GENELLENEBİLİRLİK KURAMI VE PUANLAYICILAR ARASI GÜVENİRLİK İÇİN ÖRNEK BİR UYGULAMA

Dr. Hakan Atılgan
Ege Üniversitesi

Özet

Bu çalışmada; güçlü bir istatistiksel temele sahip, farklı birçok ölçme durumu için esnek bir alternatif olarak olası bütün hata kaynaklarını birlikte değerlendirerek ölçmenin güvenirliğinin belirlenmesini sağlayan bir yaklaşım olan G-kuramına bir giriş yapılarak, temelleri vurgulanmış ve kuramının klasik test kuramına göre avantajları açıklanmıştır. Ayrıca, farklı ve çok hata kaynaklı bir ölçme durumu olarak, puanlayıcıların ölçme sürecine katıldığı hipotetik bir örnekle; ölçüt dayanaklı ölçmeler için Phi ve norm dayanaklı ölçmeler için G katsayılarının elde edilerek kullanılması gösterilmiştir.

Anahtar Sözcükler

Genellenebilirlik kuramı, puanlayıcılar arası güvenirlik, ölçme, değerlendirme.

GENERALIZABILITY THEORY AND A SAMPLE APPLICATION FOR INTER-RATER RELIABILITY

Dr. Hakan Atılgan
Ege University

Abstract

The aim of this study is to introduce the G-theory, its essentials, and advantages over the Classical Test Theory. The G-theory, a flexible alternative for various measurement cases, has a very strong statistical base, and is an approach allowing determining reliability of measurement by taking into account all potential sources of error. Therefore the use of it by obtaining Phi coefficients for criterion-referenced testing, and G coefficients for norm-referenced testing has been demonstrated through a hypothetical example, in which raters get involved with measurement process, as a measurement case possessing various and multiple sources of error.

Keywords

Generalizability theory, inter-rater reliability, measurement, evaluation.

GİRİŞ

Bir ölçme durumunda, elde edilen gözlenen puanlara hata karışması nedeniyle ölçülen özelliğe ait gerçek değerin doğrudan elde edilmesi olanaklı değildir. Ölçme işlemi ile elde edilen gözlenen puan, gerçek puan ve hata puanından oluşur. Ölçmede amaç olabildiğince gerçek puana yakın ölçme sonuçları elde edebilmektir. Ölçme yoluyla elde edilen gözlenen puanlarla, ölçülen özelliğin gerçek değerine ulaşılması beklenilir. Ölçme yoluyla gerçek puana ulaşılması demek, ölçme ile elde edilen puanların hatasız olması anlamına gelir. Ölçme işleminde kullanılan ölçme aracı ne kadar hassas olursa olsun, bütün ölçme sonuçlarına farklı kaynaklardan karışan hataların olması kaçınılmazdır. Bu nedenle, ölçülen özelliğin gerçek değerine ölçme yoluyla doğrudan ulaşılması söz konusu değildir. Oysa, ölçme çabalarının temel amacı; ölçülen özelliğin gerçek değerinin elde edilebileceği ölçme araçlarının üretilebilmesi ve ölçme sonuçlarından elde edilen puanların olabildiğince hatalardan arınık hale getirilebilmesidir. Ölçme sonuçlarına dayalı olarak verilen kararların doğruluğu; ölçme sonuçlarının hatalardan arınıklığı ve ölçme aracının ölçülmek istenilen niteliği başka niteliklerden arınık olarak ölçebilmesiyle olanaklıdır. Ölçme sonuçlarının tesadüfi hatalardan arınıklığının derecesi güvenirlik olarak adlandırılmaktadır. Klasik Test Kuramı güvenirlik katsayısını paralel iki ölçme arasındaki korelasyon katsayısı olarak tanımlar (Lord ve Novic, 1968; Baykul, 2000).

Cronbach, Rajaratnam ve Gleser (1963), aynı gözlemlerin paralel testlerin bir setinden daha çoğuna ait olduğu düşünülebileceği, bu nedenle aynı gözlemin birden çok güvenirlik katsayısına sahip olabileceğini tartışmışlardır. Nitekim alt testlerden oluşan bir ölçmede iç tutarlılık güvenirliği düşük olma eğilimindeyken, test tekrar test ya da paralel formlar güvenirliği yüksek olabilmektedir. Bu çelişki ve sınırlılığın temel nedeni; ölçme sonuçlarına karışan hataların klasik test kuramında sadece bir kaynaktan gelen hatalar olarak ele alınmasıdır. Nitekim, klasik test kuramının güvenirlik hesaplama yöntemleri, güvenirliğin anlamına göre ve ele alınan hata kaynağına göre farklılıklar gösterir (Lord ve Novic, 1968; Baykul, 2000). Test tekrar test yönteminde bir testin aynı koşullarda farklı zamanlarda uygulanmasından elde edilen sonuçların benzer olması beklenilir. Bu nedenle test tekrar test yöntemi ile hesaplanan güvenirlik ölçme aracının zamandan zamana kararlı ölçmeler yapabilme derecesi olduğundan hata kaynağı zaman olarak ele alınır. Klasik test kuramında kullanılan paralel formlar yöntemi diğer bir hesaplama yöntemidir. Paralel formlar yöntemi ile bir birinin paraleli olan ölçme araçlarının birbirleri ile tutarlı sonuçlar verip vermedikleri incelenir. Paralel formlarla elde edilen güvenirlik katsayısı tutarlılık anlamındadır. Bu yöntemle güvenirlik belirlemede birbirinin paraleli olan formlardan gelen hatalar söz konusu olduğundan formlar hata kaynağıdır. Diğer yandan, iç tutarlılık anlamında güvenirliğin hesaplanmasında kullanılan yöntemlerde ise maddelerin birbirleri ile ve testin bütünü ile ilişkisi dikkate alınır. Bu nedenle iç tutarlılık

güvenirliğinde maddeler hata kaynağıdır (Shavelson ve Webb, 1991; Crocker ve Algina, 1986; Nunnally ve Bernstein, 1994; Brennan, 2001).

Yukarıda da belirtildiği aynı ölçmeye ilişkin olarak, klasik test kuramının farklı güvenirlik yöntemleri ile farklı anlamlarda güvenirlik katsayıları hesaplanabilir. Ancak aynı ölçme için farklı anlamlarda ve farklı yöntemlerle ile hesaplanan güvenirlik katsayıları farklılıklar gösterebilmektedir. Klasik test kuramının farklı yöntemleriyle farklı anlamlarda elde edilen güvenirlik katsayılarının birbirinden farklı olmasından hareketle, Cronbach ve arkadaşları, Genellenebilirlik (G) kuramını ortaya atmışlardır. G-kuramı esnek bir alternatif olarak; puanlayıcı, zaman, test formu, madde, görev gibi bir ölçme içinde yer alabilen bütün potansiyel hata kaynaklarını eş zamanlı değerlendiren bir yaklaşımdır. G-kuramı davranış ölçmede güvenirliğin değerlendirilmesini, güvenilir gözlemlerin tasarlanmasını, araştırılmasını ve kavramlaştırılmasını sağlayan, istatistiksel bir kuramdır ve varyans analizine (ANOVA) dayalıdır. G-kuramı, günümüzde hala yaygın kullanılan klasik test kuramının gerçek puan modelinin sınırlılıklarına olan tepkilerden hareketle Cronbach, Gleser, Nanda ve Rajaratnam (1963-1972) tarafından ortaya atılmıştır (Allan, 1990; Shavelson ve Webb, 1991; Brennan, 2001).

AMAÇ

Bu çalışma ile ölçmenin güvenirliğinin belirlenmesinde G-kuramı yaklaşımına bir giriş yapılmasına, temellerinin vurgulanılmasına ve kuramının klasik test kuramına göre avantajlarının açıklanılmasına çalışılmıştır. Ayrıca, farklı ve çok hata kaynaklı bir ölçme durumu olarak, puanlayıcıların ölçme sürecine katıldığı bir örnekle; ölçüt dayanaklı ölçmeler için Phi ve norm dayanaklı ölçmeler için G katsayılarının elde edilerek, kullanılmasının gösterilmesi amaçlanmıştır.

Temel Kavramlar

G-kuramında, bir ölçmedeki potansiyel hata kaynakları (maddeler, puanlayıcılar, zaman, formlar vs.) *değişkenlik kaynağı* (*facet*) olarak adlandırılır. Her bir değişkenlik kaynağı kendi içinde farklı düzeylerden oluşur. G-kuramında değişkenlik kaynaklarının bu düzeyleri *koşul (condition)* olarak adlandırılır. G-kuramındaki değişkenlik kaynağı (facet) ve koşul varyans analizi literatüründeki faktör (factor) ve düzey (level) kavramlarına karşılık gelir (Crocker ve Algina, 1986; Shavelson ve Webb, 1991; Brennan, 2001). Örneğin bir ölçmede madde değişkenlik kaynağı ise, madde sayısı bu değişkenlik kaynağının düzeyi olur.

Bir ölçmede değişkenlik kaynaklarının seçilmiş koşullarına karşılık, bütün olası koşullar *kabul edilebilir gözlemlerin evreni* (universe of admissible observation) olarak tanımlanır. Başka bir ifadeyle, kabul edilebilir gözlemlerin evreni bir testte kullanılabilecek olası gözlemlerin tümü olarak tanımlanır. Genellenmek istenilen

bir değişkenlik kaynağının koşulları da *genellemenin evreni* (universe of generalization) olarak adlandırılır (Shavelson ve Webb, 1991; Brennan, 2001).

G-kuramında iki tür çalışma yer alır: *genellenebilirlik* (G) *çalışması* (generalizability study) ve *karar* (K) *çalışması* (decision study). *G-çalışması*; olası bütün hata kaynaklarını birlikte analiz ederek, hata kaynaklarının etkilerini ortaya koymak ve kabul edilebilir gözlemlerin evrenini tanımlamak için yapılır (Shavelson ve Webb, 1991). G-çalışmasının amacı, ölçmedeki değişkenlik kaynakları hakkında olabildiğince bilgi sağlayarak, ölçme desenine karar verilmesini ve ölçme araçlarının geliştirilmesine veya sonraki kullanımlarda değişkenlik kaynaklarından gelen hataları azaltılmasına kaynaklık etmektir.

G-çalışmasından elde edilen bilgilerle yapılan *K-çalışmasının* amacı ise, bir ölçmedeki hataları en aza indirmenin alternatiflerini araştırarak, ölçmenin en uygun desenine ulaşmaktır (Shavelson ve Webb, 1991). K-çalışması yoluyla, her bir değişkenlik kaynağının (madde, puanlayıcı vs.) koşullarının sayısının artırılması ya da azaltılması yoluyla ölçme hatası ve güvenirlikteki artma/azalma belirlenebilmektedir. Bu sayede istenilen düzeyde bir güvenirliğe ulaşmak için, değişkenlik kaynaklarının koşullarının en uygun sayısına ulaşılarak, ölçme aracının geliştirilmesi ya da ileriki uygulamalarda ölçmenin nasıl olması gerektiğine karar verilmesine olanak sağlanır.

G-kuramında; değişkenlik kaynağının sayısına bağlı olarak desenin oluşturulmasının yanı sıra, çaprazlanmış (crossed) ya da yuvalanmış (nested) olmak üzere veri yapısına bağlı iki tür desen bulunmaktadır. Ölçmedeki değişkenlik kaynağının bütün koşulları diğer bir değişkenlik kaynağın bütün koşullarını gözlüyorsa *çaprazlanmış* olarak adlandırılır ve değişkenlik kaynakları arasına "x" işareti konularak gösterilir. Bir değişkenlik kaynağın bütün koşulları diğer bir değişkenlik kaynağın bütün koşulları tarafından gözlemlenmiyor, bir değişkenlik kaynağının sadece bazı koşulları diğer bir değişkenlik kaynağın bazı koşullarınca gözlemleniyorsa *yuvalanmış* (nested) olarak adlandırılır ve iki değişkenlik kaynağı arasına "*:*" işareti konularak gösterilir (Shavelson ve Webb, 1991; Brennan, 2001).

G-kuramı; eğitim ve psikolojide güvenirlik belirlemede, bağıl (relative ya da Norm-reference) ve mutlak (absolute ya da criterion reference) değerlendirme olmak üzere iki tür karar vermenin söz konusu olduğunu dikkate alır. Bu nedenle, G-kuramıyla bağıl değerlendirmeler için genellenebilirlik (generalizability) katsayısı (G ya da $E\rho^2$), ve mutlak değerlendirmeler için ise güvenirlik (dependability) katsayısı (Φ ya da Phi) olmak üzere iki farklı katsayı ayrı ayrı hesaplanır (Shavelson ve Webb, 1981; Crocker ve Algina, 1986; Brennan, 2001; Goodwin, 2001; Shavelson, 2003).

G-Kuramının Avantajları

Bir anlamda klasik test kuramının (KTK) uzantısı olduğu söylenilen G-kuramının birçok avantajı bulunmaktadır. Bu avantajlardan belli başlıcaları özetlenerek aşağıda sıralanmıştır (Shavelson ve Webb, 1991; Brennan, 2001):

1. G-kuramı bir ölçme durumunda yer alan bütün potansiyel hata kaynaklarını birlikte ve eşzamanlı olarak değerlendirerek, tek bir hata kaynağını değerlendiren modellere göre daha kapsamlı bir güvenirlik kestirimi yapılmasına olanak sağlar. Oysa KTK'da sadece bir hata kaynağının bir defada değerlendirilmesi olanaklıdır.
2. G-kuramı ölçmenin güvenirliğinin belirlenmesinde hem mutlak hem de bağıl değerlendirmeler için katsayılar üretebilmektedir. Oysa KTK'da sadece bağıl değerlendirme için güvenirlik hesaplanır.
3. G-kuramı; KTK'nın aksine, güvenirlik çalışmalarında sadece ölçmedeki hatalara kaynaklık eden değişkenlik kaynaklarını değil, aynı zamanda bu değişkenlik kaynaklarının ortak etkilerini de dikkate alır.
4. Alternatif K-çalışmaları; çok sayıda hata kaynağının analizi ile en uygun güvenirliğe ulaşılması için, değişkenlik kaynaklarının koşullarının sayısının belirlenmesine olanak sağlar. Böylelikle istenilen düzeyde güvenirlik için her bir değişkenlik kaynağının (madde, puanlayıcı, zaman vs.) sayısının belirlenmesi G-kuramı ile olanaklıdır. Oysa KTK sadece bir değişkenlik kaynağının (madde) sayısının güvenirliğin artırılması için ne olabileceğini Sperman-Brown formülü ile hesaplayabilir.
5. G-kuramı geçerlik ve güvenirlik arasındaki geleneksel farklılığı da bir ölçüde ortadan kaldırmaktadır. G-kuramında kabul edilebilir gözlemlerin evreninden alınan örneklemin evrene genellenebilirliği test edildiğinden, kapsam geçerliğinin de bir kanıtı olarak kabul edilebilmektedir.

G-kuramının uygulanmalarının tanımlayıcı olabilmesi amacıyla, hipotetik bir veri ile G ve K çalışmalarının yapılmasına ilişkin bir örnek aşağıda verilmiştir.

Örnek

Örnek olarak 10 adaya uygulanmış dört görevden oluşan bir testin, dört bağımsız puanlayıcı tarafından puanlanmasına ilişkin hipotetik bir veri Çizelge 1'de verilmektedir.

Çizelge 1. Örnek veri

	Puanlayıcı	**1**				**2**				**3**				**4**			
Birey	Görev	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1		0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3		1	0	0	1	1	0	0	1	1	0	0	1	0	0	0	1
4		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5		1	0	0	1	1	0	0	0	1	0	0	0	1	0	0	0
6		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7		1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
8		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9		1	1	1	0	1	1	1	0	1	1	0	1	1	1	1	0
10		1	0	0	1	1	0	0	0	1	0	0	0	1	0	0	1

Bu çalışmanın amacı gereği hipotetik olarak alınan verilerin örneklemi küçük tutulmuştur (10 birey, 4 madde ve 4 puanlayıcı). Örnekteki verilerde; her bireye aynı dört madde yöneltildiği ve bu maddeler için dört bağımsız puanlayıcının bütün bireyleri dört madde üzerinden puanladığı görülmektedir. Bu durumda bireyler ölçmenin amacı olduğundan değişkenlik kaynağı (facet) olarak ele alınmazlar. Diğer yandan maddeler ve puanlayıcılar bireylerin evren puanlarının (KTK'ında gerçek puana karşılıktır) doğrulukla kestirilmesini etkileyebileceklerinden birer değişkenlik kaynağı olarak dikkate alınır. Bu durumda Çizelge 1'de verilen örnek veriler için iki-değişkenlik kaynaklı ifadesi kullanılır. Her bir bireye aynı dört madde yöneltildiğinden ve dört puanlayıcının bütün bireyleri aynı dört madde üzerinden puanladığından G-kuramının çaprazlanmış deseni söz konusu olur. Bireylere yöneltilen dört maddenin ölçülmesi amaçlanan özelliğin ölçümü için kullanılabilecek kabul edilebilir maddelerin evreninden tesadüfi olarak çekildiği ve puanlayıcıların da bu testte bireyleri puanlayabilecek puanlayıcıların evreninden örneklendiği düşünüldüğünde, desenin tesadüfi (random) etki modeli olması gerekir. Bu durumda G-kuramının iki-değişkenlik kaynaklı çaprazlanmış tesadüfi etki deseni söz konusu olur ve bu desen bireyler "b", maddeler "m" ve puanlayıcılar "p" ile gösterilmek üzere *b* x *m* x *p* olarak sembolize edilir.

G-kuramı *b* x *m* x *p* deseninden; bireyler, maddeler, puanlayıcılar olmak üzere üç ana etki, birey-madde, birey-puanlayıcı, madde-puanlayıcı ortak etkileri ve kalan etki (b x m x p, e) olmak üzere yedi varyans bileşeni hesaplanır. Bu varyans bileşenlerinin hesaplanmasında geleneksel ANOVA eşitlikleri kullanılır (Shavelson ve Webb 1991; Brennan 2001, Atılgan 2004). Çizelge 2'de kestirilen varyans bileşenleri ve toplam varyans içindeki payları verilmiştir.

Çizelge 2. Kestirilen varyans bileşenleri ve yüzdeleri (ANOVA tablosu)

*Varyans Kaynağı***	*Kareler Toplamı*	*Serbestlik Derecesi*	*Kareler Ortalaması*	*Kestirilen Varyans Bileşeni*	*Varyans Yüzdesi (%)*
b	21, 975	9	2, 44167	0, 13534	51, 99
m	2, 625	3	0, 87500	0, 01451	5, 57
p	0, 075	3	0, 02500	0*	0, 00
b x *m*	7, 875	27	0, 29167	0, 06049	23, 24
b x *p*	0, 925	27	0, 03426	0*	0, 00
m x *p*	0, 475	9	0, 05278	0, 00031	0, 12
b x *m* x *p, e*	4, 025	81	0, 04969	0, 04969	19, 09

* Negatif varyans bileşenleri sıfır olarak alınmıştır (Brennan, 2001).

** *b*: Birey, *m*: Madde, *p*: Puanlayıcı

Çizelge 2'de yer alan varyans bileşenleri ve toplam varyans içindeki paylarından hareketle varyans kaynaklarına göre bağıl olarak Shavelson ve Webb (1991) yaklaşımı ile aşağıdaki gibi yorumlanabilir.

1. Bireyler (*b*) ana etkisi için kestirilen varyans bileşeni en büyük varyans payına sahiptir. Bu durum bireylerin ölçülen özellikleri bakımından ayrılabildiklerini gösterir. Bireyler için kestirilen varyans bileşeni KTK'daki gerçek puan varyansına karşılık olarak evren puanı varyansıdır. Bu nedenle bu varyansın büyük olması gerekir.

2. Madde (*m*) ana etkisi için kestirilen varyans bileşeni bazı maddelerin diğerlerine göre güçlük düzeylerinin farklılaştığını göstermektedir.

3. Puanlayıcılar (*p*) ana etkisi puanlayıcıların bütün bireyler boyunca yaptıkları puanlamaların katılık/cömertlik düzeylerinin farklılaşıp farklılaşmadığını gösterir. Örneğimizde puanlayıcıların bütün bireyler için eşit katılık/ cömertlikte puanlama yaptıkları söylenebilir.

4. Birey-madde ortak etkisi (*b* x *m*) belli bir bireyin bağıl konumunun bir maddeden diğerine değişip değişmediğini gösterir. Örneğimizde birey-madde ortak etkisinin en büyük ikinci varyans bileşeni olması nedeniyle bireylerin bir maddeden diğer maddeye bağıl konumlarının değiştiği söylenebilir.

5. Birey-puanlayıcı ortak etkisi (*b* x *p*) belli bir puanlayıcının belli bir bireyi diğer puanlayıcılara göre daha katı/cömert puanlayıp puanlamadığını gösterir. Çizelge 2'de görüldüğü gibi bu varyans bileşenin sıfır olması hiçbir puanlayıcının hiçbir bireyi diğer puanlayıcılara göre daha katı/cömert puanlamadıklarını göstermektedir.

6. Madde-puanlayıcı ortak etkisi (*m* x *p*) puanlayıcıların puanlamalarının bir görevden diğerine kararlı olup olmadığını göstermektedir. Çizelge 2'de verilen örnekte madde-puanlayıcı ortak etkisi için kestirilen varyans bileşenin sı-

fıra çok yakın olması nedeniyle puanlayıcıların bir görevden diğerine puanlamalarında kararlı oldukları söylenebilir.

7. Kalan etki varyansı (*b* x *m* x *p, e*) birey-madde-puanlayıcı ortak etkisi ve/veya tesadüfi hatalardan oluşur. Örnekte üçüncü büyüklükteki varyansa sahip olduğundan kalan etkinin büyük olduğu söylenebilir.

G-kuramı yukarıda da belirtildiği gibi sadece G-çalışmasıyla elde edilen varyansların bağıl büyüklüklerine dayalı olarak ana ve ortak etkilerin yorumlanmasını sağlamaz, aynı zamanda güvenirlik için G ve Phi katsayılarının hesaplanmasını da sağlar. Ayrıca alternatif K-çalışmaları ile puanlayıcı ve/veya madde sayılarının arttırılıp azaltılması sonucunda G ve Phi katsayılarındaki değişmeleri ortaya koyar. Bu yolla ölçmenin psikometrik özelliklerinden ödün vermeden en ekonomik ve verimli ölçme deseninin belirlenmesi için de kullanılır (Shavelson ve Webb, 1991; Brennan, 2001).

G-kuramında bağıl değerlendirmeler için genellenebilirlik (generalizability) katsayısı (*G* ya da $E\rho^2$) hesaplanır. Bu amaçla öncelikle bağıl hata varyansının belirlenmesi gereklidir. Bağıl değerlendirme üç tür hata kaynağından etkilenir, (a) birey-madde ortak etkisi, (b) birey-puanlayıcı ortak etkisi ve (c) kalan varyans. Bağıl hata terimi "δ" sembolü ile gösterilmek üzere bağıl hata varyansı;

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

(1)

eşitliği ile hesaplanır. G katsayısı ise;

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

(2)

eşitliği ile hesaplanır. Çizelge 2'den kestirilen varyans bileşenlerini kullanarak, bağıl hata varyansı;

$$\sigma_{\delta}^{2} = \frac{0,06049}{4} + \frac{0}{4} + \frac{0,04969}{16} = 0,018228$$

olarak hesaplanır. Bağıl hata varyansı kullanılarak ise G katsayısı;

$$G = \frac{0,13534}{0,13534 + 0,018228} = 0,88$$

olarak bulunur.

G-kuramında mutlak değerlendirmeler için Phi (Φ) (dependability) katsayısı hesaplanır. Bu amaçla öncelikle mutlak hata teriminin belirlenmesi gereklidir. Mutlak değerlendirmenin etkilendiği varyans bileşenleri; (a) madde (b) puanlayıcı ana etkileri, (c) madde-puanlayıcı, (d) birey-madde, (e) birey-puanlayıcı ortak etkileri ve (f) kalan varyanstır. Mutlak hata terimi "Δ" sembolü ile gösterilmek üzere mutlak hata varyansı;

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

(3)

eşitliği ile gösterilir. Eşitlik 3'den yararlanılarak Phi katsayısı;

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

(4)

eşitliği ile hesaplanabilir. Çizelge 2'den kestirilen varyans bileşenlerini kullanarak, mutlak hata varyansı;

$$\sigma_{\Delta}^{2} = \frac{0,01451}{4} + \frac{0}{4} + \frac{0,06049}{4} + \frac{0}{4} + \frac{0,00031}{16} + \frac{0,04969}{16} = 0,021875$$

olarak hesaplanır. Mutlak hata varyansı kullanılarak Phi katsayısı;

$$\Phi = \frac{0,13534}{0,13534 + 0,021875} = 0,86$$

olarak bulunur.

Eşitlik (1) ile tanımlanan bağıl hata ve eşitlik (3) ile tanımlanan mutlak hata formüllerinde paydada bulunan madde sayısı (n_m) ve puanlayıcı sayısı (n_p) sayıları yerine sonsuz sayıda alternatif madde ve puanlayıcı sayıları yazılabilir. Böylelikle madde ve puanlayıcı sayısındaki artma ve azalmalara bağlı olarak G ve Phi katsayıları da değişir. G-kuramı, bu yolla psikometrik özellikleri en uygun düzeye çıkarmak ve/veya bu özelliklerden taviz vermeden test uzunluğu ve puanlayıcı sayılarının en verimli ve ekonomik hale getirilmesine olanak sağlar.

Çizelge 3. Alternatif K-çalışmaları

$n_{m'}$	*2*	*6*	*2*	*6*	*2*	*4**	*6*	*2*	*6*	*2*	*6*
$n_{p'}$	2	2	3	3	4*	4*	4*	5	5	6	6
b	.135	.135	.135	.135	.135	.135	.135	.135	.135	.135	.135
m	.007	.002	.007	.002	.007	.004	.002	.007	.002	.007	.002
p	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
b x *m*	.030	.010	.030	.010	.030	.015	.010	.030	.010	.030	.010
b x *p*	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
m x *p*	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
b x *m* x *p, e*	.012	.004	.008	.003	.006	.003	.002	.005	.002	.004	.001
σ^2_δ	.043	.014	.039	.013	.037	.018	.012	.035	.012	.034	.012
σ^2_Δ	.050	.017	.046	.015	.044	.022	.015	.043	.014	.042	.014
G	.760	.905	.778	.913	.788	.881	.918	.794	.920	.797	.922
Phi	.730	.890	.747	.899	.756	.861	.903	.761	.905	.765	.907

* Orijinal madde ve puanlayıcı sayıları

np': Puanlayıcı sayısı nm': Madde sayısı

Çizelge 1'de verilen ölçme durumu için farklı puanlayıcı ve madde sayılarıyla yapılan alternatif K-çalışmaları sonucunda elde edilen mutlak ve bağıl hata varyansları ile G ve Phi katsayıları Çizelge 3'de örnek olarak verilmiştir.

SONUÇ

KTK halen günümüzde popüler olmakla birlikte, potansiyel hata kaynaklarının birden fazla olması durumunda güvenirliğin hesaplanmasının tek bir analizle yapılmasına ve ortak tek bir güvenirlik katsayısı üretilmesine olanak sağlamamaktadır. G-kuramı ise, olası hata kaynaklarının tamamını bir analizle, tek bir çalışmayla belirleyebilen, kapsamlı tek bir güvenirlik katsayısını bağıl ve mutlak değerlendirme için ayrı ayrı kestirebilen bir kuramdır. Özellikle potansiyel hata kaynaklarının birden çok olduğu ölçme durumları için G-kuramının KTK'na göre oldukça güçlü bir alternatif olduğu görülmektedir. Bu bağlamda G-kuramı, puanlayıcılar arası güvenirliğin hesaplanmasında korelasyonel teknikler ve Kappa istatistiğine göre daha güçlü ve kapsamlı bir alternatif olabilmektedir. Aynı şekilde KTK'nın test-tekrar test güvenirliği yöntemi ile iç-tutarlılık güvenirlik yöntemlerini de birleştirebilmektedir.

G-kuramı; bir ölçme durumunda, ölçmede yer alan değişkenlik kaynaklarının (facet) kabul edilebilir gözlemlerinin evreninden örneklenen ve koşul olarak adlandırılan sayılarının istenilen G ve Phi katsayıları için ne olması gerektiğini belirlenmesinde kullanılabilmektedir. Bu yolla istenilen psikometrik özelliklerde ve verimli ölçmelerin yapılması için ölçme aracında kaç maddenin bulunması,

kaç puanlayıcının puanlama yapması gerektiğine vb. karar verilmesi olanaklı olduğundan G-kuramı önemli bir alternatif olabilmektedir.

G-kuramı; geleneksel olarak geçerlik ve güvenirlik arasındaki farklılığı kısmen de olsa ortadan kaldırdığı için belli ölçüde de olsa, geçerlik çalışmaları için ayrılan zaman ve çabadan tasarruf edilmesine olanak sağlayabilmektedir.

Sonuç olarak; güçlü bir istatistiksel temeli olan G-kuramının, birçok ölçme deseni için ölçmenin psikometrik özelliklerinin belirlenmesinde ve ölçme araçlarının geliştirilmesinde klasik test kuramının yerine kullanılmasının uygun olabileceği söylenebilir.

KAYNAKÇA

Atılgan, H. (2004). Genellenebilirlik kuramı ve çok değişkenlik kaynaklı rasch modelinin karşılaştırılmasına ilişkin bir araştırma. (Yayınlanmamış Doktora Tezi) Ankara: Hacettepe Üniversitesi, Ankara.

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, yıl 5, sayı 18, 28-40.

Allal, L. (1990). Generalizability Theory. Edited by Walberg, J. H. ve Haertel, D. G. The International Encyclopedia of Educational Evaluation. Pergamon Pres, p. 274-279.

Baykul, Y. (2000). Eğitimde ve psikolojide ölçme: klasik test teorisi ve uygulaması. Ankara: ÖSYM.

Brennan, R. L. (2003). Coefficients and indices in generalizability theory. CASMA Center for advanced studies in Measurement and Assessment. Research Report No 1.

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

Brennan, R. L. (2001). Manual for mGENOWA version 2.1, Iowa Testing Programs Occasional Papers, Number 50, Iowa :College Education The University of Iowa.

Brennan, R. L. ve Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. Educational and Psychological Measurement, 41, 687-699.

Crocker, L ve Algina, J. (1986). Introduction to classical and modern test theory. Belmont CA :Wadsworth Group/Thomson Learning Inc.

Cronbach, L. J. Rajaratnam, N. ve Gleser, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137-163.

Goodwin, L. D. (2001). Interrater agreement and reliability. Measurement in Physical Education and Exercise science, 5(1), 13-14.

Lane, S. ve Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoring system for sample essays. Applied Measurement in Education, 2(3), 195-205.

Lee, G. ve Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test score composed of testlets. Applied Measurement in Education, 12(3), 237-255.

Lee, Y., Kantor, R. ve Mollaun, P. (2002). Score dependability of the writing and speaking section of new TOEFL. Educational Testing Service.

Lord, F. M. ve Novick, R. M. (1968). Statisttical theories of mental test scores. California: Addison-Wesley Publishing Company.

Lynch, B. K. ve McNamara, T. F. (1998). Using G-theory and many-facet rasch measurement in the development of performance assessments of the ESL speaking skills of imigrants. Language Testing, 15 (2) 158-180.

Nunnally, J. C. ve Bernstein, I. H. (1994). Psychometric theory. 3rd Editions, McGraw-Hill Inc.

Shavelson, R. J., ve Diğerleri. (1990). Generalizability of job performance measurement: marine corps rifleman. Military Psychology, 2 (3), 129-144.

Shavelson, R. J. ve Webb, M. N. (2003). Generalizability theory. ed. Kempf-Leonard, Kimberly. Encyclopedia of Social Measurement, San Diego: Academic Pres.

Shavelson, R. J ve Webb, M. N. (1991). Generalizability theory a prime. California: SAGE Publication, Inc.

Thorndike, R., L. (1990). Reliability. Edited by Walberg, J. H. & Haertel, D. G. The International Encyclopedia of Educational Evaluation. Pergamon Pres, p. 260-273.

VanLeeuwen, D. M. (1997). Assessing reliability of measurements with generalizability theory: an application to inter-rater reliability. Journal of Agricultural Education, Vol. 38, No. 3.

YAZAR HAKKINDA

Yrd. Doç. Dr. Atılgan, doktora eğitimini Hacettepe Üniversitesi Eğitim Bilimleri Bölümü Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda tamamlamıştır. Halen, Ege Üniversitesi'nde Eğitimde Ölçme ve Değerlendirme Anabilim dalı öğretim üyesidir. Çalışma ve araştırma konuları; Genellenebilirlik kuramı, Klasik test kuramı, Çok Yüzeyli Rasch Modeli, test-madde yanlılığı ve ölçek geliştirmedir.

İletişim adresi: Hakan Atılgan

Ege Üniversitesi, Eğitim Fakültesi

Eğitim Bilimleri Bölümü

35100 Bornova / İzmir

Telefon: 0232.3434000/5268

E-posta: hakan.atilgan@ege.edu.tr

ABOUT THE AUTHOR

Asst. Prof. Dr. Atılgan completed his Ph.D. study in Hacettepe University Educational Sciences, Educational Measurement and Evaluation Department. He is a professor of educational measurement and evaluation at Ege University. His researches and research areas include Generalizability theory, classical test theory, Many-Facets Rasch Measurement, test-item bias and scale development.

Correspondence Address: Hakan Atilgan

Ege University, Faculty of Education

Department of Educational Science

35100 Bornova / Izmir / Turkey

Phone: +90.232.3434000 / 5268

Email: hakan.atilgan@ege.edu.tr