Eğitim ve Bilim
2011, Cilt 36, Sayı 162

Education and Science
2011, Vol. 36, No 162

Rasgele Veriler Üzerinde Genellenebilirlik Kuramı ve Klasik Test Kuramı'na Göre Güvenirliğin Karşılaştırılması

The Comparison of Reliability According to Generalizability Theory and Classical Test Theory on Random Data

Neşe GÜLER*

Sakarya Üniversitesi

Öz

Bu çalışmada Genellenebilirlik Kuramı (GK) ile Klasik Test Kuramının (KTK) benzer ve farklı yönleri açıklanarak verilen farklı bir örnek ile tasvir edilmeye çalışılmıştır. Daha önce yapılan çalışmalardan farklı olarak burada yer alan örnek, 125 öğrencinin 18 maddeye verdiği cevapları 4 farklı puanlayıcı puanlamış düşüncesiyle tamamen rasgele oluşturulmuş verilere dayanmaktadır. Oluşturulan bu rasgele verinin GK ve KTK'ya dayalı güvenirlik sonuçları hesaplanarak tartışılmıştır. Değişkenlik kaynağının maddeler olduğu tek değişken kaynaklı çapraz desen (b x m) için hesaplanan G katsayısı ile Cronbach α değerleri her bir puanlayıcı için ayrı ayrı hesaplanmış ve çok düşük değerler elde edilmiştir. Değişkenlik kaynağının maddeler ve puanlayıcılar olduğu tümüyle çapraz desen (b x m x p) için GK'ya dayalı G-katsayısı ve Φ-katsayısı sırasıyla .457 ve .456 olarak hesaplanmıştır.

Anahtar Sözcükler: Genellenebilirlik Kuramı, genellenebilirlik katsayısı, phi katsayısı, güvenirlik.

Abstract

In this study, a framework was tried to be presented by focusing on the similarities and differences of generalizability theory (GT) and classical test theory (CTT). Although this example consists of 125 students, 18 items, and 4 raters, all data was obtained completely in a random way. This completely random data which was different than any data previously used in the research available was used to assess and compare the GT and CTT. The results reflects that the Cronbach's α and G-coefficients are very low and the same for a single facet design (s x i) and G-coefficient and Φ coefficient were obtained as .457 and.456, respectively for two facet design (s x i x r) which was not examined by CTT.

Keywords: Generalizability theory, classical test theory, generalizability coefficient, phi coefficient, reliability

Summary

Purpose: Like other disciplines, taking right decisions depends on reliable and valid measurement results. For this reason, studying reliability and validity of the measurement tools is one of the most important subjects for researchers. When the results are obtained from any measurement procedure, it is inevitable that there is more or less some sort of a measurement error. Also, often times, there is more than one source which cause the error. Some early performance based and behavioral studies argued that GT is more appropriate for examining reliability and validity than CTT by eliminating traditional discrepancy between reliability and validity in CTT (Volpe, McConaughy and Hantze, 2009; Güler, 2009). In GT studies, it is possible to indicate how validly a measurement can be interpreted as a representative of a certain set of possible measures (Bergeron and others, 2008). Thus, when there is any measurement procedure which involves large number of variations, such as raters, items, occasions and others, GT occurs as a preferable

* Yrd. Doç. Dr. Neşe GÜLER, Sakarya Üniversitesi, Eğitim Fakültesi, Ölçme ve Değerlendirme ABD gnguler@gmail.com

theory for researchers. In the literature, most generalizability studies are related to investigating the reliability of the real data sets obtained from different measurement tools (i.e. multiple choice item test, performance assessment etc.). In this study, reliability of the data which is randomly produced both by CTT and GT are compared. In other words, similarities and differences in both theories are aimed to be detected.

Method: In this study, the tables which includes 18x4 cells (18 lines and 4 coloums) were handed out to 125 students. The students were asked to fill in the tables with the numbers from 0 to 5 randomly. Each table represented students' (p) grades which were obtained from 4 different raters' (r) grades to 125 students' responses to the 18 items (i). Thus, fully crossed design p x i x r was used in this study. Because the students were considered as the objects of measurement in this study, students' variations from other possible variation sources such as the raters, items and their interactions are kept separately. First, according to the classical test theory, internal consistency values (Cronbach's α) were calculated for each rater separately. Corresponding to this, for each rater variance components and G-coefficients were found for single-facet in which there is only item facet fully crossed design (p x i). Second, according to GT variance components were estimated and G and Φ coefficients were calculated for two-facet in which there are item and rater facets fully crossed design (p x i x r). The statistical analyses were made by using the SPSS 16.0 and the SPSS syntax which was developed by Musquash ve O'Connor (2006).

Results: According to CTT and GT, for each rater, the Cronbach's α values and G- coefficients of scores which were obtained from 125 students' response to 18 items were calculated and G-coefficients and the internal consistency coefficients were found as exactly the same. For rater 1, 2, 3 and 4 these coefficients were -0.216, -0.041, -0.237, -0.372, respectively. At the second step, according to two facets fully crossed design (p x i x r) G-study the variance values were calculated. Relating to fully crossed design G-coefficients was found as .457 and Φ coefficient was found as .456.

Conclusion and Discussion: This study illustrated the similarities and differences of GT and CTT. Although both theories gave the same results for single facet design, GT is superior to classical test theory since there is more than one facet design which has multiple sources of measurement error variance and their interactions. So, because GT is easy to understand and calculate it is more preferable especially for single facet design. However, when there are more than one facet GT cannot give information about multiple sources of errors simultaneously at a time. Thus, because GT considers multiple sources of measurement error variance and their interaction effects clearly, it is suggested that we use it when there are more than one facet. Also, like the previous studies, both theories give same results for reliability when there is only one source of error. Unlike the previous studies, in this study, both theories revealed unreliability of the completely random data.

Giriş

Diğer bilim dallarında da olduğu gibi psikoloji ve eğitim bilimlerinde yapılan araştırmalarda ölçme sonuçlarının güvenilirliği en önemli konular arasında yer almaktadır. Yapılan bir ölçmeden elde edilen sonuçlarda az ya da çok miktarda hata bulunması kaçınılmazdır. Hatta bu hataya sebep olan birden fazla kaynak da bulunabilmektedir. Bir cetvelle bir uzunluğun ölçüldüğü bir durumda, elde edilen ölçme sonucunda cetvelden, ölçmeyi yapan bireyden, ölçmenin yapıldığı ortamın ses, ışık vb. koşullarından ortaya çıkabilecek hata/hataların bulunması olasıdır. Mümkün olan tüm cetvellerle tüm bireylerin aynı uzunluğu ölçerek elde edeceği sonuçların ortalaması, bu uzunluğun "ideal" ölçüsü olarak kestirilebilir (Brennan, 1992). Bu tür bir ölçme durumunda, ölçme sonuçlarındaki farklılığı oluşturan kaynağın cetvelden değil de bireylerden kaynaklandığı tahmin ediliyorsa, ölçme sonuçlarının ortalamasının "ideal"den uzaklaşmasına sebep olacak hata kaynağı cetvelden çok bireyler olacaktır.

Yukarıda verilen basit bir ölçme durumundan çok daha karmaşık olan davranış bilimlerine ilişkin ölçme sonuçlarının olabildiğince tesadüfi hatalardan arınık,bir başka deyişle güvenilir

olmasını sağlamak da bir o kadar karmaşık olmaktadır. Literatüre bakıldığında ölçme sonuçlarının güvenirliğinin çalışıldığı başlıca üç kuram bulunmaktadır: 1. Klasik Test Kuramı, 2. Genellenebilirlik Kuramı ve 3. Madde Tepki Kuramı. Son iki kuram özellikle son yıllarda pek çok araştırmacı tarafından önerilmesine rağmen halen Klasik Test Kuramı (KTK) en çok kullanılan kuram olma unvanını kaybetmemştir (Musquash ve O'Connor, 2006). Diğer kuramların dayandığı matematiksel ifadelerin anlaşılmasının daha güç olması ve kullanımlarındaki karmaşıklık KTK'nın neden daha çok tercih edildiğine bir cevap olmaktadır. KTK'da yer alan temel varsayımlardan biri olan gözlenen puanın (X), gerçek puan (T) ve hata puanından (E) oluşması (X=T + E) denkleminden de anlaşılacağı üzere ölçme sonuçlarına karışan birden fazla hata kaynağı KTK'da aynı anda tek bir çalışmayla ele alınamaz (Güler, 2009; Baykul, 2000). KTK ile karşılaştırıldığında Genellenebilirlik Kuramı (GK); 1. Test-tekrar test güvenirliği, iç tutarlılık ve puanlayıcılar arası güvenirlik gibi farklı hata kaynaklarına dayalı hesaplanan güvenirlik katsayılarının tek bir çalışmayla aynı anda hesaplanmasını sağlar, 2. Sadece her bir hata kaynağı değil hata kaynaklarının etkileşimlerinin etkisinin de kestirilebilmesine imkan tanır, 3. Farklı karar (K) çalışmaları ile istenilen güvenirliğe sahip olunabilecek değişkenlik kaynaklarının düzeylerinin belirlenmesine yardımcı olur (örneğin; yapılan çalışmada kaç madde kullanılsaydı ya da kaç puanlayıcı kullanılsaydı istenilen güvenirliğe ulaşılabilirdi vb.), 4. Sadece göreceli kararlar için değil mutlak kararlar için de güvenirliğin kestirilmesine olanak tanır (Shavelson ve Webb, 1991; Yin ve Shavelson, 2008). Bunlara ek olarak, performansın ölçülmesine dayalı ve davranış bilimlerinde yapılan çalışmalar, KTK'daki güvenirlik ve geçerlik arasındaki farklılığı GK'nın ortadan kaldırarak tek bir çalışma ile her ikisinin birlikte incelenebildiğini ortaya koymuştur (Güler, 2009; Volpe, McConaughy ve Hintze, 2009). GK'ya dayalı çalışmalarda evreni temsil eden eldeki veri setinin olası diğer ölçme durumlarına geçerli olarak nasıl genellenebileceğini yorumlamak mümkündür (Bergeron ve diğerleri, 2008). Tüm bunlar göz önüne alındığında özellikle birden fazla değişkenlik kaynağının etkin olduğu çok sayıda puanlayıcının kullanıldığı ya da birden fazla kez ölçmenin yapıldığı durumlarda GK, KTK'na tercih edilebilir bir kuram olarak karşımıza çıkmaktadır.

Genellenebilirlik Kuramı'na Genel Bir Bakış

Genellenebilirlik Kuramı bir grup bireyden - hatta bazen sadece tek bir bireyden (Lei, Smith ve Suen, 2007) - elde edilen ölçme sonuçlarının, bu sonuçların elde edildiği belirli sayıdaki maddelerin , puanlayıcıların ya da durumların çok daha ötesine genellenebilmesi amacını taşır (Brennan, 1992; Shavelson ve Webb, 1991). Belirli bir örneklemden elde edilen verilerin çok daha geniş olası gözlemlerin elde edilebileceği evrende yapılabilmesine ve böylelikle güvenirlik ya da genellenebilirlik başlığı altında yorumlanabilmesine imkan tanır (Kane, 2002). Yapılan çalışmada yer alan durumun ötesinde olası tüm gözlem koşullarının ve değişkenlik kaynaklarının yer aldığı kabul edilebilir gözlemlerin bulunduğu kapsam bütününe GK'da "evren" (universe) adı verilir. Böylece GK, evrene ilişkin doğru kestirimlerde bulunulması durumunda güvenilir sonuçlara ulaşılabileceğini ifade ederek güvenirlik ile geçerlik arasındaki geleneksel farklılığı da ortadan kaldırmış olur (Güler, 2009; Allal ve Cardinet, 1997). GK'da ölçme sürecinde yer alan maddeler, puanlayıcılar ya da ölçme durumları vb.nin her birine değişkenlik kaynağı (facet) adı verilir. Bu değişkenlik kaynaklarının sayısı da değişkenlik kaynaklarının "düzey"i (level) olarak ifade edilir. Her bir değişkenlik kaynağının düzeyi sonsuz büyüklükte olabilir. Ölçme durumlarında asıl ilgilenilen değişkenliği ortaya çıkaran bireyler, öğrenciler vb. ise değişkenlik kaynağı olarak değil de gerçek, sistematik değişkenliği oluşturan "ölçme objesi" (object of mesaurement) olarak adlandırılır (Musquash ve O'Connor, 2006; Kieffer, 1998). Ancak ölçme objesinin her zaman bireylerden oluşması gibi bir zorunluluk da bulunmamakta; madde, durum vb. değişkenlik kaynakları da çalışmanın doğasına uygun olarak ölçme objesi olabilmektedir (Brennan, 1992). Ölçme objesinin olası tüm ölçme durumlarında elde edilebilecek değerlerinin ortalamasına "evren puanı" adı verilir. Evren puanı, araştırmacının asıl ilgilendiği gerçek değişimi yansıtır ve KTK'daki gerçek puan varyansına benzer yorumlanır (Kieffer, 1998).

GK'da yer alan değişkenlik kaynakları sabit (fixed) ya da tesadüfi (random) olarak ele alınabilir.

Eğer yapılan çalışmada yer alan bir değişkenlik kaynağı olası tüm durumlara genellenmek isteniyorsa, bu değişkenlik kaynağı tesadüfi değişkenlik kaynağı olarak yorumlanabilir. Yukarıda verilen cetvel örneği tekrar ele alınırsa; şayet ölçülmek istenilen uzunluğun olası tüm cetveller ve tüm bireyler tarafından ölçüldüğünde ideal değerinin ne olacağı kestirilmeye çalışılıyorsa, burada ölçme aracı da ölçmeyi yapan bireyler de birer tesadüfi değişkenlik kaynağı olacaktır. Ancak belirli tek bir cetvel var ve istenilen, bu cetvelle olası tüm bireylerin yapacağı ölçme ile uzunluğun ideal değeri kestirilecekse, burada ölçme aracı sabit bir değişkenlik kaynağı, ölçme yapan bireyler ise tesadüfi olarak ifade edilecektir. Kısacası sabit bir değişkenlik kaynağı ölçme yapılan durumla sınırlı kalacak, daha geniş bir evrene genellenmesi istenmeyecektir. Değişkenlik kaynağının sabit olduğu durumda hata kaynağı azalacak, ölçmenin kesinliği artacak, ancak ölçme sonuçlarının genellenmesine ilişkin yorum yapmak zorlaşacaktır (Brennan, 1992).

GK'da, KTK'da olmayan farklı karar verme durumları için farklı güvenirlik katsayıları elde etmek mümkündür. KTK'da elde edilen güvenirlik katsayında olduğu gibi göreceli kararlar için elde edilen genellenebilirlik katsayısının yanı sıra KTK'da dikkate alınamayan mutlak kararlar için güvenirlik katsayısı da hesaplanabilmektedir. Böylece GK, bir bireyin ölçme durumundaki diğer bireylere göre durumunun ne olduğunun dikkate alındığı göreceli kararların yanı sıra daha katı olan bireyin diğer bireylerden bağımsız belirli bir puana (kesme puanı gibi) göre daha yüksek ya da daha düşük puan alıp almadığının yorumlanmasına ilişkin göreceli kararların alındığı durumlarda da güvenirliğin hesaplanmasına olanak tanır (Musquash ve O'Connor, 2006). Hem genellenebilirlik (G) katsayısı hem de Φ katsayısı 0 ile 1 arasında değerler alır. Φ katsayısı G katsayısına göre daha katı bir değerdir.Tek değişkenlik kaynaklı tamamıyla çaprazlanmış desenlerde elde edilen G katsayısı KTK'da yer alan Cronbach α katsayısına benzer yorumlanır (Musquash ve O'Connor, 2006; Sudweeks, Reeve ve Bradshaw, 2005).

GK'da yer alan çalışmalar, çaprazlanmış (crossed) ya da yuvalanmış (nested) desenlerden oluşabilir. Eğer çalışmada yer alan bir değişkenlik kaynağının tüm düzeyleri diğer değişkenlik kaynağının tüm düzeylerinde bulunuyorsa, bu çalışma desenine tümüyle çaprazlanmış desen adı verilir. Örneğin; bir sınıfta yer alan tüm öğrenciler (b) bir testteki tüm maddeleri (m) yanıtlıyor ve tüm öğrencilerin tüm maddeleri aynı puanlayıcılar (p) tarafından puanlanıyorsa, bu desen tümüyle çapraz desen olarak ifade edilir. Çapraz desen "x" sembolü ile gösterilir. Verilen örnekteki çapraz desenin gösterimi "b x m x p" şeklinde olacaktır. Diğer yandan, bir değişkenlik kaynağının bir düzeyi diğerinin sadece tek bir düzeyinde bulunuyor diğerlerinde yer almıyorsa, bu çalışmada da yuvalanmış desen söz konusu olmaktadır. Örneğin, bir yazılı sınavda her bir öğrenci farklı bir maddeyi (m) cevaplıyor ve her öğrencinin (b) cevabı farklı bir puanlayıcı (p) tarafından değerlendiriliyorsa, bu çalışmada yuvalanmış desen kullanılmış demektir. Yuvalanmış desen ":" sembolü ile gösterilir. Verilen örnekteki yuvalanmış desenin gösterimi "b : m : p" şeklindedir. Bazı çalışmalrda ise hem çapraz hem yuvalanmış desen bir arada bulunmaktadır ki bu tür desenlere karışık (mixed) desen adı verilir (Brennan, 1992; Shavelson ve Webb, 1991). Burada ifade edilen tüm bu desenlerde yapılan çalışmalarda GK kullanılabiliyor olsa da tüm değişkenlik kaynaklarına ilişkin kestirimlerde bulunabilmek adına, mümkün olduğu durumlarda tümüyle çaprazlanmış desenlerin kullanılması GK çalışmalarında bir avantaj sağlamaktadır (Kieffer, 1998).

GK'da güvenirliğin araştırılmasında iki çalışma yer almaktadır: 1. Genellenebilirlik çalışması (G-çalışması), 2. Karar çalışması (K-çalışması) . G çalışması tüm değişkenlik kaynaklarına ilişkin aynı anda ve birlikte ANOVA yöntemi ile kestirilmlerde bulunabilmeyi sağlar (Güler, 2009; Atılgan, 2005). G-çalışmasıyla kestirilen değerler sonraki D-çalışması aşamasında kullanılır. G-çalışmasından elde edilen sonuçlar kullanılarak K-çalışmasıyla belirli amaçlar için hatanın en aza indirgenebileceği durumlar kestirilmeye çalışılır. K- çalışmasıyla varılan sonuçlar da araştırmacının madde, puanlayıcı ya da gözlem sayılarını değiştirdiğinde nasıl sonuçlara varabileceği hakkında kestirimlerde bulunmasına yardımcı olur (Volpe, McConaughy ve Hintze, 2009). K- çalışması bir anlamda KTK'da yer alan Spearman Brown formülünün kullanıldığı amaca benzer yorumlanabilir (Musquash ve O'Connor, 2006). Spearman Brown formülüyle, ölçmenin yapıldığı ölçme aracında yer alan madde sayısındaki değişime göre güvenirliğin

kestirimi mümkün olabilmektedir. K-çalışmasında ise bu kestirim sadece madde sayısıyla kısıtlı kalmayıp, aynı anda tek bir çalışma ile ölçme durumunda yer alan tüm değişkenlik kaynaklarının düzeylerinin farklılaşması durumunda güvenirliğin yani genellenebilirlik ve Φ katsayısının alabileceği değerlerin kestirilmesine imkân tanır. Böylece K-çalışmaları en etkili ölçme durumlarının ve güvenirliğin kestirilmesine yardımcı olur (Lee ve Fitzpatrick, 2003).

Literatürde yer alan Genellenebilirlik Kuramı'na ilişkin pek çok çalışma farklı ölçme araçlarının (çoktan seçmeli testler, performans değerlendirme vb.) güvenirliklerini ya da genellenebilirliklerini belirlemek amacıyla yapılmıştır (Atılgan, 2008,2005; Yelboğa ve Tavşancıl, 2010; Yelboğa, 2008; Güler ve Gelbal, 2010; Güler, 2009; Volpe, McConaughy ve Hintze, 2009). Tüm bilimsel araştırmalarda kullanılan ölçme araçlarının güvenilirliğinin yüksek olması istenir. Kullanılan ölçme araçları da olabildiğince yüksek düzeyde güvenilir olacak şekilde geliştirilmeye çalışılır. Bu sebepledir ki literatürde, Klasik Test kuramına göre Cronbach Alfa katsayısı ve Genellenebilirlik Kuramı'na göre G katsayısı olabildiğince güvenilir elde edilmeye çalışılan ölçme sonuçları üzerinde karşılaştırılmıştır ve bu katsayıların eşit ya da eşite yakın değerler verdiği gözlenmiştir. Bu durumda akla şöyle bir soru gelebilir: 'Eğer elde edilen ölçme sonuçları gerçek puandan çok ölçme hatası içeriyorsa, bir başka deyişle güvenirliği oldukça düşükse bu veriler üzerinden Klasik Test Kuram'ına ve Genellenebilirlik Kuramı'na göre elde edilecek güvenirlik katsayıları da birbirine eşit olacak mıdır?' Bu çalışmanın amacı bu soruyu cevaplayacak şekilde, tamamıyla gelişigüzel türetilmiş, ölçme hatasının yüksek, güvenirliğin düşük olduğu verilerin güvenirlik düzeylerini Klasik Test Kuramı ve Genellenebilirlik Kuramı'na göre araştırarak, elde edilen sonuçları karşılaştırmaktır. Başka bir deyişle, olumsuz bir durumda her iki kuramın bu olumsuzluğu ortaya çıkarmadaki benzer ve farklı yönlerini ortaya koyabilmek amaçlanmıştır.

Yöntem

Bu çalışmada 125 öğrenciye, 18 satır ve 4 sütun olmak üzere, 18 x 4 hücreden oluşan tablolalar dağıtılmıştır. Öğrencilerden tüm hücreleri 0 ile 5 arasındaki sayılarla tamamen rasgele olarak doldurmaları istenmiştir. Böylece her bir öğrencinin tablosu, o öğrencinin 18 maddeye verdiği cevapların 4 farklı puanlayıcı tarafından puanlanmış olduğu varsayılarak SPSS'e girilmiştir. Böylece 125 öğrenci (b) 18 madde (m) ve 4 puanlayıcının (p) yer aldığı tümüyle çapraz desen (b x m x p) kullanılmıştır. Yapılan çalışmaların genelinde olduğu gibi bu çalışmada da ölçmenin hedefi "öğrenciler" olarak düşünüldüğünden öğrenciler ölçme objesi olarak kabul edilmiştir. Böylece öğrencilerden kaynaklanan değişim, ölçmenin olası hata kaynaklarından (maddeler, puanlayıcılar ve bunlar arası etkileşim) ayrı tutulmuştur.

Çalışmada her bir puanlayıcıdan elde edilen puanların ayrı ayrı Klasik Test Kuramı'na göre değerleri hesaplanmıştır. Buna karşılık olarak da yine her bir puanlayıcıdan elde edilen puanlar için tek değişkenlik boyutun maddeler olduğu b x m çapraz desene göre G katsayıları hesaplanmış ve Cronbach α değerleri ile karşılaştırılmıştır.

Ayrıca Genellenebilirlik Kuramı'na göre iki değişkenlik boyutunun birlikte yer aldığı (maddeler ve puanlayıcılar) tümüyle çapraz desende (b x m x o) her bir değişkenlik kaynağının ayrı ayrı ve birbirleriyle etkileşimlerinin varyansları hesaplanmış, G ve Φ katsayıları bulunmuştur. Verilerin analizinde ve Genellenebilirlik Kuramı için Musquash ve O'Connor (2006) tarafından SPSS için oluşturulmuş yazılım kullanılmıştır.

Bulgular

Genellenebilirlik Kuramı'na göre 125 öğrencinin 18 maddeye verdiği cevapların her bir puanlayıcı tarafından puanlandığı varsayımına dayalı oluşturulan rasgele verilerin tek değişkenlik kaynağına göre (maddeler) b x m çapraz desenine göre elde edilen varyans bileşenlerinin kestirimi Tablo 1'de verilmiştir.

Tablo 1.

Tek Değişkenlik Kaynaklı b x m Çapraz Desenine Göre ANOVA Tablosu Ortalama Kare (Mean Square) Değerleri

Varyans Kay.	Sd	1 Puanlayıcı	2 Puanlayıcı	3 Puanlayıcı	4 Puanlayıcı
b	124	2.550	2.628	2.344	2.070
m	17	2.328	2.704	2.030	3.514
b x m	2108	3.101	2.735	2.900	2.840

Tablo 1'deki değerler kullanılarak varyans bileşenleri aşağıda verilen formüller (Brennan, 1992) yardımıyla kestirilerek Tablo 2'de yer alan değerler elde edilmiştir.

$$\sigma_b^2 = \frac{MS_b - MS_{bm}}{n_m}$$

$$\sigma_m^2 = \frac{MS_m - MS_{bm}}{n_b}$$

$$\sigma_{bm}^2 = MS_{bm}$$

Tablo 2.

Tek Değişkenlik Kaynaklı b x m Çapraz Desenine Kesitirilen G Çalışması Varyans Değerleri

Varyans Kay.	Sd	1 Puanlayıcı	2 Puanlayıcı	3 Puanlayıcı	4 Puanlayıcı
b	124	-.03061	-.0059	-.03089	-.0428
m	17	-.006184	-.000248	-.00696	.00539
b x m	2108	3.101	2.735	2.900	2.840

Yukarıdaki tablo değerleri;

Hata varyansı: $\sigma^2(\delta) = \frac{\sigma_{bm}^2}{n_m}$

Genellenebilirlik katsayısı: $E_b^2 = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2(\delta)}$

Yukarıdaki formüller kullanılarak Tablo 3'te yer alan her bir puanlayıcı için tek değişkenlik kaynaklı bxm çapraz deseninde kestirilen genellenebilirlik katsayılarına ulaşılmıştır. Tablo 3'te aynı zamanda çalışmada var olduğu düşünülen dört ayrı puanlayıcının her biri için 125 öğrencinin 18 maddesine verdiği puanlar üzerinden elde edilen Cronbach α değerleri de yer almaktadır.

Tablo 3.

Her Bir Puanlayıcının puanlarına İlişkin Klasik Test Kuramı'na Göre Güvenirlik Katsayıları ve Genellenebilirlik Katsayıları

	1. Puanlayıcı	2. Puanlayıcı	3. Puanlayıcı	4. Puanlayıcı
Genel.kats.	*-0.216*	*-0.041*	*-0.237*	*-0.372*
Cronbach α	*-0.216*	*-0.041*	*-0.237*	*-0.372*

Tablo 3'ten görüleceği üzere, tek değişkenlik kaynaklı tümüyle çapraz desen için kestirilen genellenebilirlik katsayısı değerleri Klasik Test Kuramı'na göre elde edilen Cronbach α değerleri ile tamamıyla birbirine eşit çıkmıştır. Böylece alanyazında yer aldığı gibi bu iki katsayının eşitliği, tamamıyla rasgele elde edilmiş veriler için de sağlanmıştır.

Çalışmanın ikinci aşamasında bxpxm iki değişkenlik kaynaklı tümüyle çapraz desene ilişkin G çalışması yapılarak varyans bileşenleri Tablo 4'te verilen eşitlikler yardımıyla kestirilmiştir.

Tablo 4.

İki Değişkenlik Kaynaklı Tesadüfi Desen İçin Varyans Bileşenlerinin Kestirilmesi

Varyans Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	Kestirilen Varyans Bileşenleri
Öğrenci (b)	SS_b	n_b-1	$MS_b = SS_b / n_b$-1	$\sigma^2_{(b)}$
Puanlayıcı (p)	SS_p	n_p-1	$MS_p = SS_p / n_p$-1	$\sigma^2_{(p)}$
Madde (m)	SS_m	n_m-1	$MS_m = SS_m / n_m$-1	$\sigma^2_{(m)}$
b x p	SS_{bp}	(n_b-1)(n_p-1)	$MS_{bp} = SS_{bp} / n_{bp}$-1	$\sigma^2_{(bp)}$
b x m	SS_{bm}	(n_b-1)(n_m-1)	$MS_{bm} = SS_{bm} / n_{bm}$-1	$\sigma^2_{(bm)}$
p x m	SS_{pm}	(n_p-1)(n_m-1)	$MS_{pm} = SS_{pm} / n_{pm}$-1	$\sigma^2_{(mp)}$
b x p x m, e	$SS_{bpm,e}$	(n_b-1)(n_p-1)(n_m-1)	$MS_{bpm,e} = SS_{bpm,e} / n_{bpm,e}$-1	$\sigma^2_{(bmp)}$

Genellenebilirlik çalışmalarındaki analizlerinin temeli random-etki faktöriyel ANOVA'ya dayalı olmasına rağmen genellenebilirlik kuramının hipotez testiyle bir ilişkisi bulunmadığından F ve p değerleri tabloda yer almamaktadır (Shavelson ve Webb, 1991; Brennan, 2001). Tablo 4'teki eşitliklere ilişkin varyans bileşenlerinin kestirilen değerleri Tablo 5'te verilmiştir.

Tablo 5.

İki Değişkenlik Kaynaklı b x m x p Tümüyle Çapraz Desen İçin Varyans Bileşenleri Kestirim Değerleri

Varyans Kaynağı	Sd	Kareler Ortalaması	Varyans	Varyans Yüzdeler
b	124	4.087	.034	.012
m	17	2.953	.001	.000
p	3	2.235	.000	.000
bxm	2108	2.723	.000	.000
bxp	372	1.835	.000	.000
mxp	51	2.541	.000	.000
bxmxp	6324	2.951	2.951	.988

Tablo 5'in ilk iki sütununda sırasıyla her bir varyans kaynağına ilişkin serbestlik derecesi, kareler ortalaması değerleri bulunmaktadır. Son iki sütunda ise varyans değerleri ile yüzdeleri yer almaktadır. Bu sütundaki değerlerinin çoğunluğunun "0" değeri aldığı görülmektedir. Kestirilen varyans bileşenlerinin negatif değerler alması durumunda, bu değerlerin sıfır olarak alınması önerilmektedir (Shavelson, Webb & Rowley, 1989). Genellenebilirlik çalışmalarında bireylere ilişkin varyansın ne kadar büyük olması isteniyorsa, hata varyansı olarak kabul edilne residual varyansın da olabildiğince küçük olması istenir. Bu değer, öğrencilerin puanları arasındaki farklılığın madde ve puanlayıcılardan kaynaklandığını ifade ettiği gibi çalışma deseninde yer almayan başka faktörlerden meydana gelen değişkenliğin de söz konusu olabileceğini göstermektedir. Bu tabloda yer alan verilerden de anlaşılacağı üzere, Genellenebilirlik Kuramı'nın bir avantajı olarak, araştırmacı toplam varyansın ne kadarının hangi kaynaktan ya da hangi kaynakların etkileşiminden ortaya çıktığını açıkça görebilmektedir. Bu şekildeki ayrıntılı bir bilgiye güvenirlik kestirimindeki diğer yaklaşımlarda rastlanılmamaktadır (Goodwin, 2001).

Her bir değişkenlik kaynağı ve bunlar arasındaki etkileşimin toplam varyanstaki payını belirlemenin yanı sıra, Genellenebilirlik Kuramı'na dayalı çalışmalarda, Klasik Test Kuramı'ndaki güvenirlik katsayısına benzer yorumlanan G katsayısının değerini hesaplamak mümkündür.

Göreceli karar için hesaplanan G katsayısı her bir ölçme objesinin, değişkenlik kaynağındaki aldığı ham puanın ne kadar yüksek olduğu değil, diğer ölçme objelerinin puanlarının sıralaması arasındaki yerine bağlı olarak hesaplanır. Daha önce de belirtildiği gibi, G katsayısı Klasik Test Kuramı'ndaki güvenirlik katsayısı olarak yorumlanır ve 0 ile 1 arasında değer alır. Bu katsayı, G çalışmasında yer alan değişkenlik kaynakları üzerinden puanların genellenebilme ya da güvenirlik düzeyini göstermektedir. Genellenebilirlik Kuramında, Klasik Test Kuramı'ndan farklı olarak bir de mutlak karar için phi (güvenirlik-dependability) katsayısı da hesaplanmaktadır. Phi katsayısı, çok daha katı bir değer olup, hem ölçme objelerinin puanları sıralamasındaki tutarlılığın derecesini hem de ham puanların değerlerine bağlı tutarlılığın derecesini ortaya koyar. Tablo 6'da bu katsayıların hesaplanmasında kullanılan eşitlikler görülmektedir.

Tablo 6.

İki Değişkenlik Kaynaklı Tesadüfi Desen İçin Hata ve Güvenirlik Kestirimleri

Mutlak hata $(\sigma^2_{(\Delta)})$	Bağıl hata $(\sigma^2_{(\delta)})$	Φ Katsayı	G-katsayı
$\frac{\sigma^2_x}{N_x}+\frac{\sigma^2_v}{N_v}+\frac{\sigma^2_{bx}}{N_x}+\frac{\sigma^2_{bv}}{N_v}+\frac{\sigma^2_{xv}}{N_xN_v}+\frac{\sigma^2_{bxv}}{N_xN_v}$	$\frac{\sigma^2_{bx}}{N_x}+\frac{\sigma^2_{bv}}{N_v}+\frac{\sigma^2_{bxv}}{N_xN_v}$	$\frac{\sigma^2_b}{\sigma^2_b+\frac{\sigma^2_x}{N_x}+\frac{\sigma^2_v}{N_v}+\frac{\sigma^2_{bx}}{N_x}+\frac{\sigma^2_{bv}}{N_v}+\frac{\sigma^2_{xv}}{N_xN_v}+\frac{\sigma^2_{bxv}}{N_xN_v}}$	$\frac{\sigma^2_b}{\sigma^2_b+\frac{\sigma^2_{bx}}{N_x}+\frac{\sigma^2_{bv}}{N_v}+\frac{\sigma^2_{bxv}}{N_xN_v}}$

Tablo 6'daki eşitliklerden de görüleceği üzere, mutlak hata bağıl hatadan daha büyük bir değere sahiptir. Buna bağlı olarak phi katsayısı genellenebilirlik katsayısından daha küçük bir değer alır. Buna göre, bxmxr tümüyle çapraz desenine ilişkin elde edilen G katsayısı .457 ve Phi katsayısı .456'dır.

Sonuç

Bu çalışma ile GK ile KTK'nın benzer ve farklı yönleri açıklanarak verilen örnekle de tasvir edilmeye çalışılmıştır. Aynı zamanda GK ile ilgili temelde bilinmesi gerekenler alanyazın çerçevesinde özetlenerek konu hakkında bilgi edinmek isteyenlerin genel bir bakış açısı oluşturabilmesi amaçlanmıştır. Tek değişkenlik kaynaklı desenler için her iki kurama göre hesaplanan güvenirlik değerleri alanyazında da belirtildiği gibi bu çalışmada yer alan tamamen rasgele oluşturulmuş güvenilir olmayan veriler için de aynı değerlerde çıkmıştır. Puanlayıcı, durum vb. gibi birden fazla değişkenlik kaynağının bulunduğu durumlara ilişkin yapılacak güvenirlik çalışmalarında tek bir seferde her bir değişkenlik kaynağı ve bunların etkileşimlerine dayalı güvenirliği belirleyebilmemizi sağlayarak GK'nın KTK'ya olan avantajı vurgulanmıştır. Böylece hem anlaşılmasının hem de hesaplanmasının daha kolay olması sebebiyle tek değişkenlik kaynaklı desenlerde KTK daha tercih edilebilirken, değişkenlik kaynağının fazla olduğu durumda pek çok bilginin bir arada elde edildiği GK önerilebilir. Ayrıca diğer çalışmalardan farklı olarak burada yer alan örnek veri, tamamen gelişi-güzel oluşturulan (ölçme hatasının yüksek, güvenirliğin düşük olmasının beklendiği) ölçme sonuçlarına dayandırılmış ve her iki kurama göre elde edilen değerler verilerin güvenilir olmadığını teyit ederek, negatif bir durum için de alanyazını destekler sonuçlar vermiştir. Çalımada yer alan veriler, belirli sayıda öğrenci, madde ve puanlayıcıdan oluşacak şekilde sınırlı kalmıştır. Bu sınırlılığın giderilebilimesi için, farklı farklı sayıda öğrenci, madde ve puanlayıcıların yer alacağı simulasyon çalışmalarının yapılması önerilebilir.

Teşekkür

Genellenebilirlik Kuramı ile Klasik Test Kuramı'nın karşılaştırılmasını tamamen rastgele oluşturulmuş (güvenilir olmayan) bir veri seti üzerinden örneklendirmemi sağlayan hocam Sayın Prof. Dr. Yaşar BAYKUL'a vermiş olduğu değerli önerisinden dolayı teşekkürlerimi bir borç bilirim.

Kaynakça

Allal, L. ve Cardinet, J. (1997). Generalizability Theory. *Educational Research Methodology and Measurement an International Handbook.* (Second Editon). Editor Keeves, J. P. Cambridge, UK.

Atılgan, H. (2008) Using Generalizability Theory to Assess the Score Reliability of the Special Ability Selection Examinations for Music Education Programs in Higher Education. *International Journal of Research & Method in Education*, 31: 1, 63 - 76.

Atılgan, H. (2005). A Sample Application for Generalizability Theory and Interrater Reliability. *Eğitim Bilimleri ve Uygulama*, 4 (7).

Baykul, Y. (2000). *Measurement in Education and Psychology: Classical Test Theory and Its Application*.

Bergeron, R., Floyd, R. G., McCormack, A. C., & Farmer, L. W. (2008). The generalizability of externalizing behavior composites and subscale scores across time, rater, and instrument. *Journal of Educational Measurement*. 37, 1, 91-108.

Brennan, R. L. (2001). *Generalizability Theory*.Iowa City, Iowa: ACT Publications.

Brennan, R. L. (1992). *Elements of Generalizability Theory*. New York: Springer-Verlog.

Goodwin, L. D. (2001). Inter-rater Agreement and Reliability. *Measurement in Psychical Education and Exercises Science*, 5 (1), 13-14.

Güler, N., & Gelbal, S. (2010).Studying Reliability of Open Ended Mathematics Items According to Classical Test Theory and Generalizability Theory. *Educational Sciences: Theory & Practice*. 989-1019, 10(2), Spring 2010.

Güler, N. (2009). Generalizability Theory and Comparison of the Results of G and D Studies Computed by SPSS and GENOVA Packet Programs. *Education and Science*, volume 34, no 154.

Kane, M. (2002). Inferences About Variance Components and Reliability-Generalizability Coefficients in the Absence of Random Sampling. *Journal of Educational Measurement*. 39, 2, 165-181.

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

Lee, G. & Fitzpatrick, A. R. (2003). The Effects of a Student Sampling Plan on Estimates of the Errors for Students Passing Rates. *Journal of Educational Measurement*. 40, 1, 17-28.

Lei, P., Smith, M. ve Suen, H. K. (2007). The Use of Generalizability Theory to Estimate Data Reliability in Single-Subject Observational Research. *Psychology in the Schools*. 44, 5.

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

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

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

Sudweeks, R. R., Reeve, S. and Bradshaw, W. S. (2005). A Comparison of Generalizability Theory and Many Facet Measurement in An Anlaysis of College Sophomore Writing. *Assessing Writing*. 9, 239-261.

Volpe, R. J., McConaughy, S. H., & Hintze, J. M. (2009). Generalizability of Classroom Behavior Problem and On-Task Scores from the Direct Observation Form. *School Psychology Review*, 38, 3

Yelboğa, A. ve Tavşancıl, E. (2010). Klasik Test ve Genellenebilirlik Kuramı'na Göre Güvenirliğin Bir İş Performansı Ölçeği Üzerinde İncelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri*. *10*(3). 1825-1854.

Yelboğa, A. (2008). Güvenirliğin Değerlendirilmesinde Genellenebilirlik Kuramı'nın Kullanılması: Endüstri ve Örgüt Psikolojisinde Bir Uygulama. *Psikoloji Çalışmaları Dergisi.* 28, 35-54.

Yin, Y. & Shavelson, R. J. (2008). Application of Generalizability Theory to Concept Map Assessment Research. *Applied Measurement in Education*. 21, 273-291.