmeli test olduğu görülmüştür.

Aktaş (2013) aynı performans görevinin farklı sayıda puanlayıcılar tarafından üç farklı teknikle puanlanmasından elde edilen puanların güvenirliklerinin, Genellenebilirlik Kuramına göre incelenmesi adlı çalışmasında Genellenebilirlik Kuramı güvenirlik kestirimlerinin ortanca değerlerini incelemiştir. Beş puanlayıcının kontrol listesi kullanarak yaptıkları çalışmada, puanlayıcı sayısının ve kullanılan ölçeğin kategori sayısının arttırılmasıyla ortanca değerlerin arttığı sonucuna varılmıştır. Genellenebilirlik Kuramından elde edilen standart hatalar, puanlayıcı sayısı arttıkça azalmıştır. En düşük standart hata değerlerinin 10 puanlayıcı olması durumunda elde edildiği saptanmıştır. Genellenebilirlik Kuramı güvenirlik kestiriminin en yüksek değer aldığı durum, puanlayıcı sayısının 5 ve kategori sayısının 2 olduğu zamanda gözlenmiştir.

Araştırmanın Amacı

Araştırmanın amacı, 6. Sınıf Fen Bilimleri dersi 7. Ünitesi olan Elektriğin İletimi konu kazanımlarını içeren, araştırmacı tarafından hazırlanmış olan yapılandırılmış grid, çoktan seçmeli test ve performans görevi ölçme araçlarından elde edilen puanların birbirleri ile tutarlılığını araştırmaktır. Farklı özellikteki ölçme araçları ile aynı bireyler ve durumlarda elde edilen ölçme sonuçlarına ait hata miktarlarının her bir değişken için ayrı ayrı etkilerine bakılması amaçlanmıştır. Farklı ölçme araçları ve birden fazla hata kaynağı ele alınacağı için bu araştırma Genellenebilirlik Kuramı temel alınarak gerçekleştirilmiştir.

Alt Problemler

1. G Kuramına göre kestirilen varyanslar ve toplam varyansların açıklanan yüzdesiçoktan seçmeli test için nedir?
2. Test için yapılan K çalışması doğrultusunda maddelere ait hata varyanslarının değişimleri çoktan seçmeli test için nasıldır?
3. K çalışması sonuçlarına göre G ve Phi katsayılarının değişimleri çoktan seçmeli test için nasıldır?
4. G Kuramına göre kestirilen varyanslar ve toplam varyansların açıklanan yüzdeleri yapılandırılmış grid için nedir?
5. Yapılandırılmış grid için yapılan K çalışması doğrultusunda maddelerin hata varyansları değişimleri yapılandırılmış grid için nasıldır?
6. Yapılan K çalışmasındaki farklı senaryolara göre G ve Phi katsayılarının değişimi yapılandırılmış grid için nasıldır?
7. G Kuramına göre kestirilen varyanslar ve toplam varyansların açıklanan yüzdeleri performans görevi için nedir?

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

8. Performans görevi için yapılan K çalışması doğrultusunda maddelerin hata varyansı değişimleri performans görevi için nasıldır?
9. Yapılan K çalışmasındaki farklı senaryolara göre G ve Phi katsayılarının değişimi performans görevi için nasıldır?

Yöntem

Gerçekleşen bir olayı, bireyin ya da grubun özelliklerini var olduğu şekilde tanımlayan, durumu nicel ya da nitel yönden ortaya koymaya yarayan araştırma türüne betimsel araştırma denir (Karasar, 1998). Betimsel araştırmalar ne ve nasıl sorularına sistematik olarak cevap vererek olay ve durumların detaylı olarak betimlenmesi amacıyla yapılır (Başol, 2008). Bu araştırma, Genellenebilirlik Kuramı ile ölçme araçlarının güvenirliğinin belirlenmesi amacıyla yapıldığından bir durum belirleme çalışması olduğu için betimsel araştırma niteliğindedir.

Çalışma Grubu

Araştırma evrenini Akdeniz bölgesinde öğrenim gören altıncı sınıf öğrencileri oluşturmaktadır. Araştırmanın çalışma grubu ise, 2014-2015 eğitim öğretim yılında, Antalya'da ortaokula devam eden 6. Sınıf düzeyinden rastgele seçilen iki sınıftaki 40 öğrenciden oluşmaktadır. Çalışma grubu büyüklüğünün, literatürdeki genellenebilirlik çalışmaları dikkate alınarak, yeterli olduğu düşünülmektedir. Çalışma grubu öğrencilerinin 23'ü kız 17'si erkektir. Bahsi geçen iki sınıfta Elektriği İletelim ünitesi üç haftalık bir süreçte araştırmanın birinci yazarı rehberliğinde işlenerek araştırma verileri bu süreçte toplanmıştır.

Veri Toplama Araçları

Araştırmada altıncı sınıf Fen Bilimleri dersi Elektriğin İletimi ünitesi kazanımlarını içeren, öğrenci performanslarını değerlendirmeyi amaçlayan, üç farklı ölçme aracı araştırmacılar tarafından geliştirilmiştir. Bu ölçme araçları, performans görevi, yapılandırılmış grid ve çoktan seçmeli testtir. "Direnci neyin değiştirebileceğini biliyorum" adlı performans görevine ait yönerge ve performans görevinin değerlendirilmesinde kullanılabilecek kontrol listesi ödeve eklenmiştir. Ayrıca ödevin raporunun nasıl yazılacağını gösteren bir taslak verilmiştir. Performans görevinde öğrencilerin deney düzenekleri kurmaları istenmiştir. Kontrol listesi deney düzeneğinin nasıl kurulacağını sırasıyla belirten ifadelerden oluşturulmuştur. Değerlendirilmesi kontrol listesindeki adımların doğruluğuna göre yapılmıştır. Kavram yanılgılarını, yanlış ve eksik öğrenmeleri ortaya koyması yönüyle bilinen yapılandırılmış grid (Başol, 2016) bu yönüyle öğretime katkı getireceği düşünülerek mevcut çalışmada da tercih edilmiştir. Yapılandırılmış grid 16 resimli kutucuk ve 14 soru maddesinden oluşmaktadır. Yapılandırılmış gridin değerlendirilmesi formül kullanılarak gerçekleştirilmiştir. Yapılandırılmış gridi puanlayıcıların etkili olarak puanlayabilmeleri amacıyla, maddelerin ve resimlerin anlaşılabilirliğinden emin olmak için farklı bir grup öğrenciye ön uygulama yapılmıştır.Otuz maddeden oluşan çoktan seçmeli test geliştirilmiştir. Test maddelerinin geçerli ve güvenirliğini artırmak için ön uygulama yapılmıştır. Uygulama sonuçlarına göre yirmi maddeden oluşan çoktan seçmeli testin nihai hali belirlenmiştir. Üç farklı puanlayıcı her ölçme aracının değerlendirmesini ayrı ayrı yapmıştır.

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

Performans Ölçülmesinde Kullanılan Çoktan Seçmeli Testin İncelenmesi

Öğrencilere çoktan seçmeli testin 20 maddelik nihai hali uygulanmıştır. Testin madde analiz sonuçları Tablo 1'de, testten elde edilen betimsel istatistikler Tablo 2' de gösterilmiştir.

Madde No	Madde Güçlük İndeksi (pj)	Madde Ayırt Edicilik İndeksi(rj)
1	.78	.43
2*	.95	.07
3	.60	.79
4*	.90	.21
5	.60	.57
6	.63	.48
7	.50	.62
8*	.93	.21
9	.57	.86
10	.85	.36
11*	.35	.06
12*	.85	.13
13*	.35	.21
14	.63	.40
15	.82	.43
16*	.90	.14
17	.72	.50
18	.80	.42
19*	.53	.08
20*	.45	.23

Not: * Madde ayırt edicilik indeksi .20'nin altında olanlar toplam puana eklenmemiştir.

Tablo 2. Çoktan Seçmeli Testin Nihai Uygulamasına Ait Betimsel İstatistik

Öğrenci Sayısı (N)	40
Madde Sayısı (K)	20
Aritmetik Ortalama (X)	68.5
Varyans (s^2)	10.360
Standart Sapma (s)	3.219
Ortanca	70.0
En Düşük Puan (Min.)	25
En Büyük Puan (Max.)	95
Çarpıklık	-.381
Basıklık	-.142
KR 20	.684

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

Yirmi maddeden oluşan çoktan seçmeli test 40 öğrenciye uygulanmıştır. Testten alınan en yüksek puan 95 en düşük puan ise 25'tir. Testin aritmetik ortalaması 68.5; medyanı 70; standart sapması 3.21; varyansı 10.36 şeklindedir. Puanların çarpıklık katsayısı -.38 basıklık katsayısı ise -.14 olarak hesaplanmıştır. Bu değerler elde edilen puan dağılımının normale yakın olduğunu göstermiştir. Testin güvenirliği içtutarlılık hesaplama yöntemlerinden KR-20 hesaplanarak incelenmiş .68 ile .70 sınır değerine yakın olduğu bulunmuştur.

İşlem

Ünite, araştırmanın birinci yazarının rehberliğinde üç haftalık bir sürede işlenmiştir. Öncelikle öğrencilere performans görevi yönergesiyle verilip ünite sonuna kadar süre verilmiştir. Hazırlanan ölçme araçlarının üçü de aynı kazanımları ölçmeye yönelik olarak hazırlanmıştır. Ders işleme sürecinde deney yöntemi kullanılmıştır. Sınıfa getirilen malzemelerle öğrencilerin deney düzenekleri kurmaları ve gözlem yapmaları sağlanmıştır. Soru-cevap yöntemiyle öğrencilerin eksik bilgileri tamamlanmıştır. Ünite sonunda yapılandırılmış grid, bir diğer derste ise çoktan seçmeli test uygulanmıştır. Son olarak istenilen tarihte performans görevleri teslim alınmıştır. Üç farklı puanlayıcıher üç ölçme aracının sonuçlarını farklı zamanlarda değerlendirmişlerdir.

Verilerin Analizi

İkinci araştırmacının hazırladığı çoktan seçmeli testten elde edilen verilerin madde istatistikleri analizinde TAP.exe (Brooks ve Johanson, 2003), betimsel istatistiklerde SPSS 22 (IBM Inc, NY, 2013) ve Genellenebilirlik analizinde EduG6.1-e (Quebec, Canada, 2012)programları ile gerçekleştirilmiştir. Çalışmada çaprazlanmış desen uygulanmıştır. Çaprazlanmış (crossed) desen; bir faktörün bütün koşullarının diğer değişkenlik kaynağının bütün koşulları ile gözlemlenmesiyle oluşan desendir.

Bulgular

***Alt problem 1*: "**Performansın ölçülmesine yönelik G kuramına göre kestirilen varyanslar ve toplam varyansların açıklama yüzdeleri nelerdir?"

Çoktan seçmeli test için birey(b) ve madde(m) değişkenlerinin çaprazlanmış desen değişimi ve varyans kaynaklarının oranlarını belirlemek için Genellenebilirlik (G) çalışmasına ait sonuçlar Tablo 3'de verildiği şekildedir.

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

Varyans Kaynağı	Sd	Kareler Toplamı	Kareler Ortalaması	Varyans	%
B	39	20.838	.534	.018	8.4
M	19	27.313	1.437	.032	14.5
Bm	741	124.836	.168	.168	77.1
Toplam					100

Tablo 3'e göre birey (b) ana etkisi için kestirilen varyans bileşeninin toplam varyansın % 8,4' ünü açıkladığı görülmektedir. Tek değişkenli desenle yapılan incelemede bireyler için kestirilen varyans bileşeni, toplam varyans içinde en düşük paya sahip olan varyans bileşenidir. Genellenebilirlik Kuramı çalışmalarında, birey ana etkisi evren puanı varyansı olarak değerlendirilir. Ölçülen özellik bireyler arası farklılaşmayı ifade eder (Shavelson ve Webb, 1991; Brennan 2001). Sonuç olarak bireyler arası farklılıkların az olduğu şeklinde yorumlanır.

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

Ancak Genellenebilirlik çalışmalarında bireyler için kestirilen varyansın toplam varyans içindeki oranının en büyük olması istenir. Ölçme ile elde edilen sonuçlar bireyler arası farklılıkların ortaya çıkarılabildiğinin bir göstergesidir (Güler, 2008).

Madde (m) ana etkisi için tek değişkenli desenle yapılan G çalışmasında kestirilen varyans bileşeni (.032) toplam varyansın % 14.5' ini açıklamaktadır. Madde güçlüklerinin birbirinden farklılığına işaret etmektedir. Madde ana etkisinin varyans bileşeni büyüklüğün, toplam varyans değişkeni büyüklüğünde ikinci orana sahiptir. Bu değer, birey etkisinden daha fazladır.

Birey x madde ortak etkisi (.168) toplam varyansın % 77.1' ini açıklamaktadır. Birey x madde ortak etkisi tek değişkenli desenle yapılan G çalışmasında elde edilen en büyük varyans değeridir. Bu durum; bu ölçme için birey x madde ortak etkisinden kaynaklanan farklılığın büyük olduğuna, belli bireylerin bağıl durumlarının bir maddeden diğerine çok farklılaştığı anlamına gelir. Ayrıca birey x madde varyans değerinin büyük olması birey ve madde ortak etkisi veya tesadüfü hataların büyük olabileceği anlamına gelebilir.

Alt problem 2: **"**Performansın ölçülmesine yönelik madde sayılarının arttırılıp azaltılması senaryolarına göre hesaplanan hata varyansları nelerdir?"

Yapılan çoktan seçmeli test için 20 madde ve bu madde sayısının azaltılıp arttırılması durumunda hata varyanslarının değişimlerini belirlemek amacıyla G kuramı çalışmasıyla yapılan K çalışmasına göre hesaplanan hata varyansları değişimleri Tablo 4' te verilmiştir

Tablo 4. Performansın Ölçülmesine İlişkin Yapılan K Çalışması ile Madde Sayıları Senaryolarına Göre Bağıl ve Mutlak Hata Varyansları

Madde Sayısı	Bağıl	Mutlak
16	.010	.012
18	.009	.011
20	**.008**	**.010**
22	.007	.009
24	.007	.008

Tablo 4' te çoktan seçmeli test için madde sayılarının arttırılıp azaltılması durumlarına bağlı olarak hesaplanan bağıl ve mutlak hata varyansları verilmiştir. Buna göre, çoktan seçmeli 20 maddelik nihai testten elde edilen bağıl hata varyansı .008 ve mutlak hata varyansı .010 olarak kestirilmiştir. Bu değerler incelendiğinde, asıl uygulamadan elde edilen mutlak hata varyansı, bağıl hata varyansından yüksektir. Madde sayıların arttırılıp azaltılması ile yapılan senaryolara göre bağıl ve mutlak hata varyansları hesaplanmıştır. Madde sayısının dört azaltılması durumunda (16 madde üzerinden) yapılan hesaplamada bağıl hata varyansı .010 ve mutlak hata varyansı .012 olarak kestirilmiştir. Bu senaryo çalışması üzerinden elde edilen hata varyanslarında; mutlak hata varyansının bağıl hata varyansından yüksek olduğu gözlenmiştir. Madde sayısının iki azaltılması durumunda (18 madde üzerinden) yapılan hesaplamalarda bağıl hata varyansı .009 ve mutlak hata varyansı .011 olarak kestirilmiştir. Bu durumda madde sayısı azaldıkça, bağıl hata varyansının mutlak hata varyansına yaklaştığı görülmüştür. Sonuç olarak, senaryo çalışması üzerinden elde edilen hata varyanslarında; mutlak hata varyansının bağıl hata varyansından yüksek bulunmuştur.

Madde sayısının iki arttırılması durumunda (22 madde üzerinden) yapılan hesaplamalarda bağıl hata varyansı .007 ve mutlak hata varyansı .009 olarak kestirilmiştir. Sonuç olarak yine elde edilen hata varyanslarında; mutlak hata varyansının bağıl hata varyansından yüksek olduğu gözlenmiştir. Son olarak madde sayısının dört arttırılması durumunda (24 madde üzerinden) yapılan hesaplamalarda bağıl hata varyansı .007 ve mutlak hata varyansı .008 olarak kestirilmiştir.

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

Bu senaryo çalışmasında da diğer iki senaryodaki gibi mutlak hata varyansının bağıl hata varyansına göre biraz daha yüksek olduğu gözlenmiştir.

Tablo 4' e göre çoktan seçmeli test maddelerinin arttırılıp azaltılmasına göre bağıl ve mutlak hata varyansları değişim göstermektedir. Madde sayılarının arttırılması ile hem bağıl hem de mutlak hata varyanslarının azalma gösterdiği ortaya çıkmıştır. Ayrıca kestirilen hata varyanslarından mutlak hata varyansının, bağıl hata varyansına göre yapılan tüm senaryo denemelerinde daha yüksek değere sahip olduğu gözlenmiştir. Bu sonuçlara göre madde sayısı arttıkça tahminin güvenirliğinin artacağı sonucuna varılabilir.

Alt problem 3: **"**Performansın ölçülmesine yönelik madde sayılarının arttırılıp azaltılması senaryolarına göre hesaplanan G ve Φ katsayıları nelerdir?"

Performansın ölçülmesinde kullanılan çoktan seçmeli test için 20 madde ve madde sayısının arttırılıp azaltılması durumlarında G Kuramı çalışması ile yapılan K çalışması sonucu elde edilen G ve Φ katsayıları Tablo 5' te verilmiştir

Tablo 5. Performansın Ölçülmesine İlişkin Yapılan K Çalışması ile Madde Sayıları Senaryolarına Göre G ve Phi Katsayıları

Madde Sayısı	G	Φ
16	.634	.593
18	.661	.621
20	**.684**	**.646**
22	.704	.667
24	.722	.686

Tablo 5' 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 sonuçlarında; G katsayısı .684 ve Φ katsayısı .646 olarak, madde sayısının dört azaltılması durumunda (16 madde üzerinden) G katsayısı .634 ve Φ katsayısı .593 olarak, madde sayısının iki azaltılması durumunda (18 madde üzerinden) G katsayısı .661 ve Φ katsayısı .621 olarak kestirilmiştir.

Madde sayılarının arttırılması üzerinden yapılan analiz sonuçlarına göre; madde sayısının 2 arttırılması durumunda (22 madde üzerinden) G katsayısı .704 ve Φ katsayısı .667 olarak, madde sayısının 4 arttırılması durumunda (24 madde üzerinden) G katsayısı .722 ve Φ katsayısı .686 olarak kestirilmiştir.

Tablo 5 incelendiğinde, madde sayısının azaltılması durumlarında Φ katsayısı ve G katsayılarının azaldığı, madde sayısının arttırıldığı durumlarda ise Φ katsayısı ve G katsayılarının arttığı gözlemlenmiştir.

Performansın Ölçülmesinde Kullanılan Yapılandırılmış Gridin İncelenmesi

Performansın ölçülmesine yönelik uygulanan yapılandırılmış grid 16 kutucuk ve 14 maddeden oluşmaktadır. Uygulanan sınav üç farklı puanlayıcı tarafından puanlanmış ve puanlayıcılar üzerinden elde edilen verilerle işlemler gerçekleştirilmiştir. Yapılandırılmış gride yönelik puanlayıcılardan elde edilen puanlara ait betimsel istatistikler Tablo 6' da verilmiştir.

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

Tablo 6. Performansın Ölçülmesinde Kullanılan Yapılandırılmış Grid için 3 Puanlayıcıya Ait Betimsel İstatistikler (N=40)

İstatistikler	1. Puanlayıcı	2. Puanlayıcı	3. Puanlayıcı
Ortalama	76.02	76.25	75.77
Ortanca	78.50	79	78.50
Tepe Değer	64	79	80
Std. Sapma	10.74	10.62	10.90
Varyans	115.51	112.91	118.89
Çarpıklık	-.39	-.47	-.52
Basıklık	-.48	-.32	-.16
Minimum	49	49	47
Maksimum	94	94	94
Ranj	45	45	47
α güvenirliği	.98	.96	.95

Tablo 6 incelendiğinde, 40 öğrencinin 14 madde üzerinden aldıkları puanlara ilişkin en yüksek ortalama 76.25 şeklindedir ve ikinci puanlayıcıya aittir. En düşük ortalama ise 75.77 ile üçüncü puanlayıcıya aittir. Birinci puanlayıcı 76.02 ortalama değeri ile bu iki değer arasında yer almıştır. Birinci, ikinci ve üçüncü puanlayıcıya ilişkin ortanca değer aritmetik ortalamadan yüksektir ve puanlayıcılara ait puanların sola çarpık dağılım gösterdiği gözlenmiştir. Bu durum çarpıklık katsayılarının birinci, ikinci ve üçüncü puanlayıcıya ait puan değerleri için negatif çıkmasıyla da görülmektedir.

Basıklık katsayısına bakıldığında; her üç puanlayıcıya ait puan değerlerinin 0'dan küçük çıkması puanların normalden daha basık dağılım gösterdiğini ortaya koymaktadır. Puanlayıcıların verdikleri puanlara ait Cronbach Alfa (α) güvenirlik katsayıları birbirine yakın ve yüksek değerlerdir. Puanlayıcıların 14 madde üzerinden verdikleri puanlar arasındaki korelasyon değerleri Tablo 7' de verilmiştir.

Tablo 7. Puanlayıcıların 14 Maddeye Verdikleri Puanlar Arasındaki Korelasyon Katsayıları

	1. Puanlayıcı	2. Puanlayıcı	3. Puanlayıcı
1. Puanlayıcı	-	.996	.991
2. Puanlayıcı		-	.991

Tablo 7 incelendiğinde, puanlayıcıların 14 madde üzerinden verdikleri puanların birbirleri ile olan korelasyon katsayılarının yüksek olduğu gözlenmektedir. Elde edilen bu değerlere göre puanlayıcılar arasındaki uyumun mükemmele yakın olduğundan söz edilebilir. Özellikle birinci ve ikinci puanlayıcılar arasındaki korelasyon katsayısının oldukça yüksek olduğu görülmüştür.

Alt problem 4: "Performansın ölçülmesine yönelik G Kuramına göre kestirilen varyanslar ve toplam varyansların açıklama yüzdeleri nelerdir?"

Performansın ölçülmesine yönelik hazırlanan 14 maddelik yapılandırılmış grid ölçme aracının G çalışması ile elde edilen varyanslarını ve varyans yüzdelerini hesaplamak için tümüyle çaprazlanmış b x m x p modeli uygulanmıştır. Ölçmenin uygulandığı 40 öğrenci, 14 madde ve 3 puanlayıcıdan oluşan verilerde iki yüzeyli çaprazlanmış desenle 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 8' de verilmiştir

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

Tablo 8. İki Yüzeyli G Çalışması Sonucunda Ölçmenin Kestirilen Varyansları ve Toplam Varyansı Açıklama Oranları

Varyans Kaynağı	Sd	Kareler Toplamı	Kareler Ortalaması	Varyans	%
b	39	1796.454	46.062	1.083	11.2
m	13	1747.033	134.387	-.329	.0
p	2	2.832	1.416	-.300	.0
bm	507	2356.395	4.647	.135	1.4
bp	78	11.977	.153	-.291	.0
mp	26	4511.934	173.535	4.232	43.7
bmp	1014	4299.922	4.240	4.240	43.8
Toplam					100

Tablo 8'e göre, birey (b) ana etkisi için kestirilen varyans bileşenini (1.083) toplam varyansın % 11.2' sini açıklamaktadır. İki yüzeyli çaprazlanmış desenle bireyler için kestirilen varyans bileşeni, toplam varyans içinde en yüksek üçüncü sıraya sahiptir. Madde (m) ana etkisi için kestirilen varyans bileşeni iki yüzeyli çaprazlanmış desenle yapılan G çalışmasına göre kestirilen varyans bileşeni eksi değer aldığı için (-.329) toplam varyansı açıklama yüzdesi içinde (%0) bir etkiye sahip olmadığı görülmüştür.

Puanlayıcı ana etkisinin G çalışması ile kestirilen varyans bileşeni eksi değer aldığı için (-.300) toplam varyansı açıklama yüzdesi içinde (%0) bir etkiye sahip olmadığı görülmüştür. Varyans analizinde bazı durumlarda varyansın negatif değer aldığı görülür. Negatif kestirimler, ölçme modelinin hatalı belirlenmesinden ya da örnekleme hatasından kaynaklanabilir. Negatif kestirimler görece büyük olduğunda, ölçme modelinin hatalı olduğunun bir göstergesi olduğu düşünülür. Güler ve ark (2012)'ye göre bu gibi durumlarda farklı ölçme modelleri denemek sorunu çözebilir. Negatif kestirimler görece küçük olduğunda ise (sıfıra yakın), bunun sebebi çok büyük evreni temsil eden oldukça küçük örneklem sayısı alınması olabilir (Güler ve ark, 2012). Shavelson ve Webb (2005) böyle bir durumda dört çözüm önerisinde bulunmuşlardır. İlk olarak Cronbach, Gleser, Nanda ve Rajaratnam (1972) negatif varyans değerinin sıfır alınmasını önermişler, ikinci olarak Brennan (2001) beklenen ortalama kareler eşitliğinde negatif varyanslar olduğu gibi kullanılarak negatif varyansların sıfır alınmasını, üçüncü olarak Shavelson ve Webb (2005) Bayesian metot kullanılarak tahmin edilen varyans için en küçük değerin sıfır olarak girilmesini, son olarak Searle (1987) maksimum olabilirlik modeli kullanılmasını önermiştir (Akt. Shavelson ve Webb, 2005).

Puanlayıcı etkisinin iki yüzeyli çaprazlanmış desenle yapılan G çalışması ile kestirilen varyans oranının çok düşük olması, puanlayıcıların tüm bireyler için yaptıkları puanlamalar arasında bir fark bulunmadığı, puanlamalar arasında tutarlılık olduğu şeklinde yorumlanabilir.

Birey x madde (bm) ortak etkisi (.135) toplam varyansın % 1.4' ünü açıklamaktadır. Birey x madde ortak etkisi iki yüzeyli çaprazlanmış desenle kestirilen en yüksek dördüncü değere sahip varyans değeridir. Bu durum birey ve maddeler arasında büyük farklılıklar olmadığını göstermektedir.

Birey x puanlayıcı (bp) ortak etkisi (-.291) ile toplam varyansın % 0' ını açıklamaktadır. Bu sonuç iki yüzeyli çaprazlanmış desene göre birey x puanlayıcı ortak etkisinden kaynaklanan bir farklılığın olmadığı anlamına gelir.

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

Madde x puanlayıcı (mp) ortak etkisi (4.232) iki yüzeyli çaprazlanmış desenle kestirilen en yüksek ikinci değere sahip varyans değeridir. Bu değer toplam varyansın % 43.7'sini açıklamaktadır. Bu değer, puanlayıcıların verdikleri puanların maddelere göre büyük farklılık gösterdiğini ifade etmektedir.

Birey x madde x puanlayıcı (artık) ortak etkisi varyans bileşeni de (4.240) toplam varyansın % 43.8' ini açıklamaktadır. Bu oran varyans değerleri arasında en yüksek 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 5: "Performansın ölçülmesine yönelik madde sayılarının arttırılıp azaltılması senaryolarına göre hesaplanan mutlak ve bağıl hata varyansları nelerdir?"

Performansının ölçülmesi için yapılan 16 kutucuk ve 14 maddelik yapılandırılmış gridin üç puanlayıcı tarafından değerlendirilmesi yapılmıştır. Madde ve puanlayıcı sayılarının arttırılıp azaltılmasına göre yapılan K çalışması sonucu kestirilen mutlak ve bağıl hata varyansları Tablo 9'da verilmiştir.

Tablo 9. Performansın Ölçülmesine İlişkin Yapılan K Çalışması ile Madde ve Puanlayıcı Sayıları Senaryolarına Göre Bağıl ve Mutlak Hata Varyansları

Madde Sayıları	Puanlayıcı Sayıları					
	3		4		5	
	Mutlak	Bağıl	Mutlak	Bağıl	Mutlak	Bağıl
10	.296	.154	-	-	-	-
12	.246	.129	-	-	-	-
14	**.211**	**.110**	.161	.085	.130	.070
16	.185	.096	**-**	**-**	-	-

Tablo 9 incelendiğinde, performansın ölçülmesinde kullanılan 16 kutucuk ve 14 maddelik yapılandırılmış grid ve üç puanlayıcıya göre elde edilen mutlak hata varyansı .211 ve bağıl hata varyansı .110 olarak kestirilmiştir. Madde sayısının iki azaltılarak 12 madde üzerinden; 3 puanlayıcı için kestirilen mutlak hata varyansı .246 ve bağıl hata varyansı .129 olarak kestirilmiştir.

Madde sayısının 4 azaltılarak 10 madde ve puanlayıcı sayısının 3 olması durumunda mutlak hata varyans .296 ve bağıl hata varyansı .154 olarak kestirilmiştir. Madde sayısının 2 artırılarak 16 madde ve puanlayıcı sayısının 3 olması durumunda mutlak hata varyans .185 ve bağıl hata varyansı .096 olarak kestirilmiştir. Madde sayısının aynı kalması ve puanlayıcı sayılarının değişimlerine göre hata varyansları incelendiğinde; madde sayısı 14 iken puanlayıcı sayısının bir arttırılıp dört olması durumunda mutlak hata varyansı .161 ve bağıl hata varyansı .085, puanlayıcı sayısının iki arttırılıp beş olması durumunda mutlak hata varyansı .130 ve bağıl hata varyansı .070 olarak kestirildiği görülmüştür.

Tablo 9' da yer alan değerler incelendiğinde madde sayısının ve puanlayıcı sayılarının artması durumlarında mutlak ve bağıl hata varyans değerlerinin giderek azaldığı açıkça görülmektedir. Tüm madde ve puanlayıcı sayılarının değişimi senaryolarına göre kestirilen mutlak ve bağıl hata varyansı değerleri incelendiğinde bağıl hata varyans değeri mutlak hata varyans değerine göre daha azdır.

Alt problem 6: "Performansın ölçülmesine yönelik madde sayılarının arttırılıp azaltılması senaryolarına göre hesaplanan G ve Φ katsayıları nelerdir?"

Uygulanan yapılandırılmış gride 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 10' da verilmiştir.

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

Tablo 10. Performansın Ölçülmesine İlişkin Yapılan K Çalışması ile Madde ve Puanlayıcı Sayıları Senaryolarına Göre Phi ve G Katsayıları

Madde Sayıları	**Puanlayıcı Sayıları**					
	3		**4**		**5**	
	G	**Φ**	**G**	**Φ**	**G**	**Φ**
10	.874	.785	-	-	-	-
12	.893	.814	-	-	-	-
14	**.907**	**.836**	.926	.870	.939	.892
16	.917	.854	-	-	-	-

Tablo 10'da; iki yüzeyli çaprazlanmış desenle yapılan ölçme sonuçlarına göre 14 madde ve üç puanlayıcıya göre kestirilen G katsayısının .907 ve Φ katsayısının da .836 olduğu görülmüştür. Kestirilen katsayı değerlerine göre G katsayısı Φ katsayısından daha yüksektir. Madde sayısının dört azaltılması ile 10 madde üzerinden kestirildiğinde; puanlayıcı sayısının üç olması durumunda G katsayısı .874 ve Φ katsayısı .785 olmuştur. Madde sayısının iki azalması ile 12 madde üzerinden puanlayıcı sayısının üç olduğu durumda G katsayısı .893 ve Φ katsayısı .814 olarak kestirilmiştir. Madde sayısının iki artırılması ile 16 madde üzerinden, puanlayıcı sayısının ise üç olması durumunda G katsayısı .917 ve Φ katsayısı .854 olarak kestirilmiştir. Madde sayısı 14 iken puanlayıcı sayısının dört olması durumunda G katsayısı .926 ve Φ katsayısı .870, puanlayıcı sayısının beş olması durumunda G katsayısı .939 ve Φ katsayısı .892 olarak kestirilmiştir.

Tablo 10' a göre 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

Performansın Ölçülmesinde Kullanılan Performans Görevinin İncelenmesi

Performansın ölçülmesine yönelik verilen performans görevi 10 maddelik bir kontrol listesiyle üç puanlayıcı tarafından değerlendirilmek üzere oluşturulmuştur. Performans görevine yönelik puanlayıcılardan elde edilen puanlara ait betimsel istatistikler Tablo 11' de verilmiştir.

Tablo 11. Performansın Ölçülmesinde Kullanılan Performans Görevi için 3 Puanlayıcıya Ait Betimsel İstatistikler (N=40)

	PUANLAYICILAR		
İstatistikler	1. Puanlayıcı	2. Puanlayıcı	3. Puanlayıcı
Ortalama	84.25	76.00	82.25
Ortanca	90.00	80.00	90.00
Tepe Değer	100	80	100
Std. Sapma	18.79	19.32	20.93
Varyans	353.26	373.33	438.39
Çarpıklık	-1.38	-.85	-1.79
Basıklık	1.41	.13	3.02
Minimum	30	30	20
Maksimum	100	100	100
α güvenirliği	.89	.94	.94

Tablo 11 incelendiğinde, 40 öğrencinin 10 madde üzerinden aldıkları puanlara ilişkin en yüksek ortalama birinci puanlayıcıya aittir ve 84.25 şeklindedir. En düşük ortalama ise 76.0 ile ikinci puanlayıcıya aittir. Üçüncü puanlayıcı 82.25 ortalama değeri ile bu iki değer arasında yer almıştır. Birinci, ikinci ve üçüncü puanlayıcıya ilişkin ortanca değer aritmetik ortalamadan

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

yüksektir ve puanlayıcılara ait puanlamanın sola çarpık dağılım gösterdiği gözlenmiştir. Bu durum çarpıklık katsayılarının birinci, ikinci ve üçüncü puanlayıcıya ait puan değerleri için negatif çıkmasından da anlaşılabilir.

Basıklık katsayısına bakıldığında; her üç puanlayıcıya ait puan değerlerinin 0'dan büyük çıkması puanların normalden daha sivri dağılım gösterdiğini ortaya koymaktadır. Puanlayıcıların verdikleri puan değerlerine ait Cronbach Alfa (α) güvenirlik katsayıları birbirine yakın ve yüksek değerlerdir. Puanlayıcıların 10 madde üzerinden verdikleri puanlar arasındaki korelasyon değerleri Tablo 12' de verilmiştir.

Tablo 12. Puanlayıcıların 10 Maddeye Verdikleri Puanlar Arasındaki Korelasyon Katsayıları

	1. Puanlayıcı	2. Puanlayıcı	3. Puanlayıcı
1. Puanlayıcı	-	.818	.907
2. Puanlayıcı		-	.891

Tablo 12 incelendiğinde, puanlayıcıların 10 madde üzerinden verdikleri puanların birbirleri ile olan korelasyon katsayılarının yüksek olduğu gözlenmiştir. Elde edilen bu değerlere göre puanlayıcılar arasındaki uyumun yüksek olduğundan söz edilebilir. Özellikle birinci ve üçüncü puanlayıcılar arasındaki korelasyon katsayısının oldukça yüksek olduğu görülmektedir.

Alt problem 7: "Performansın ölçülmesine yönelik G Kuramına göre kestirilen varyanslar ve toplam varyansların açıklama yüzdeleri nelerdir?"

Performansın ölçülmesine yönelik hazırlanan performans görevinin 10 maddelik kontrol listesinin G çalışması ile elde edilen varyanslarını ve varyans yüzdelerini hesaplamak için tümüyle çaprazlanmış b x m x p modeli uygulanmıştır. Ölçmenin uygulandığı 40 öğrenci, 10 madde ve 3 puanlayıcıdan oluşan verilerde iki yüzeyli çaprazlanmış desenle 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 13' te verilmiştir.

Tablo 13. İki Yüzeyli G Çalışması Sonucunda Ölçmenin Kestirilen Varyansları ve Toplam Varyansı Açıklama Oranları

Varyans Kaynağı	Sd	Kareler Toplamı	Kareler Ortalaması	Varyans	%
b	39	41.516	1.064	.033	20.6
m	9	12.250	1.361	.009	5.8
P	2	.561	.280	.0003	.2
bm	351	42.816	.121	.001	1.1
bp	78	3.438	.044	-.007	.0
mp	18	3.655	.203	.002	1.3
bmp	702	81.678	.116	.116	70.9
Toplam					100

Tablo 13 incelendiğinde, birey (b) ana etkisi için kestirilen varyans bileşenini (.033) toplam varyansın % 20.6' sını açıklamaktadır. İki yüzeyli çaprazlanmış desenle bireyler için kestirilen varyans bileşeni, toplam varyans içinde en yüksek ikinci sırada paya sahiptir. Madde (m) ana etkisi için kestirilen varyans bileşeni iki yüzeyli çaprazlanmış desenle yapılan G çalışmasına göre kestirilen varyans bileşeni (.009) toplam varyansı açıklama yüzdesi içinde %5.8' lik bir etkiye sahiptir. İki yüzeyli çaprazlanmış desenle maddeler için kestirilen varyans bileşeni, toplam varyans içinde en yüksek üçüncü sırada paya sahiptir. Puanlayıcı ana etkisinin G çalışması ile kestirilen varyans bileşeni (.0003) ile toplam varyansı açıklama yüzdesi % .2'lik bir etkiye sahiptir. Puanlayıcı etkisinin iki yüzeyli çaprazlanmış desenle yapılan G çalışması ile kestirilen varyans oranının çok

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

düşük olması, puanlayıcıların tüm bireyler için yaptıkları puanlamalar arasında bir fark bulunmadığının, puanlamaların tutarlı olduğuna işaret etmektedir.

Birey x madde (bm) ortak etkisi (.001), toplam varyansın % 1.1'ini açıklamaktadır. Birey x madde ortak etkisi iki yüzeyli çaprazlanmış desenle kestirilen en yüksek beşinci değere sahip varyans değeridir. Birey x puanlayıcı (bp) ortak etkisi eksi değer aldığı için (-.007) toplam varyansın % 0' ını açıklamaktadır. Birey x puanlayıcı etkisinin iki yüzeyli çaprazlanmış desene göre birey x puanlayıcı ortak etkisinden kaynaklanan bir farklılığın olmadığı yorumu yapılabilir. Madde x puanlayıcı (mp) ortak etkisi (.002) toplam varyansın % 1.3' ünü açıklamaktadır. İki yüzeyli çaprazlanmış desenle bireyler için kestirilen varyans bileşeni, toplam varyans içinde en yüksek dördüncü sırada paya sahiptir.

Birey x madde x puanlayıcı (artık) ortak etkisi varyans bileşeni ise (.11635) toplam varyansın % 70.9' unu açıklamaktadır. Bu oran varyans değerleri arasında en yüksek 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 8: "Performansın ölçülmesine yönelik madde sayılarının arttırılıp azaltılması senaryolarına göre hesaplanan mutlak ve bağıl hata varyansları nelerdir?"

Performansının ölçülmesi için verilen performans görevinin 10 maddelik kontrol listesiyle üç puanlayıcı tarafından değerlendirilmesiyle veriler elde edilmiştir. Madde ve puanlayıcı sayılarının arttırılıp azaltılmasına göre yapılan K çalışması sonucu kestirilen mutlak ve bağıl hata varyansları Tablo 14'te verilmiştir.

Tablo 14. Performansın Ölçülmesine İlişkin Yapılan K Çalışması ile Madde ve Puanlayıcı Sayıları Senaryolarına Göre Bağıl ve Mutlak Hata Varyansları

Madde Sayıları	Puanlayıcı Sayıları					
	3		4		5	
	Mutlak	Bağıl	Mutlak	Bağıl	Mutlak	Bağıl
10	.005	.040	.004	.003	.003	.002
12	.004	.003	-	-	-	-
14	.003	.002	-	-	-	-
16	.003	.002	-	-	-	-

Tablo 14 incelendiğinde, performansın ölçülmesinde kullanılan performans görevi kontrol listesinin 10 maddesi ve üç puanlayıcıya göre elde edilen mutlak hata varyansı .005 ve bağıl hata varyansı .040 olarak kestirilmiştir. Madde sayısının aynı kalması ve puanlayıcı sayılarının değişimlerine göre hata varyansları incelendiğinde; madde sayısı 10 iken puanlayıcı sayısının bir arttırılıp dört olması durumunda mutlak hata varyansı .004 ve bağıl hata varyansı .003, puanlayıcı sayısının iki arttırılıp beş olması durumunda mutlak hata varyansı .003 ve bağıl hata varyansı .002 olarak kestirilmiştir.

Madde sayısının iki arttırılarak 12 madde üzerinden; 3 puanlayıcı için kestirilen mutlak hata varyansı .004 ve bağıl hata varyansı .003 olarak kestirilmiştir. Madde sayısının 4 arttırılarak 14 madde üzerinden ve puanlayıcı sayısının 3 olması durumunda ise mutlak hata varyansdeğeri .003 ve bağıl hata varyansı .002 olarak kestirilmiştir. Madde sayısının 6 arttırılarak 16 madde üzerinden ve puanlayıcı sayısının 3 olması durumunda mutlak hata varyansı .003 ve bağıl hata varyansı .002 olarak kestirilmiştir.

Tablo 14 incelendiğinde madde sayısının ve puanlayıcı sayılarının artması durumlarında mutlak ve bağıl hata varyans değerlerinin giderek azaldığı görülmektedir. Tüm madde ve puanlayıcı sayılarının değişim senaryolarına göre, kestirilen mutlak ve bağıl hata varyansı değerleri

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

incelendiğinde bağıl hata varyans değeri mutlak hata varyans değerine göre daha az olduğu görülmüştür.

Alt problem 9: "Performansın ölçülmesine yönelik madde sayılarının arttırılıp azaltılması senaryolarına göre hesaplanan G ve Φ katsayıları nelerdir?"

Uygulanan performans görevi kontrol listesine 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 15' de verilmiştir.

Tablo 15. Performansın Ölçülmesine İlişkin Yapılan K Çalışması ile Madde ve Puanlayıcı Sayıları Senaryolarına Göre Phi ve G Katsayıları

Madde Sayıları	Puanlayıcı Sayıları					
	3		4		5	
	G	Φ	G	Φ	G	Φ
10	**.892**	**.866**	.916	.889	.930	.903
12	.908	.885	**-**	**-**	-	-
14	.920	.899	-	-	-	-
16	.930	.910	-	-	-	-

Tablo 15 incelendiğinde; iki yüzeyli çaprazlanmış desenle yapılan ölçme sonuçlarına göre 10 madde ve üç puanlayıcıya göre kestirilen G katsayısı .892, Φ katsayısı da .866 olarak kestirilmiştir. Sonuçlara göre G katsayısı Φ katsayısından daha yüksektir.

Madde sayısının iki arttırılması ile 12 madde üzerinden kestirilen katsayılar incelendiğinde; puanlayıcı sayısının üç olması durumunda G katsayısı .908 ve Φ katsayısı .885 çıkmıştır. Madde sayısının dört arttırılması ile 14 madde üzerinden puanlayıcı sayısının üç olması durumunda G katsayısı .920 ve Φ katsayısı .899 olarak kestirilmiştir. Madde sayısının altı arttırılması ile 16 madde üzerinden puanlayıcı sayısının üç olması durumunda G katsayısı .930 ve Φ katsayısı .910 olarak kestirilmiştir.

Madde sayısı 10 iken puanlayıcı sayısının dört olması durumunda G katsayısı .916 ve Φ katsayısı .889 olarak, puanlayıcı sayısının beş olması durumunda G katsayısı .930 ve Φ katsayısı .903 olarak kestirilmiştir. Tablo 15' e göre 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 yüksektir. 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 değerler almıştır.

Sonuçlar ve Tartışma

Bu araştırma, ortaokul altıncı sınıf düzeyindeki öğrencilerin performanslarının belirlenmesinde kullanılan farklı ölçme araçlarından alınan puanlar incelenerek yapılmıştır. Çalışmada G Kuramı kullanılarak analizler yapılmıştır.

Araştırma bulgularına göre, bireylerin çoktan seçmeli testten, yapılandırılmış grid ve performans görevinden aldıkları notların paralellik gösterdiği gözlenmiş olup en yüksek notların performans görevinden alındığı görülmüştür. Standart sapma değerlerinin değişimine baktığımızda yapılandırılmış grid, performans görevi ve çoktan seçmeli test için her bir puanlayıcının vermiş olduğu puanlar kendi içerisinde birbirlerine yakın bulunmuştur. Fakat performans görevinin standart sapma değeri yapılandırılmış gridin ve çoktan seçmeli testin standart sapma değerinden daha yüksektir.

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

Çoktan seçmeli test, yapılandırılmış grid ve performans görevi sırasında aritmetik ortalama, ortanca ve standart sapma değerleri giderek artış göstermiştir. Buna göre başarı puanları açısından en yüksek puan performans görevi olurken sonrasında yapılandırılmış grid ve çoktan seçmeli test gelmektedir.

Alanda yapılan çalışmalar incelendiğinde farklı sınav türlerinin karşılaştırıldığı ve üzerinde G kuramı çalışması yapılan araştırmalara fazla rastlanmamıştır. Genellikle performans değerlendirilmesinde puanlayıcıların birbirleri ile tutarlılığının incelendiği ve farklı desenlerin birbiriyle karşılaştırmalarının yapıldığı araştırmalar bulunmaktadır.

Çoktan seçmeli test, yapılandırılmış grid ve performans görevinde kullanılan kontrol listesinde madde sayısının artması sonucu güvenirlik değerinin arttığı bulgularda gözlenmiştir. Puanlayıcı ve madde sayısının artması araştırmanın güvenirliği açısından önemli bir özelliktir. Klasik bir ölçme aracı olmasına rağmen puanlanmasının, kapsam geçerliliğin sağlanmasının kolay olması nedeniyle en çok tercih edilen ölçme aracı olmaya devam edeceği görülmektedir.

Performans görevi bireysel farklılıkları göz önüne seren en etkili ölçme aracı olarak bulunmuştur. Performans görevi öğrencilerin yaratıcılıklarını geliştirmesine fırsat sunmaktadır. Bu yönüyle okullarda performans görevi uygulamasından vazgeçilmiş olmasının doğru bir karar olmadığı söylenebilir.

Yapılandırılmış grid uygulamasında ve puanlanmasında hala sorunlar olduğu söylenebilir. Puanlamanın çok zaman alması öğretmenlerin tercih etmediği bir ölçme aracı olmasına neden olmaktadır. Değerlendirilmesi öğretmenlerin gözünü korkutmakta ve güvenilir sonuçlar elde etmemize engel olabilmektedir.

Çoktan seçmeli testler kullanım kolaylığı ve güvenirlikleri nedeniyle en çok tercih edilen ölçme aracı olma özelliğini korumaktadır. Her üç sınav türü içinde güvenirlik çalışması yapılmış ve güvenirlik G Kuramı ile Genellenebilirlik katsayısı hesaplanmıştır. Araştırma sonucunda hesaplanan değerlere göre yapılandırılmış grid G katsayıları oldukça yüksek bulunmuştur.

Kaynaklar

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.

Aktaş, M. (2013). Aynı performans görevinin farklı sayıda puanlayıcılar tarafından üç farklı teknikle puanlanmasından elde edilen puanların güvenirliklerinin Genellenebilirlik Kuramına göre incelenmesi. Yayımlanmamış Yüksek Lisans Tezi, Mersin Üniversitesi, Mersin.

Başol, G. (2008). *Bilimsel araştırma süreci ve yöntem*. İçinde Kılıç, O. veCinoğlu M. (Ed.), *Bilimsel araştırma yöntemleri*, Bölüm 5, İstanbul: LisansYayıncılık.

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*. USA: Springer-Verlag New York Inc.

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

Büyükturan, E., Demirtaşlı N.(2012). Çoktan seçmeli testler ile yapılandırılmış gridlerin, psikometrik özellikler açısından karşılaştırılması., *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *4588*(1), 395-415.

Crocker, L., &Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Harcourt Brace Javanovich College Publishers, USA.

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

Doğan, S. (2012). Kavram haritası ve yapılandırılmış grid tekniğinin çoktan seçmeli testlerle karşılaştırılması. Yayımlanmamış Yüksek Lisans Tezi. Mersin Üniversitesi, Mersin.

Eason, S. H. (1989). Why Generalizability Theory yields better results than Classical Test Theory. *Mid-South Educational Research Association Annual Meeting:* 8-10 November 1989- Little Rock, AR.

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

Eroğlu, M. G. (2010). Kavram haritası ve yapılandırılmış grid ile elde edilen puanların geçerlilik ve güvenirliklerinin incelenmesi. Yayımlanmamış Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara.

Güler, N. (2008). Klasik Test Kuramı, Genellenebilirlik Kuramı ve Rasch Modeli üzerine bir araştırma. Doktora Tezi, Hacettepe Üniversitesi, Ankara.

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

Karasar, N. (1998). *Araştırmalarda rapor hazırlama yöntemi.* Ankara: Pars Matbaacılık.

Kieffer, K. M. (1998). *Why Generalizability Theory is essential and Classical Test Theory often inadequate?* Paper presented at the annual meeting of the Southwestern Psychological Association. New Orleans, LA. USA.

Lord, F. M. &Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Welsley Publishing Company.

McMillian, J. H. (1997). Classroom assessment: Principles and practice for effective instruction. Needham Heights, MA: Allyn and Bacon.

Nalbantoğlu, F. Gelbal S. (2011). Iletişim becerileri istasyonu örneğinde Genellenebilirlik Kuramıyla farklı desenlerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi.* 41:509-518.

Oosterhof, A. (1999). *Developing and using classroom assessments.* Upper Saddle River, NJ: Prentice Hall.

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

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

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

IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Thompson, B. (2003). Score reliability: Contemporary thinking on reliability issues. Sage Publications.

Extended English Summary

Introduction

Learning theories lead to development of various teaching models. Every new challenge proposes new ways of understanding and affects other parts of education system as well. This has led to some changes in understanding and evaluation, as well. Certain decisions, e.g. pass/fail, are made according to the performance of a student. In order to make accurate and on-the-spot decisions, continuous and reliable measurements are needed. Many factors such as measuring instrument, student, scorer, timing, motivation, and environment affect student scores. For a reliable measurement process, all sources of error that can occur must be minimized.

Apart from traditional measurement assessment tools in education, alternative assessment tools have also started to be used in many ways. How effective is the use of different measurement tools in determining student capacity? Does each assessment tool give reliable results? Which instrument gives more reliable results? The answer to these questions is important in this regard.

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

In the case of reliability, in addition to the Classical Test Theory, the Generalizability Theory (G Theory) has begun to be used in recent years. It is based on the analysis of variance (ANOVA). The Generalizability Theory is a theory consisting of a conceptual framework and method that enables many sources of error that may arise during the measurement process.

Generalizability (G) Theory is a "modern" measurement theory. It has superior properties compared to the Classical Test Theory. If a test is only applied once and is scored by only one scorer, it is not possible to obtain completely reliable results. It is unlikely that this score will match with the individual's overall score, on different test forms, and with scores from multiple scorers. Individuals can have different scores from different scorers and different test forms in different situations. The answer to the question "What is the most important source of this variability or mistakes?" can be explained by the Generalizability Theory.

The concept of reliability in Generalizability Theory can be applied to any simple or complex project that researcher wants to generalize. In Generalizability Theory, sources of error, that are called facet (sources of variability), are scorer, item or time.

In Turkey, the studies on Generalizability Theory usually focus on performance measurement processes, scorers and classical measurement tools. Considering the studies on the field, the Theory of Generalizability allows us to examine the effects of many sources of error all together. With this aspect, Generalizability Theory has emerged as a preferred theory in reliability studies. It is compared with the Classical Test Theory in many researches. In this study, it is aimed to examine the consistency of the scores obtained from different measurement tools with the Generalizability Theory.

Method

The main purpose of the research is to compare the results of the G and K studies of a multiple choice test, a structured grid, and a performance task by the use of the Generalizability Theory. The research is important because it aims to reveal the reliability of several measurement instruments through the Generalizability Theory. With this work, it would be possible to show whether the errors involved in the scores obtained from different instruments on the same topic would differ or not, and it could also be possible to decide which measuring instrument is more reliable. Our study is important because there was no research comparing the reliability of the instruments used in the current study through the Generalizability Theory.

The population is composed of middle school student, sixth graders who were studying in the Mediterranean region of Turkey during the 2014-2015 academic years. The sample of the research consisted of 40 students, 23 girls and 17 boys, randomly selected from two schools located in Antalya, a city located in the Mediterranean region. In the study, a multiple-choice test, a structured grid and performance task, all on the same topic, developed by the researchers, were applied to the students. Three scorers took part in the evaluation of the structured grid and performance task. At the end, G and K studies were carried out on the measurement results.

Results and Discussion

In the study, reliability studies were performed for all three assessment tools and reliability indices were calculated for Classic Test Theory and G Theorem. According to the results, Cronbach's Alpha (α) and G coefficients were very high and close to each other. Kuder-Richardson 20 reliability was .68, The G coefficient was .68 and the Phi (Φ) coefficient was .64. α value, the G coefficient, and the Phi (Φ) coefficient for the structured grid was .96, .91 and is .83, respectively. When the performance task results were examined, α value found as .89, G coefficient was .89, and Phi (Φ) coefficient was .86. According to the results of the decision studies, it was seen that the variation of the G and Phi coefficients for all three measuring instruments were parallel.

Başol, G., & Şevran, B. (2019). Bir genellenebilirlik kuramı analizi: Değerlendirmede kullanılan ölçme araçlarının güvenirliği "Haydi Elektriği İletelim". *Journal of Human Sciences*, *16*(1), 370-391. doi:10.14687/jhs.v16i1.5590

The scorers evaluated the structured grid and performance task and the correlation between the scorers was examined. According to the results, their scores were consistent with each other. As the number of items in the multiple choice test increased, so did the reliability of the result. Also as more criterions in the control list included, we found that more reliable scores were obtained. In line with the literature, we found that increasing the number of points and items is an important feature in terms of the reliability of the research.

In our research, performance task was the most effective means of measuring individual differences. On the other hand we know that Turkish Ministry of Education abandoned the use of performance tasks on the base that parents were doing the job. We surely say that performance task is a reliable and effective way helping students practice what they learned through the class. Therefore, we suggest its use more often in Science education.

We found that there were problems with the application and scoring of structured grid. The confusion in the scoring seems to frighten teachers and prevent them from using it. As a result, multiple-choice tests remain the most preferred measurement tool because of their ease of use and reliability.