Ocak 2017 Cilt:25 No:1 Kastamonu Eğitim Dergisi 361-372

Genellenebilirlik Kuramında Tümüyle Çaprazlanmış ve Maddelerin Puanlayıcılara Yuvalandığı Desenlerin Karşılaştırılması

Comparing Fully Crossed and Nested Designs Where Items Nested in Raters in Generalizability Theory

Celal Deha Doğan
Ankara Üniversitesi, Eğitim Bilimleri Fakültesi, Ölçme ve Değerlendirme Bölümü, Ankara.

Hatice Özlem Anadol
TOBB Ekonomi ve Teknoloji Üniversitesi Yabancı Diller Bölümü, Ankara.

İlk Kayıt Tarihi:20.03.2016 ***Yayına Kabul Tarihi:25.04.2016***

Özet

Bu çalışmada İngilizce Kompozisyon Yazma Becerisinin puanlanması sürecinde tümüyle çaprazlanmış desenin (bxpxm) ve maddelerin puanlayıcıya yuvalandığı ancak bireylerin maddeler ve puanlayıcılar ile çaprazlanmış olduğu desenin (bx(m:p)) kullanıldığı durumlarda elde edilen G ve Phi katsayılarının karşılaştırılması amaçlanmıştır. Çalışmaya bir vakıf üniversitesi hazırlık okulunda öğrenim gören ve 3 puanlayıcı dahil olmuştur. Çalışma sonucunda tümüyle çaprazlamış desen ile elde edilen G ve Phi katsayıları daha yüksek çıkmıştır. Değişkenlik kaynaklarına göre varyans bileşenleri incelendiğinde birey ana etkisine ilişkin varyans tümüyle çaprazlanmış desen için daha yüksekken kalan etkiye ilişkin varyans değeri daha düşüktür. Bu bulgular sınıf içi uygulamalarda tümüyle çaprazlanmış desenin daha güvenilir sonuçlar verdiğini göstermektedir. Bu bağlamda sınıf içi uygulamalarda pratik koşullar sağlandığında tümüyle çaprazlanmış desenin kullanılması önerilmektedir.

Anahtar Kelimeler: *Puanlayıcılar arası Güvenirlik, Genellenebilirlik Kuramı, Tümüyle Çaprazlanmış Desenler, Yuvalanmış Desenler.*

Abstract

This study aims to investigate the comparison of G and Phi coefficients calculated both by a fully crossed (bxpxm) design (b: person; p: rater; m:item) and a nested (bx(m:p)) design in which items are nested in raters but individuals are crossed with items and raters in the process of grading English Composition Writing Skill. The study consists of students who attend a private university and 3 raters. According to the results, G and Phi coefficients of the fully crossed design were higher. In terms of sources of variance, while person has the highest percentage in total variance in fully crossed design, occasion has relatively low percentage. Findings indicate that fully crossed design yields more reliable results in classroom practices. In this respect, fully crossed design is recommended for classroom practices.

Keywords: *Interater Reliability, Generalizability Theory, Fully crossed designs, Nested designs.*

1. Giriş

Sınıf içi ölçme ve değerlendirme uygulamalarında öğrencilerin üst düzey zihinsel süreçlerin ölçülmesinde objektif olarak puanlanabilen çoktan seçmeli testlerin yanı sıra yanıtını öğrencinin yapılandırdığı açık uçlu soruların, performans görevlerinin kullanımı yerinde olacaktır. Yanıtı öğrenci tarafından yapılandırılan soruların ve performans görevlerinin objektif testlere kıyasla eksik yanı puanlama güvenirliğidir. Puanlama sürecinde puanlayıcıların subjektif görüşleri devreye girebileceğinden puanlayıcılar arası güvenirlik bu tür ölçme ve değerlendirme etkinliklerinde daha büyük öneme sahiptir.

Yanıtı öğrencinin yapılandırdığı ölçme ve değerlendirme etkinliklerinde puanlama sürecine subjektif görüşlerin girmemesi ve daha nesnel puanlama yapılabilmesi için yaygın olarak kullanılan araçlardan birisi de dereceli puanlama anahtarlarıdır. Dereceli puanlama anahtarları (DPA) yanıtı öğrencinin yapılandırdığı, doğru yanıtın derecelerinin olduğu ve birden fazla doğru yanıtın olabileceği sınav türlerinin puanlanmasında kullanılan araçlardır. Dereceli puanlama anahtarları öğrencilere yaptıkları çalışmaların hangi ölçütlere göre değerlendirileceğini ve performanslarının hangi puana denk geleceğini gösteren puanlama araçlardır (Goodrich, 1996). Bu araçlar aynı zamanda öğretmenlerin puanlamaları daha nesnel yapmalarına da katkı getirmektedir.

Alan yazında bütünsel ve analitik olmak üzere iki tür DPA'dan bahsedilir. Bütünsel DPA'larda öğrencinin gösterdiği performansın bütününe tek bir puan verilmektedir ve her düzeyde performansın kalitesini belirleyen tanımlamalar bulunmaktadır. Bu tip puanlama anahtarları öğrenci performansındaki bazı küçük hataların göz ardı edilebileceği ve performansın bütününe odaklanıldığı durumlarda kullanılmaktadır (Arter ve McTighe, 2001; Kutlu, Doğan ve Karakaya 2014). Daha yaygın olarak kullanılan analitik DPA'lar öğrenci performansının çeşitli boyutlarındaki başarı düzeyleri hakkında bilgi veren puanlama araçlarıdır. Böylece öğrencinin belli bir alandaki güçlü ve zayıf yönleri hakkında bir profil sunabilmektedir (Gronlund 1998). Bu çalışmada öğrencilerin İngilizce kompozisyon yazma becerilerinin değerlendirilmesinde analitik DPA'dan faydalanılmıştır.

Her ne kadar kullanılan DPA'ların puanlayıcılar arası uyumu artıracağı ve objektif puanlamaya katkı getireceği beklense de yanıtı öğrenci tarafından yapılandırılan sınavların güvenirliğinin test edilmesi gerekmektedir. Bu süreçte klasik test kuramına, genellenebilirlik kuramına ve madde tepki kuramına dayalı yöntemler ile belirlemeler yapılabilir. Klasik test kuramında puanlayıcılar arası uyumun test edilmesi sürecinde sınıf içi korelasyon katsayısından, Cohen'nin ve Fleiss'in kapa katsayılarından, Kendall'ın W katsayısından ve puanlayıcılar arası uyum düzeyinden faydalanılabilir. Madde tepki kuramına dayalı olarak ise Çok Yüzeyli Rasch Modeli kullanılabilir. Diğer madde tepki kuramı modellerinde olduğu gibi çok yüzeyli rasch modelde de sınanın örtük yeterliliği, sınanan yanıtlarının olasılığı olarak kullanılmaktadır (Macmillan, 2010). Bu model, puanlayıcı katılığı, görev güçlüğü ve sınanan performansını etkileyen herhangi diğer değişkenlik kaynaklarının güçlüklerine ait parametreleri modele ekleyip, ölçmenin bu değişkenlik kaynaklarını tanımlayarak kestiren, temel rasch modelin bir uzantısıdır (Iramaneerat, Myford, Yudkowsky, 2008).

Yanıtı öğrenci tarafından yapılandırılan açık uçlu soruların puanlanmasında puanlayıcılar önemli bir hata kaynağı olmakla beraber ölçme sonuçlarına karışacak farklı hata kaynakları da mevcuttur. Ancak klasik test kuramına dayalı güvenirlik hesaplama yöntemlerinde bir hata kaynağı dikkate alınır. Bu durum farklı varyans kaynaklarına dayalı güvenirliklerin aynı anda kestirilmesine olanak vermez. Genellenebilirlik kuramı, puanlayıcı zaman, ölçme formu, görevler ya da maddeler gibi çeşitli değişkenlik kaynaklarından gelebilecek hataları birlikte ve eş zamanlı olarak değerlendirerek tek bir güvenirlik katsayısının hesaplanmasına olanak verir (Güler, Gelbal 2010a). Bu çalışmada ilgili verilerin çözümlenmesi sürecinde genellenebilirlik kuramından faydalanılmıştır.

Genellenebilirlik kuramında değişkenlik kaynaklarının sayısına da bağlı olarak tek yüzeyli veya çok yüzeyli evrenler üzerinden çalışmalar gerçekleştirilebilir (Güler, Uyanık ve Teker, 2012). Değişkenlik kaynaklarından bir tanesi, genellikle birey değişkenlik kaynağı, ölçme objesi olarak ele alınır. Örneğin birey, madde ve puanlayıcı değişken kaynaklarının yer aldığı bir durum, birey ölçme objesi olarak alındığında geriye iki yüzey kaldığı (madde, puanlayıcı) için iki yüzeyli evren olarak adlandırılır.

İki yüzeyli evrenlerde oluşturulabilecek farklı desenler bulunmaktadır. Tüm bireylerin tüm maddeleri yanıtladığı ve tüm puanlayıcıların tüm bireyleri ve maddeleri puanladığı desen tümüyle çaprazlanmış desen (bxmxp) olarak adlandırılır. Ancak pratik nedenlerden dolayı her koşulda tümüyle çaprazlanmış desenlerin kullanılması mümkün olamamaktadır. Bu durumda yuvalanmış desenlerden faydalanılmaktadır. Farklı yuvalanmış desenler yer almakla beraber pratik koşullar gereği en sık kullanılanlardan bir tanesi de bx(m:p) desenidir. Bu desende her puanlayıcı testte yer alan farklı maddeleri puanlarken tüm bireyler tüm maddeleri yanıtlamakta ve tüm puanlayıcılar tüm bireyleri puanlamaktadır. Başka bir ifade ile maddeler puanlayıcılar içine yuvalanmışken bireyler maddelerle ve puanlayıcılarla çaprazlanmıştır. Oluşturulabilecek farklı yuvalanmış desenler olmakla beraber bu çalışma yukarıda açıklanan iki desen ile sınırlandırılmıştır.

Açık uçlu soruların puanlanmasında tümüyle çaprazlanmış desenlerin kullanımı pratik nedenlerden dolayı bazen mümkün olmamaktadır. Ancak bahsi geçen yuvalanmış desenin kullanıldığı durum ile tümüyle çaprazlanmış desenin kullanıldığı durumda elde edilen güvenirlik katsayılarının belirlenmesi ve karşılaştırılması uygulamaya yönelik önemli katkılar sağlayacaktır.

Alan yazında yapılan çalışmalar incelendiğinde açık uçlu veya performansa dayalı sınavların güvenirliğinin belirlenmesinde klasik test kuramına, genellenebilirlik kuramına ve çok yüzeyli rasch modeline dayalı yöntemlerin kullanıldığı çalışmalar yer almaktadır (Linarce & Wright & Lunz, 1990; Engelhard, 1994; Lynch & McNamara, 1998; Nakamura, 2000; Engelhard & Myford, 2003; Sudweeks & Reeve & Bradshaw, 2004; Kan, 2005-1; Kan, 2005-2; Iramaneerat & Yudkowsky & Myford & Downing, 2008 ; Iramaneerart & Myford & Yudkowsky, 2009; Güler & Gelbal 2010a,2010b; Akın & Baştürk 2010-2012; Stenlund, 2013; Parlak ve Doğan 2014; Özel ve Acar 2014; Doğan & Yosmaoğlu 2015; Büyükkıdık ve Anıl 2015).

Alan yazındaki öne çıkan çalışmalar incelendiğinde genellenebilirlik kuramında

iki yüzeyli evrenler için tümüyle çaprazlanmış ve yuvalanmış desenlerin uygulandığı durumların karşılaştırıldığı çalışmalar oldukça azdır (Yılmaz ve Gelbal, 2011).

Bu bağlamda araştırmanın genel amacı İngilizce kompozisyon yazma becerisinin puanlanması sürecinde tümüyle çaprazlanmış ve maddelerin puanlayıcılara yuvalanmış olduğu desenlerin kullanıldığı durumlarda elde edilen G ve Phi katsayılarının karşılaştırılmasıdır. Bu genel amaç doğrultusunda aşağıdaki sorulara yanıt aranacaktır.

1. Tüm değişkenlik kaynaklarının çaprazlanmış olduğu tümüyle çaprazlanmış desen (bxmxp) sonucunda elde edilen

a. G ve Phi katsayıları nedir?

b. Varyans bileşenleri nedir?

2. Maddelerin puanlayıcılara yuvalanmış olduğu yuvalanmış desen (bx(m:p)) sonucunda elde edilen

a. G ve Phi katsayıları nedir?

b.Varyans bileşenleri nedir?

3. Her iki desenden elde edilen varyans bileşenleri, G ve Phi katsayıları farklılaşmakta mıdır?

2. Yöntem

Araştırma Modeli

Bu araştırmada İngilizce Kompozisyon Yazma becerisinin puanlanması sürecinde tümüyle çaprazlanmış modelin (bxpxm) ve maddelerin puanlayıcıya yuvalandığı modelin (bx(m:p)) kullanıldığı durumlarda elde edilen G ve Phi katsayılarının karşılaştırılması amaçlanmıştır. Bu nedenle çalışma temel araştırma niteliğindedir. Karasar (2005) Bilgi edinmeye, kuram (teori) geliştirmeye ya da var olan kuramları sınamaya yönelik çalışmalar, temel çalışmaları temel araştırma olarak tanımlamaktadır.

İşlem

Araştırmada iki farklı desen oluşturulmuştur. Öncelikle tüm bireylerin tüm maddeleri yanıtladığı, tüm puanlayıcıların da tüm birey ve maddeleri puanladığı tümüyle çaprazlanmış desen oluşturulmuştur. Bu bağlamda 3 puanlayıcı 6 maddenin tümünü ve toplam 123 bireyin tamamını puanlamış, tüm bireyler ise 6 maddenin hepsini yanıtlamıştır.

Akabinde ise maddelerin puanlayıcılara yuvalandığı ancak bireylerin maddeler ve puanlayıcılar ile çaprazlanmış olduğu (bx (m:p)) yuvalanmış desen oluşturulmuştur. Bu süreçte tüm bireyler tüm maddeleri yanıtlamış ve tüm puanlayıcılar tüm bireyleri puanlamıştır. Ancak 1 numaralı puanlayıcı sadece 1. ve 2. maddeleri, 2 numaralı puanlayıcı sadece 3. ve 4. maddeleri 3 puanlayıcı ise sadece 5. ve 6. maddeleri yanıtlamışlardır. Her iki desen oluşturulurken de aynı 3 puanlayıcı görev almıştır.

Çalışma Grubu

Araştırmanın çalışma grubunu 2015-2016 öğretim yılında bir vakıf üniversitesi İngilizce hazırlık bölümünde öğrenim gören 123 öğrenci ve aynı kurumda görev yapan 3 okutmandan oluşmaktadır. Katılımcılar araştırmaya gönüllük esasına dayalı olarak dahil olmuşlardır. Araştırmada puanlayıcı olarak görev alan 3 okutman 5 senedir ilgili kurumda çalışmaktadırlar ve öğrenci kompozisyonlarının dereceli puanlama anahtarı kullanılarak puanlanması sürecinde deneyime sahiptirler.

Verilerin Toplanması

Araştırmada gerekli verilerin toplanması sürecinde öncelikli olarak öğrencilere yurt dışında öğrenim görme konusunda İngilizce en az 150 sözcükten oluşan bir kompozisyon yazmalarına yönelik bir ödev verilmiştir. Öğrenciler ilgili çalışmayı 2 ders saati içerisinde okulda gerçekleştirmişlerdir. Akabinde öğrenci çalışmaları 6 ölçüte (Dilbilgisi Kullanımı, Kelime Kullanımı, Konu Bütünlüğü ve Bağlantı, Cümleler Arası Geçiş, Örneklendirme, Sonuç Cümlesi) sahip İngilizce kompozisyon yazma analitik dereceli puanlama anahtarı kullanılarak puanlanmıştır. Dereceli puanlama anahtarında yer alan ölçütler 1-3 arasında puanlanmaktadır.

Verilerin Çözümlenmesi

İlgili verilerin çözümlenmesi sürecinde hem tümüyle çaprazlanmış desen hem de yuvalanmış desenlerde bulunan temel ve ortak etkiye sahip değişkenlik kaynakları için kareler ortalamaları, varyans bileşenleri ve yüzdeleri hesaplanmıştır. Akabinde göreli / mutlak hata varyansları ve G ve Phi katsayıları her iki desen içinde ayrı ayrı hesaplanmıştır. Gerekli hesaplamaların yapılması sürecinde EduG 6.1 paket programından faydalanılacaktır.

3. Bulgular

Bu bölümde tümüyle çaprazlanmış desen ve yuvalanmış desen için bulgular sırasıyla sunulmuştur.

Tümüyle Çaprazlanmış Desen (bxmxp) için Bulgular

123 birey, 6 madde ve 3 puanlayıcı için oluşturula tümüyle çaprazlanmış desene ilişkin G çalışması neticesinde elde edilen varyans bileşenleri çizelge 1 de sunulmuştur. Çizelgede "b" sembolü birey, "m" sembolü madde ve p sembolü puanlayıcı değişkenlik kaynaklarını ifade etmektedir.

Çizelge 1. Tümüyle çaprazlanmış (bxmxp) desen için elde edilen varyans bileşenler

Varyans Kaynağı	Toplam Kareler	sd	Kareler Ortalaması	Varyans	%
B	372.5919	122	3.05403	0.15015	34.3
M	168.8702	5	33.77416	0.09111	208
P	2.86540	2	1.43270	0.00162	0.4
bxm	127.0187	610	0.20823	0.01873	4.3
bxp	72.02349	244	0.29518	0.02386	5.5
mxp	0.97200	10	0.09720	-0.00045	0.0
Bxmxe	185.4725	1220	0.15203	0.15203	34.7
Toplam	929.81391	2213			

İlgili değerler incelendiğinde tümüyle çaprazlanmış desenin kullanıldığı durumda birey (b) ana etkisi için kestirilen varyans bileşeninin (0.15) toplam varyans içerisinde ikinci en yüksek paya (%37.7) sahip olduğu görülmektedir. Bu durum yapılan ölçme işlemi ile elde edilen boyutta bireyler arası farklılıkların belirlenebildiği şeklinde yorumlanabilir.

Madde ana etkisi incelendiğinde ise kestirilen varyans bileşeninin ise (0.09) toplam varyans içerisinde üçüncü en yüksek paya sahiptir ve toplam varyansın %20 .8'inin açıklamaktadır. Bu bulgu maddelerin güçlük düzeylerinin farklılık gösterdiğine işaret etmektedir.

Puanlayıcı (p) ana etkisi incelendiğinde ise oldukça küçük bir varyans bileşeninin (0.0016) kestirildiği görülmektedir. Puanlayıcı ana etkisi toplam varyansın %0.4 gibi çok küçük bir kısmını açıklamaktadır. Bu bulgular ışığında puanlayıcıların tüm bireyler boyunca yaptıkları puanlamaların katılık ve cömertlik düzeylerinin farklılaşmadığı belirtilebilir.

Birey-madde ortak etkisi (bxm) incelendiğinde elde edilen varyans bileşeninin (0.018) toplam varyansın %4.3'ünü açıkladığı görülmektedir. Bu bulgu maddelerin güçlük düzeylerinin bireyden bireye az da olsa bir farklılık gösterdiğini işaret etmektedir.

Birey -puanlayıcı ortak etkisi (bxp) için hesaplanan varyans bileşeninin (0.023) toplam varyansın % 5.5'ini açıkladığı görülmektedir. Bu bulgu puanlayıcıların verdikleri puanların bireyden bireye az da olsa farklılık gösterdiği şeklinde yorumlanabilir.

Madde - puanlayıcı ortak etkisi için hesaplanan varyans bileşeninin (-0.0004) negatif değer aldığı ve toplam varyans içerisinde en küçük paya sahip olduğu belirlenmiştir. Cronbach ve arkadaşları (1972) negatif varyans değerlerinin sınıf alınmasının önermiştir. Brennan (1983) ise varyans bileşenlerinin hesaplanması sürecinde negatif varyans değerlerinin kullanılması gerektiğini ancak hesaplama işlemleri tamamlandıktan sonra negatif değerlerin yerine sıfır atanması gerektiğini vurgulamıştır. Bu çalışmada Brennan'nın önerdiği yol dikkate alınmıştır. Güler,Uyanık ve Teker (2012) Brenan'ın yönteminin varyans bileşenlerinin yanlı kestirimini engelleyebileceğini belirtmekle beraber her iki yönteminde belirli sınırlıkları olduğunu vurgulamışlardır.

Bu bağlamda Madde puanlayıcı ortak etkisi için elde edilen varyans bileşenin toplam varyansa herhangi bir katkı yapmadığı belirtilebilir. Bu bulgu puanlayıcıların verdikleri puanların maddelere göre farklılık göstermediğine işaret etmektedir.

Kalan etki değişkenlik kaynağı (bxmxp,e) incelendiğinde en yüksek varyans bileşenine (.0152) sahip olduğu belirlenmiştir. Kalan etki bileşeni toplam varyansın 34.7'sini açıklamaktadır. Bu ölçme sürecine bulgu bireyler, maddeler ve puanlayıcı arası etkileşim ile çalışmada ölçülmeyen tesadüfi hataların dahil olduğunu göstermektedir.

Yukarıda belirtilen tümüyle çaprazlanmış desen için hesaplanan güvenirlik katsayıları çizelge 2'de sunulmuştur.

Çizelge 2. bxmxp deseni G ve Phi katsayıları

N_{birey}=123	
N_{madde}=6	
$N_{puanlayıcı}$=3	
G katsayısı	.88
Phi katsayısı	.81

Tümüyle çaprazlanmış desen için elde edilen G ve Phi katsayılarına bakıldığında göreli hata varyansına dayalı olarak hesaplanan G katsayısının .88 mutlak hata varyansına datalı olarak hesaplana Phi katsayısının ise .81 olduğu belirlenmiştir. Göreli hata varyansından farklı olarak mutlak hata varyansı hesaplanırken tüm bileşenler (ana ve ortak etkiler) dikkate alındığından daha büyük bir değere sahip olmakta ve dolayısı ile Phi katsayısı her zaman G katsayısından düşük çıkmaktadır. Güvenirlik katsayılarının 0 ile 1 arasında değiştiği ve .70 ve üzeri katsayıların kabul edilebilir olduğu dikkate alındığında her iki güvenirlik katsayısının da kabul edilebilir sınırlar içerisinde olduğu belirtilebilir.

Maddelerin Puanlayıcılara Yuvalandığı ((bx (m:p)) Desen için Bulgular

123 birey, 6 madde ve 3 puanlayıcı için maddelerin puanlayıcılara yuvalandığı ancak bireylerin maddeler ve puanlayıcılar ile çaprazlanmış olduğu (bx (m:p)) yuvalanmış desenden elde edilen varyans bileşenleri çizelge 3'te sunulmuştur.

Çizelge 3. Maddelerin puanlayıcılara yuvalandığı ((bx (m:p)) desen için elde edilen varyans bileşenleri

Varyans Kaynağı	Toplam Kareler	sd	Kareler Ortalaması	Varyans	%
b	120.9865	122	0.99169	0.13146	30
p	8.49051	2	4.24526	-0.04291	0.0
m:p	44.31707	3	14.77236	0.11869	27.1
bxp	49.50949	244	0.20291	0.01446	3.3
bxm:pe	63.68293	366	0.17400	0.17400	39.7
Toplam	286.9865	737			

Çizelge 4 incelendiğinde birey değişkenlik kaynağına ilişkin varyans bileşeni sahip olduğu 0.13'lik değer ile ikinci en yüksek varyans değerine sahiptir. Bireylere ait

varyans bileşeni toplam varyansın %30'unu açıklamaktadır. Bu durum yapılan ölçme işlemi ile elde edilen boyutta bireyler arası farklılıkların belirlenebildiği şeklinde yorumlanabilir. Ancak tümüyle çaprazlanmış desende bireylere ait varyans bileşeni değerinin daha yüksek olduğu ve toplam varyansın %34'ünü açıkladığı görülmektedir. Bu bağlamda tümüyle çaprazlanmış desenin bireyler arasındaki farklılığı maddelerin puanlayıcılara yuvalandığı desene kıyasla daha iyi ortaya çıkardığı belirtilebilir.

Puanlayıcı ana etkisi için hesaplanan varyans bileşeninin (-0.042) negatif değer aldığı ve varyans bileşenin toplam varyansa herhangi bir katkı yapmadığı ifade edilebilir. Elde edilen bu bulgu puanlayıcıların puanlamaları arasındaki değişkenliğin hemen hemen hiç olmadığı şeklinde yorumlanabilir.

Madde puanlayıcı ortak etkileşimi için varyans değeri (.118) toplam varyansın %27.1'ini açıklamaktadır. Bu değer üçüncü en büyük varyans değeridir ve her bir puanlayıcının puanladığı maddelerden alınan puanlar arasında farklılıklar olduğunu göstermektedir.

Birey puanlayıcı ortak etkisini içi elde edilen varyans bileşeni 0.01'dir ve toplam varyansın yaklaşık % 3'ünü açıklamaktadır. Bu bulgu maddelerin güçlük düzeylerinin bireyden bireye az da olsa farklılık gösterdiği şeklinde yorumlanabilir.

Kalan etki değişkenlik kaynağı (bxm:p,e) incelendiğinde en yüksek varyans bileşenine (.0174) sahip olduğu belirlenmiştir. Kalan etki bileşeni toplam varyansın 39.7'sini açıklamaktadır. Bu ölçme sürecine bulgu bireyler, maddeler ve puanlayıcı arası etkileşim ile çalışmada ölçülmeyen tesadüfi hataların dahil olduğunu göstermektedir. Tümüyle çaprazlanmış desen ile karşılaştırıldığında maddelerin puanlayıcılara yuvalandığı desende kalan etki varyansının daha yüksek olduğu ve dolayısı ile ölçme sürecine tesadüfi hataların daha çok karıştığı belirtilebilir.

Yukarıda belirtilen maddelerin puanlayıcılara yuvalandığı yuvalanmış desen için hesaplanan güvenirlik katsayıları çizelge 4'te sunulmuştur.

Çizelge 4. bx (m:p) deseni için G ve Phi katsayıları

N_{birey}=123	
N_{madde}=6	
$N_{puanlayıcı}$=3	
G katsayısı	.80
Phi katsayısı	.71

Maddelerin puanlayıcılara yuvalandığı, bireylerin maddeler ve puanlayıcılar ile çaprazlanmış olduğu (bx (m:p)) yuvalanmış desen için elde edilen güvenirlik katsayıları incelendiğinde G katsayısının .80 Phi katsayısının ise .71 olduğu görüşmektedir. Güvenirlik katsayılarının 0 ile 1 arasında değiştiği ve .70 ve üzeri katsayıların kabul edilebilir olduğu dikkate alındığında her iki güvenirlik katsayısının da kabul edilebilir sınırlar içerisinde olduğu belirtilebilir. Ancak tümüyle çaprazlanmış desenden elde edilen G ve Phi katsayılarının, maddelerin puanlayıcılara yuvalandığı desendekinden daha yüksek olduğu belirtilebilir.

4. Sonuç ve Tartışma

Araştırma kapsamında öğrencilerin İngilizce kompozisyon yazma becerilerinin değerlendirildiği 123 birey, 6 madde ve 3 puanlayıcının yer aldığı bir durumda, tümüyle çaprazlanmış desen ve maddelerin puanlayıcılara yuvalandığı desen için elde edilen varyans bileşenleri ve güvenirlik katsayıları karşılaştırılmıştır.

Elde edilen bulgular neticesinde birey ana etkisine ilişkin varyans bileşeninin tümüyle çaprazlanmış desende daha yüksek olduğu belirlenmiştir (bxmxp deseni için: .15 - bx(m:p) deseni için .13). Başka bir ifadeyle tümüyle çaprazlanmış desende bireylere ilişkin var olan farklılıkları daha iyi ortaya çıkartıldığı belirtilebilir. Birey değişkenlik kaynağı ölçme objesi olan öğrencilerin maddelerden aldıkları farklı puanlardır ve yüksel varyans değerine sahip olması istendik bir durumdur (Güler, Uyanık ve Teker, 2012).

Kalan etki değişkenlik kaynağına bakıldığında her iki desen içinde en yüksek varyans değerine sahip olduğu görülmektedir. Benzer konuda yapılan çalışmalarda da benzer bulgulara ulaşılmıştır (Yılmaz ve Başusta, 2015; Güler ve Teker, 2015). Ancak maddelerin puanlayıcılara yuvalandığı desen için elde edilen varyans bileşeni ve toplam varyans içinde açıkladığı oran tümüyle çaprazlanmış desenden elde edilen değerlere göre daha yüksektir. Bu durumda maddelerin puanlayıcılara yuvalandığı desen kullanıldığında tümüyle çaprazlanmış desene kıyasla ölçme sürecine daha çok sistematik olmayan hata karıştığı şeklinde yorumlanabilir.

Her iki desen için elde edilen güvenirlik katsayıları dikkate alındığında tümüyle çaprazlanmış desen için hesaplanan G ve Phi katsayılarının (G:.88 – Phi: .81) maddelerin puanlayıcılar içine yuvalandığı desendekilere kıyasla (G: .80, Phi: .81) daha yüksek olduğu görülmektedir. Bu bağlamda İngilizce kompozisyon becerisinin analitik dereceli puanlama anahtarı ile puanlamasında tümüyle çaprazlanmış desen kullanıldığı duruma maddelerin puanlayıcılara yuvalandığı desene kıyasla daha güvenilir sonuçlar elde edilmektedir.

Yanıtını öğrencinin yapılandırdığı ve farklı doğru yanıtların bulunabileceği tür ölçme araçlarından elde edilen sonuçların güvenirliği objektif testlere kıyasla daha problematiktir. Bu süreçte farklı hata kaynaklarının sürece aynı anda dahil edilmesine olanak veren Genellenebilirlik Kuramı önemli avantaj sağlamaktadır. Ancak bazı durumlarda pratik nedenlerden dolayı yuvalanmış desenlerin kullanımı zorunlu olmaktadır. Bu çalışmada elde edilen bulgular tümüyle çaprazlanmış desen kullanıldığı durumda maddelerin puanlayıcılara yuvalandığı desene kıyasla daha güvenilir ölçümlerin yapılabileceğini ortaya koymaktadır. Bu bağlamda sınıf içi yapılan ölçme ve değerlendirme uygulamalarında daha güvenilir sonuçlar elde etmek için tümüyle çaprazlanmış desenin kullanılması önerilmektedir.

Yılmaz ve Gelbal (2011) çalışmalarında tümüyle çaprazlanmış desen ile (bxpxm) bireylerin puanlayıcılara yuvalandığı desen ((b:p)xm) ile karşılaştırmışlar ve bireylerin puanlayıcılara yuvalandığı desende G ve Phi katsayılarının tümüyle çaprazlanmış desendekinden daha yüksek kestirildiğini belirtmişlerdir. Bu sonuç araştırma bulguları ile çelişmektedir. Bu noktada büyük gruplarda gerçekleştirilen çalışmalarda yu-

valanmış desenlerin zaman işgücü ve ekonomiklik açısından tümüyle çaprazlanmış desenlere kıyasla daha avantajlı olduğu belirtilebilir. Maddelerin veya bireylerin dönüşümlü olarak puanlanması, puanlayıcıların daha az yorulmasına ve sağlıklı değerlendirme yapmasına katkı sağlayabilir. Büyük gruplarda gerçekleştirilen çalışmalarda tüm puanlayıcıların tüm birey ve maddeleri puanladığı tümüyle çaprazlanmış desen kullanıldığında, puanlayıcıların yorgunluktan veya dikkatsizlikten dolayı hatalı değerlendirmeler yapmaları söz konusu olabilir. Ancak büyük bir grup üzerinde gerçekleştirilmeyen sınıf içi uygulamalarda tümüyle çaprazlanmış desenin kullanılması daha güvenilir sonuçlar elde edilmesine katkı sağlayacaktır.

Özellikle geniş ölçekli sınavlarda da kullanılma sıklığı artan açık uçlu soruların puanlanması sorunu gündemdedir. Bu tip durumlarda tümüyle çaprazlanmış desenin kullanılması katılımcı sayısı çok fazla olduğundan mümkün olmamaktadır. Ancak farklı yuvalanmış desenlerin hangisinin (maddelerin puanlayıcılara yuvalandığı, puanlayıcıların bireylere yuvalandığı vb.) daha yüksek güvenirlik düzeyini sağladığına yönelik çalışmaların yapılması benzer konuları çalışacak araştırmacılara önerilmektedir.

5. Kaynakça

Akın, Ö.,&Baştürk, R. (2010). Assessment of research assignment by many-facet Rasch measurement approach. *Journal of Measurement and Evaluation in Education and Psychology, 1*(1), 51–57.

Akın, Ö.,&Baştürk, R. (2012). The evaluation of the basic skills in violin training by many-facetrasch model. *Pamukkale Üniversity Journal of Education, 31*, 175–187.

Arter, J. A. Ve Mctighe, J. (2001). *Scoring Rubrics in TheClassroom:Using Peformance Criteria for Assessing and Improving Student Performance*, Thousand Oaks, CA: Corvin Press

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: ACT, Inc.

Büyükkıdık, S., & Anıl, D. (2015). Investigation of reliability in generalizability theory with different designs on performance based assessment. *Education and Science, 40*(177), 285–296.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioural measurements: Theory of generalizability of scores and profiles. New York: Wiley.

Doğan, C. D., &Yosmaoğlu, B. (2015). The effect of the analytical rubrics on the objectivity in physiotherapy practical examination.*TurkiyeKlinikleri Journal of Sports Science, 7*(1), 9–15.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112.

Engelhard, G., &Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English Literature and composition program with a many-faceted Rasch model. *ETS Research Report Series* (1), i-60.

Goodrich, H. (1996).*Students Self Assessment: At theintersection of metacognition and authentic assessment***.** Doctoraldisertation. Cambrdige, MA: HarwardUniversity

Güler, N., &Gelbal, S. (2010a). Studying reliability of open-ended mathematics items according to the classical test theory and generalizability. *Educational Sciences: Theory & Practice, 10*(2), 989–1019.

Güler, N., &Gelbal, S. (2010b). A study based on the classical test theory and many facet Rasch model. *Egitim Eurasian Journal of Educational Research, 38*, 108–125.

Güler, N, Uyanık,G. K., Teker, G. T. (2012). Genellenebilirlik Kuramı. Pegem Akademi, Ankara Türkiye

Grounlond, N. E. (1998).**Assessment of Student Achievement.** USA: By Allyn & Bacon Viacom Company

Iramaneerat, C., Yudkowsky, R., Myford, C. M., & Downing, S. M. (2008). Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances in Health Sciences Education*, *13*(4), 479–493.

Iramaneerat, C., Myford, C. M., Yudkowsky, R., & Lowenstein, T. (2009). Evaluating the effectiveness of rating instruments for a communication skills assessment of medical residents. *Advances in Health Sciences Education*, *14* (4), 575–594.

Kan, A. (2005a). The effect of using grading scale and response key to (same) grader's reliability. *Eurasian Journal of Educational Research,* 166–167.

Kan, A. (2005b). The effect of using grading scale and response key to (different) grader's reliability. *Eurasian Journal of Educational Research,* 207–219.

Kutlu, Ö., Doğan, D. ve Karakaya, İ. (2014). Ölçme ve Değerlendirme: Performansa ve Portfolyaya Dayalı Durum Belirleme. Ankara: Pegem Akademi Yayıncılık.

Linacre, J. M., Wright, B. D., &Lunz, M. E. (1990). *A facets model for judgmental scoring*. MESA Memo, 61. Chicago, IL: MESA.

Lynch, B. K.,&McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speakingskills of immigrants. *Language Testing*, *15*(2), 158-180.

Macmillan, P. D. (2000). Classical, generalizability, andmultifacetedRaschdetection of interratervariability in large, sparsedatasets. *TheJournal of experimentaleducation*, *68*(2), 167-190

Nakamura, Y. (2002). Teacher assessment and peer assessment in practice. *Educational Studies,*44., 203-215

Özel, S. & Acar, T. (2014, 11-13 Haziran).Okullarda Sınıf İçi Ölçmelerde G Katsayısı, IV. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde Sözlü Bildiri olarak sunulmuştur, Hacettepe Üniversitesi.

Parlak, B., & Doğan, N. (2014). Comparison of answer key and scoring rubric for the evaluation of student performances. H*acettepe.University Journal of Education 29*(2), 189–197.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*(3), 239–261.

Stenlund, T. (2013). Agreement in assessment of prior learning related to higher education: An examination of inter-rater and intra-rater reliability. *International Journal of Lifelong Education 32*(4), 535–547.

Yılmaz, N.F, Gelbal, S. (2011). İletişim becerileri istasyonu örneğinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H.U Journal of Education).*41 (2011) 509-518

Extended Abstract

The overall objective of the study is to make a comparison of G and Phi coefficients of fully crossed design and nested design in which items are nested in raters but individuals are crossed with items and raters. Answers in response to the questions below were searched in accordance with this overall objective:

1. What are the G and Phi coefficients and variance components obtained from fully crossed design in which all variance components are crossed?

2. What are the G and Phi coefficients and variance components obtained from nested

design in which items are nested in raters but individuals are crossed with items and raters?

3. Do variance components, G and Phi coefficients obtained from both of the designs differ?

This study aims to investigate the comparison of G and Phi coefficients calculated both by fully crossed (bxpxm) design (b: person; p: rater; m:item) and nested (bx(m:p)) design in which items are nested in raters but individuals are crossed with items and raters in the process of grading English Composition Writing Skill; therefore, this study is a basic research.

Two different designs were constructed in this research. Initially, a fully crossed design in which all individuals responded to all items and all raters graded all items and individuals was constructed. In this respect, 3 raters graded all 6 items and all 123 individuals, and all individuals responded to all 6 items.

Subsequently, a nested (bx (m:p)) design in which items are nested in raters, but individuals are crossed with items and raters was constructed. In this process, all individuals responded to all items and all raters graded all individuals. However, first rater graded only first and second items, second rater graded only third and fourth items and third rater graded only fifth and sixth items. The same raters took part in grading sessions of both designs. The research was conducted on 123 students who attend English preparatory school of a private university and 3 instructors in 2015-2016 academic year.

In the process of data collection, students were assigned to write an English composition about studying abroad in at least 150 words. Students completed their tasks within 2 class hours at school. After collecting data, student compositions were graded with regards to 6 criteria (Grammar, Vocabulary, Cohesion and Coherence, Transition, Evidence and Examples and Concluding Sentence) by an analytic scoring rubric for English composition writing skill.

Within the scope of the research, variance components and G coefficients of fully crossed and nested design in which items were nested in raters were compared in the case of 123 individuals, 6 items and 3 raters.

In the process of data analysis, variance components were calculated for variance sources which have basic and common effects. After that, G and Phi coefficient were calculated for both of the designs separately. EduG 6.1software was used to make all the required calculations.

Obtained results indicate that variance of the person main effect is higher in fully crossed design (.15 for the bxmxp design and .13 for the bx(m:p) design). Namely, it can be stated that the existing differences between individuals are better observed with fully crossed design.

It can be seen that variance of the residual effect has the highest value for both fully crossed design and nested design However, the variance component and the ratio to the total variance obtained with the design in which items are nested in raters are higher than those obtained with the fully crossed design. It can be interpreted that with the nested design, more non-systematic error is involved in comparison to fully crossed design.

Considering the reliability coefficients obtained with both of the designs, it can be seen that the calculated G and Phi coefficients (G:.88 and Phi: .81) of fully crossed design is higher than those of nested design (G: .80, Phi: .81). Hence, for the grading of the English composition writing skill with an analytical scoring rubric, fully crossed design yields more reliable results compared to the nested design.

The results obtained in this study indicate that it is possible to obtain more reliable results when fully crossed design is employed, in comparison to the nested design case. Therefore, it is suggested to employ fully crossed design in order to obtain more reliable results in classroom testing and measurement applications.