Trakya Üniversitesi Eğitim Fakültesi Dergisi
2017, Cilt 7, Sayı 1, 1-9

Geliş Tarihi: 16.08.2016 *Yayına Kabul Tarihi: 11.10.2016*

İngilizce Konuşma Sınavından Elde Edilen Verilerin Güvenirliğinin Genellenebilirlik Kuramı ile Belirlenmesi

The Determination of an English Speaking Exam's Data Reliability Using Generalizability Theory

Meltem ACAR GÜVENDİR[1], Emre GÜVENDİR[2]

Öz: Bu araştırmanın amacı, İngilizce konuşma sınavının güvenilirliğini Genellenebilirlik (G) kuramı ile hesaplamak ve olası hata kaynaklarının neler olabileceğini belirlemektir. Çalışmanın evrenini Trakya Üniversitesi İngiliz Dili Eğitimi birinci sınıfta öğrenim görmekte olan 50 öğrenci oluşturmaktadır. Çalışmada yer alan 50 öğrencinin konuşma becerileri beş kriter doğrultusunda iki puanlayıcı tarafından değerlendirilmiştir. Araştırmada, Trakya Üniversitesi Eğitim Fakültesi, İngiliz Dili Eğitimi Anabilim Dalı 2015-2016 akademik yılı İngilizce konuşma becerileri dönem sonu sınavına ait veriler kullanılmıştır. Verilerin analizi alt amaçlara göre genellenebilirlik kuramıyla belirlenen desen doğrultusunda yapılmıştır. Araştırmada 50 öğrencinin (ö) 5 kriter (k) doğrultusunda 2 puanlayıcı (p) tarafından puanlanması ile ö x k x p deseni için genellenebilirlik (G) ve karar (K) çalışması yapılmıştır. Bu çalışmada G kuramı ve karar (k) çalışması çerçevesinde yapılan analizler sonucunda ele alınan konuşma sınavı için hata payının en düşük olduğu durum öğrenci sayısı sabit tutulduğunda kriter ve puanlayıcı sayısının arttırılmasıdır. Çalışma, konuşma sınavı puanlarının G kuramı kullanılarak güvenilirliğinin belirlenmesi açısından alanyazına bir örnek teşkil edecektir.
Anahtar Sözcükler: *Konuşma becerisi; genellenebilirlik kuramı; güvenirlik, İngilizce; puanlayıcı.*

Abstract: Using Generalizability theory, this study examined the dependability of an English speaking exam's data and determined potential error resources. The universe of the study includes 50 freshman students who study English Language Teaching at Trakya University. English speaking skills of these students were graded by two raters according to five criteria. The study used the data obtained from the final English speaking exam that the participants took during the 2015-2016 year at Trakya University. For data analysis, a model generated by G theory was used. According to the results of G and decision (D) study, if the number of students is kept constant and the numbers of criteria and raters are increased, the lowest error value will be obtained. The study will serve as a sample in terms of using G theory to examine the reliability of speaking exams data.
Key Words: *Speaking skill; generalizability theory; reliability; English; rater.*

1. GİRİŞ

Yabancı bir dili öğrenmenin en temel unsurlarından biri o dili etkin bir biçimde konuşabilmektir. Ancak yabancı dilde etkin bir konuşur konumuna gelmek hem zorlu bir tecrübe, hem de uzun yıllar alabilecek bir süreçtir (Hughes, 2011). Yabancı bir dili konuşma aşamasında var olan tümevarımsal süreçler uygun sözcük seçimi, telaffuz özellikleri, hedeflenen anlamı aktaracak yapısal kalıplar, sözel olmayan unsurların anlama etkisinin tespiti ve söylem yapılarının anlaşılmasını içerirken, tümdengelimsel süreçler ise içerik bilgisi, kültürel bilgi, konuşmacının etkileşimdeki rolü, bireylerarası ilişkiler ve koşulların uygunluğu gibi unsurları içerir (Saville-Troike, 2012). Bunlara ek olarak öğrencilerin konuştukları kişilerin aktardıklarını anlamaları ve bunlara uygun bir şekilde cevap verip iletişimin gerektirdiği koşulları sağlamaları gerekmektedir (Kormos, 2006). Öğrencilerin erek dili planlama, işleme ve üretme becerileri etkileşim aşamasında anlık olarak test edilir. Karşılıklı etkileşim esnasında bu aşamaları gerçekleştirmek için sahip olunan zamanın kısıtlılığından kaynaklanan koşullar konuşma becerisinin genellikle öğrenciler tarafından daha zorlayıcı olarak algılanmasına neden olur (Wang, 2014).

Konuşma, güvenilir bir şekilde değerlendirilebilecek en zor beceridir (Bachman & Palmer, 1981). Bir yabancı dil konuşurunun dilsel üretimlerinin başarılı olup olmadığının belirlenmesi aşamasında birçok etmen etkileşim halinde bulunur. Öğrencilerin konuşma düzeyleri genellikle yüz yüze etkileşim esnasında ortaya konulan anlık performanslar ışığında bir veya birkaç puanlayıcı tarafından değerlendirilir. Yapılan değerlendirmeyi etkileyen etmenlerden bazıları puanlayıcıların odaklandığı öğeler (örn. telaffuz, akıcılık

[1] Yrd. Doç. Dr., Trakya Üniversitesi, meltemacar@gmail.com
[2] Yrd. Doç. Dr., Trakya Üniversitesi, emreguvendir@gmail.com

vb.), öğrencinin dil düzeyi, puanlayıcılar ve öğrencinin birbirlerini tanıma düzeyleri, puanlayıcıların ve öğrencinin kişisel özellikleri, etkileşim esnasında ele alınan konular ve yöneltilen sorular, öğrenciye sunulan konuşma görevleri ve bu görevlerin tamamlanması için verilen zamandır (Luoma, 2004). Bu öğelerden kaynaklanabilecek sorunları aşabilmek ve yapılacak olan sınavın daha güvenilir olmasını sağlayabilmek için sınav esnasında öğrencilere verilen konuşma görevlerinin dikkatli planlanması (Foster & Skehan, 1999), puanlamayı yapacak kişilerin izlenecek adımlar konusunda eğitilmesi ve performanslarının incelenmesi (McNamara, 1996; O'Sullivan, 2000), sınav puanlama formlarının oluşturulması (Fulcher, 1996; North, 1995) ve sınavın kaydedilip yapılan değerlendirmelerin tekrar gözden geçirilmesi (Luoma, 2004) önerilmektedir. Ancak tüm bu adımlar izlense de gerçekleştirilen sınavdan elde edilen puanların güvenilirliğini sınav bittikten sonra hesaplamak önem taşımaktadır.

Bu durumda süreçte yapılan ölçme sonuçlarının güvenirliğinin ve geçerliğinin belirlenmesi önemlidir. Ölçme sürecinde kullanılan ölçme araçlarının güvenirliği ve geçerliğinin değerlendirme sonuçlarının doğruluk derecesini arttırmak için yüksek olması beklenir (Alharby, 2006). Güvenirlik, ölçme sonuçlarının hatasızlık derecesidir (Baykul, 2000) ve aynı zamanda, yapılan ölçmelerin tutarlılık derecesidir. Ölçme sonuçlarının güvenirliğinin belirlenmesinde Klasik Test Kuramı (KTK), Madde Tepki Kuramı (MTK) ve Genellenebilirlik Kuramı (G kuramı) olmak üzere üç kuramdan yararlanılabilir. Bu üç kuramın içerisinde en yaygın olarak kullanılan KTK'dir. Güler (2011), bunu diğer kuramların dayandığı matematiksel ifadelerin anlaşılmasının daha güç olmasına ve kullanımlarındaki karmaşıklığa bağlamaktadır.

Klasik Test Kuramının dayandığı varsayımlardan biri gözlenen puanın, gerçek ve hata puanlarının toplamından oluşmasıdır. Bu varsayım KTK'nin temel denklemlerinden olan "X=T+E" denklemiyle gösterilir (Baykul, 2000). X gözlenen puanı, T gerçek puanı, E ise hata puanını simgeler. Bu denklem, bir çalışmada ölçme sonuçlarına karışan birden fazla hata kaynağının KTK ile ele alınamayacağını göstermektedir (Baykul, 2000; Güler, 2011). Bir diğeri ise tekrarlı ölçmelerle tahmin edilen bireylerin gerçek puanlarının birbirinden bağımsızlığı varsayımıdır. Güvenirlik katsayısı, gözlenen puanların varyansı ile gerçek puanların varyansının oranı olarak tanımlanır. Gözlenen puan, gerçek puan varyansı ve hata varyansı olarak iki farklı varyans kaynağından meydana gelir. Gerçek puan varyansı dışındaki tüm varyanslar, farklı varyans kaynaklarından oluşmaktadır. Genel olarak güvenirlik, hata kaynaklarına göre farklı olabilir. Ancak KTK kullanımında tek bir hata kaynağı ele alınır. Örneğin, birden çok puanlayıcının bulunduğu bir sınavda KTK sadece puanlayıcılar arası tutarlılıktan kaynaklanan hatayı verir. Böyle bir durumda farklı hata kaynaklarından (örn. sınavda verilen görevler, ele alınan konular, öğrenci seviyesi vb.) ve bu kaynakların etkileşiminden doğacak hatalar ele alınmaz.

Testlerden elde edilen puanların güvenirliğinin belirlenmesinde kullanılan diğer bir kuram olan MTK'de ise ölçme hataları her birey için ayrı ayrı kestirilebilir ve her bir madde ve yetenek düzeyi için güvenirlik madde ve test bilgi fonksiyonu şeklinde hesaplanır. Ancak KTK'de ölçme hataları tüm grup için hesaplanabilir ve güvenirlik, testi alan grubun puan dağılımı için tek bir değer olarak hesaplanır (Nartgün, 2002). Dolayısıyla KTK'nin araştırmacılara vereceği güvenirlik sonucu sınırlı olacaktır.

Güvenirlik hata kaynağına göre tutarlılık ve kararlılık gibi farklı isimlerle ifade edilir. Ölçmede benzer özellikleri puanlayan çoklu puanlayıcılardan elde edilen puanlar arasındaki tutarlılık anlamındaki güvenirlik, puanlayıcılar arası güvenirlik olarak belirtilir. Bu durumda hata kaynağı sadece puanlayıcılardır. Puanlayıcılar arası güvenirlik, önemli bir hata kaynağı olmasına rağmen, farklı hata kaynaklarından ve bunlar arasındaki etkileşimden kaynaklanan hata kaynakları da vardır. Fakat çoklu hata kaynakları ve bunlar arasındaki etkileşim KTK tarafından ele alınmamaktadır. Ayrıca klasik test kuramı, öğrencilerin mutlak değerlendirmelerini ölçmeye olanak vermemektedir. Bu yüzden, farklı varyans kaynakları arasındaki etkileşimi belirlemek KTK ile olanaksızdır.

Genellenebilirlik Kuramı, gözlenen puanlardaki tutarsızlık kaynaklarının miktarını belirleyen ve davranışsal ölçmenin güvenirliğini değerlendiren istatistiksel bir kuramdır (Cronbach vd., 1972, akt. Brennan, 2001). Shavelson ve Webb (1991) G kuramını, davranışsal ölçme güvenirliğinin istatistiksel bir kuramı olarak tanımlamışlardır. G kuramı, ölçme sonuçları için kapsamlı bir çerçeve ve istatistiksel yollar sunar. Aynı zamanda, bu kuram test puanlarının ve puanlayıcılar arasındaki tutarlılığın bir ölçüsüdür (Brennan vd., 2003). Kuramın temelinde varyans analizi yatar. Gözlenen puanlar, ölçme sonuçlarının farklı varyans kaynaklarına ayrılması ile gerçek puanlara genellenebilir. Bu kuram KTK'nin sınırlılıklarına bir cevap olarak ortaya çıkmıştır (Shavelson & Webb, 1991).

Genellenebilirlik Kuramı, hata kaynaklarını ele alış biçimi bakımdan KTK'den ayrılır (Cronbach, Linn, Brennan, & Haertel, 1995). G kuramı hatayı farklı hata kaynaklarına böler ve hatayı çoklu değişkenlik kaynaklarıyla birlikle ele alır (Shavelson & Webb, 1991). Bu yüzden, birden fazla güvenirlik katsayısı

genellenebilirlik kuramı ile hesaplanabilir. G kuramı güvenirlik yöntemlerinin tümünü içerir (Eason, 1989) ve bu kuramda mutlak ve bağıl değerlendirmeler arasında bir fark vardır (Brennan, 2001). G kuramı, güvenirlik ve geçerlik arasındaki farkı da ortadan kaldırır. G kuramının KTK'den bir farkı da karar çalışmasıdır. Karar çalışması en etkili ve daha güvenilir sonuç elde etmek için kullanılır. Özetle G kuramı, bir analizde çoklu varyans kaynaklarını ele alır ve varyans kaynağının büyüklüğünü belirler. Ayrıca bu kuram, mutlak ve bağıl kararlar ile iki farklı güvenirlik katsayısı hesaplama imkânı sunar. Son olarak kuram, karar çalışmasıyla ölçme hatalarını en aza indirgeyen ölçmelere olanak tanır.

Alanyazın incelendiğinde, G kuramının kullanılarak konuşma becerilerinin ölçüldüğü puanların güvenirliği belirlenmiş ve farklı karar çalışmaları yürütülmüştür (Atılgan & Tezbaşaran, 2005; Bachman, Lynch, & Mason, 1995; Han, 2016; Nalbantoğlu Yılmaz & Gelbal, 2011; Sato, 2012; Srikaew, Tangdhanakanond, & Kanjanawasee, 2015; Vafaee & Yaghmaeyan, 2015). Bachman, Lynch ve Mason (1995), İspanyolca konuşma testinden dil bilgisi puanlarının güvenirliği üzerindeki görev ve puanlayıcı etkisini G kuramı ile incelemişlerdir. Sato (2012), G kuramını kullanarak konuşma becerisi test puanları için konuşmanın detaylandırması ile dilbilimsel ölçütün göreli katkısını incelemiştir. Atılgan ve Tezbaşaran (2005), G kuramı ve alternatif karar çalışmaları ile senaryolar ve gerçek durumlar için elde edilen G ve Phi katsayılarını karşılaştırmışlardır. Araştırma sonuçlarına göre, puanlayıcı sayısı düşürüldüğünde daha küçük G ve Phi katsayısı elde edilmiştir. Srikaew, Tangdhanakanond ve Kanjanawasee (2015), Taylandlı öğrencilerin İngilizce konuşma testinden aldıkları puanların güvenirliğini çok değişkenli G kuramına göre belirlemişlerdir. Onlara göre, akıcılık görevi puanların en yüksek varyans bileşeni olarak elde edilmiştir. Vafaee ve Yaghmaeyan (2015), konuşma testinden elde edilen puanların güvenirliğini tek değişenli G kuramına göre ve teste ilişkin dört ölçekli analitik rubrikten elde edilen puanların güvenirliğini de çok değişkenli G kuramına göre belirlemişlerdir. Ele aldıkları puanlayıcı ve görev sayılarına göre testi alan bireylerin yüksek düzeyde uyumlu ölçmelerini elde etmelerine karşın görev ve puanlayıcı sayısındaki artışın daha yüksek G ve Phi katsayısı sunacağına da vurgu yapmışlardır. Han (2016), İngilizce ve Çince tercüman sertifika puanlarının güvenirliğini G kuramı ile belirlemiştir. Ona göre ek görevler, güvenirliğin arttırılması için ekstra puanlayıcı kullanımından daha etkili bir sonuç verecektir. Nalbantoğlu Yılmaz ve Gelbal (2011), iletişim becerileri istasyonu örneğinde G kuramıyla farklı desenleri karşılaştırmışlardır. Araştırma sonucunda her iki desende kestirilen varyans değerleri birbirleriyle paralellik göstermiştir. Sonuç olarak, puanlayıcıların belli sayıdaki öğrencileri dönüşümlü olarak puanlamasının zaman, iş gücü ve ekonomik açıdan daha uygun olduğu sonucuna ulaşılmıştır.

Alanyazına bakıldığında özellikle Türkiye'deki yabancı dilde konuşma becerilerini ölçen sınavlardan elde edilen puanların güvenirliğinin G kuramı ele alınarak belirlendiği çalışmalara rastlanılmamıştır. Bu yüzden bu araştırmanın amacı, yabancı dilde gerçekleştirilmiş bir konuşma sınavından elde edilen puanların güvenilirliğini G kuramı ile hesaplamak ve olası hata kaynaklarının neler olabileceğini belirlemektir. Bu doğrultuda çalışma, konuşma sınavlarının güvenilirliğinin G kuramı kullanılarak belirlenmesi açısından alanyazına bir örnek teşkil edecektir.

2. YÖNTEM

2.1. Araştırmanın Modeli

Araştırma, G kuramı ile lisans düzeyinde yapılan İngilizce konuşma sınavından elde edilen puanların güvenirliğininin belirlenmesi olduğundan betimsel bir araştırma niteliği taşımaktadır.

2.2. Evren

Araştırmada örneklem seçimine gidilmemiştir. 2015-2016 öğretim yılı Trakya Üniversitesi Eğitim Fakültesi İngiliz Dili Eğitimi programının birinci sınıfında öğrenim görmekte olan ve İngilizce konuşma dersini alan öğrenciler araştırmanın evrenini oluşturmaktadır. Çalışmada yer alan 50 öğrencinin konuşma becerileri beş kriter doğrultusunda iki puanlayıcı tarafından değerlendirilmiştir. Araştırmada öğrencilerin konuşma becerilerini belirlenen beş kritere göre puanlayan puanlayıcılardan biri Trakya Üniversitesi Eğitim Fakültesi İngiliz Dili Eğitimi programında yardımcı doçent doktor, diğeri ise aynı programda öğretim görevlisi olarak çalışmaktadır.

2.3. Verilerin Toplanması

Araştırmada, Trakya Üniversitesi Eğitim Fakültesi, İngiliz Dili Eğitimi Anabilim Dalı 2015-2016 akademik yılına ait İngilizce konuşma becerileri dersi dönem sonu sınavına ait veriler kullanılmıştır.

Konuşma becerileri sınavında öğrencilere daha önceden verilen konuşma konularından kendilerine uygun olanı seçmeleri istenmiştir. Sınavdan önce öğrenci belirlemiş olduğu konuyu puanlayıcılara vermiş ve puanlayıcılar bu konu doğrultusunda konuşma ortamı yaratmışlardır. Konuşma ortamı kayıt altına alınmıştır. Sınav esnasında her bir öğrenciye eşit süre verilmiş ve bu sırada iki puanlayıcı bulunmuştur. Bu puanlayıcılar, İngiliz Dili Eğitimi Anabilim Dalı tarafından konuşma sınavında kullanılmak üzere beş kritere göre hazırlanan konuşma becerisi değerlendirme formunu kullanarak, öğrencilerin konuşma süreçlerini her bir kriter için 20 puan üzerinden puanlamışlardır. Konuşma becerisi değerlendirme formunda yer alan beş kriter; doğruluk, akıcılık, anlaşılabilir telaffuz, sözcük ve ifade kullanımı, görev başarısı şeklindedir.

Tüm veriler elde edildikten sonra G kuramında kullanılmak üzere araştırmacılar tarafından bir desen tasarlanmıştır. Bu desen, öğrenci (ö), kriter (k) ve puanlayıcı (p) değişkenleri olmak üzere öğrencilerin aynı konuşma becerileri değerlendirme formundaki kriterler doğrultusunda puanlayıcılar tarafından öğrencilerin konuşma süreçlerinin puanlamasıyla oluşturulmuş ö x k x p desenidir.

2.4. Veri Analizi

Verilerin analizi araştırma amacına göre G kuramıyla belirlenen desen doğrultusunda yapılmıştır. Araştırmada 50 öğrencinin (ö) 5 kriter (k) doğrultusunda 2 puanlayıcı (p) tarafından puanlanması ile ö x k x p deseni için genellenebilirlik ve karar çalışması yapılmıştır. Araştırmada kullanılan ö x k x p (ö: öğrenci, k: kriter ve p: puanlayıcı) deseni için oluşturulmuş veri yapısı örneği aşağıda gösterilmiştir.

	K1		**K2**		**K3**		**K4**		**K5**	
	P1	**P2**	**P1**	**P2**	**P1**	**P2**	**P1**	**P2**	**P1**	**P2**
Ö1	X	X	X	X	X	X	X	X	X	X
Ö2	X	X	X	X	X	X	X	X	X	X
Ö3	.	.	.	.	.	.	.	.	.	.
Ö..	.	.	.	.	.	.	.	.	.	.
Ö..	.	.	.	.	.	.	.	.	.	.
Ö..	.	.	.	.	.	.	.	.	.	.
Ö50	X	X	X	X	X	X	X	X	X	X

Şekil 1. ö x k x p deseni veri yapısı örneği

Şekil 1'de gösterildiği gibi ö x k x p deseninde 50 öğrenci 5 kriter doğrultusunda 2 puanlayıcı tarafından birlikte puanlanmıştır. G kuramı ile desene ait varyans bileşenlerinin kestirilmesi, değişkenlerin toplam varyansı açıklama oranlarının hesaplanmasında ve karar çalışmalarının yapılmasında GENOVA programı kullanılmıştır. Program Brennan (2001) tarafından G kuramı analizleri için geliştirilmiş olup, araştırmacıların tanımladığı değişkenlik kaynakları ve bu değişkenlik kaynaklarıyla oluşturduğu desen için genellenebilirlik ve karar çalışmalarının yapılmasına olanak sağlamaktadır.

3. BULGULAR

Konuşma sınavında, 50 öğrencinin konuşma becerileri 5 kriter doğrultusunda 2 puanlayıcı tarafından puanlanmıştır. Değişkenlerin (öğrenci, kriter ve puanlayıcı) tamamı çaprazlanmış, puanlayıcıların her bir öğrenciyi aynı değerlendirme formunu kullanarak puanlamasıyla oluşturulan ö x k x p (ö: öğrenci, k: kriter ve p: puanlayıcı) ile G çalışması yapılmıştır. G çalışması sonucunda ö x k x p deseni ile kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri Tablo 1'de verilmiştir.

Tablo 1. Konuşma sınavı verileri için varyans bileşenlerinin ANOVA tahmini

Değişimin Kaynağı	sd	Ortamalama kareler	Tahmin edilen varyans bileşenleri	Toplam varyansın yüzdesi
Öğrenci (Ö)	49	76.77	7.07	60
Kriter (K)	4	12.52	.00	0
Puanlayıcı (P)	1	206.08	.76	7
ÖK	196	2.70	.00	0
ÖP	49	6.24	.67	6
KP	4	13.71	.22	2
ÖKP	196	2.88	2.88	25

Tablo 1'de verilen ö x k x p desenine ait G çalışması sonucunda kestirilen varyans ve toplam varyansı açıklama yüzdeleri incelendiğinde, öğrenci (Ö) ana etkisine ait varyans bileşeninin toplam varyansın %60'ını açıkladığı görülmektedir. Öğrencilere ait varyans bileşeni öğrencilerin konuşma sınavına ait başarılarının nasıl değiştiğinin bir tahminini verir. Tablo 1'de görüldüğü gibi öğrencilere ait varyans bileşeni (σ^2(ö)=7.07) yüksek çıkmıştır. Öğrencilere ait varyans bileşeninin toplam varyansı açıklama oranının yüksek olması, öğrencilerin konuşma becerisi bakımından farklılaştığı, grubun ölçülen özellik bakımından farklı olduğu şeklinde yorumlanabilir.

Kriter (K) ana etkisi için kestirilen varyans bileşeni, toplam varyansın % 0'ını açıklamaktadır. Kriter etkisine ait varyans bileşeni toplam varyans içinde en düşük varyans değerine (σ^2(k)=.00) sahiptir. Kriter etkisi için kestirilen varyans bileşeninin düşük olması, kriterlerin öğrenciler tarafından yapılabilme durumları arasında farklılık olmadığını gösterir.

Puanlayıcı (P) ana etkisine ait varyans bileşeninin toplam varyansın %7'sini açıkladığı görülmektedir. Puanlayıcılara ait varyans bileşeni puanlayıcıların konuşma sınavına ait verdikleri puanların nasıl değiştiğinin bir tahminini verir. Tablo 1'de görüldüğü gibi puanlayıcıya ait varyans bileşeni (σ^2(p)=.76) küçük çıkmıştır. Puanlayıcılara ait varyans bileşeninin toplam varyansı açıklama oranının küçük olması, puanlayıcıların konuşma sınavına ilişkin verdikleri puanlar bakımından az bir farklılaşmanın olduğu şeklinde yorumlanabilir.

Öğrenci x Kriter (ÖK) ortak etkisine ait varyans bileşeni toplam varyansın %0'ını açıklamaktadır. Öğrenci x kriter etkileşimi öğrencilerin konuşma becerilerinin kriterlere göre farklılık gösterip göstermediğini verir (Shavelson & Webb, 1991). Tablodan da görüldüğü gibi öğrenci x kriter etkileşimi toplam varyans içinde en düşük varyans değerine sahiptir. Bu durum öğrencilerin konuşma becerilerinin bir kriterden diğerine farklılık göstermediği şeklinde yorumlanabilir.

Öğrenci x Puanlayıcı (ÖP) ortak etkisine ait varyans bileşeni toplam varyansın %6'sını açıklamaktadır. Öğrenci x puanlayıcı etkileşimi öğrencilerin konuşma becerilerinin puanlayıcılara göre farklılık gösterip göstermediğini verir (Shavelson & Webb, 1991). Öğrenci x puanlayıcı etkileşiminin toplam varyans içindeki yeri düşüktür. Bu durum öğrencilerin konuşma becerilerinin bir puanlayıcıdan diğerine az bir farklılık gösterdiği şeklinde yorumlanabilir.

Kriter x Puanlayıcı (KP) ortak etkisine ait varyans bileşeni toplam varyansın %2'sini açıklamaktadır. Buna göre, Kriter x puanlayıcı etkileşimi puanlayıcıların bir kriterden diğer kritere kararlı puanlama yapıp yapmadıklarını gösterir (Shavelson & Webb, 1991). Kriter x puanlayıcı etkileşiminin toplam varyans içindeki yeri düşüktür. Puanlayıcıların bir kriterden diğerine kararlı puanlama yaptığı söylenebilir.

Öğrenci x Kriter x Puanlayıcı (ÖKP) ortak etkisine ait varyans bileşeni artık varyans olarak adlandırılır. Tablo 1'de görüldüğü gibi artık varyans bileşeni toplam varyansın %25'ini açıklama oranı ile büyüklük bakımından değişkenler arasında ikinci sıradadır. Artık varyans bileşeninin büyük çıkması öğrenci, kriter ve puanlayıcı ortak etkileşimi ve/veya tesadüfi hata kaynaklarının büyük olabileceğinin göstergesi olabilir.

Tüm değişkenlerin çaprazlandığı ö x k x p deseninde öğrenciler ölçmenin nesnesi (object of measurement) olarak belirlenip kriter ve puanlayıcı sayılarının artırılıp azaltılmasıyla yapılan senaryolar için kestirilen genellenebilirlik katsayısı (G), Phi katsayısı, bağıl hata varyansı ve mutlak hata varyanslarına ait değerler Tablo 2'de gösterilmiştir.

Tablo 2. ö x k x p desenine ait karar çalışması ile kriter ve puanlayıcı sayılarının değiştirilmesiyle oluşturulmuş senaryolara göre G ve Phi katsayıları

Ö	K	P	$\hat{\sigma}^2_{\text{bağıl}}$	$\hat{\sigma}^2_{\text{mutlak}}$	$\hat{p}^2$	$\hat{\phi}$
50	8	4	.26	.45	.97	.94
50	5	4	.31	.51	.96	.93
50	2	4	.53	.74	.93	.91
50	8	2	.52	.91	.93	.89
50	**5**	**2**	**.62**	**1.02**	**.92**	**.87**
50	8	1	1.03	1.81	.87	.79
50	1	3	1.18	1.51	.86	.82
50	5	1	1.25	2.05	.85	.78
50	1	1	3.55	4.52	.67	.61

Araştırmada 50 öğrencinin 5 kritere göre 2 puanlayıcı tarafından aldığı puanlara yönelik yapılan karar çalışmasında G katsayısı .92, Phi katsayısı .87 olarak hesaplanmıştır. Tablo 2'deki verilere göre öğrenci sayısı (nö=50) sabit tutulup kriter ve puanlayıcı sayısı azaltıldığında (nk=1, np=1) G katsayısı .67, Phi katsayısı .61, öğrenci sayısı ve kriter sayısı sabit tutulup, puanlayıcı sayısı arttırıldığında (nö=50, nk=5, np=4) G katsayısı .96, Phi katsayısı .93, azaltıldığında (nö=50, nk=5, np=1) G katsayısı .85, Phi katsayısı .78'dir. Öğrenci sayısı sabit tutulduğunda, kriter sayısı azaltılıp, puanlayıcı sayısı arttırıldığında (nö=50, nk=2, np=4; nö=50, nk=1, np=3) G katsayısı sırasıyla .93 ve .86, Phi katsayısı sırasıyla .91 ve .82'dir. Öğrenci ve puanlayıcı sayısı sabit tutulup, kriter sayısı arttırıldığında (nö=50, nk=8, np=2) G katsayısı .93, Phi katsayısı .89, puanlayıcı sayısı azaltılıp, kriter sayısı arttırıldığında (nö=50, nk=8, np=1) G katsayısı .87, Phi katsayısı .79; ancak hem puanlayıcı hem de kriter sayısı arttırıldığında (nö=50, nk=8, np=4) G katsayısı .97, Phi katsayısı ise .94 olarak bulunmuştur. Bu durumda elde edilen G ve Phi katsayıları karar çalışmasına göre en yüksek değerlerdir. Buna göre konuşma becerisinin ölçüldüğü durumlarda öğrenci sayısı sabit tutulduğunda kriter ve puanlayıcı sayısı arttırıldığında güvenirlik değeri yükselmektedir.

Tablo 2'den öğrenci sayısı sabit tutulup, kriter ve puanlayıcı sayılarının arttırıldığı durumda (nö=50, nk=8, np=4) bağıl ($\hat{\sigma}^2_{\text{bağıl}}$=.26) ve mutlak hata ($\hat{\sigma}^2_{\text{mutlak}}$=.45) varyansları en düşüktür. Buna karşın öğrenci sayısı sabit tutulup, kriter ve puanlayıcı sayılarının azaltıldığı durumda (nö=50, nk=1, np=1) ise bağıl ($\hat{\sigma}^2_{\text{bağıl}}$=3.55) ve mutlak hata ($\hat{\sigma}^2_{\text{mutlak}}$=4.52) varyansları en yüksektir. Böylece puanlayıcı ve kriter sayısının artırılmasına göre bağıl ve mutlak hata varyanslarının azaldığı görülmektedir. Yapılan tüm bu açıklamalar ışığında sınavda kullanılan kriter ve puanlayıcı sayısının artırılması güvenirliği artırmaktadır.

4. TARTIŞMA ve SONUÇ

Araştırmada İngilizce konuşma sınavından elde edilen puanların güvenirliği G kuramı ile hesaplanmış ve olası hata kaynaklarının neler olabileceği belirlenmiştir. Çok etmenli yapısı ve olası hata durumlarının fazlalığı konuşmanın güvenilir bir şekilde değerlendirilebilecek en zor becerilerin başında gelmesine yol açmaktadır. Dolayısıyla konuşma sınavı öncesi ve sınav esnasında izlenecek adımları doğru bir şekilde izleyip, sınav sonunda hata kaynaklarını tespit etmek daha güvenilir sınavlar gerçekleştirmek açısından önem teşkil etmektedir. Bu çalışmada, G kuramı çerçevesinde yapılan analizler sonucunda ele alınan konuşma sınavı için hata payının en düşük olduğu durum öğrenci sayısı sabit tutulduğunda kriter ve puanlayıcı sayısının arttırılmasıdır. G kuramı, tek bir analizde çoklu varyans kaynaklarını ele alarak varyans kaynağının büyüklüğünü belirlemiştir. Özellikle güvenirlik düzeyinin düşük çıktığı konuşma sınavlarında sınavı gerçekleştirilenlerin hatanın hangi değişkenlerden ve hangi değişkenlerin etkileşiminden doğduğunu bilmeleri önem taşımaktadır. Bu doğrultuda tek bir hata kaynağına yoğunlaşmak ve atılacak adımları bu hata kaynağına göre belirlemek yapılacak olan konuşma sınavlarından elde edilen puanların güvenirliğini yükseltmek adına sınırlı bilgiler sunacaktır.

Mevcut çalışmada ele alınan konuşma sınavında ikiden fazla puanlayıcının bulunması ve beşten fazla kriterin belirlenmesi sınavdan elde edilen puanların güvenirliğini arttırmak adına önemli bir adım olarak belirlenmiştir. Bu çalışmada 2 puanlayıcı ve 5 kriterle elde edilmiş olan G ve Phi katsayısı da düşük değildir; ancak karar çalışması sonucunda puanlayıcı ve kriter sayısının fazla olduğu durumda

güvenirliğinin daha yüksek çıktığı gözlendiğinden, bu türde yapılacak olan sınavlarda, şayet mevcut personel sayısı uygunsa, ikiden fazla puanlayıcının olması elde edilen puanların güvenirliği açısından olumlu olacaktır. Ancak puanlayıcı ve kriter sayısının arttırılması iş yükünü de beraberinde getirecektir. Srikaew, Tangdhanakanond ve Kanjanawasee (2015) ve Vafaee ve Yaghmaeyan (2015), İngilizce konuşma becerisi, Gebril (2010) ise İngilizce yazma becerisi üzerindeki çalışmalarında benzer şekilde görev ve puanlayıcının arttırıldığı durumlarda G ve Phi katsayısının arttığını gözlemlemişlerdir. Atılgan ve Tezbaşaran da (2005) benzer şekilde puanlayıcı sayısının artırıldığı durumlarda yüksek G ve Phi katsayılarının elde edildiğine vurgu yapmışlardır. Srikaew, Tangdhanakanond ve Kanjanawasee (2015) ise bu artırım durumda, yapılacak ölçmelerin de artacağına dikkat çekerek, G ve Phi katsayılarının yükselmesi için görev sayısının artırılmasının, ek olarak puanlayıcılar bulmak ve onları eğitmekten daha az maliyetli olacağını önermişlerdir. Benzer şekilde Han (2016), puanların güvenirliğini yükseltmek için görev sayısının artırılmasının puanlayıcı sayısının artırılmasından daha etkili sonuçlar vereceğini savunmuştur. Nalbantoğlu Yılmaz ve Gelbal (2011), puanlayıcı sayısının artırıldığı durumda G ve Phi katsayılarının yükseldiğini belirtmişler; ancak puanlayıcı sayısını artırmanın her zaman, her durumda mümkün olmayacağını da eklemişlerdir. Bu durumda Vafaee ve Yaghmaeyan (2015), görev sayısının artırılmasının yanı sıra puanlayıcıların eğitilmesi gerektiği üzerinde durmuştur

İlhan ve Çetin (2014), performans değerlendirmeye karışan puanlayıcı etkilerini azaltmanın yollarından biri olarak puanlayıcı eğitimlerini önermişlerdir. Onlara göre, puanlayıcı eğitimleri ile değerlendirmede kullanılacak puanlama ölçeklerinin puanlayıcılara tanıtılması, değerlendirme işlemine yönelik örnek uygulamaların yaptırılması ve böylelikle puanlayıcılar arasındaki ortak bir anlayışın oluşmasının amaçlanması gerektiğine vurgu yapmışlardır. Puanlayıcı eğitimleri; puanlayıcı hatası eğitimleri, performans boyutları eğitimi, davranış gözlem eğitimi ve referans çerçevesi eğitimi olmak üzere dört başlıkta ele alınmalıdır.

Ayrıca çalışmada konuşma dersini yürüten öğretim üyesi tarafından belirlenmiş olan beş kritere göre puanlayıcılar puanlama yaptığından bu haliyle konuşma dersinde uygulanan formata sadık kalınmıştır. Ancak konuşma sınavı yürütecek olan öğretim üyeleri ya da araştırmacılar bu beş kriterin yanı sıra O'Sullivan'ın (2008) belirttiği gibi; söylem yönetimi, etkileşimsel iletişim, öte dil kullanımı gibi birçok konuşma sınavında göz önünde bulundurulan kriterleri de kullanabilirler. Gerçekleştirilen bu çalışmada kriter sayısının arttırılmasının güvenirlik açısından olumlu sonuçlar doğuracağı görülmüştür. Dolayısıyla, konuşma sınavları için oluşturulmuş olan puanlama formlarındaki kriterlerin gözden geçirilmesi önem taşımaktadır.

Çalışma sonuçları genel olarak ele alındığında, G kuramının konuşma sınavlarını kurgulayan diğer eğitmenlere daha güvenilir sınavlar hazırlama ve değişik senaryoları ele alan güvenilirlik hesaplamaları yapma konusunda olanaklar sunduğu görülmektedir. Dolayısıyla, sınav sonuçlarının güvenilirliğinin düşük çıktığı durumlarda, buna neden olan koşulların ve unsurların neler olduğunun G kuramı ile tespit edilmesi, sınav uygulayıcılarının tek bir hataya yoğunlaşmasına ve benzer hataları tekrarlamasına olanak tanımayacaktır. Gerçekleştirilen bu çalışma yabancı dilde konuşma sınavından elde edilen puanların güvenirliğini etkileyen etmenlerin G kuramı vasıtasıyla belirlenmesi açısından alanyazına bir katkı sunmaktadır.

5. KAYNAKLAR

Alharby, E. R. (2006). A *comparison between two scoring methods, holistic vs. analytic using two measurement models, the generalizability theory and the many facet rasch measurement within the context of performance assessment*. The Pennsylvania State University.

Atılgan, H., & 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,* 28-41.

Bachman, L., & Palmer, A. (1981). The construct validation of the FSI oral interview. *Language Learning*, *31*, 67-86.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12,* 238-257.

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

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

Brennan, R. L. Yin, P., & Kane, M. T. (2003). Methodology for examining the relability of group mean difference scores. *Journal of Educational Measurement, 40*(3), 207-230.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995). Generalizability analysis for educational assessments. *Evaluation Comment*, Summer, (pp. 1-29). Los Angeles: UCLA Center for the Study of Evaluation and The National Center for Research on Evaluation, Standards and Student Testing.

Eason, S. H. (1989). *Why generalizability theory yields better results than classical test theory*. Mid- South Educational Research Association Annual Meeting: 8-10 November 1989- Little Rock, AR.

Foster, P. & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research*, *3* (2), 15-47.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*, 208-38.

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing, 15*, 100-117.

Güler, N. (2011). *Eğitimde ölçme ve değerlendirme*. Ankara: PegemA Yayıncılık.

Han, C. (2016). Investigating score dependability in English/Chinese interpreter certification performance testing: A generalizability theory approach. *Language Assessment Quarterly*, *13*(3), 186-201.

Hughes, R. (2011). *Teaching and researching: speaking*. New York: Routledge.

İlhan, M., & Çetin, B. (2014). Performans değerlendirmeye karışan puanlayıcı etkilerini azaltmanın yollarından biri olarak puanlayıcı eğitimleri: Kuramsal bir analiz. *Journal of European Education*, *4*(2), 29-38.

Kormos, J. (2006). *Speech production and second language acquisition*. New Jersey: Lawrence Erlbaum Associates.

Luoma, S. (2004). *Assessing speaking*. United Kingdom: Cambridge University Press.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

Nalbantoğlu Yılmaz, F.& Gelbal, S. (2011). İletişim Becerileri İstasyonu Örneğinde Genellenebilirlik Kuramıyla Farklı Desenlerin Karşılaştırılması.*Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *41*, 509-518.

Nartgün, Z. (2002). *Aynı tutumu ölçmeye yönelik likert tipi ölçek ile metrik ölçen madde ve ölçek özelliklerinin klasik test kuramı ve örtük özellikler kuramına göre incelenmesi*. Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System*, *23*, 445-65.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System*, *28*, 373-86.

O'Sullivan, B. (2008). Notes on assessing speaking. Cornell University Language Resource Center. Retrieved from http://www.lrc.cornell.edu/events/past/2008-2009/papers08/osull1.pdf

Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, *29*(2), 223-241.

Shavelson, J.R., & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park. CA: sage Publication.

Srikaew, D., Tangdhanakanond, K., & Kanjanawasee, S. (2015). English speaking skills assessment for grade 6 Thai students: an application of multivariate generalizability theory. *International Journal of Psychology: A Biopsychosocial Approach*, *16*, 47-66.

Saville-Troike, M. (2012). *Introducing second language acquisition*. United Kingdom: Cambridge University Press.

Wang Z. (2014). On-line time pressure manipulations: L2 speaking performance under five types of planning and repetition conditions. In P. Skehan (Eds.), *Processing perspectives on task performance* (pp. 27-62). Amsterdam: John Benjamins.

Vafaee, P., & Yaghmaeyan, B. (2015). Providing Evidence for the Generalizability of a Speaking Placement Test Scores. *Iranian Journal of Language Testing*, *5*(2), 78-95.

EXTENDED ABSTRACT

Introduction

Generalizability theory (G theory) focuses on multiple variance resources and determines the size of variance resources. Furthermore, G theory calculates two different reliability coefficients using absolute and relative decisions. The theory enables measurements which minimize errors using decision (D) study. Using G theory, this study examines the reliability of an English speaking exam's data and determines potential error resources. This study will contribute to the literature in terms of using G theory to examine the reliability of speaking exams data.

Method

The study aims to determine the reliability of an English speaking exam data which makes it a descriptive study. The universe of the study includes 50 freshman students who study English Language Teaching at Trakya University. Speaking skills of these students were scored according to five criteria by two raters, who work as an assistant professor and lecturer at Department of Foreign Languages/Trakya University. The study used the data obtained from the final English speaking exam that the participants took during the 2015-2016 academic year at Trakya University. Raters scored the students' speaking performance by using a form which involved five criteria. These criteria are accuracy, fluency, pronunciation, vocabulary and expression use, and task achievement. The researchers made a model which is in the form of s (student) x c (criteria), and x r (rater). Data analysis was done with this model according to the G theory. G study and D study were done for s x c x r.

Result and Discussion

In this study, G and Phi coefficients were found as .92 and .87 for the data which was obtained from 50 students, 5 criteria, and 2 raters. When the number of student is constant and the numbers of criteria and rater is low (nc=1, nr=1), G and Phi coefficients are .67 and .61. When the numbers of student and criteria are constant and the number of rater is higher (nc=5, nr=4), G and Phi coefficients are .96 and .93. When the numbers of student and criteria are constant and the number of rater is lower (nc=5, nr=1), G and Phi coefficient are .85 and .78. When the number of students is kept constant, the number of criteria is decreased, the number of raters is increased (nc=2, nr=4; nc=1, nr=3), G coefficients are .93 and .86, Phi coefficients are .91 and .82. When the number of raters is higher (nc=5, nr=4), G and Phi coefficient are .96 and .93. When the numbers of student and rater are constant and the number of criteria is higher (nc=8, nr=2), G and Phi coefficients are .93 and .89. When the number of raters is lower and the number of criteria is higher (nc=8, nr=1), G and Phi coefficients are .87 and .79. When both the numbers of rater and criteria increase, G and Phi coefficients are .97 and .94. This result gives the highest coefficient according to the D study. Furthermore, when the number of rater and criteria increases, the relative and absolute error variances decrease.

In this study, according to results of G and D study, when the number of students is constant and numbers of criteria and rater are increased, error is the lowest. If the researchers use G theory to examine the reliability of speaking exam data, they could see different error resources and obtain reliable results and do reliable evaluation.

G theory determines the size of variance resources using multiple variance resources with one analysis. If a speaking exam has low reliability, the researchers/raters are supposed to know from which variables errors arise. Consequently, focusing on one error resource and determining next steps to be taken according to only one error resource will be limited for increasing reliability of data. For the current research, the ideal scenario was having two raters and five criteria in order to maximize the reliability coefficients of the exam. For other exams, G theory could be used to identify the best scenarios and prepare reliable speaking tests.