Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 41: 509-518 [2011]

İLETİŞİM BECERİLERİ İSTASYONU ÖRNEĞİNDE GENELLENEBİLİRLİK KURAMIYLA FARKLI DESENLERİN KARŞILAŞTIRILMASI*

COMPARISON OF DIFFERENT DESIGNS IN ACCORDANCE WITH THE GENERALIZABILITY THEORY IN COMMUNICATION SKILLS EXAMPLE

Funda NALBANTOĞLU YILMAZ**, Selahattin GELBAL***

ÖZET: Araştırmanın amacı; genellenebilirlik kuramına göre performans puanlamada öğrencilerin birden fazla puanlayıcı tarafından birlikte ve dönüşümlü olarak puanlanmasıyla oluşturulan desenlerden elde edilen G ve K çalışmaları sonuçlarını karşılaştırmaktır. Araştırmanın çalışma grubunu, 2007- 2008 öğretim yılı Hacettepe Üniversitesi Tıp Fakültesi üçüncü sınıf öğrencilerinden tesadüfi olarak seçilen 48 öğrenci ve üç puanlayıcı oluşturmaktadır. Araştırmada puanlayıcıların öğrencileri aynı iletişim becerileri değerlendirme formuyla birlikte ve dönüşümlü puanlamasıyla oluşturulan ö x g x p ve (ö:p)xg desenleri (ö: öğrenci, g: görev, p: puanlayıcı) kullanılmıştır. Analizler sonucunda her iki desenle kestirilen varyans değerlerinin birbirleriyle paralellik gösterdiği, yapılan karar çalışmaları sonucunda her iki desendeki G ve Phi katsayıları arasında çok fark olmamakla birlikte (ö:p)xg deseninde katsayıların daha büyük çıkma eğiliminde olduğu görülmektedir. Böylece puanlayıcıların belli sayıdaki öğrencileri dönüşümlü olarak puanlamasının zaman, iş gücü ve ekonomiklik açısından daha uygun olduğu sonucuna varılmıştır.

Anahtar sözcükler: genellenebilirlik kuramı, performansın ölçülmesi

ABSTRACT: The aim of this study is to compare the results of G and D studies that were obtained by design that formed jointly and alternatively scoring of students by more than one rater in performance assessment in terms of generalizability theory. 48 students that were chosen randomly from the third-class degree students of the Medical Faculty of Hacettepe University at 2007-2008 academic years and 3 raters constitute the study group. In this study, G and D studies that done for s x t x r and (s: r) x t designs (s:student, t:task, r:rater) are compared. As a result of study, it is observed that variances that were estimated for variables in both designs are parallel to each others. Decision studies are no such a big difference. In this way scoring certain number of students alternately is much more convenient in time, labor and economy.

Keywords: generalizability theory, performance measurement

1. GİRİŞ

Büyük bir hızla gelişen bilgi ve teknoloji dünyasında değişimlere ayak uydurabilecek, yeni bilgileri kullanabilecek, araştırıp sorgulayabilecek ve öğrendikleriyle yeni bilgiler üretebilecek bireyler yetiştirmek önemlidir. Bu anlamda bireylerin çevresel, sosyal, evrensel ve bilimsel gelişimlere ulaşmasında eğitim temel yapıdır.

Miller (1990), tıp eğitiminde öğrenci başarılarını belirlemek için dört düzeyde tanımladığı bir öğrenme piramidi önermiştir (Şekil 1).

Şekil 1. Miller Piramidi (Vleuten, 2000)

* Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü 'nde Prof. Dr. Selahattin Gelbal danışmanlığında yapılan yüksek lisans tezinin bir kısmıdır.

** Eğitim Öğretim Planlamacısı, Nevşehir Üniversitesi, e-posta: fundan@nevsehir.edu.tr
*** Prof. Dr., Hacettepe Üniversitesi, e-posta: gelbal@hacettepe.edu.tr

Miller (1990), bu piramidi en alt basamağında öğrenci konunun "ne olduğunu bilir", onun üst basamağında "öğrenci nasıl olduğunu bilir", daha üst basamağında "öğrenci nasıl olduğunu gösterir" ve en üst basamakta ise "yapar" şeklinde tanımlamıştır (Norcini, 2003; Özvarış ve Sayek 2005; Vleuten, 2000). Bu durumda Miller piramidinin üst kısımları, yani öğrencilerin bir duruma ait becerileri yapıp göstermesi performans değerlendirmeyi açığa çıkarmaktadır.

Performans, "bir öğrenme görevine yönelik çabalar ve ortaya konulan ürün" (Büyüköztürk, 2007), performans değerlendirme ise "belirlenmiş bir görevi yürütmek için gerekli olan bilgi ve becerilerin yapay ya da gerçek yaşam durumlarında uygulanmasını içeren değerlendirme" (Arias, 2010); "gözlem ve kanıya dayanan değerlendirme" (Palm, 2008) olarak tanımlanmaktadır.

Performansın ölçülmesinde öğrenciler kendilerine verilen performans görevlerini yerine getirebilmek için ön bilgilerini sorgular, kullanır veya çeşitli araştırmalar yapar (Bekiroğlu, 2008). Böylelikle performans ölçülürken öğrencilerin öğrenmeleri pekiştirilmiş, öğrencilerin bilgiyi nasıl anladıkları ve bu bilgiyi nasıl kullandıkları belirlenmiş olur (Brualdi, 1998).

Öğrencilerin bir duruma ait performanslarının değerlendirilmesinde ölçülecek performans durumuna ait davranışların gözlenmesi gerekmektedir. Bu nedenle performans sınavlarını diğer sınavlardan ayıran önemli özelliklerden biri de öğrencilere bilginin uygulanmasını gösterebilme fırsatı vermesidir (Brualdi, 1998). Bu açıdan bakıldığında öğretmen öğrencinin performansını gözlemlerken öğrencinin neleri bilip neleri bilmediğini, kavramsal yanlışlıklarının olup olmadığını, varsa neler olduğunu belirleyebilir (Moskal, 2003).

Performans değerlendirmede öğrenciyi güvenilir puanlamak oldukça önemlidir. Puanlayıcılar performansın ölçülmesinde güvenirliği etkileyen önemli hata kaynağı olmakla birlikte ölçümlere karışan diğer değişkenlik kaynakları ve puanlayıcıların birey, görev gibi birçok değişkenlik kaynaklarıyla etkileşimleri de güvenirlik için önemlidir (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001). Bu nedenle güvenirlik belirlenirken birçok değişkenlik kaynağından gelen hatalar ve bu değişkenlik kaynaklarının birbirleriyle etkileşimleri sonucu çıkabilecek hatalar da dikkate alınmalıdır. Güvenirliğin ölçülmesinde değişkenlik kaynaklarını ve bu kaynaklar arasındaki etkileşimleri de bir arada değerlendirebilen yöntemlerden biri de Genellenebilirlik Kuramı'dır.

1.1. Genellenebilirlik Kuramı

Genellenebilirlik kuramı, davranış ölçümlerinin güvenilirliğinin değerlendirilmesini, gözlenen puanlardaki tutarsızlık kaynaklarının miktarının belirlenmesini sağlayan istatistiksel bir kuramdır (Brennan, 2001). Shavelson ve Webb (1991) ise genellenebilirlik kuramını, davranışsal ölçümlerin güvenilirliğinin istatistiksel bir teorisi olarak tanımlamaktadır. Genellenebilirlik kuramı kapsamlı bir çerçeve ve ölçme sonuçları için istatistiksel yollar ileri sürer. Aynı zamanda bu kuram test puanlarının ve puanlayıcılar arası tutarlılığın da bir ölçüsüdür (Brennan, Yin & Kane, 2003). Genellenebilirlik (G) kuramının temeli varyans analizi (ANOVA) üzerine kurulmuştur. Varyans analiziyle toplam varyans desendeki bağımsız değişkenlere bölünür. Böylece ölçme sonuçları farklı varyans kaynaklarına ayrılarak bireylerin ya da objelerin gözlenen puanlarının evren puanlarına (gerçek puanlarına) genellenebilmesi sağlanır.

Çalışmada kullanılan veriler OSCE (Objektif Yapılandırılmış Klinik Sınav) sınavının iletişim becerileri istasyonundan alınmıştır. OSCE sınavı ve iletişim becerileri istasyonu hakkındaki bilgiler aşağıda verilmiştir.

1.2. Osce Sınavı

OSCE (Objective Structured Clinical Examination) sınavı üst düzey düşünme becerilerini ölçmede, birden fazla istasyondan oluşan ve her bir istasyonda farklı klinik becerilerin önceden belirlenen kriterler doğrultusunda puanlayıcılar tarafından değerlendirildiği performans sınavıdır. Bu sınav, 1977 yılında İskoçya Dundee Üniversitesi'nde Ronald Harden tarafından cerrahi bölümünde yapılan sınavlar için kullanılmıştır (Elçin, Odabaşı ve Sayek, 2005).

Hacettepe Üniversitesi Tıp Fakültesi'nde OSCE uygulamaları 2004-2005 öğretim yılından itibaren Dönem IV Genel Cerrahi staj sınavında, Dönem I-III İyi Hekimlik Uygulamaları programının yıl sonu değerlendirmesinde kullanılmaktadır (Elçin, Odabaşı ve Sayek, 2005). Sınav verilerinin alındığı Dönem III OSCE sınavı, farklı klinik becerilerin ölçüldüğü 6 istasyondan oluşmaktadır. Çalışma kapsamında kullanılan iletişim becerileri istasyonu standart hasta görüşmelerinin yapıldığı, öğrencilerin hasta ile görüşme süreçlerinin değerlendirildiği bir istasyondur. Bu istasyonda öğrencilerin hasta ile görüşmesiyle ilgili temel iletişim becerileri (görüşmeyi başlatma, bilgi toplama, ilişkiyi kurma, açıklama ve planlama, görüşmeyi bitirme) değerlendirilmektedir (Elçin, Odabaşı ve Sayek, 2005). İletişim becerileri istasyonunda öğrencilerin performanslarının değerlendirilmesinde bir hastalık öyküsü üzerinde eğitilmiş, rol yapma yeteneği yüksek standart hastalar kullanılmaktadır.

Öğrencilerin hasta ile görüşme süreçleri tıp eğitiminin önemli bir parçasıdır. İletişim becerileri ile yapılan çalışmalarda hekimlerin hastaları dinlemede çok zaman ayırmadıkları, bununla birlikte hasta yakınmalarının bir kısmının açıklanamadığı, hastaların hekimlerin konuşma dilini anlamadıklarını, bu nedenlerle de doğru ilaç kullanımının azaldığı, hasta memnuniyetsizliklerinin arttığı gibi sonuçlar elde edilmiştir (Kurtz, 1998 akt. Elçin ve ark. 2007). Bu açıdan bakıldığında doktor adaylarının hastalarla görüşmesinde iletişim becerilerinin kazandırılması ve değerlendirilmesi büyük önem kazanmaktadır.

Bilindiği gibi performans değerlendirmede birden fazla puanlayıcının kullanılması puanlama güvenirliğini olumlu yönde etkilemektedir. Fakat OSCE sınavı gibi öğrencilerin birden fazla performans durumlarının ölçüldüğü, öğrenci sayısının fazla olduğu sınavlarda öğrenci performanslarını değerlendirmek sınavda kullanılan puanlayıcılar için iş gücü ve zaman açısından bazı sınırlılıklar taşımaktadır. Bu nedenle araştırmanın amacı; genellenebilirlik kuramına göre performans puanlamada öğrencilerin birden fazla puanlayıcı tarafından birlikte ve dönüşümlü olarak puanlanmasıyla oluşturulan desenlerden elde edilen G çalışması ve bu desenlerle yapılan karar çalışmaları sonuçlarını karşılaştırmaktır. Bu amaçla araştırmanın çok sayıda öğrencinin ve birden fazla puanlayıcının kullanıldığı performans sınavlarında kısa zamanda güvenilir sonuçlar elde etmek için puanlayıcıların nasıl kullanılması gerektiği sorusuna cevap teşkil etmesi beklenmektedir. Aynı sınav için oluşturulmuş farklı desenlerden hangisinin kullanılmasının daha uygun ve bilgi verici olduğu belirlenerek araştırmanın uzun zaman, maliyet ve iş gücü gerektiren performans sınavlarına katkı sağlaması beklenmektedir.

Ayrıca genellenebilirlik kuramı birçok alanda önemli bilgiler sağlamasıyla birlikte Türkiye'de çok yaygın olarak kullanılmamaktadır. Bu nedenle araştırma genellenebilirlik kuramının kullanımını ve değişkenlerin yuvalanmış veya çapraz olarak tasarlanması sonucu oluşturulan desenlerin kullanılabilirliğini örnekleme amacı da taşımaktadır. Tüm bu amaçlar doğrultusunda "genellenebilirlik kuramına göre performans puanlamada öğrencilerin birden fazla puanlayıcı tarafından, birlikte ve dönüşümlü olarak puanlamalarıyla oluşturulan farklı desenlerin genellenebilirlik (G) ve karar (K) çalışmaları sonuçları nasıldır?" genel araştırma sorusu doğrultusunda aşağıdaki alt sorulara yanıt aranmıştır.

a. ö x g x p ve (ö:p) x g desenleriyle yapılan G çalışmaları sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri nasıldır?

b. ö x g x p ve (ö:p) x g desenlerinde puanlayıcı ve öğrenci sayılarının artırılıp azaltılmasıyla yapılan karar çalışmalarından elde edilen G ve Phi katsayılarının karşılaştırılması nasıldır?

2. YÖNTEM

2.1. Çalışma Grubu

Araştırmanın çalışma grubunu, 2007- 2008 öğretim yılı Hacettepe Üniversitesi Tıp Fakültesi üçüncü sınıf öğrencilerinden tesadüfi olarak seçilen 48 öğrenci oluşturmaktadır. Çalışmada belirlenen 48 öğrencinin iletişim becerileri 15 görev doğrultusunda 3 puanlayıcı tarafından değerlendirilmiştir. Araştırmada öğrencilerin iletişim becerileri istasyonundaki hasta görüşmelerinin puanlanmasında

görev alan puanlayıcılar tıp eğitimi alanında olup; bir araştırma görevlisi, bir doktor ve bir doçentten oluşmaktadır.

2.2. Çalışma Verileri

Araştırmada, Hacettepe Üniversitesi Tıp Fakültesi, Tıp Eğitimi ve Bilişimi Anabilim Dalı 2007-2008 akademik yılı OSCE final sınavına ait iletişim becerileri istasyonundan elde edilen veriler kullanılmıştır.

İletişim becerileri istasyonunda öğrenciler aynı senaryo üzerinden belirlenen bir hastayla görüşme yapmaktadır. İstasyonda her bir öğrenciye hasta ile görüşmeleri için eşit süre verilmektedir. Sınav esnasında, iletişim becerileri istasyonunda tek puanlayıcı bulunmaktadır. Bu puanlayıcı Tıp Bilişimi ve Eğitimi Anabilim Dalı tarafından OSCE sınavında kullanılmak üzere hazırlanan iletişim becerisi değerlendirme formunu kullanarak, öğrencilerin hasta ile görüşme süreçlerini gözlendi (1)-gözlenmedi (0) şeklinde puanlamıştır. Ayrıca iletişim becerileri değerlendirme formunda 15 görev bulunmaktadır.

Öğrencilerin iletişim becerileri istasyonundaki hasta görüşmeleri sınav esnasında kayıt altına alınmıştır. Araştırmada kullanılan diğer puanlayıcılar sınav esnasında değerlendirme yapan puanlayıcının puanladığı öğrencileri aynı iletişim becerileri değerlendirme formuyla kamera kayıtlarını izleyerek birbirlerinden bağımsız olarak puanlamıştır.

Tüm veriler elde edildikten sonra çalışmada kullanılmak üzere araştırmacı tarafından iki farklı senaryo üzerinden genellenebilirlik kuramı analizleri için iki desen tasarlanmıştır. Bu desenlerden ilki, öğrenci (ö), görev (g) ve puanlayıcı (p) değişkenleri olmak üzere öğrencilerin aynı iletişim becerileri değerlendirme formundaki görevler doğrultusunda puanlayıcılar tarafından hasta ile görüşme süreçlerinin puanlamasıyla oluşturulmuş ö x g x p desenidir. İkincisi ise aynı hasta görüşme verileri kullanılarak, her bir puanlayıcının araştırmaya katılan öğrencilerden sadece bir kısmını puanlamasıyla öğrenci ve puanlayıcı değişkenlerinin yuvalanmış, görevlerin ise bu değişkenlerle çaprazlandığı (ö:p) x g desenidir.

2.3. Verilerin Analizi

Araştırmanın ilk kısmında 48 öğrencinin (ö) 15 görev (g) doğrultusunda 3 puanlayıcı (p) tarafından puanlanması ile ö x g x p deseni için genellenebilirlik (G) ve karar (K) çalışması yapılmıştır. Araştırmanın ikinci kısmında ilk desende kullanılan verilerle 3 puanlayıcıdan her birinin bu kez 16'şar öğrenciyi 15 görev doğrultusunda puanlanmasıyla oluşturulmuş (ö:p) x g deseni için de genellenebilirlik ve karar çalışması yapılmıştır. Her iki desenden elde edilen sonuçlar karşılaştırılmıştır.

Genellenebilirlik kuramı ile desenlere ait varyans bileşenlerinin kestirilmesinde, değişkenlerin toplam varyansı açıklama oranlarının hesaplanmasında ve her bir desen için karar çalışmalarının yapılmasında EduG 3.07 programı kullanılmıştır.

3. BULGULAR VE YORUM

Elde edilen bulgular ve yorumlar G çalışması ve karar çalışması sonuçları olarak aşağıda verilmiştir.

3.1. ö x g x p ve (ö:p) x g Desenlerinden Elde Edilen G Çalışması Sonuçları

Aynı verilerle farklı senaryo durumuna göre oluşturulmuş ö x g x p ve (ö:p) x g desenlerine ait varyans ve toplam varyansları açıklama yüzdeleri Tablo 1'de gösterilmiştir.

Tablo 1: ö x g x p ve (ö:p) x g Desenlerinden Elde Edilen Varyans ve Toplam Varyansları Açıklama Yüzdeleri

	ö x g x p			(ö:p) x g	
Varyans Kaynağı	**σ^2**	**%**	**Varyans Kaynağı**	**σ^2**	**%**
ö	0.00266	2.1			
g	0.04705	37.1	**g**	0.05109	41.3
p	0.00114	0.9	**p**	0.00094	0.8
öp	0.00230	1.8	**ö : p**	0.00239	1.9
ög	0.02008	15.8	**gp**	0.00507	4.1
gp	0.00808	6.4	**gö : p**	0.06409	51.9
ögp	0.04542	35.8			
Toplam	0.12673	100	**Toplam**	0.12358	100

ö: öğrenci, g:görev ve p:puanlayıcı (ng=15, np=3, ö x g x p; nö= 48 ve (ö: p) x g; nö:p=16)

Tablo 1'deki verilere göre görevlere ait varyans bileşenlerinin toplam varyansı açıklama oranı, ö x g x p deseninde %37.1 ve (ö:p) x g deseninde %41.3 olarak hesaplanmıştır. Görevlere ait varyans bileşenlerinin yüksek çıkması her iki desende de görevlerin zorluk-kolaylık bakımından farklılaştığını göstermektedir. Ayrıca Tablo 1 incelendiğinde (ö:p)x g deseninde kestirilen görev ana etkisine ait varyans bileşeninin, ö x g x p deseninde kestirilen varyans bileşenine göre daha büyük kestirildiği görülmektedir. Böylece (ö:p)xg deseniyle görevlerin daha iyi ayırt edildiği yorumu yapılabilir.

Her iki desenle kestirilen puanlayıcı ana etkisine ait varyans bileşeni incelendiğinde, her iki desende de puanlayıcı etkisine ait varyans bileşeninin sıfıra oldukça yakın olduğu bulunmuştur. Bu bağlamda, puanlayıcıların her iki desende de puanlama bakımından farklılığa neden olmadığı, puanlayıcıların öğrencileri tutarlı puanladığı yorumu yapılabilir. Ayrıca Tablo 1'de ö x g x p ve (ö:p)xg desenlerinde görev x puanlayıcı ortak etkileşimine ait varyans bileşenlerinin (ö x g x p deseninde σ^2(gp)=0.00808 ve (ö:p)xg deseninde σ^2(gp)=0.00507), sırasıyla toplam varyansın %6.4 ve %4.1'ini açıkladığı görülmektedir. Böylece her iki desendeki görev x puanlayıcı ortak etkileşimleri arasında da çok büyük farklılığın olmadığı, puanlayıcıların bir görevden diğerine kararlı puanlama yaptıkları söylenebilir.

Puanlayıcıların aynı kontrol listesini kullanarak öğrencileri dönüşümlü puanlamasıyla tasarlanan (ö:p) x g deseninde öğrencilerin puanlayıcılarla yuvalandığı (ö:p) değişkene ait varyans bileşeni (σ^2(ö:p)= 0.00239) oldukça küçüktür. (ö:p) x g deseninde öğrenci ana etkisi ve öğrenci x puanlayıcı ortak etkileşimleri ayrı değerlendirilmek yerine, (ö:p) değişkeni altında ortak değerlendirilmiştir. Bir başka deyişle σ^2(ö:p)= σ^2(ö,öp)= σ^2(ö)+ σ^2(öp)' dir (Brennan, 2001). Buna rağmen tüm puanlayıcıların aynı kontrol listesiyle bütün öğrencileri puanlamasıyla oluşturulan ö x g x p deseninde öğrenci ana etkisi (ö) ve öğrenci x puanlayıcı ortak etkileşimi (öp) ayrı ayrı hesaplanmış ve incelenmiştir.

Tablo 1'deki desenler için ö, öp ve ö:p varyans bileşenleri incelendiğinde her iki desende de öğrencilerin iletişim becerileri bakımından farklılaşmadığı ve öğrencilerin durumlarının bir puanlayıcıdan diğerine değişmediği yorumu yapılabilir. Ayrıca ö x g x p deseninde öğrenci (%2.1) ve öğrenci x puanlayıcı ortak etkileşimi (%1.8) küçük çıkmasına rağmen öğrenci-görev (σ^2(ög)) ortak etkisi (%15.8) büyük çıkmıştır. Bu öğrencilerin durumlarının bir görevden diğerine farklılaştığının bir göstergesidir.

Tablo 1 incelendiğinde tüm değişkenlerin çaprazlandığı ö x g x p deseninde artık varyansın (σ^2(ögp)), toplam varyansı açıklama oranının %35.8 olduğu, öğrencilerin puanlayıcılarla yuvalandığı, görevlerin öğrenciler ve puanlayıcılar ile çaprazlandığı (ö:p)xg deseninde ise artık varyansın (σ^2(gö:p), toplam varyansın %51.9'unu açıkladığı görülmektedir. Tablo 1'den görüldüğü gibi her iki desende de artık varyans yüksek çıkmış olmasına rağmen (ö:p) x g deseninde bu oran daha yüksek çıkmıştır. Bu durumun (ö:p) x g deseninde, tüm değişkenlerin çaprazlandığın ö x g x p deseninden farklı olarak

öğrenci x görev ortak etkisine ait varyans bileşeninin σ^2(ög) artık varyansa dahil edilmiş olmasından kaynaklandığı söylenebilir.

3.2. ö x g x p ve (ö:p) x g Desenleri İle Senaryolara Göre Yapılan K Çalışmalarından Elde Edilen G ve Phi Katsayılarının Karşılaştırılması

ö x g x p ve (ö:p) x g desenleri ile öğrenci ve puanlayıcı sayılarının artırılıp azaltılmasıyla yapılan karar çalışması ile elde edilmiş G ve Phi katsayılarına ait veriler Tablo 2'de gösterilmiştir.

Tablo 2: ö x g x p ve (ö:p) x g Desenlerine Ait K Çalışması ile Puanlayıcı ve Öğrenci Sayılarının Değiştirilmesiyle Oluşturulmuş Senaryolara Göre G ve Phi Katsayıları

		ö x g x p					**(ö:p) x g**	
n_p	$n_ö$	**G**	**Phi**	n_p	$n_{ö:p}$	$n_ö$	**G**	**Phi**
2	**24**	0.88984	0.87776	**2**	**12**	**24**	0.90756	0.89850
2	**48**	0.90510	0.89396	**2**	**24**	**48**	0.92961	0.92094
2	**72**	0.91030	0.89949	**2**	**36**	**72**	0.93720	0.92867
3	**24**	0.91873	0.90946	**3**	**8**	**24**	0.92138	0.91459
3	**48**	**0.93208**	**0.92384**	**3**	**16**	**48**	**0.94412**	**0.93785**
3	**72**	0.93662	0.92874	**3**	**24**	**72**	0.95195	0.94587
4	**24**	0.93389	0.92618	**4**	**6**	**24**	0.92845	0.92286
4	**48**	0.94619	0.93954	**4**	**12**	**48**	0.95154	0.94654
4	**72**	0.95036	0.94408	**4**	**18**	**72**	0.95950	0.95471

ö: öğrenci, g: görev ve p: puanlayıcı

ö x g x p deseninde öğrenciler tüm puanlayıcılar tarafından aynı görevler doğrultusunda değerlendirilmektedir. Araştırmada kullanılan veriler ışığında 48 öğrenciden her birinin 3 puanlayıcı tarafından 15 görev doğrultusunda puanlanmasıyla elde edile G katsayısı 0.93208, Phi katsayısı ise 0.92384 olarak kestirilmiştir. Aynı verilerle farklı senaryo durumuna göre oluşturulmuş (ö:p) x g deseninde ise ilk desende kullanılan puanlayıcıların her biri, bu kez farklı öğrencileri aynı görevler doğrultusunda puanlamıştır. Böylece (ö:p) x g deseni ile 48 öğrencinin 3 puanlayıcı tarafından fakat her bir puanlayıcının 48 öğrenciden sadece 16'sını 15 görev doğrultusunda puanlamasıyla elde edilen G katsayısı 0.94412, Phi katsayısı ise 0.93785 olarak kestirilmiştir. Aynı verilerle, puanlayıcıların öğrencileri farklı puanlamasıyla oluşturulmuş desenlerden ö x g x p deseni ile yapılan karar çalışmasıyla kestirilen G katsayısı, (ö:p) x g deseni ile kestirilen G katsayısından 0.01204 ve Phi katsayısı ise (ö:p) x g deseni ile kestirilen Phi katsayısından 0.01401 daha küçük hesaplanmıştır. Böylelikle; aynı veriler kullanılarak farklı senaryo durumuna göre oluşturulmuş iki desenden, öğrencilerin puanlayıcılarla yuvalandığı (ö:p) x g deseninde G ve Phi katsayılarının daha yüksek kestirildiği yorumu yapılabilir.

Tablo 2'den ö x g x p deseninde araştırmada kullanılan öğrenci sayısı sabit tutularak ($n_ö$=48) puanlayıcı sayısı azaltıldığında (n_p=2) G katsayısının 0.90510, puanlayıcı sayısı artırıldığında (n_p=4) ise G katsayısının 0.94619 olduğu görülmektedir. Aynı şekilde (ö:p) x g deseni için toplam öğrenci sayısı ($n_ö$=48) sabit tutularak puanlayıcı sayısı azaltıldığında (n_p=2, $n_{ö:p}$=24) G katsayısının 0.92961, puanlayıcı sayısı (n_p=4, $n_{ö:p}$=12) artırıldığında ise G katsayısının 0.95154 olduğu görülmektedir. Bu verilerden hareketle öğrenci sayısı sabit tutularak puanlayıcı sayısının azaltılması ve artırılması durumlarında G katsayısının (ö:p) x g deseninde tüm değişkenlerin çaprazlandığı desene göre daha yüksek kestirildiği söylenebilir.

Tablo 2'den araştırmada kullanılan $n_ö$=48, n_p=3 durumuna göre puanlayıcı sayısının azaltılıp artırılması senaryoları için Phi katsayıları incelenirse, ö x g x p deseninde öğrenci sayısı sabit tutularak ($n_ö$=48) puanlayıcı sayısı azaltıldığında (n_p=2) Phi katsayısının 0.89396, puanlayıcı sayısı artırıldığında (n_p=4) ise Phi katsayısının 0.93954 olarak hesaplandığı görülmektedir. Aynı şekilde (ö:p)xg deseni için toplam öğrenci sayısı ($n_ö$=48) sabit tutularak puanlayıcı sayısı azaltıldığında (n_p=2, $n_{ö:p}$=24) Phi katsayısı 0.92094, puanlayıcı sayısı (n_p=4, $n_{ö:p}$=12) artırıldığında ise Phi katsayısı 0.94654 olarak hesaplanmaktadır. Böylece toplam öğrenci sayısı sabit tutularak puanlayıcı sayısının azaltılması ve artırılması durumlarında Phi katsayısının (ö:p) x g deseninde ö x g x p desenine göre daha yüksek hesaplandığı söylenebilir.

Tablo 2'deki verilerden bu kez puanlayıcı sayısı sabit tutularak (n_p=3) öğrenci sayısı azaltıldığında ($n_ö$=24) ö x g x p deseninde G katsayısının 0.91873, öğrenci sayısı artırıldığında ($n_ö$=72) ise G katsayısının 0.93662 olduğu görülmektedir. Aynı durumda (ö:p) x g deseninde ise puanlayıcı sayısı sabit tutularak (n_p=3) puanlanan öğrenci sayısı ($n_{ö:p}$=8, $n_ö$=24) azaltıldığında G katsayısının 0.92138, puanlanan öğrenci sayısı ($n_{ö:p}$=24, $n_ö$=72) artırıldığında ise G katsayısının 0.95195 olduğu görülmektedir. Aynı senaryolar için Phi katsayısı incelenirse, ö x g x p deseninde puanlayıcı sayısı sabit tutularak (n_p=3) öğrenci sayısı azaltıldığında ($n_ö$=24) Phi katsayısı 0,90946, öğrenci sayısı artırıldığında ($n_ö$=72) ise Phi katsayısı 0,92874 olarak hesaplanmaktadır. Aynı durumda (ö:p) x g deseninde puanlayıcı sayısı sabit tutularak öğrenci sayısı ($n_{ö:p}$=8, $n_ö$=24) azaltıldığında Phi katsayısı 0.91459, artırıldığında ($n_{ö:p}$=24, $n_ö$=72) ise Phi katsayısı 0.94587 olarak hesaplanmaktadır. Böylece puanlayıcı sayısı sabit tutularak öğrenci sayısının azaltılması ve artırılması durumlarında G ve Phi katsayılarının (ö:p) x g deseninde tüm değişkenlerin çaprazlandığı desene göre daha yüksek çıkma eğiliminde olduğu söylenebilir.

Her iki desen içinde puanlayıcı ve öğrenci sayısının (n_p=4, $n_ö$=72) artırılması durumunda ise ö x g x p deseninde G katsayısı 0.95036 ve Phi katsayısı 0.94408, (ö:p) x g deseninde ise G katsayısı 0.95950, Phi katsayısı 0.95471 olarak hesaplanmaktadır. Böylece öğrenci ve puanlayıcı sayıları artırıldığında G ve Phi katsayılarının (ö:p) x g deseninde daha yüksek kestirildiği görülmektedir.

Genel olarak Tablo 2 incelendiğinde, her iki desende tüm senaryo durumlarına göre G katsayılarının Phi katsayılarından daha büyük kestirildiği, puanlayıcı ve öğrenci sayılarının artırılmasının her iki desende de G ve Phi katsayılarını artırdığı görülmektedir. En önemlisi de farklı puanlayıcı ve öğrenci senaryolarına göre ö x g x p deseni ile kestirilen G ve Phi katsayılarının (ö:p) x g deseni ile kestirilen G ve Phi katsayılarına göre daha küçük çıkma eğiliminde olduğu görülmektedir.

Tüm bulgulardan hareketle her iki deseninde G çalışması sonucunda değişkenlik kaynaklarına ait paralel sonuçlar ürettiği, desenlere ait K çalışmaları sonuçlarının da çok fazla farklılık göstermediği tespit edilmiştir.

4. SONUÇ VE ÖNERİLER

Bu bölümde, araştırmanın amaçları doğrultusunda elde edilen bulgulara dayalı olarak sonuçlara ve önerilere yer verilmiştir.

ö x g x p ve (ö:p) x g desenlerinden G çalışmasıyla elde edilen varyans ve toplam varyansı açıklama oranları karşılaştırıldığında;

1. Görev ana etkisine ait varyans bileşenleri her iki desende de yüksek çıkmıştır. Böylece görevlerin her iki desende de zorluk-kolaylık bakımından farklılaştığı sonucuna varılmıştır.

2. Her iki desende de puanlayıcılara ait varyans değerlerinin toplam varyansı açıklama oranları düşük çıkmıştır. Bu durumda puanlayıcıların tutarlı puanlama yaptıkları sonucuna varılmıştır.

3. ö x g x p ve (ö:p) x g desenlerinde görev x puanlayıcı (gp) ortak etkileşimleri için kestirilen varyans değerleri arasında çok büyük farklılık yoktur. Her iki desende de, puanlayıcıların bir görevden diğerine tutarlı puanlama yaptığı sonucuna varılmıştır.

4. (ö:p) x g deseninde öğrenci ana etkisi ve öğrenci x puanlayıcı ortak etkileşimleri ayrı ayrı hesaplanmak yerine, (ö:p) değişkeni altında ortak hesaplanmıştır. Buna rağmen ö x g x p deseninde

öğrenci ana etkisi (ö) ve öğrenci x puanlayıcı ortak etkileşimine (öp) ait varyans değerlerinin ayrı ayrı hesaplandığı sonucu çıkartılmıştır.

5. ö x g x p deseni için ö ve öp, (ö:p) x g deseni içinse ö:p varyans bileşenleri incelendiğinde her iki desende de öğrencilerin iletişim becerileri bakımından farklılaşmadığı ve öğrencilerin durumlarının bir puanlayıcıdan diğerine değişmediği sonucuna varılmıştır.

6. Öğrenci x görev (ög) ortak etkileşimi (ö:p) x g deseninde artık varyansa ilave edilmiş olmakla birlikte bu değer ö x g x p deseninde ayrı bir varyans kaynağı olarak incelenmiştir. Böylece ö x g x p deseninde öğrencilerin bir görevden diğerine değişiklik gösterip göstermediğiyle ilgili yorum yapabilmenin mümkün olduğu sonucuna varılmıştır.

7. Her iki desende de artık varyans yüksek çıkmış olmasına rağmen (ö:p) x g deseninde artık varyans daha yüksek bulunmuştur. Bu durumun (ö:p) x g deseninde, tüm değişkenlerin çaprazlandığın ö x g x p deseninden farklı olarak öğrenci x görev ortak etkisine ait varyans bileşeninin (σ^2(ög)) artık varyansa dahil edilmiş olmasından kaynaklandığı söylenebilir.

Sonuç olarak ö x g x p ve (ö:p) x g desenleri ile yapılan analizler sonucunda her iki desende değişkenler için kestirilen varyans değerlerinin birbirleriyle paralellik gösterdiği, farklılık olarak ö x g x p deseninin öğrenci x görev ortak etkileşimi için de bilgi ürettiği görülmektedir.

ö x g x p ve (ö:p) x g desenlerinde öğrenci ve puanlayıcı sayılarının artırılıp azaltılmasıyla yapılan karar çalışmalarında;

1. Öğrenci sayısı sabit tutularak puanlayıcı sayısının azaltılması ve artırılması durumlarında her iki desenle kestirilen G ve Phi katsayıları arasında çok fark olmamakla birlikte, bu katsayıların (ö:p)xg deseninde daha yüksek kestirilme eğiliminde olduğu,

2. Puanlayıcı sayısı sabit tutularak öğrenci sayısının azaltılması ve artırılması durumlarında her iki desenle kestirilen G ve Phi katsayıları arasında çok fark olmamakla birlikte, bu katsayıların (ö:p)xg deseninde daha yüksek çıkma eğiliminde olduğu,

3. Araştırmada kullanılan verilere göre öğrenci ve puanlayıcı sayılarının birlikte artırılması durumunda G ve Phi katsayılarının (ö:p)xg deseninde daha yüksek kestirildiği sonuçlarına varılmıştır.

Yapılan karar çalışmalarıyla her iki desende de puanlayıcı sayısı artırıldığında G ve Phi katsayıları artmaktadır. Fakat puanlayıcı sayısını artırmak her zaman, her durumda mümkün olmamaktadır. Özellikle puanlayıcıların öğrencilerle çaprazlandığı, yani sınava katılan öğrencilerin her birinin bütün puanlayıcılar tarafından puanlaması durumunda puanlayıcı sayısını artırmak zaman ve iş gücü açısından ekonomik olmamaktadır.

Sonuç olarak, puanlayıcı ve öğrenci sayılarının artırılmasının her iki desende de G ve Phi katsayılarını artırdığı, en önemlisi de puanlayıcıların öğrencileri dönüşümlü puanladığı (ö:p) x g deseni ile kestirilen G ve Phi katsayılarının ö x g x p deseni ile kestirilen G ve Phi katsayılarına göre daha büyük çıkma eğiliminde olduğu görülmektedir.

Sınavda puanlayıcıların öğrencileri puanlamada aralarında farklılıklar olmadığı, puanlayıcılar arasında tutarlılık olduğu bulunmuştur. Her iki desenle yapılan karar çalışmaları sonucunda öğrencilerin puanlayıcılarla çapraz tasarlandığı ve öğrencilerin puanlayıcılarla yuvalandığı desenlerde G ve Phi katsayıları arasında çok fark olmamakla birlikte (ö:p) x g deseninde katsayıların daha büyük çıkma eğiliminde olduğu görülmektedir. Bu sonuçlardan hareketle OSCE sınavına çok sayıda öğrencinin katılması ve sınavın birden fazla istasyondan oluşması nedeniyle öğrencilerin puanlayıcıların hepsi tarafından tek tek puanlanması yerine puanlayıcıların belli sayıdaki öğrencileri dönüşümlü olarak puanlamasının zaman, iş gücü ve ekonomiklik açısından daha uygun olduğu sonucuna varılmıştır. Ayrıca yapılan karar çalışmaları sonucunda puanlayıcı sayısının artırılması ve puanlayıcıların puanladığı öğrenci sayısının artması da güvenirliği olumlu yönde etkilemektedir. Fakat sınav öğrencilerin puanlayıcılar tarafından dönüşümlü puanlanması şeklinde organize edildiğinde puanlayıcı sayısını artırmak G ve Phi katsayılarında çok az bir değişiklik yapmaktadır. Bu nedenle sınav, puanlayıcıları puanlamada yormayacak ve sıkmayacak şekilde her bir puanlayıcıya çok sayıda öğrenci verilmeden öğrenci sayısına göre puanlayıcı sayısının belirlenmesi uygun görülmüştür.

Bu sonuçlardan hareketle aşağıdaki önerilerde bulunulabilir.

1. Çok sayıda öğrencinin bulunduğu performans sınavlarında puanlayıcılar arası tutarlılık sağlandığında her bir puanlayıcının sınava katılan bütün öğrencileri puanlaması yerine puanlayıcıların belli sayıdaki öğrencileri dönüşümlü olarak puanlaması zaman ve iş gücü açısından daha uygundur.

2. Puanlayıcıların öğrencileri dönüşümlü olarak puanlayabilmesi için puanlayıcılar arası tutarlılığın olmasına dikkat edilmelidir. Puanlayıcılar arası tutarlılığın olmadığı bir durumda puanlayıcıların öğrencileri dönüşümlü puanlaması güvenilir sonuçlar vermeyebilir.

KAYNAKLAR

Arias, R. M. (2010). Performance Assessment. Papeles del Psicologo, 31(1), 85-96.

Atılgan, H. (2004). *Genellenebilirlik Kuramı Ve Çok Değişkenlik Kaynaklı Rasch Modelinin Karşılaştırılmasına İlişkin Bir Araştırma.* Doktora Tezi, Hacettepe Üniversitesi, Ankara.

Bekiroğlu, F. (2008). Performansa Dayalı Ölçümler: Teori ve Uygulama. *Türk Fen Eğitim Dergisi,* 5(1), 113- 131.

Brennan, R. L. (2000). Performance Assessments From The Perspective of Generalizability Theory. *Applied Psychological Measurement,* 24(4), 339- 353.

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

Brennan, R. L., Yin, P., & Kane, M. T. (2003). Methodology for Examining The Reliability of Group Mean Difference Scores. *Journal of Educational Measurement*, 40(3), 207-230.

Brualdi, A. (1998). *Implementing Performance Assessment in The Classroom.* Practical Assessment, Research & Evaluation, 6(2). Erişim: 29 Ocak 2009, http://PAREonline.net/getvn.asp?v=6&n=2.

Büyüköztürk, Ş. (2007). Performansa Dayalı Durum Belirleme nedir? *İlköğretmen Dergisi*, 8, 28-32.

Chang, L., & Hocevar, D. (2000). Models of Generalizability Theory in Analyzing Existting Faculty Evaluation Data. *Applied Measurement In Education*, 13(3), 255-275.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.

Elçin, M., Odabaşı, O. ve Sayek, İ. (2005). Yapılandırılmış Objektif Klinik Sınavlar. *Hacettepe Tıp Dergisi,* 36, 1-2.

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

Kane, M. T. (2003). Generalizability Theory. *International Journal of Testing*, 3(1), 95-100.

Lei, P., Smith, M., & Suen, H. K. (2007). The Use of Generalizability Theory To Estimate Data Reliability in Single Subject Observational Research. *Psychology in the Schools*, 44(5), 433-439.

Miller, G. E. (1990). The Assessment of Clinical Skills, Competence, Performance. *Academic Medicine*, 65, 68-67.

Moskal, B. M. (2003). *Recommendations for Developing Classroom Performance Assessments and Scoring Rubrics.* Practical Assessment, Research & Evaluation, 8(14). Erişim: 29 Ocak 2009, http://PAREonline.net/getvn.asp?v=8&n=14.

Norcini, J. (2003). Work Based Assessment. *ABC of Learning and Teaching in Medicine*, 326(7392), 753-755.

Önal, İ. (2005). *İlköğretim Fen Bilgisi Öğretiminde Performans Dayanıklı Durum Belirleme Uygulaması Üzerine Bir Çalışma.* Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara.

Özvarış, Ş. B. ve Sayek, İ. (2005). Tıp Eğitiminde Değişim. *Hacettepe Tıp Dergisi*, 36, 65-74.

Palm, T. (2008). Performance Assessment and Authentic Assessment: A Conceptual Analysis of the Literature. *Practical Assessment, Research & Evaluation,* 13(4), 1-11.

Shavelson, J. R., & Webb N. M. (1991). *Generalizability Theory :A Primer***.** Newbury Park. CA: Sage Publications.

Vleuten, C. (2000). Validity of Final Examinations in Undergraduate Medical Training. *Education and Debate*, 312, 1217- 1219.

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

Extended Abstract

It is important to assess students reliability in performance assessment. Besides; rater agreement is important error for performance assessment, other facets (student, task and interaction them) are important for reliability. So in this study we use Generalizability Theory to assess a lot of facets and interaction them for reliability.

The "Objective Structured Clinical Examination" (OSCE) used to final exam for IV. term General Surgical Exam and I-III. term medicine practice in the Medical Faculty of Hacettepe University at 2004-2005 academic years (Elçin, Odabaşı ve Sayek, 2005). Third- class degree students' OSCE exam is six stations which assess different clinical skills. In this station, student comunicate with patient on a standart medical history and rater asses them according to communication skills. Standard patients used in communication station for assess students performances. Prior to application, they trained with a medical history.

It is known, using one more rater is better for score reliability in performance assessment. But like OSCE exam which student assesed for a lot of performance tasks and contain a lot of students is hard for raters. This exams a lot of limitedness like time, workforce etc. So the main goal of the study is to compare the results of Generalizability (G) and Decision (D) studies that were obtained by design that formed jointly and alternatively assessment of students by more than one rater in performance assessment in terms of generalizability theory. With this aim; the research is supposed to explain how must use rater for reliability results in the performance exams which used one more students and raters.

Besides Generalizability Theory uses a lot of researches, it doesn't use often in our country. So this research is an example of Generalizability Theory' using, nested and crossed designs.

In the direction of this aims the following questions look for an answer. How do the results of G and D studies that were obtained by design that formed jointly and alternatively assessment of students by more than one rater in performance assessment according to generalizability theory?

In this research, 48 students that were chosen randomly from the third-class degree students of the Medical Faculty of Hacettepe University at 2007-2008 academic years constitute the study group. Also three rater took in charge in scoring the patient discourses in communication skills station.

In this study, several Generalizability and Decision studies were done for s x t x r and (s: r) x t designs (s: student, t: task, r: rater) that formed by raters asses students by means of the same communication skills form jointly and alternately. 48 students assessed by tree raters in s x t x r designs. In other words each rater gived points all of the students. And in (s: r) x t designs each rater gived points only sixteen each. In total 48 students assessed by tree rater. Besides, G and D studies that done for the two designs are compared at the last part of the study. The Decision study is used to examine alternative ways of the number of raters so that cost and time expenditure can be minimized in future Objective Structured Clinical Exams.

The result of research is as in the following;

1. In accordance with the analysis done by s x t x r and (s: r) x t designs, it is observed that variance rates that were estimated for variables in both designs are parallel to each others.

2. In the exam, it is found that there is no difference between raters in scoring the students.

3. In both designs there is no difference between students comunication skills and performances.

4. As a result of decision studies that were done by each two designs, it is observed that there are no such a big difference between G and Phi factors; but in (s: t) x r designs these factors tend to be higher.

5. When the number of students is stable and the number of raters is decrease or increase, there is no difference G and Phi coefficients in both designs.

By considering the results of findings, the following advice may say.

G and Phi coefficients were little change, when the OSCE exam organize like scoring the students alternately and the number of raters was increased. And so the number of raters was thought according to the number of students. As a result, in performance exams that include so many students, when there is a consistence between raters, scoring certain number of students alternately is much more convenient in time, labor and economy rather than scoring students particularly by the all graders. And before scoring certain number of students alternately, rater agrement must be provided.