Eğitim ve Bilim
2009, Cilt 34, Sayı 154

Education and Science
2009, Vol. 34, No 154

Genellenebilirlik Kuramı ve SPSS ile GENOVA Programlarıyla Hesaplanan G ve K Çalışmalarına İlişkin Sonuçların Karşılaştırılması

Generalizability Theory and Comparison of the Results of G and D Studies Computed by SPSS and GENOVA Packet Programs

Neşe GÜLER**
Sakarya Üniversitesi

Öz

Genellenebilirlik kuramında, birden fazla kaynaktan meydana gelen hataların her birinin ve etkileşimlerinin büyüklüklerini aynı anda tek bir analizle kestirmek mümkündür. Genellenebilirlik kuramı klasik test kuramını da kapsayan, onun uzantısı olan bir kuram niteliği taşımaktadır. Genellenebilirlik kuramına ilişkin analizler genellikle GENOVA paket programıyla yapılmıştır. Ancak bu programın kullanımının zor ve karmaşık olması, genellenebilirlik çalışmalarının yapılmasındaki en büyük sınırlılık olarak araştırmacıların karşısına çıkmaktadır. Musquash ve O'Connor (2006), genellenebilirlik kuramına ilişkin tüm analizlerin yapılabileceği bir SPSS programı geliştirmişlerdir. Bu çalışmada, genellenebilirlik kuramına ve terminolojisine ilişkin genel bir bakış açısı oluşturulmaya çalışılmıştır. Ayrıca, genellenebilirlik kuramına bağlı genellenebilirlik (G) ve karar (K) çalışmalarında elde edilen genellenebilirlik ve güvenirlik katsayılarının yukarıda ifade edilen iki farklı paket programıyla elde edilen değerleri bir arada sunulmuştur.

Anahtar Sözcükler: Genellenebilirlik kuramı, GENOVA, genellenebilirlik katsayısı, phi katsayısı, güvenirlik.

Abstract

In generalizability theory, it is possible to estimate one reliability value according to several different sources of measurement error and their interaction into one study. However, in classical test theory, different reliability values can be obtained for situations in more than one source of variance. Generalizability theory is an extension of the classical test theory which considers multiple sources of measurement error simultaneously. Analyses related to generalizability theory were generally done with GENOVA computer packet program. Because using the program is hard and complex to understand, it is a major problem in studies related to generalizability theory. Musquash and O'Connor (2006) described to use SPSS (also SAS and MATLAB) for conducting the analyses about generalizability theory. In this study, generalizability and dependability coefficients for both generalizability (G) study and decision (D) studies were presented by using both GENOVA and SPSS computer packet programs. This study also provides an illustrative example and comparison of the results on both programs.

Keywords: Generalizability theory, GENOVA, generalizability coefficient, phi coefficient, reliability

* Yrd. Doç. Dr. Neşe GÜLER, Sakarya Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Hendek, gngguler@gmail.com

Summary

Purpose: Nunally (1982) said that: "Science is interested in repeatable experiments. If the results of experiments are influenced by random error, the results are not completely repeatable. Thus, science is limited by reliability of the measurement tools and scientists who use these measurement tools." As herein is understood, it is one of the most important necessities to be able to get reliable data for a scientific study to be completed. In classical test theory, in order to measure different reliability coefficients, different sources of variability are calculated. On the other hand, in generalizaibility theory it is possible that researchers obtain only one reliability coefficient when considering multiple sources of variability simultaneously. Research literature about generalizability theory shows that analyses related to generalizability theory were generally done with GENOVA computer packet program. Because using of this program is hard and complex to understand, it is a major problem for doing studies related to generalizability theory. Musquash and O'Connor (2006) described to use SPSS (also SAS and MATLAB) for conducting the analyses about generalizability theory. In this study, based on generalizability theory generalizability and dependability coefficients were presented by using both GENOVA and SPSS computer packet programs. This study also provides an illustrative example and comparison of the results based on both programs.

Method: In this study, 18 open-ended mathematic items were used. Responses were graded by 4 different graders by means of a six-point scale. The participants were 203 8th and 9th grade elementary school students. Reliability of students' grades was analyzed with generalizability theory. Because all students responded to the all of the same items and all of the responds were graded by same 4 graders, the design in the study was considered as a fully crossed design (student x item x rater).The object of measurement was defined as the students and the universe of admissible observations as the items on the test and the four graders. Amount of variance components and their interactions, G and phi coefficients were reported by using GENOVA and SPSS programs. Different D-studies were conducted and for each D-study, G and phi coefficients were calculated according to GENOVA and SPSS programs.

Results: G-coefficient is interpreted as reliability coefficient in the classical test theory and its value is between 0 and 1. This coefficient infers the generalizability or the reliability of the grades over the all facet in the G study. In addition to this, in the generalizability theory, unlike the classical test theory, phi coefficient can be calculated for absolute decisions. For the absolute decisions, absolute error and phi coefficient found by SPSS were 0.133 and 0.897 respectively; by GENOVA were 0.142 and 0.892. For the relative decisions, relative error and G coefficient calculated by SPSS were 0.105 and 0.917 respectively; by GENOVA were 0.109 and 0.914. According the results, the biggest difference between the results obtained by SPSS ve GENOVA programs was 0.009 for the absolute error. The difference for phi coefficients was 0.005 and for G coefficients was 0.003. Also different D-studies were conducted. According to the results of the D-studies, the biggest difference between the values obtained by SPSS and GENOVA programs was 0.005 for the phi coefficient.

Conclusion and Discussion: Generalizability theory is one of the most important extension of the classical test theory and *it should be especially used when there is more than one grading condition* (Nunally ve Bernstein, 1994).

The most important reason why the researchers abstain from studying generalizability theory is that the most popular generalizability theory program GENOVA is a complex and hard to understand program. Musquash and O'Connor (2006) described simple and understandable programs for conducting analyses for encouraging the use of generalizability theory techniques. In addition to information obtained by GENOVA, this program provides graphics as visual information for the absolute and relative errors and the G and phi coefficients. The results of the study also support that this program can be alternative for studying generalizability theory. Furthermore, because in this study only items and graders were considered as facets, it can be suggested that different studies be conducted for different purposes and more facets might be useful for future studies.

Giriş

Nunally'nin (1982) de ifade ettiği gibi, *"Bilim tekrarlanabilir deneylerle ilgilenmektir. Eğer deneylerden elde edilen ölçme sonuçları tesadüfi ölçme hatalarından etkileniyorsa, sonuçlar tamamıyla tekrarlanabilir değildir. Böylece bilim, ölçme araçlarının ve bu ölçme araçlarını kullanan bilim adamlarının güvenirliği ile sınırlı kalmaktadır."*

Bu ifadeden de anlaşılacağı üzere, güvenilir ölçme sonuçları elde etmek bilimin en önemli gerekliliklerinden biri olmaktadır. Yapılacak tüm bilimsel araştırmalarda güvenirliğin belirlenmesi araştırmacılar için oldukça önem taşımaktadır. Güvenirlik nasıl tanımlanır? Ölçmenin doğası gereği ölçme sonuçlarında iki tür varyans (değişim) söz konusu olmaktadır (Kieffer, 1998; Shavelson, Webb ve Rowley, 1989): sistematik ve sistematik olmayan (random) varyans. Sistematik varyans gerçek varyans olup ölçmeden meydana gelen değişimi gösterir ve ölçmenin her tekrarlanışında ortaya çıkabilecek değişimdir. Gerçek varyans olması beklenen, istenen varyanstır. Sistematik olmayan varyans ise, ölçmeden ya da ölçmenin yapıldığı örneklemden kaynaklanan, ölçmenin tekrarlanmasıyla tekrar elde edilemeyecek değişimlerdir. Sistematik olmayan varyans, istenmeyen varyans ya da hata varyansı olarak da tanımlanır. Yapılan ölçmelerde, toplanan ölçme sonuçlarında bu tür hataların olabildiğince azaltılması ya da yok edilmesi istenir.

Güvenirlik kavramı doğrudan, sistematik olmayan bu hata kaynakları ile ilgilidir ve bu bağlamda ölçmede, "ölçme aracının güvenirliği" ifadesi pek doğru bir kullanım olmamaktadır. Bunun yerine "belirli bir ölçme aracıyla elde edilen puanların güvenirliği" ifadesi çok daha yerinde ve doğru olacaktır. Çünkü ölçme aracıyla yapılan bir ölçmeden elde edilen sonuçlar güvenilir olabilirken, aynı araçla yapılan başka bir ölçme sonuçları güvenilir olmayabilir. Böylece günvenirlik en öz ve genel olarak "ölçme sonuçlarının tesadüfi hatalardan arınık olma derecesi" olarak tanımlanabilir (Baykul, 2000).

Klasik test kuramında, güvenirlik gerçek puan varyansının (sistematik varyans), gözlenen puan varyansına oranı olarak ifade edilir. Buradaki gözlenen varyans, gerçek puan varyansı ile hata varyansı (sistematik olmayan varyans) toplamına eşittir. Sonuç olarak hata varyansı ve buna bağlı olarak güvenirlik, elde edilecek ölçmeden ölçmeye değişim gösterecektir. Test tekrar test güvenirliğinde puanların iki zaman arasındaki değişimi hata olarak ele alınırken, madde örnekleminden ortaya çıkabilecek değişim ele alınmaz. İç tutarlılık katsayısı hesaplanırken ise madde örnekleminden dolayı meydana gelen değişim hata kaynağı olarak düşünülürken, bu kez zamana bağlı oluşabilecek değişim hata kaynağı olarak düşünülmez. Buradan da anlaşılacağı üzere, klasik test kuramında güvenirlik, hata kaynağına bağlı olarak farklılık göstermektedir. Yapılan bir ölçmenin sonuçlarına ilişkin farklı değişkenlik kaynaklarına göre farklı güvenirlik katsayıları hesaplanır. Özellikle hesaplanan bu güvenirlik değerleri arasındaki farklılık oldukça fazla olduğunda, hangi güvenirlik değerinin göz önünde bulundurulacağı oldukça karmaşık bir hal alabilir. Örneğin, bir testin test tekrar test güvenirliğinin yüksek, birinci uygulamanın puanlarına ilişkin iç tutarlılık katsayısının oldukça düşük, ikinci uygulamanın puanlarına ilişkin iç tutarlılık katsayısının ise orta değerler aldığını düşünelim. Bu durumda araştırmacı "hangi güvenirliğe göre" yorum yapacağına karar vermekte zorlanacaktır (Kieffer, 1998).

Genellenebilirlik kuramında klasik test kuramının aksine, aynı anda birden fazla hata kaynağı bir arada göz önünde bulundurularak tek bir güvenirlik değerine ulaşmak mümkün olmaktadır. Shavelson ve Webb'e (1991) göre, genellenebilirlik kuramı dört farklı açıdan klasik test kuramının daha genişletilmiş bir halidir: 1. Genellenebilirlik kuramı, çoklu varyans kaynaklarını tek bir analizde ele alır. 2. Her bir varyans kaynağının büyüklüğünün belirlenmesini sağlar. 3. Hem bireylerin performanslarına dayalı bağıl kararlar hem de bireylerin performanslarıyla ilgili mutlak kararlar alınmasına ilişkin iki farklı güvenirlik katsayısının (sırasıyla; G katsayısı ve phi katsayısı) hesaplanmasına olanak tanır. 4. Belirli bir amaca bağlı olarak, ölçme hatasının en aza indirgenebileceği ölçmelerin düzenlenmesine (Karar "K" çalışmaları) imkân tanır.

Genellenebilirlik kuramı, belirli bir ölçme sonucuna ya da gözlenen puana değil ölçme sonuçlarının belirli bir örnekleminden çok daha geniş olan evrenine nasıl genellenebileceğine

odaklanmaktadır. Genellenebilirlik evreni, ölçülmek istenilen özelliğe ilişkin tüm ölçme sonuçlarını kapsar. Bir puanın kullanışlılığı, o puanın genellenebilirlik evreni adı verilen daha geniş durumlara doğru bir şekilde nasıl genellenebileceğine bağlıdır. Bir bireyden elde edilebilecek ideal ölçme sonucu, bireyin, tüm kabul edilebilir gözlemleri üzerinden elde edilecek olan puanlarının ortalamasıdır ve bu değere "evren puanı" adı verilir. (Shavelson, Webb ve Rowley, 1989). Araştırmacının asıl amacı, eldeki örneklemden evrene genelleme yapabilmektir. Klasik test kuramının temel kavramlarından biri olan "güvenirlik" yerine, genellenebilirlik kuramında daha geniş ve esnek bir kavram olan "genellenebilirlik" kullanılmaktadır. Böylece klasik test kuramında sorulan "Gözlenen puanlar ne doğrulukta gerçek puanı yansıtmaktadır?" sorusu yerine genellenebilirlik kuramında "Gözlenen puanlar, bireylere ilişkin davranışların, belirlenen evrendeki tüm durumlara doğru bir şekilde genellenmesine ne şekilde izin verecektir?" sorusuna cevap aramaya dönüşmektedir.

Genellenebilirlik kuramını, varyans analizinin (ANOVA) dayandığı mantığa bağlı olarak açıklamak mümkündü. Bir araştırmacı nasıl ANOVA ile olası önemli bağımsız değişkenlerin büyüklüğünü kestirmek ve belirlemek istiyorsa, genellenebilirlik kuramında da bu kez ölçmedeki olası önemli hata kaynaklarının büyüklüğünü kestirmek ve belirlemek istemektedir. ANOVA çalışmasında araştırmacı, bağımsız değişkenleri manipule edip diğer değişkenlerin bir kısmını kontrol altına alabilir, bir kısmını da göz ardı edebilir. Ölçme durumunda da aynı şekilde ölçmedeki belirli hata kaynaklarını manipule edip diğerlerini kontrol altına almak ya da göz ardı etmek mümkündür (Shavelson, Webb ve Rowley, 1989).

Genellenebilirlik kuramı ile klasik test kuramı arasındaki ilişki faktöriyel ve basit ANOVA arasındaki ilişkiye paralellik göstermektedir. Basit ANOVA'da varyans bileşenleri değişkenler arası (between) ve değişken içi (within) olarak iki bölümde isimlendirilir. Bunlardan ilki, bir gruptan diğerine farklılaşan desen faktörüyle ilişkilendirilebilir ve sistematik varyans olarak düşünülebilir. Diğeri ise, araştırmada asıl odaklanılan gruptaki farklılıktan meydana gelen değişim olup bu değişim tesadüfi (random) hata olarak yorumlanabilir. Benzer şekilde, klasik test kuramında da bir gerçek puan varyansı bir de hata varyansı bulunmaktadır. Buradaki ilk varyans bileşeni ölçülen bireyler arasındaki doğal farklılıktan meydana gelen değişimi gösterip sistematik varyans olarak ele alınırken, diğeri gerçek puan varyansı ile ilişkili olmayan sistematik olmayan varyans olarak düşünülür (Kieffer, 1998).

Basit ANOVA faktöriyel ANOVA'ya göre daha az soruya cevap bulabilir ve daha az etkindir. Basit ANOVA yerine faktöriyel ANOVA'da, çoklu varyans bileşenleri bulunmaktadır. Hem her bir faktör için hem de bu faktörlerin etkileşimleri için ayrı ayrı varyans kaynakları ve bir de hata kaynağı mevcuttur. Benzer şekilde, klasik test kuramının genişletilmiş hali olan genellenebilirlik kuramı da ölçmedeki değişime sebep olan çoklu kaynaklar hakkında bilgi edinmemize olanak tanır. Klasik test kuramında varyans bileşenleri sadece iki kaynaktan elde edilirken, genellenebilirlik kuramında varyans bileşenleri, ölçmenin objesi olan bireylere ilişkin sistematik varyans, çoklu hata kaynakları ve bunlar arasındaki etkileşime karşılık gelen pek çok kaynaktan oluşmaktadır (Kieffer, 1998; Crocker ve Algina, 1986).

Genellenebilirlik kuramında yer alan çoklu hata kaynakları bir örnek üzerinden açıklanabilir. Bir başarı testinin iki farklı puanlayıcı tarafından puanlandığı bir durumda kestirilebilecek hata kaynağı ile aynı testin paralel formlarından elde edilen puanlara ilişkin kestirilen hata kaynağı aynı olmayacaktır. Klasik test kuramında bu hata kaynaklarını aynı anda kestirmek mümkün olmamaktadır. Böylece örneğin bir araştırmacı, altı öğrencinin on beş maddelik bir başarı testine verdikleri cevapların iki farklı puanlayıcı tarafından puanlanması halinde her bir puanlayıcının verdiği puan için iç tutarlılık katsaysını mı, yoksa puanyıcılar arası tutarlılığını mı hesaplaması gerektiğine karar vermelidir. Klasik test kuramında aynı anda her iki durumun ayrı ayrı ve birlikte ölçme güvenirliğine etkisini kestirebilmesi mümkün değildir. Ayrıca "İstenilen güvenirliğe ulaşabilmek için puanlayıcı sayısını mı, madde sayısını mı yoksa her ikisini birlikte mi artırması uygun olacaktır." sorusuna da cevap bulması olası değildir.

Genellenebilirlik kuramı, davranış bilimlerinde yer alan bu tür örneklerdeki problemlere çözüm sunmaktadır. Örnekte yer alan puanlayıcılar, maddeler ya da zaman gibi hata kaynaklarına

genellenebilirlik kuramında değişkenlik kaynağı (facet) adı verilir. Değişkenlik kaynağı, ölçme hatasının olası kaynağı olarak tanımlanabilir. Böylece değişkenlik kaynağıyla ilişkili bir varyans istenen bir varyans olmayıp bu tür bir varyansın olabildiğince küçük olması istenir (Alharby, 2006). Değişkenlik kaynaklarının düzeyleri koşullar (conditions) olarak adlandırılır. Örneğin, "puanlayıcılar" bir değişkenlik kaynağı ise birinci puanlayıcı, ikinci puanlayıcı, üçüncü puanlayıcı vb.nin her biri bir koşuldur. Genellikle var olan bir değişkenlik kaynağının olası koşullarının sonsuz büyüklükte olduğu varsayılır. Araştırmada yer alan gözlemlerin yapıldığı örneklemin yerine geçebilecek olası gözlemlerin tümüne *"kabul edilebilir gözlemlerin evreni (the universe of admissible observation)" adı verilir. "Genellenebilirlik evreni (the universe of generalization)"* ise araştırmacının genellemek istediği koşulların tümüdür. Pek çok ölçme durumunda bireyler ya da öğrenciler ölçmenin objesi (the object of measurement) olarak ele alınırlar. Bir başka deyişle bireyler, istenilen kararların alınacağı ölçmenin hedefleri durumundadırlar. Bireyler arası farklılıklar doğaldır ve sistematiktir. Bu sebeple bireylere bağlı varyans istenilen bir durum olduğu için bireyler değişkenlik kaynağı olarak düşünülmez. Evren puanı amaçlanan değişkenlik kaynakları için kabul edilebilir gözlemlerin evreninden elde edilecek tüm puanların ortalaması olan ölçme puanı olarak tanımlanır. Evren puanı varyansı, klasik test kuramında yer alan "gerçek puan varyansı"na benzer.

Genellenebilirlik kuramına bağlı yapılan çalışmalardaki ölçme sonuçları çapraz (crossed) ya da yuvalanmış (nested) olabilmektedir. Tamamıyla çapraz desenlerde, ölçme objesi olan bireyler tüm değişkenlik kaynaklarının tüm durumlarında puanlanmıştır. Örneğin, tüm öğrenciler (b) testte yer alan tüm maddeleri (m) cevaplandırmış ve iki farklı puanlayıcı (p) tüm öğrencilerin tüm cevaplarını puanlamışsa, öğrenciler, maddeler ve puanlayıcılara göre tümüyle çaprazlanmış bir desen (b x m x p) söz konusu olacaktır. Yuvalanmış desende ise, bireyler değişkenlik kaynaklarının sadece bir durumu için puanlanmışken diğer durumlarda puanları bulunmamaktadır. Örneğin, bir testte yer alan her bir maddeyi (m) farklı bir öğrenci (b) cevapladığında ve her bir öğrencinin cevabını farklı bir puanlyıcının (p) puanladığı durumdaki desenler yuvalanmış desen (b : m : p) olarak tanımlanır. Genelllenebilirlik kuramında yapılan çalışmalardaki ölçme sonuçlarının bir kısmının çapraz, bir kısmının ise yuvalanmış olduğu karışık (mixed) desenler de mümkündür (Güler, 2008). Ancak tüm olası değişkenlik kaynaklarına ilişkin hatanın kestirimine imkân verdiğinden dolayı G çalışmalarında tümüyle çapraz desenler daha çok tercih edilmektedir (Musquash ve O'Connor, 2006).

Genelllenebilirlik kuramında söz konusu olan değişkenlik kaynakları sabit ya da rasgele olarak tanımlanabilmektedir. Bir değişkenlik kaynağında yer alan tüm durumlar, o değişkenlik kaynağında yer alabilecek olası tüm diğer durumlar ile değiştirilebilir olma özelliğine sahipse, bu değişkenlik kaynağı rasgele olarak tanımlanır (Kieffer, 1998). Örneğin, yapılan bir matematik sınavında yer alan maddeler yine aynı alanda yapılacak bir sınavda olabilecek diğer maddeler ile değiştirilebilir nitelikteyse, bu durumda çalışmada kullanılan maddeler rasgele olarak ele alınır. Rasgele durumların söz konusu olduğu değişkenlik kaynaklarına bağlı yapılan çalışmalar, araştırmacıya o değişkenlik kaynağı için tüm durumların yer aldığı evrene genelleme yapabilme olaslığı sağlar. Aksine, araştırmacı yaptığı çalışmada yer alan değişkenlik kaynağına bağlı sadece belirli durumlarla ilgileniyorsa, diğer durumlara genelleme yapmak gibi bir amacı yoksa, bu durumda ele alınan değişkenlik kaynağı sabit olarak tanımlanır (Crocker ve Algina, 1986). Sabit değişkenlik kaynaklarının bulunduğu çalışmalarda araştırmacının genelleme yapması doğru olmayacaktır (Kieffer, 1998).

Genelllenebilirlik kuramında, klasik test kuramından farklı olarak iki ayrı hata varyansı bulunur. Bu farklılık, genelllenebilirlik kuramında hem bağıl hem de mutlak olmak üzere iki ayrı anlamda karar vermenin mümkün olmasından kaynaklanmaktadır. Bağıl 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. Bu katsayı, klasik kuramdaki güvenirlik katsayısına benzemektedir. Mutlak karar için hesaplanan G katsayısı ise çok daha katı bir değer olup, hem ölçmenin objesine ilişkin

(genellikle bireylerdir) puanların sıralamasındaki tutarlılığın derecesini hem de ham puanların tutarlılığın derecesini ortaya koyar. Belirli bir kesme puanının üzerindeki puanın önem taşıdığı performans ölçümlerinde (örneğin, ehliyet sınavlarında, uzmanlık sınavlarında vb.) mutlak G katsayısı tercih edilebilir (Lee ve Frisbie, 1999; Brennan, 1992). Elde edilen puanların puan sıralamasındaki yerinin önem taşıdığı durumlarda bağıl G katsayısını kullanmak uygun olacaktır. Bağıl ve mutlak kararlar için hesaplanan G katsayılarındaki karışıklığı ortadan kaldırmak üzere, bağıl kararlar için hesaplanan değere G katsayısı, mutlak kararlar için hesaplanan değere phi (fi) katsayısı ya da güvenirlik (dependability) katsayısı adı verilmektedir. Bağıl kararlar için kestirilen G katsayısında yer alan varyans bileşenleri, bireyler (ölçme objesi) ile diğer değişkenlik kaynaklarının etkileşimi ve hata olarak tanımlanan değişkenlik kaynaklarının bileşiminin etkileşiminden oluşmaktadır. Örneğin, birey x madde x puanlayıcı (b x m x p) çapraz deseninde bağıl karar için varyans bileşenleri: σ^2_{bm}, σ^2_{bp} ve $\sigma^2_{bmp,e}$ olmaktadır. Puanın diğer puanları içindeki sıralamasının önem taşımadığı, ancak gözlenen puanın büyüklüğüne dayalı kararlar vermek için hesaplanan G katsayısıda (mutlak karar için) yer alan varyans kaynakları sadece bireyler ile değişkenlik kaynaklarının etkileşimini değil, değişkenlik kaynaklarının kendilerinden ve birbirleriyle etkileşimlerinden meydana gelen değişimleri de içermektedir. Örneğin, birey x madde x puanlayıcı (b x m x p) çapraz deseninde mutlak karar için varyans bileşenleri: σ^2_{m}, σ^2_{p}, σ^2_{bm}, σ^2_{bp}, σ^2_{mp} ve $\sigma^2_{bmp,e}$ şeklindedir (Shavelson ve Webb, 1991).

Genellenebilirlik kuramını uzantısı olduğu klasik test kuramından ayıran bir özelliğinin, bir seferde sadece bir hata kaynağını değil, pek çok hata kaynağını birlikte ele alabildiği yukarıda açıklanmıştı. Buna göre, genellenebilirlik kuramında güvenirlik kestiriminde bulunulurken farklı hata kaynakları aynı anda göz önünde bulundurulmaktadır. Genellenebilirlik kuramı, madde ya da puanlayıcı gibi değişkenlik kaynaklarının sayısının daha az ya da daha fazla olduğu durumlarda güvenirliğin nasıl olacağına ilişkin kestirimlerde bulunabilmesine izin verir. Bu açıdan bakıldığında, genellenebilirlik kuramı, klasik test kuramında yer alan Spearman-Brown formülünün bir uzantısı olarak da yorumlanabilir. Genellenebilirlik kuramında, değişkenlik kaynaklarının farklı sayılarına ilişkin yapılan güvenirlik kestirimleri çalışmasına karar "K" çalışması (D study) adı verilmektedir (Goodwin, 2001).

Genellenebilirlik kuramına ilişkin analizler, alanyazına bakıldığında, genellikle GENOVA paket programıyla yapılmıştır. Ancak bu programın kullanımının zor ve karmaşık olması, genellenebilirlik çalışmalarının yapılmasındaki en büyük sınırlılık olarak araştırmacıların karşısına çıkmaktadır. Musquash ve O'Connor'ın 2006 yılında yayımlanan makalelerinde, genellenebilirlik kuramına ilişkin tüm analizlerin yapılabileceği SPSS ve SAS programları geliştirmişler ve programın nasıl kullanılacağını açıklamışlardır. Bu çalışmada, maddeler ve puanlayıcılar gibi çoklu varyans kaynaklarının bir arada bulunduğu ölçme durumlarında kullanılmasının önem taşıdığı genellenebilirlik kuramına bağlı elde edilen genellenebilirlik ve güvenirlik katsayılarının hesaplanmasına ilişkin GENOVA ve Musquash ve O'Connor (2006)'ın geliştirdikleri SPSS paket programlarının sonuçları bir arada sunulmaktır. Gerçek bir örnek üzerinden madde ve puanlayıcı değişkenlik kaynaklarının ele alındığı GENOVA ve SPSS analizleriyle elde edilen hem orjinal çalışma hem de farklı karar çalışmalarına ilişkin genellenebilirlik ve güvenirlik katsayıları ve elde edilen sonuçların karşılaştırılması yapılmıştır.

Yöntem

Bu çalışmada, 203 öğrencinin 18 açık uçlu matematik maddesine verdikleri cevaplara ilişkin dört puanlayıcının puanları güvenirliği genellenebilirlik kuramı ile incelenmiştir. Her bir öğrenci tüm maddeleri cevaplamış ve tüm öğrencilerin cevapları dört puanlayıcı tarafından puanlanmış olduğu için değişkenlik kaynaklarına ilişkin tümüyle çapraz desen (öğrenci x madde x puanlayıcı) kullanılmıştır. Ölçmenin hedefi "öğrenciler" olarak düşünüldüğünden, öğrenciler ölçme objesi olarak kabul edilmiştir. Böylece öğrencilerden meydana gelen değişim, ölçmenin olası hata kaynaklarından (maddeler, puanlayıcılar ve bunlar arası etkileşim) ayrı tutulmuştur.

Her bir değişkenlik kaynağının ayrı ayrı ve birbirleriyle etkileşimlerinin varyansları hesaplanmış, genellenebilirlik (G) ve phi katsayıları bulunmuştur. Farklı karar çalışmalarında elde edilen genellenebilirlik (G) ve phi katsayıları her iki programda hesaplanarak karşılaştırılmıştır.

Bulgular

Her bir değişkenlik kaynağı ve bunlar arasındaki etkileşimlerin varyanslarının nasıl hesaplandığı Tablo 1'de verilmiştir.

Tablo 1.

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	$\sigma^2(bpm)$

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 1'deki eşitliklere ilişkin GENOVA ve SPSS paket programlarına dayalı sonuçlar Tablo 2'de görülmektedir.

Tablo 2.

İki Değişkenlik Kaynaklı Tesadüfi Desen İçin SPSS ve GENOVA Programlarına Göre Varyans Bileşenleri Kestirim Değerleri

Vary. Kay.	Sd	Kareler Ortalaması		Varyans		Varyans Yüzdeleri	
		SPSS	*GENOVA*	*SPSS*	*GENOVA*	*SPSS*	*GENOVA*
b	202	*91.127*	91.940	*1.160*	1.167	*32.6*	32.5
m	17	*141.026*	154.089	*0.159*	0.176	*4.5*	4.9
p	3	*297.608*	331.019	*0.075*	0.089	*2.1*	2.5
bm	3434	*6.754*	7.017	*1.553*	1.643	*43.6*	45.8
bp	606	*1.381*	1.333	*0.047*	0.049	*1.3*	1.4
mp	51	*5.744*	4.838	*0.026*	0.022	*0.7*	0.6
bmp	10.302	*0.544*	0.444	*0.544*	0.444	*15.3*	12.4

Tablo 2'de görüleceği üzere, son sütunda toplam varyans yüzdeleri yer almaktadır. Bu sütundaki değerlere göre, öğrencilere ilişkin varyans yüzdeleri SPSS'te %32.6; GENOVA'da ise %32.5 olarak bulunmuştur. Diğer taraftan, madde ve puanlayıcı değişkenlikleri toplam varyansın sırasıyla SPSS sonuçlarına göre %4.5 ve %2.1'ini; GENOVA sonuçlarına göre ise %4.9 ve %2.5'ini açıklamaktadır. Bu yüzdeler, her iki programdan elde edilen sonuçlar birbirine çok yakındır. Varyansın yaklaşık %33'ünün 18 madde ve 4 puanlayıcı üzerinden öğrenciler arasındaki değişiklikten kaynaklandığını göstermektedir (Bu istenilen bir durumdur; öğrencilerin matematik başarıları arasındaki farklılıkları ifade etmektedir). Bundan sonraki üç değer ise varyans bileşenlerinin ikili etkileşimlerinin varyans yüzdelerini vermektedir. SPSS ve GENOVA sonuçlarına göre sırasıyla; toplam varyansın %43.6 ve %45.8'ini öğrenci-madde (bxm) etkileşimi, %1.3 ve 1.4'ü öğrenci-puanlayıcı (bxp) etkileşimi, %0.7 ve %0.6'sı madde-puanlayıcı (mxp) etkileşimi açıklamaktadır. Tablo 1'in son sütununda yer alan son

değerler de üç değişkenlik kaynağının birlikte etkileşimi (bxmxp) olan belirlenemeyen (tesadüfi) hata kaynağının toplam varyanstaki yüzdesini göstermektedir. SPSS'e göre bu değer %15.3, GENOVA'ya göre %12.4'tür. Genellenebilirlik çalışmalarında bu varyansın 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 kuramının 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 kuramına dayalı çalışmalarda, klasik test kuramındaki 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 kuramındaki 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 kuramından 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 3'te bu katsayıların hesaplanmasında kullanılan eşitlikler görülmektedir.

Tablo 3.

İki Değişkenlik Kaynaklı Tesadüfi Desen İçin Hata ve Güvenirlik Kestirimleri

Mutlak hata $(\sigma^2_{(\Delta)})$	Bağıl hata $(\sigma^2_{(\delta)})$	Phi katsayısı (Φ)	G-katsayısı(G)
$\frac{\sigma^2_a}{N_a}+\frac{\sigma^2_m}{N_m}+\frac{\sigma^2_{ba}}{N_a}+\frac{\sigma^2_{bm}}{N_m}+\frac{\sigma^2_{am}}{N_aN_m}+\frac{\sigma^2_{bam}}{N_aN_m}$	$\frac{\sigma^2_{ba}}{N_a}+\frac{\sigma^2_{bm}}{N_m}+\frac{\sigma^2_{bam}}{N_aN_m}$	$\frac{\sigma^2_b}{\sigma^2_b+\frac{\sigma^2_a}{N_a}+\frac{\sigma^2_m}{N_m}+\frac{\sigma^2_{ba}}{N_a}+\frac{\sigma^2_{bm}}{N_m}+\frac{\sigma^2_{am}}{N_aN_m}+\frac{\sigma^2_{bam}}{N_aN_m}}$	$\frac{\sigma^2_b}{\sigma^2_b+\frac{\sigma^2_{ba}}{N_a}+\frac{\sigma^2_{bm}}{N_m}+\frac{\sigma^2_{bam}}{N_aN_m}}$

Tablo 3'teki 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. Tablo 3'teki eşitliklere ilişkin SPSS ve GENOVA programlarıyla elde edilen sonuçlar Tablo 4'te görülmektedir.

Tablo 4.

İki Değişkenlik Kaynaklı Tesadüfi Desen İçin SPSS ve GENOVA Programlarına İlişkin Sonuçlar

	Mutlak hata	Bağıl hata	Phi katsayısı	G-katsayısı
SPSS	*0.133*	*0.105*	*0.897*	*0.917*
GENOVA	0.142	0.109	0.892	0.914

Tablo 4'te mutlak karar için mutlak hata ve phi katsayısı değerleri sırasıyla SPSS programında 0.133 ve 0.897 iken GENOVA programında 0.142 ve 0.892 olarak bulunmuştur. Göreceli karar için bağıl hata ve phi katsayısı değerleri ise sırasıyla SPSS programında 0.105 ve 0.917 iken GENOVA programında 0.109 ve 0.914 olarak elde edilmiştir. Buradan da görüleceği üzere, SPSS ve GENOVA programlarından elde edilen değerler arasındaki en büyük farklılık 0.009 gibi oldukça küçük bir değerle mutlak hatadadır. Phi katsayısındaki farklılık 0.005 ve G katsayısındaki farklılık 0.003 olarak görülmektedir.

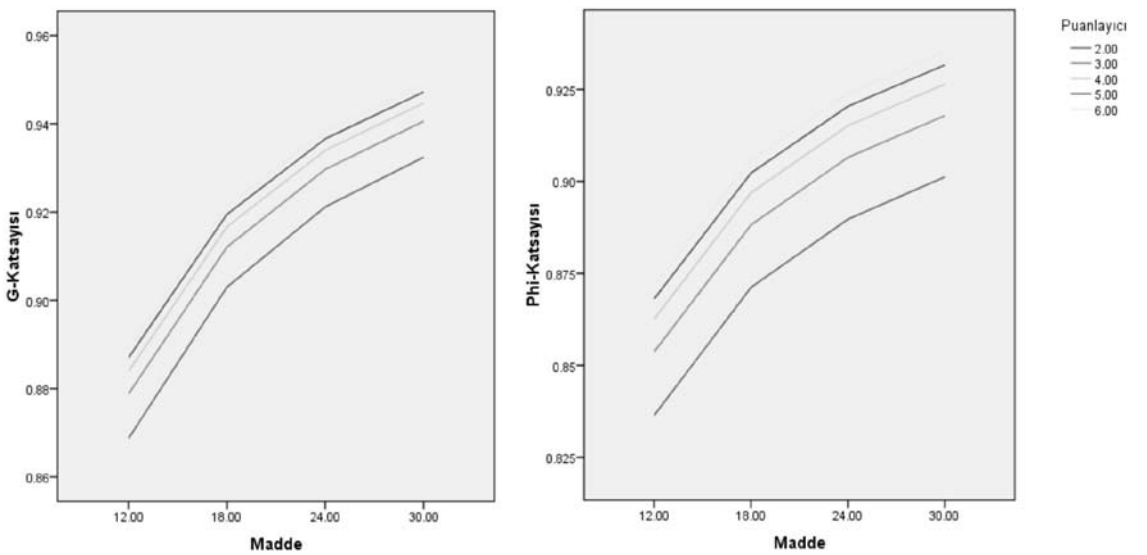
Tablo 3'teki eşitliklerin paydalarında yer alan puanlayıcı ve madde sayıları değiştirilerek farklı K çalışmaları yapılabilir. K çalışmalarındaki hesaplamalar, klasik test kuramındaki Spearman-Brown formülüne benzer düşünülebilir (Goodwin, 2001). Ancak K çalışmalarındaki hesaplamalar tek bir değişkenlik kaynağı (madde) için değil, farklı değişkenlik kaynakları (puanlayıcı ya da zaman gibi) için de yapılabilir. Bir başka deyişle, asıl çalışmada yer alan madde ve puanlayıcı sayıları azaltılarak ya da artırılarak elde edilecek G katsayılarını ve phi katsayılarını hesaplamak mümkündür. Bu çalışmada yer alan farklı K çalışmaları için SPSS ve GENOVA programlarıyla elde edilen sonuçlar Tablo 5'te görülmektedir.

Tablo 5.

Farklı K Çalışmaları İçin SPSS ve GENOVA Programlarına İlişkin Sonuçlar

	puanlayıcı sayısı=4, madde sayısı=24		puanlayıcı sayısı=6, madde sayısı=18	
	Phi katsayısı	G katsayısı	Phi katsayısı	G katsayısı
SPSS	*0.915*	*0.934*	*0.906*	*0.921*
GENOVA	0.910	0.932	0.901	0.918

Tablo 5'te madde sayısının üçte bir ve puanlayıcı sayısının ikide bir artırılmasına ilişkin yapılan K çalışmalarında hesaplanan phi ve G katsayılarının değerleri verilmiştir. Puanlayıcı sayısı aynı kalarak, madde sayısının üçte bir artırılmasıyla (puanlayıcı sayısı=4, madde sayaıs=24) elde edilen phi ve G katsayıları sırasıyla SPSS programında 0.915 ve 0.934 iken GENOVA programında 0.910 ve 0.932 olarak bulunmuştur. Madde sayısı aynı kalarak, puanlayıcı sayısının ikide bir arttırılmasıyla (puanlayıcı sayısı=6, madde sayısı=18) elde edilen phi ve G katsayıları sırasıyla SPSS programında 0.906 ve 0.921 iken GENOVA programında 0.901 ve 0.918 olarak elde edilmiştir. Buradan da görüleceği üzere, SPSS ve GENOVA programlarından elde edilen değerler arasındaki en büyük farklılık 0.005 gibi oldukça küçük bir değerle her iki K çalışması için de phi katsayılarında gözlenmektedir. G katsayılarının farklılığının ise 0.002 ve 0.003 olduğu görülmektedir. Ayrıca SPSS programında yapılan K çalışmalarında, farklı sayılardaki değişkenlik boyutları için hesaplanan bağıl ve göreceli hata değerlerine ve phi ve G katsayılarına ilişkin grafiksel gösterim mümkündür. Ancak bu tür bir grafiksel gösterim GENOVA programı ile elde edilememektedir. Şekil 1'de farklı K çalışmaları için SPSS programında elde edilen phi ve G katsayılarının grafikleri verilmiştir.



Şekil 1. Farklı K Çalışmaları İçin SPSS Programında Elde Edilen G ve Phi Katsayılarına İlişkin Grafikler

Sonuç

"Genellenebilirlik kuramı klasik test kuramının en önemli uzantılarından biri olup, özellikle birden fazla puanlamanın yapıldığı durumlarda daha çok kullanılmalıdır" (Nunally ve Bernstein, 1994).

Bu çalışmada da görüleceği gibi, genellenebilirlik kuramı, birden fazla puanlayıcının bulunduğu ölçme durumlarında her bir değişkenlik kaynağına (birey "öğrenci", puanlayıcı, madde) ve bunlar arasındaki etkileşime ilişkin ayrıntılı bilgi veren bir yaklaşımdır. Ancak böylesine açıklayıcı bilgi veren bu kuramın pratikte kullanımının zor ve anlaşılmasının güç olması, tercih edilmemesindeki en büyük sınırlılıktır (Musquash ve O'Connor, 2006; Goodwin, 2001; Shavelson, Webb ve Rowley, 1989). Genelllenebilirlik kuramının kullanımı için en çok bilinen GENOVA programının anlaşılmasının ve kullanılmasının karmaşık oluşu özellikle ülkemizde genelllenebilirlik kuramıyla yapılan çalışmaların oldukça az olmasını açıklayan en önemli sebeplerden biridir (Atılgan, 2008; Yelboğa, 2007; Atılgan, 2005; Atılgan ve Tezbaşaran, 2005; Atılgan, 2004). Musquash ve O'Connor (2006) tarafından geliştirilen genellenebilirlik kuramının kullanılmasına izin veren SPSS programı oldukça pratik bir alernatif olarak karşımıza çıkmaktadır. GENOVA'daki tüm bilgileri sağlayan, hatta bu bilgilere ek olarak mutlak ve bağıl hata kaynakları ile G ve phi katsayıları için ayrı ayrı grafiksel sonuçlar veren, kullanımı kolay bu program, tercih edilebilir bir alternatif olarak önerilebilir. Çalışmadan elde edilen sonuçlardan görüleceği üzere, iki programa ilişkin bulunan sonuçların birbirine nerdeyse eşit değerler vermesi de bu öneriyi destekler niteliktedir. Bu küçük farklılıklar, ondalıklı sayıların virgülden sonrasında yapılan yuvarlama işlemlerinden kaynaklanabilir. Brennan'nın (1992) belirttiği gibi, genellenebilirliğe ilişkin yapılan analizler çok sayıda hesaplama gerektirmektedir ve her bir hesaplamada sayılarda yapılan yuvarlama işlemi sonucu kestirilen varyans bileşenlerinin değerleri değişime uğrayabilmektedir.

Bu çalışmada değişkenlik kaynağı olarak sadece birey, puanlayıcı ve madde alınmıştır. Farklı ve daha fazla değişkenlik kaynaklarının bulunduğu çalışmalarla da iki programdan elde edilen sonuçların karşılaştırılmasının faydalı olabileceği düşünülmektedir. Ayrıca, çalışmada tümüyle çaprazlanmış desen (bxmxp) kullanılmıştır; yuvalanmış desen ya da karışık (hem çaprazlanmış hem yuvalanmış) desenin bulunduğu durumlara ilişkin benzer bir çalışmanın yapılması da önerilebilir.

Kaynakça

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 Asssessment*. Yayınlanmamış doktora tezi. The Pennsylvenia State University.

Atılgan, H. (2005). Genellenebilirlik Kuramı ve Puanlayıcılar Arası Güvenirlik İçin Örnek Bir Uygulama. *Eğitim Bilimleri ve Uygulama*, 4 (7).

Atılgan, H. ve Tezbaşaran, A. A. (2005). Genellenebilirlik kuramı alternatif karar çalışmaları ile senaryolar ve gerçek durumlar için elde edilen G ve Phi katsayılarının tutarlılığının incelenmesi, Eğitim Araştırmaları Dergisi, Eurasian Journal of Educational Research, Yıl.5 Sayı.18

Atılgan, H. (2004). *Genellenebilirlik Kuramı ve Çok Değişkenlik Kaynaklı Rasch Modelinin Karşılaştırılmasına İlişkin Bir Araştırma*. Yayımlanmamış doktora tezi. Hacettepe Üniversitesi: Ankara.

Baykul, Y., Gelbal, S ve Kelecioğlu, H. (2003). *Anadolu Lisesi Öğretmenleri İçin Eğitimde Ölçme ve Değerlendirme.* Milli Eğitim Basımevi, İstanbul.

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

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

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

Goodwin, L. D. (2001). Interrater Agreement and Reliability. *Measurement in Psychical Education and Exercises*

Science, 5 (1), 13-14.

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

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. ve Frisbie, D. A. (1999). Estimating Reliability Under a Generalizability Theory Model for Test Scores Composed of Testlets. *Applied Measurement in Education*. 12, 3, 237-255.

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

Nunally, J. C. & Bernstein, I. H. (1994). *Psychometric Theory* (3.baskı). New York: Mc-Graw-Hill.

Nunally, J. C. (1982). Reliability of Measurement. *Encylopedia of Educational Research*. (5.baskı) Editor H.E. Mitzel. New York.

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

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

Yelboğa, A. (2007). *Klasik Test Kuramı ve Genellenebilirlik Kuramına Göre Güvenirliğin Bir İş Performansı Ölçeği Üzerinde İncelenmesi*. Yayımlanmamış doktora tezi. Ankara Üniversitesi, Ankara.