GEFAD / GUJGEF 37(2): 591-610 (2017)

Genellenebilirlik Kuramında Gerçekleştirilen Karar Çalışmaları Ne Kadar Kararlı?*

How Consistent Are Decision Studies in G Theory?

Ömer KAMIŞ[1], Celal Deha DOĞAN[2]

[1] *Ankara Üniversitesi Eğitim Bilimleri Enstitüsü Eğitimde Ölçme ve Değerlendirme A.B.D., okamis@ankara.edu.tr*

[2]*Ankara Üniversitesi Eğitim Bilimleri Fakültesi Eğitimde Ölçme ve Değerlendirme A.B.D., ddogan@ankara.edu.tr*

Makalenin Geliş Tarihi: 11.10.2016 ***Yayına Kabul Tarihi: 21.03.2017***

ÖZ

Bu araştırmanın amacı, Genellenebilirlik Kuramı karar çalışmalarından elde edilen G ve Phi katsayıları ile gerçek durumlardan (karar çalışmasındaki yüzeylerin evrenden seçkisiz olarak seçilemediği) elde edilen G ve Phi katsayılarını karşılaştırmaktır. Araştırma temel araştırma türündedir. Çalışma grubunu 84 lisans öğrencisi ile 3 yüksek lisans öğrencisi ve 1 öğretim üyesi oluşturmuştur. Veri toplama aracı olarak araştırmacılar tarafından geliştirilen başarı testi kullanılmıştır. Verilerin analizinde Edu G 6.1-e programı kullanılmıştır. Bulgular incelendiğinde gerçek durumlarda elde edilen ve farklı karar çalışmaları sonucunda kestirilen G ve Phi katsayılarının birbirlerine yakın değerler almalarına rağmen farklılaştıkları görülmüştür. G kuramının kullanıldığı araştırmalarda yapılan karar çalışmalarında kestirilen G ve Phi katsayıları yorumlanırken, eğer evrenden seçkisiz olarak puanlayıcı seçimi pratik olarak mümkün değilse bu katsayıların gerçek durumdaki değerleriyle aynı olmayabileceğinin bilinmesi ve yorumlamada bu duruma dikkat edilmesi önerilir.

***Anahtar Sözcükler**: Genellenebilirlik Kuramı, Karar Çalışması, G Katsayısı, Phi Katsayısı.*

ABSTRACT

The aim of this research is to compare the G and Phi coefficients obtained by D studies in G theory and the actual cases for the same conditions of similar facets in D studies (when they are not randomly selected from the universe). This research is a theoretical research. The study group consists of 84 university students, 3 graduate students and 1 academician. The relevant data were gathered from the achievement test in Research Methodology Courses developed by the

*Bu çalışmanın bir bölümü, Akdeniz Üniversitesinde düzenlenen 5. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde sözlü bildiri olarak sunulmuştur.

researchers. Edu G 6.1-e programme was used to analyze the data. According to the research findings, although G and Phi coefficients obtained by D studies and actual cases were close to each other, they differ. The findings revealed that there is no systematic relationship between G and Phi coefficients obtained by D studies and actual cases. If it is not possible to select the raters randomly from the universe, G and Phi coefficients obtained by D studies may not be the same as they are in actual cases. Therefore, researchers should be careful while interpreting G and Phi coefficients in D studies.

Keywords: *Generalizability Theory, Decision Study, G Coefficient, Phi Coefficient.*

GİRİŞ

İnsanlar günlük hayatlarında çeşitli sebeplerle kararlar alırlar. Bu kararların bazılarında ölçme araçlarından yararlanırlar. Alınan kararların isabetli olabilmesi için ölçme araçlarında bazı özellikler aranır. Bu özelliklerden en önemli ikisi geçerlik ve güvenirliktir. Geçerlik, bir ölçme aracının amaca hizmet etme derecesi iken güvenirlik, ölçme aracının hatasızlık derecesi olarak tanımlanabilir (Messick, 1995; akt: Köse, 2012).

Baykul (2010)' a göre güvenirliğin duyarlılık, kararlılık ve tutarlılık gibi anlamları vardır. Duyarlılık, ölçme aracının veya ölçme sonuçlarının biriminin büyüklüğü ile ilgiliyken, kararlılık bir özelliğin aynı araçla birden çok defa ölçülmesi sonucu birbirine yakın sonuçların alınması anlamındadır. Tutarlılık ise, testi oluşturan maddelerin testin bütünüyle olan tutarlılığıdır.

Crocker ve Algina (1986)'ya göre güvenirlik, kararlılık anlamında, aynı bireylerin benzer koşullar altında tekrar test edildiğinde aynı sonuçların elde edilmesidir. İşe vuruk terim olarak güvenirlik, bireylerin sapma puanlarının veya z puanlarının aynı testin tekrarlanan veya paralel (eşdeğer form) uygulamaları sonucu nispeten tutarlı kalması olarak tanımlanabilir. Ölçme sonucunda elde edilen puanlara testin kapsamı, dikkatsizlik gibi farklı sebeplerden hatalar karışabilir. Ölçme hataları geniş anlamda tesadüfi ve sistematik olarak sınıflandırılabilir. Sistematik hatalar, kişilerin bazı özel karakteristiklerinden dolayı onların bireysel puanlarını tutarlı olarak etkileyen hatalarken, tesadüfi hatalar bireysel puanları tamamen şans eseri etkiler. Testi alan bireylerin puanlarını pozitif ya da negatif yönde etkileyebilir. Tesadüfi hatalar, varsayımda bulunmak, dikkatsizlik, uygulama hataları, kapsam örnekleme, puanlama hataları gibi nedenlerden kaynaklanabilir. Hem sistematik hem de tesadüfi hatalar puanların yorumlanmasında sorun oluşturur. Sistematik hatalar ölçme aracının tutarlılığı üzerinde değil test puanlarının geçerliği üzerinde etkilidir ve ölçme aracının pratik yararını zedeler. Tesadüfi hatalar ise test puanlarının hem kullanışlılığını hem de tutarlılığını azaltır.

Klasik Test Kuramı'na (KTK) göre ölçme hatası, testi alan bireylerin gözlenen puanları ile gerçek puanları arasındaki farktır. KTK' ye göre bu fark ne kadar az olursa ölçme

sonucu elde edilen puanların güvenirliği de o kadar yüksek olur. KTK' ye göre bir ölçme sonucu, içindeki tesadüfi hataların azlığı oranında güvenilirdir (Turgut ve Baykul, 2011). Klasik Test Kuramı'na göre güvenirlik tahmin yöntemleri; eşdeğer formlar yöntemi, test-tekrar test yöntemi, iki yarıya bölme yöntemi, Cronbach alpha ve Kuder-Richardson (KR20-21) formülleridir. Bu yöntemler güvenirliği katsayı olarak veren yöntemlerdir. Bunun haricinde bir de güvenirliğin puan olarak kestirimi vardır. Bu da ölçmenin standart hatası olarak isimlendirilir (Erkuş, 2003). Bu yöntemlerden eşdeğer formlar ve test-tekrar test yöntemi iki test uygulaması gerektirirken, iki yarıya bölme yönteminde testin bir kere uygulanması yeterlidir. Cronbach alpha ve KR20-21 yöntemleri ise madde kovaryanslarına dayanan yöntemlerdir.

KTK' de kullanılan bu güvenirlik tahmin yöntemlerinin her birinde ayrı bir hata kaynağı ele alınıp incelenmektedir. Örneğin eşdeğer formlar yönteminde hata kaynağı testin kapsamı olurken, test-tekrar test yönteminde hata kaynağı zaman ve test uygulama koşullarıdır. KTK'nin bu sınırlılığına karşı çoklu hata kaynaklarını tek bir seferde ayırt eden Genellenebilirlik Kuramı (GK), ilk olarak Cronbach, Rajaratman ve Gleser tarafından ortaya atılmış ardından Brennan, Shavelson ve Webb'in katkılarıyla geliştirilmiştir (Güler, Uyanık ve Teker, 2012).

G Kuramı, KTK ve varyans analizinin (ANOVA) bir devamıdır. Kabul edilebilir gözlemler evreni, genellenebilirlik (G) çalışması, genelleme evreni, karar (K) çalışması gibi kavramsal konuları içerir. Brennan (2001)'e göre G Kuramı'nın belki de en önemli yönü ve eşsiz özelliği içerdiği bu kavramsal konulardır. Genellenebilirlik analizleri, sadece çeşitli hata kaynaklarının göreceli önemini anlamak için değil aynı zamanda verimli ölçme süreçleri tasarlamada da kullanılır. G Kuramı'nın bir diğer avantajı da çok sayıda hata kaynağının tek bir analizle ayrı ayrı belirlenebilmesidir. Aynı zamanda G Kuramı ile bireyler hakkında göreli ve mutlak kararlar alınabilmektedir (Shavelson ve Webb, 1991).

G Kuramı'nda bireyler hakkında göreli ve mutlak kararlar almak için G ve Phi katsayıları hesaplanmaktadır. G katsayısı göreli kararlar, Phi katsayısı mutlak kararlar için hesaplanmaktadır. Bu katsayıları hesaplamak için ise öncelikle göreli ve mutlak hata

varyansları hesaplanır. Göreli hata varyansını hesaplamak için ölçme objesi ile ilişkili varyans değerleri kullanılırken mutlak hatayı hesaplamak için ölçme objesi dışında kalan tüm varyans değerleri kullanılır (Brennan, 2001; Shavelson ve Webb, 1991). G Kuramı'nda bir ölçme durumuna ilişkin maddeler, bireyler, puanlayıcılar, zaman gibi etkenler birer değişkenlik kaynağıdır. Genellenebilirlik Kuramı'nda G ve K çalışması olmak üzere iki ayrı çalışma yapılmaktadır. G çalışmasının amacı, ölçmedeki değişkenlik kaynaklarına ilişkin varyans değerlerini hesaplayarak mümkün olduğu kadar fazla bilgi sağlamaktır. K çalışmalarında ise G çalışmasında elde edilen bilgiler kullanılarak yeterli düzeyde güvenirliğe sahip ölçme durumları oluşturmak için değişkenlik kaynaklarının sayısının ne olması gerektiği hakkında kestirimlerde bulunulur.

K çalışmaları neticesinde elde edilen G ve Phi katsayıları gerçekleştirilecek yeni uygulamada ilgili değişkenlik kaynaklarının (puanlayıcıların, maddelerin vb.) evrenden seçkisiz olarak seçilmesini gerekli kılar. Puanlayıcıların evrenden seçkisiz seçilmesi kavramı ile ilgili çalışmayı puanlayabilecek kişilerin olduğu bir evrenin tanımlanması ve o evrenden puanlayıcıların seçkisiz bir şekilde seçilmesi kastedilmektedir. Örneğin matematik becerisini ölçmek için kurum içinde gerçekleştirilen bir sınavın puanlanması sürecinde hâlihazırda kurumda çalışan 5 puanlayıcının kullanılması G kuramında bahsedilen puanlayıcıların seçkisiz seçilmesi durumunu tam olarak karşılamamaktadır.

Puanlayıcıların evrenden seçkisiz olarak seçilmesi geniş ölçekli ölçme uygulamalarında mümkün olabilmekle beraber kurum içi ve sınıf içi ölçme uygulamalarında pratik olarak mümkün olamamaktadır. Performansa dayalı sınavları içeren tıp, sağlık, beden eğitimi gibi fakültelerde, yazma ve konuşma gibi becerilerin ölçüldüğü üniversitelerin İngilizce hazırlık okullarında birden fazla puanlayıcının olduğu ve farklı yuvalanmış desenlerin kullanılabileceği pek çok gerçek ölçme durumu söz konusudur. Alanyazındaki G ve K çalışmalarının büyük bir çoğunluğu da bu ve benzer durumlarda gerçekleştirilmektedir (Atılgan ve Tezbaşaran, 2005; Deliceoğlu ve Çıkrıkçı Demirtaşlı, 2012; Güler 2009; Lin ve Zhang, 2014; Nalbantoğlu Yılmaz ve Başusta, 2015;). Bu durumlarda K çalışması neticesinde elde edilen G ve Phi katsayılarının gerçek durumları ne düzeyde doğru kestirdiği bir soru işaretidir.

Alanyazında kurum içi veya sınıf içi uygulamalar için puanlayıcıların veya maddelerin varyans kaynağı olarak ele alındığı Genellenebilirlik Kuramı'na ilişkin gerçekleştirilen pek çok çalışma bulunmaktadır. Bu çalışmaların birçoğu G çalışması sonrası gerçekleştirilen K çalışmalarını içermektedir (Anıl ve Büyükkıdık, 2012; Arterberry, Martens, Cadigan ve Smith, 2012; Büyükkıdık ve Anıl, 2015; Güler, Eroğlu ve Akbaba, 2014; Hoyt ve Melby, 1999; Yelboğa, 2012;). K çalışması neticesinde elde edilen G ve Phi katsayılarının gerçek durumda elde edilenlerle karşılaştırıldığı sadece bir çalışmaya rastlanmıştır (Atılgan ve Tezbaşaran, 2005). Bu çalışmada bir programa öğrenci seçmek amacı ile yapılan özel yetenek seçme sınavlarının iki ardışık yıldaki verileri kullanılmış ve farklı puanlayıcı sayıları için karar çalışmalarından ve gerçek durumlardan elde edilen G ve Phi katsayıları karşılaştırılmıştır. Ancak puanlayıcıların evrenden seçkisiz seçilmesinin mümkün olmadığı gerçek durumlarda ve K çalışmaları sonucunda elde edilen G ve Phi katsayılarının karşılaştırıldığı bir çalışma bulunmamaktadır.

Kurum içi ve sınıf içi gerçekleştirilen G ve K çalışmalarında değişkenlik kaynaklarının evrenden seçkisiz seçilmesi mümkün olmamaktadır. Bu koşulun sağlanamadığı durumlarda K çalışması neticesinde elde edilen G ve Phi katsayılarının gerçek durumları ne derece doğru kestirdiğinin belirlenmesi K çalışmalarının daha doğru yorumlanmasına ve daha nitelikli uygulamalar yapılmasına katkı getirecektir. Çalışmanın ileriki kısımlarında "gerçek durum" kavramı ile değişkenlik kaynaklarının evrenden seçkisiz olarak seçilmesinin mümkün olmadığı gerçek durumlar kastedilmiştir.

Bu çalışmanın amacı, Genellenebilirlik Kuramı karar çalışmaları ile kestirilen G ve Phi katsayıları ile karar çalışmasındaki yüzeyler ve koşullara ilişkin gerçek durumlardan elde edilen G ve Phi katsayılarını karşılaştırmaktır. Bu genel amaç doğrultusunda aşağıdaki sorulara yanıt aranacaktır:

1. Puanlayıcı sayısının;
   a. Gerçekte 2 olduğu durumda elde dilen G ve Phi katsayıları nedir?
   b. Gerçekte 3 olduğu durumda yapılan karar çalışması sonucunda 2 puanlayıcı için kestirilen G ve Phi katsayıları nedir?

c. Gerçekte 4 olduğu durumda yapılan karar çalışması sonucunda 2 puanlayıcı için kestirilen G ve Phi katsayıları nedir?

2. Puanlayıcı sayısının;

a. Gerçekte 3 olduğu durumda elde dilen G ve Phi katsayıları nedir?

b. Gerçekte 2 olduğu durumda yapılan karar çalışması sonucunda 3 puanlayıcı için kestirilen G ve Phi katsayıları nedir?

c. Gerçekte 4 olduğu durumda yapılan karar çalışması sonucunda 3 puanlayıcı için kestirilen G ve Phi katsayıları nedir?

3. Puanlayıcı sayısının;

a. Gerçekte 4 olduğu durumda kestirilen G ve Phi katsayıları nedir?

b. Gerçekte 2 olduğu durumda yapılan karar çalışması sonucunda 4 puanlayıcı için kestirilen G ve Phi katsayıları nedir?

c. Gerçekte 3 olduğu durumda yapılan karar çalışması sonucunda 4 puanlayıcı için kestirilen G ve Phi katsayıları nedir?

YÖNTEM

Araştırma Tür ve Modeli

Araştırma, temel araştırma türünde betimsel tarama modeli ile tasarlanmıştır. "Temel araştırmalar, salt amacı var olan bilgiye yenilerini katmak olan araştırmalardır. Tarama modelleri ise, geçmişte ya da halen var olan bir durumu var olduğu şekliyle betimlemeyi amaçlayan araştırma yaklaşımıdır" (Karasar, 2012, s.24). Bu araştırmada puanlayıcı sayısındaki değişimin, karar çalışmalarından ve gerçek durumlardan elde edilen G ve Phi katsayılarında meydana getirdiği değişim incelenmiştir.

Araştırma kapsamında öncelikle çalışma grubundaki öğrencilere uygulanan başarı testi 4 puanlayıcı tarafından değerlendirilmiştir. Akabinde 4 puanlayıcı arasından 2 tanesinin, 3 tanesinin rastgele seçildiği ve 4 puanlayıcının tamamının kullanıldığı durumlar için G ve Phi katsayıları hesaplanmıştır. Son olarak hesaplanan bu katsayılar karar çalışmalarında 2, 3 ve 4 puanlayıcının olduğu durumlar için kestirilen katsayılarla kararlaştırılmıştır.

Çalışma Grubu

Araştırmanın çalışma grubunu Ankara Üniversitesi Eğitim Bilimleri Fakültesinin Sınıf Öğretmenliği Bölümünde öğrenim gören 84 ikinci sınıf öğrencisi ile Ankara Üniversitesi Ölçme ve Değerlendirme Anabilim Dalında görev yapan biri öğretim üyesi, üçü yüksek lisans öğrencisi olan 4 puanlayıcı oluşturmuştur.

Verilerin Toplanması

Çalışma grubunu oluşturan öğrencilere ve puanlayıcılara çalışma hakkında kısa bilgi verilmiştir. Araştırmacılar tarafından önceden hazırlanan başarı testi öğrencilere uygulanmış ardından uygulanan başarı testi 4 puanlayıcı tarafından bağımsız bir şekilde puanlanarak veriler toplanmıştır.

Bu uygulamada, her öğrenciye dağıtılan test aynı maddeleri içermiş ve 4 puanlayıcının her biri, her öğrencinin her bir maddeye verdiği yanıtları yine araştırmacılar tarafından geliştirilen dereceleme ölçeğini kullanarak puanlamıştır. Buna göre çalışmada Genellenebilirlik Kuramı'ndaki desenlerden tümüyle çaprazlanmış desen (bxmxp) kullanılmıştır.

Veri Toplama Aracı

Başarı testi, "Bilimsel Araştırma Yöntemleri" dersinde öğrenilen bilgi ve becerileri yoklamaya (ölçmeye) yönelik hazırlanan 4 adet açık uçlu sorudan oluşmuş olup araştırmacılar tarafından geliştirilmiştir. Birinci soruda öğrencilerden, verilen bir örnek durum üzerinden araştırmanın sayıltı ve sınırlılıklarını; ikinci soruda ise örnek araştırma soruları üzerinden bir araştırma probleminde olması gereken nitelikleri belirlemeleri istenmiştir. Üçüncü soruda öğrencilerden, verilen örnek duruma ilişkin hipotez yazmaları ve değişkenleri belirlemeleri, dördüncü soruda ise yarı deneysel ve gerçek deneysel desen oluşturmak için yapılması gereken eylemleri belirtmeleri istenmiştir. Başarı testindeki her bir maddeyi puanlamak için araştırmacılar tarafından geliştirilen ve beş düzeyden oluşan bir dereceleme ölçeği geliştirilmiştir. Buna göre, bir öğrencinin bu testten alabileceği en yüksek puan 20 olurken en düşük puan 4 olmaktadır.

Verilerin Analizi

Karar çalışmasından ve karar çalışmasındaki yüzey ve koşullara ilişkin gerçek durumlardan elde edilen varyans değerleri kullanılarak göreli ve mutlak hata varyansları hesaplanmıştır. Göreli hata varyansı dikkate alınarak G katsayısı, mutlak hata varyansı dikkate alınarak da Phi katsayısı hesaplanmıştır. Elde edilen verilerin çözümlenmesinde Edu G 6.1-e programı kullanılmıştır.

BULGULAR

Bu bölümde öncelikle test hakkındaki betimsel istatistiklerden ortalama ve standart sapma değerlerine kısaca değinilmiş ardından 2, 3 ve 4 puanlayıcı için tümüyle çaprazlanmış desendeki değişkenlik kaynaklarına ilişkin varyans değerleri verilmiştir. Son olarak da araştırma sorularına ilişkin bulgular sunulup kısaca yorumlanmıştır.

Puanlayıcı sayısının 2, 3 ve 4 olduğu durumlarda çalışma grubu verilerine ilişkin ortalama ve standart sapma değerleri Tablo 1'de verilmiştir. Tablo oluşturulurken öncelikle her puanlayıcı için öğrencilerin bir puanlayıcıdan toplam kaç puan aldığı belirlenmiş ardından bu puanlayıcılar üzerinden ortalama alınarak öğrencilerin toplam ortalama puanları elde edilmiştir. Gruba ait ortalama ve standart sapma değerleri bu ortalama toplam puanlar üzerinden hesaplanmıştır.

Tablo 1. Puanlayıcı Sayısının 2, 3 ve 4 Olduğu Durumlarda Çalışma Grubu Verilerine İlişkin Ortalama ve Standart Sapma Değerleri

İstatistikler	Puanlayıcı Sayısı		
	2	3	4
Ortalama	10.62	11.11	10.92
Standart Sapma	2.74	2.91	2.84

Tablo 1 incelendiğinde, 2, 3 ve 4 puanlayıcının kullanıldığı durumlar için çalışma grubundaki öğrencilerin testten aldığı puanlara ait ortalama ve standart sapma değerlerinin küçük farklılıklarla birlikte birbirine yakın olduğu görülmektedir. Buna göre,

kabaca, farklı puanlayıcı sayılarında öğrencilerin testten aldıkları puanların çok fazla değişmediği söylenebilir.

Puanlayıcı sayısının 2, 3 ve 4 olduğu durumlara ilişkin tümüyle çaprazlanmış desendeki değişkenlik kaynaklarına ait varyans değerleri, serbestlik dereceleri, kareler ortalaması ve varyans yüzdeleri Tablo 2'de verilmiştir.

Tablo 2. Puanlayıcı Sayısının 2, 3 ve 4 Olduğu Durumlarda Tümüyle Çaprazlanmış Desendeki Değişkenlik Kaynaklarına Ait Varyans Değerleri

Puanlayıcı Sayısı	Değişkenlik Kaynağı	sd	Kareler Ortalaması	Hesaplanan Varyans Değeri	%
	B	83	3.76	0.29	19.86
	M	3	42.73	0.20	13.70
	P	1	3.01	-0.02	0.00
2	BM	249	1.45	0.56	38.36
	BP	83	0.31	-0.00	0.00
	MP	3	8.72	0.10	6.85
	BMP,E	249	0.33	0.33	22.60
	B	83	6.37	0.32	20.78
	M	3	40.98	0.14	9.09
	P	2	0.84	-0.01	0.00
3	BM	249	2.42	0.69	44.80
	BP	166	0.41	0.01	0.65
	MP	6	3.79	0.04	2.60
	BMP,E	498	0.35	0.35	22.73
	B	83	8.09	0.33	21.02
	M	3	71.28	0.19	12.10
	P	3	3.54	-0.00	0.00
4	BM	249	2.86	0.62	39.49
	BP	249	0.38	0.00	0.00
	MP	9	4.27	0.05	3.18
	BMP,E	747	0.38	0.38	24.20

Tablo 2 incelendiğinde, puanlayıcı sayısının 2, 3 ve 4 olduğu durumların hepsinde de BM etkileşimli varyans bileşen değeri toplam varyans içinde en yüksek orana sahiptir. Buna göre, farklı puanlayıcı sayılarının hepsinde de maddelerin güçlük düzeylerinin bireyden bireye farklılık gösterdiği söylenebilir. Toplam varyans içinde ikinci en yüksek oran BMP,E artık bileşenine aittir. Bu duruma göre bireyler, maddeler ve puanlayıcılar arası etkileşim ile bu araştırmada ölçülemeyen sistematik veya sistematik olmayan değişkenlik

kaynaklarının olduğu söylenebilir. Birey değişkenlik kaynağına ilişkin varyans değeri ise her üç durumda da toplam varyans içinde üçüncü sıradadır. Buna göre, bireylerin birbirinden farklılık gösterdiği söylenebilir. Diğer bir deyişle yapılan ölçme işlemi bireyleri birbirinden ayırmada kısmen de olsa başarılıdır. Her üç durumda da puanlayıcı varyansının toplam varyans içindeki oranı sıfır (0.00) çıktığı için puanlayıcıların puanlamaları arasında bir farklılık olmadığı ifade edilebilir. Başka bir ifadeyle, puanlayıcılar genel olarak birbirleri ile tutarlı puanlamalar yapmışlardır.

Puanlayıcı Sayısının Gerçekte ve Karar Çalışmalarında 2 Olduğu Duruma İlişkin Bulgular

Puanlayıcı sayısının gerçekte 2 olduğu ve karar çalışmalarında 2 puanlayıcı için kestirim yapıldığı durumlarda elde edilen G ve Phi katsayıları Tablo 3'te, verilmiştir.

Tablo 3. Puanlayıcı Sayısının 2 Olduğu Durum İçin Gerçekte Elde Edilen ve Karar Çalışmalarında Kestirilen G ve Phi Katsayıları

Puanlayıcı Sayısı	Gerçek Durum		Karar Çalışması (2 puanlayıcı için kestirilen)	
	G	Phi	G	Phi
2	0.62	0.54	-	-
3	-	-	0.59	0.55
4	-	-	0.62	0.56

Tablo 3 incelendiğinde, puanlayıcı sayısının 2 olduğu gerçek durumda elde edilen G ve Phi katsayılarının sırasıyla 0.62 ve 0.54 olduğu görülmektedir. Puanlayıcı sayısının 3 olduğu durumda yapılan karar çalışması ile 2 puanlayıcı için kestirilen G ve Phi katsayıları sırasıyla 0.59 ve 0.55'tir. Bu bulguya dayalı olarak gerçek durum için elde edilen G katsayısının (0.62) karar çalışmasında kestirilen G katsayısından (0.59) yüksek olduğu belirtilebilir. Bunun yanı sıra gerçek durumda elde edilen Phi katsayısı ise (0.54) karar çalışmasında kestirilen Phi katsayısından (0.55) daha düşüktür.

Puanlayıcı sayısının 4 olduğu durumda yapılan karar çalışması sonucunda 2 puanlayıcı için kestirilen G ve Phi katsayıları ise sırasıyla 0.62 ve 0.56'dır. Buna göre karar çalışmasında elde edilen G katsayısı (0.62) gerçek durumda elde edilen (0.62) ile aynı değere sahiptir. Buna karşın karar çalışmasında elde edilen Phi katsayısı (0.56) ise gerçek durumda elde edilenden (0.54) daha yüksektir. Her iki karar çalışmasında elde edilen Phi

katsayıları gerçek durumda elde edilenden daha yüksek iken G katsayıları ise gerçek durumdakinden daha düşüktür veya eşittir.

Puanlayıcı sayısının gerçekte ve karar çalışmalarında 3 olduğu duruma ilişkin bulgular

Puanlayıcı sayısının gerçekte 3 olduğu ve karar çalışmalarında 3 puanlayıcı için kestirim yapıldığı durumlarda elde edilen G ve Phi katsayıları Tablo 4'te, verilmiştir.

Tablo 4. Puanlayıcı Sayısının 3 Olduğu Durum İçin Gerçekte Elde Edilen ve Karar Çalışmalarında Kestirilen G ve Phi Katsayıları

Puanlayıcı Sayısı	Gerçek Durum		Karar Çalışması (3 puanlayıcı için kestirilen)	
	G	Phi	G	Phi
2	-	-	0.63	0.56
3	0.61	0.57	-	-
4	-	-	0.64	0.58

Tablo 4 incelendiğinde, puanlayıcı sayısının 3 olduğu gerçek durumda elde edilen G ve Phi katsayılarının sırasıyla 0.61 ve 0.57 olduğu görülmektedir. Puanlayıcı sayısının 2 olduğu durumda yapılan karar çalışmasında 3 puanlayıcı için kestirilen G ve Phi katsayıları sırasıyla 0.63 ve 0.56'dır. Buna göre karar çalışmasında kestirilen G katsayısı (0.63) gerçek durumda elde edilenden (0.61) daha yüksektir. Ancak karar çalışmasında kestirilen Phi katsayısı ise (0.56) gerçek durumda elde edilenden (0.57) daha düşüktür.

Puanlayıcı sayısının 4 olduğu durumda yapılan karar çalışmasında 3 puanlayıcı için kestirilen G ve Phi katsayıları ise sırasıyla 0.64 ve 0.58'dir. Bu bulguya göre, karar çalışasında kestirilen G ve Phi katsayılılarının (G:0.64, Phi:0.58) gerçek durumda elde edilenden (G:0.61, Phi:0.57) daha yüksek olduğu belirtilebilir.

Puanlayıcı sayısının gerçekte ve karar çalışmalarında 4 olduğu duruma ilişkin bulgular

Puanlayıcı sayısının gerçekte 4 olduğu ve karar çalışmalarında 4 puanlayıcı için kestirim yapıldığı durumlarda elde edilen G ve Phi katsayıları Tablo 5'te verilmiştir.

Tablo 5. Puanlayıcı Sayısının 4 Olduğu Durum İçin Gerçekte Elde Edilen ve Karar Çalışmalarında Kestirilen G ve Phi Katsayıları

Puanlayıcı Sayısı	Gerçek Durum		Karar Çalışması (4 puanlayıcı için kestirilen)	
	G	Phi	G	Phi
2	-	-	0.64	0.57
3	-	-	0.62	0.58
4	0.65	0.59	-	-

Tablo 5 incelendiğinde, puanlayıcı sayısının 4 olduğu gerçek durumda elde edilen G ve Phi katsayılarının sırasıyla 0.65 ve 0.59 olduğu görülmektedir. Puanlayıcı sayısının 2 olduğu durumda yapılan karar çalışmasında 4 puanlayıcı için kestirilen G ve Phi katsayıları sırasıyla 0.64 ve 0.57'dir. Karar çalışmasında kestirilen G ve Phi katsayılarının (G:0.64, Phi:0.57) gerçek durumda elde edilenlerden (G:0.65, Phi:0.59) daha düşük olduğu belirtilebilir.

Puanlayıcı sayısının 3 olduğu durumda yapılan karar çalışmasında 4 puanlayıcı için kestirilen G ve Phi katsayıları ise sırasıyla 0.62 ve 0.58'dir. Bu bulguya göre karar çalışmasında kestirilen G ve Phi katsayılarının (G:0.62, Phi:0.58) 2 puanlayıcı için kestirilen durumda olduğu gibi gerçek durumda elde edilenlerden (G:0.65, Phi:0.59) daha düşük olduğu belirtilebilir.

Bulgular incelendiğinde puanlayıcıların evrenden seçkisiz seçilmediği gerçek durumlarda elde edilen ve farklı karar çalışmaları sonucunda kestirilen G ve Phi katsayılarının birbirlerine yakın değerler almalarına rağmen farklılaştıkları görülmektedir.

TARTIŞMA ve SONUÇ

Bu çalışmada bir ölçme durumu oluşturulmuş ve madde sayısı sabit tutularak puanlayıcı sayısının 2, 3 ve 4 olduğu durum için puanlayıcıların evrenden seçkisiz seçilmediği gerçek durumlarda elde edilen ve karar çalışmalarında kestirilen G ve Phi katsayıları karşılaştırılmaya çalışılmıştır. Yapılan analizler sonucunda gerçek durumda elde edilen ve karar çalışmalarında kestirilen değerlerin birbirine yakın olmakla birlikte farklılaştıkları belirtilebilir. Bu çalışmadaki farklıkların az olmasının nedeni olarak, puanlayıcı sayısının 2, 3 ve 4 olduğu durumlar için yapılan G çalışmalarında puanlayıcı varyansının toplam varyans içindeki yüzdesinin çok düşük (%0,00) olması gösterilebilir.

G Kuramı'nda yapılan karar çalışmalarında, puanlayıcı sayısının sürekli olarak artması ile daha yüksek G ve Phi katsayılarının elde edilebileceği öngörülmektedir. Bunun altında varyans bileşen değerlerinin sabit tutulması yatmaktadır. Bu durum puanlayıcıların evrenden seçkisiz olarak seçilebildiği durumlar için geçerli olabilir. Ancak bu çalışmanın bulgularına göre puanlayıcıların evrenden seçkisiz seçilemediği gerçek durumlarda böyle bir örüntüden bahsetmek mümkün değildir. Bu süreçte eklenecek puanlayıcının özellikleri, varyans bileşen değerlerini artıracak ya da azaltacak yönde değiştirebilmektedir.

Ayrıca çalışmada elde edilen bulgular karar çalışmalarına dayalı olarak puanlayıcıların evrenden seçkisiz seçilmediği gerçek durumda elde edilebilecek G ve Phi katsayılarının sistematik bir şekilde yordanamayacağını ortaya koymuştur. Eklenecek veya çıkarılacak puanlayıcı özellikleri bilinmediğinden karar çalışmalarında kestirilen G ve Phi katsayıları gerçek durumdakinden yüksek veya düşük çıkabilir. Aynı durum madde eklendiği ve çıkarıldığı durumlarda da söz konusu olabilir.

Atılgan ve Tezbaşaran (2005) ise yaptıkları çalışma neticesinde karar çalışmalarında mevcut puanlayıcı sayısından daha çok puanlayıcının olduğu durum için kestirilen G ve Phi katsayılarının, gerçek durumda elde edilenlerden daha yüksek olduğu, daha az puanlayıcının olduğu durum için kestirilen G ve Phi katsayılarının gerçek durumda elde edilenlerden daha düşük olduğu bulgusuna ulaşmışlardır Bu bulgu mevcut araştırma

bulguları ile örtüşmemektedir. Mevcut araştırmada gerçek durumda ve K çalışmasında elde edilen katsayılar arasında sistematik bir ilişkinin olmadığı belirlenmiştir. İki araştırma bulguları arsındaki farklılık planlayıcıların evrenden seçkisiz bir şekilde seçilip seçilmemesinden kaynaklanabilir.

Bu çalışmada gerçek durumlarda ve karar çalışmalarında elde edilen G ve Phi katsayıları birbirine yakın olmakla beraber farklılaşmaktadır. Tüm koşullarda elde edilen G ve Phi katsayıları genel olarak orta düzeyde güvenirliğe işaret etmektedir. Ancak bu ufak farklılıklar ilgili G ve Phi katsayıları sınır değerlere yakın olduğu zaman (örn. 0.80) sorun yaratabilir.

Günümüzde performansa dayalı sınavlar içeren pek çok eğitim kurumunda (tıp, sağlık, beden eğitimi, güzel sanatlar fakülteleri, üniversitelerin hazırlık okulları vb.) birden fazla puanlayıcının yer aldığı ve farklı yuvalanmış desenlerin kullanılabileceği pek çok ölçme durumu söz konusudur. Bu bağlamda pek çok G ve K çalışması da kurum içi ve sınıf içi uygulamalar için gerçekleştirilmiştir. (Can-Aran, Güler ve Senemoğlu, 2014; Büyükkıdık ve Anıl, 2015; Güler, Eroğlu ve Akbaba, 2014; Hoyt ve Melby, 1999; Nalbantoğlu Yılmaz ve Gelbal, 2011; Yelboğa, 2012). Ancak bu durumlarda gerçekleştirilen bir K çalışması sonucuna göre yeniden tasarlanacak uygulamalarda ilgili değişkenlik kaynaklarının (örneğin puanlayıcı) evrenden seçkisiz olarak seçilmesi pratik koşullardan dolayı mümkün değildir. Bu gibi durumlarda, örneğin K çalışması sonucuna göre puanlayıcının arttırılması gerekiyorsa, uygulayıcılar kurum içerisinden yeni bir puanlayıcıyı mevcut puanlayıcılara eklemekte veya yine kurum içerisinden yeni puanlayıcılar seçmektedirler. Her iki durumda da puanlayıcılar evrenden seçkisiz bir şekilde seçilememektedir. Mevcut araştırma bulguları bu durumlarda K çalışmalarının gerçek durumları sistematik bir şekilde kestiremediğini ortaya koymuştur.

Bu bağlamda K çalışması neticesinde yeni bir çalışma gerçekleştirecek uygulayıcıların bu sınırlılığın farkında olması ve K çalışmasının sonuçlarını dikkatli yorumlaması gerekmektedir. Ayrıca K çalışması gerçekleştiren araştırmacıların ilgili değişkenlik kaynaklarının evrenden seçkisiz seçilmediği durumlarda K çalışması sonuçlarının gerçek durumları doğru kestiremeyebileceğini raporlaması önerilmektedir.

Araştırmacılara puanlayıcı varyansının toplam varyans içindeki yüzdesinin %0,00'dan farklı olduğu durumlar için benzer çalışmalar yapılarak gerçek durumda elde edilen ve karar çalışmalarında kestirilen G ve Phi katsayılarının karşılaştırıldığı çalışmaları gerçekleştirmeleri önerilir. Bunun yanı sıra hem madde sayısının hem de puanlayıcı sayısının değiştiği durumlar için gerçek durumda elde edilen ve karar çalışmalarında kestirilen G ve Phi katsayılarının karşılaştırıldığı çalışmaların yapılması önerilebilir. Bu araştırmada puanlayıcılar puanlayıcı evreninden seçkisiz olarak seçilmemiştir. Araştırmacılara puanlayıcıların evrenden seçkisiz seçildiği durumlar için de gerçekte elde edilen ve karar çalışmalarında kestirilen G ve Phi katsayılarının karşılaştırıldığı çalışmaları yapmaları önerilir.

KAYNAKLAR

Algina, J. ve Crocker, L. (1986). *Introduction to classical and modern test theory*. United Satates: Cengage Learning.

Anıl, D. ve Büyükkıdık, S. (2012). Genellenebilirlik kuramında dört facetli karışık desen kullanımı için örnek bir uygulama. *Eğitim ve Psikolojide Ölçme ve Değerlendirme Dergisi, 3*(2). 291-296.

Arterberry, B. J., Martens, M. P., Cadigan, J. M. and Smith, A. E. (2012). Assessing the dependability of drinking motives via generalizability theory. *Measurement and Evaluation in Counseling and Development, 45*(4), 292-302.

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. *Eurasian Journal of Educational Research, 18*, 236-252.

Baykul, Y. (2010). *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması (2. Baskı)*. Ankara: Pegem Akademi.

Brennan, L. R. (2001). *Generalizability theory, statistics for social science and public policy*. New York: Springer-Verlag.

Büyükkıdık, S. ve Anıl, D. (2015). Performansa dayalı durum belirlemede güvenirliğin genellenebilirlik kuramında farklı desenlerde incelenmesi. *Eğitim ve Bilim, 40*(177), 285-296.

Can-Aran, Ö., Güler, N. ve Senemoğlu, N. (2014). Öğrencilerin disiplinli zihin özelliklerini belirlemede kullanılan dereceli puanlama anahtarının genellenebilirlik kuramı açısından değerlendirilmesi. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, 42*, 165-171.

Deliceoğlu, G. ve Çıkrıkçı Demirtaşlı, N. (2012). Futbol yetilerine ilişkin dereceleme ölçeğinin güvenirliğinin genellenebilirlik kuramına ve klasik test kuramına dayalı olarak karşılaştırılması. *Hacettepe Spor Bilimleri Dergisi, 23*(1), 1-12.

Erkuş, A. (2003). *Psikometri üzerine yazılar*. Ankara: Türk Psikologlar Derneği Yayınları

Güler, N. (2009). Genellenebilirlik kuramı ve spss ile genova programlarıyla hesaplanan g ve k çalışmalarına ilişkin sonuçların karşılaştırılması. *Eğitim ve Bilim, 34* (154), 93-103.

Güler, N. Eroğlu, Y. ve Akbaba, S. (2014). Genellenebilirlik kuramına göre ölçüt bağımlı ölçme araçlarında güvenirlik: yemek yeme becerileri örneğinde bir uygulama. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 14* (2), 217-232.

Güler, N., Uyanık, G. ve Teker, G. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi.

Hoyt, W. T. and Melby, J. N. (1999). Dependability of measurement in counseling psychology: an introduction to generalizability theory. *The Counseling Psychologist, 27*(3), 325-352.

Karasar, N. (2012). *Bilimsel araştırma yöntemleri.* Ankara: Nobel Yayıncılık.

Köse, İ. A. (2012). Ölçme ve değerlendirmede temel kavramlar. N. Çıkrıkçı-Demirtaşlı (Ed.). *Eğitimde ölçme ve değerlendirme* içinde (s.72-113). Ankara: Elhan Yayınları.

Lin, C. K. ve Zhang, J. (2014). Investigating correspondence between language proficiency standarts and academic content standarts: a generalizability theory study. *Language Testing, 31*(4), 413-431.

Nalbantoğlu Yılmaz, F. ve Başusta, B. (2015). Genellenebilirlik kuramıyla dikiş atma ve alma becerileri istasyonu güvenirliğinin değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6*(1), 107-116.

Nalbantoğlu Yılmaz, F. ve Gelbal, S. (2011). İletişim becerileri istasyonu ölçeğinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 41*, 509-518.

Shavelson, R. J. ve Webb, N. M. (1991). *Generalizability theory a primer*. United States of America: Sage Publications.

Turgut, M. F. ve Baykul, Y. (2011). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi.

Yelboğa, A. (2012). Genellenebilirlik kuramına göre iş performansı ölçeklerinde güvenirlik. *Eğitim ve Bilim, 37* (163), 157-164.

SUMMARY

The aim of this research is to compare the G and Phi coefficients obtained by D studies in G theory and the actual cases for the same conditions of similar facets in D studies when they are not randomly selected from the universe. When the literature was examined, very little deal of research was found comparing G and Phi coefficients obtained by D studies and actual cases.

This research is a theoretical research. The study group consists of 84 university students, 4 raters working at a state university. The relevant data were gathered from the achievement test in Research Methodology Courses developed by the researchers. Achievement test consists of four open ended questions. The achievement test was administered to 84 undergraduate students. Four raters scored the achievement test independently using a five-point graded category rating scale. G and Phi coefficients were calculated both for D studies and the actual cases for the same conditions in D studies. Then G and Phi coefficients were compared. Edu G 6.1-e programme was used to analyze the data.

According to the research findings, although G and Phi coefficients obtained by D studies and actual cases (when scorers are not selected randomly from the universe) were close to each other, they were different. The findings revealed that there is no systematic relationship between G and Phi coefficients obtained by D studies and actual cases. According to the findings in this research, it was seen that there is no systematic increase in G coefficient when the number of raters increase in actual cases. Whereas in D studies the more raters are included, the higher G and Phi coefficients are obtained. So, D studies don't predict the actual values of G and Phi coefficients. But in actual cases, addition or removal of the raters may cause the increase or decrease of the variance components values. This situation may affect the error variance in negative or positive way. Moreover, characteristics of the raters will play an important role in this process. Because the variance component values kept constant in decision study, increasing the number of raters led to obtain higher G and Phi coefficients. However, in the actual cases increasing

the number of raters led to obtain higher or lower G and Phi coefficients depending on the characteristics of the raters.

Because G and Phi coefficients predicted in D studies may not be the same with the one obtained in G studies (actual cases), researchers should be careful while interpreting results of D studies. In this study, G and Phi coefficients obtained under all conditions have small difference. But all of them indicate moderate reliability generally. However, this small difference may be important when G and Phi coefficients are close to threshold value (0.80). So, this situation will lead researcher to label a test as reliable although it was not actually and vice versa. In the condition when it is not possible to select scorers randomly from the universe, results of previous D studies may be different from the actual ones. In this case, the researchers should be careful while interpreting the results of previous D studies if they don't have opportunity to select scorers from the universe.