Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
ISSN: 1309 - 6575

Genellenebilirlik Kuramıyla Dikiş Atma ve Alma Becerileri İstasyonu Güvenirliğinin Değerlendirilmesi*

Using Generalizability Theory to Assess Reliability of Suturing and Remove Stitches Skills Station

Funda NALBANTOĞLU YILMAZ ** Bilge BAŞUSTA***

Öz
Çalışmanın amacı Tıp eğitimindeki öğrencilerin mesleki becerilerinden dikiş atma ve alma becerilerine ait performans puanlarının güvenirliğini belirlemek ve puanlama güvenirliğini öğrenci, puanlayıcı, beceri ve görev etkileşimlerini de dikkate alarak dengelenmemiş verilerde genellenebilirlik kuramıyla incelemektir. Araştırmanın çalışma grubunu Hacettepe Üniversitesi Yapılandırılmış Objektif Klinik Sınavına katılan öğrencilerden 309 öğrenci oluşturmaktadır. Öğrencilerin performans durumlarının puanlanmasında ise tıp alanından 11 puanlayıcı kullanılmıştır. Araştırma kapsamında, öğrencilerin puanlayıcılarla, görevlerin ise becerilerle yuvalandığı ve bu becerilerin tüm öğrenciler için ortak olduğu dengelenmemiş (ö:p)x(g:b) deseni (ö: öğrenci, p:puanlayıcı, g: görev ve b:beceri olmak üzere) kullanılmıştır. Sonuç olarak, puanlayıcı değişkenliği puanlama farklılığına neden olmamaktadır. Her iki beceri bakımından puanlayıcı etkisinden kaynaklı farklılıkların olmadığı, beceri ve ilgili beceriye ait görevlerin puanlayıcıdan puanlayıcıya farklılık göstermediği tespit edilmiştir. Öğrencilerin dikiş atma ve alma becerilerine ait performanslarının belirlenmesi sürecine yönelik genellenebilirlik kuramı ile elde edilen güvenirlik katsayıları ise kabul edilebilir düzeydedir.

Anahtar Kelimeler: Güvenirlik, Genellenebilirlik Kuramı, Dengelenmemiş Desen

Abstract
The aim of the study is to determine the reliability of performance scores of the medical education students in the skills by suturing and removal of suture and to investigate the scoring reliability of unbalanced data regarding students, raters, skills, tasks and their interactions by using generalizability theory. The study group consisted of 309 students who attended the Objective Structural Clinical Exam (OSCE) at Hacettepe University. 11 raters from the medical field took part in the assessment of the performance of students. Unbalanced (ö:p)x(g:b) design (ö: student, p: rater, g: task and b: skill) was used in this study. As a result, the variability of raters didn't cause any differences in scoring. There is no difference in both skills and at each task of skills originating from rater effect. Reliability coefficients obtained by generalizability theory for the performance assessment process of the students'skills of suturing and removal of suture are at the acceptable level.

Key Words: reliability, generalizability theory, unbalanced design

GİRİŞ

Yükseköğretimde bugün bilginin sadece geleneksel yöntemlerle test edilmesinden ziyade öğrenmelerin değerlendirilmesi önem kazanmıştır (Gijbels & Dochy, 2006). Öğrencilerin bilgiyi bilmelerinden daha çok bilgiyi meslekleri ile ilgili durumlarda ya da gerçek yaşamdaki problemler karşısında nasıl kullandıkları ile ilgilenilmelidir. Bu açıdan artık düşünme becerilerini kullanabilme, gerçek yaşamda karşılaşılan problemleri çözmek için bilgiyi kullanma ve performansa dayalı durum belirleme dikkati çekmektedir.

* Bu araştırma, Abant İzzet Baysal Üniversitesi'nde düzenlenen III. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde sözlü bildiri olarak sunulmuştur.

** Yrd.Doç.Dr., Nevşehir Hacı Bektaş Veli Üniversitesi, Eğitim Fakültesi, Nevşehir-Türkiye, fundan@nevsehir.edu.tr
***Yrd.Doç.Dr., Mersin Üniversitesi, Eğitim Fakültesi, Mersin-Türkiye, bilgeubasusta@mersin.edu.tr.

Performansın ölçülmesinde puanlama tutarlılığı ve güvenirlik önemli iki noktadır (Moss, 1994). Performans ölçme çalışmalarında istenmeyen birçok hata kaynağı elde edilen puanların güvenirliğini olumsuz yönde etkileyebilmektedir.

Güvenirlik daha çok klasik test kuramına dayalı yöntemlerle çalışılmasına rağmen performansın ölçülmesinde puanlayıcı değişkenliğini ve puanlama güvenirliğini belirlemede genellenebilirlik kuramı güçlü bir yaklaşım sunmaktadır (Brennan 2001; Cronbach, Gleser, Nanda, Rajaratnam, 1972).

Klasik test kuramında farklı kaynaklardan gelen hataların ölçme sonucuna karışabildiği durumlarda güvenirliği çeşitli yöntemlerle (paralel testler güvenirliği, test tekrar test, eşdeğer formlar, iç tutarlık ve puanlayıcılar arası güvenirlik gibi) birlikte değerlendirmek gerekir. Diğer bir deyişle, klasik test kuramında her hata kaynağına göre en az bir güvenirlik tahmini vardır ve bu güvenirlikler hata kaynağına bağlı olarak değişir (Eason, 1989; Suen ve Lei, 2007). Genellenebilirlik kuramının klasik test kuramına göre sağladığı avantajlardan biri ise genellenebilirlik kuramında farklı hata kaynaklarının birlikte değerlendirilebilmesidir. Böylece genellenebilirlik kuramıyla daha gerçekçi hata betimlemesi yapılarak, daha gerçekçi sonuçlara ve kararlara ulaşılabilmektedir.

Genellenebilirlik kuramı Cronbach, Gleser, Nanda ve Rajaratnam (1972) tarafından klasik test kuramına alternatif olarak geliştirilmiştir (Brennan, 2001; Shavelson ve Webb, 1991). Genellenebilirlik kuramı, gözlenen puanlardaki hata kaynaklarının miktarının belirlenmesini ve davranışın ölçülmesinde güvenirliğin belirlenmesini sağlayan ve temeli varyans analizine dayanan istatistiksel bir kuramdır (Brennan, 2001; Cronbach ve diğerleri, 1972; Shavelson ve Webb, 1991).

Genellenebilirlik kuramı ile genellenebilirlik ve karar çalışması olmak üzere iki farklı çalışma yapılabilmektedir. Genellenebilirlik çalışması ile araştırmacı, çalışma amacı doğrultusunda belirlediği varyans kaynaklarını kestirir ve yorumlar. Karar çalışmasında ise genellenebilirlik çalışmasından elde edilen kestirimler ile güvenirlik katsayıları hesaplanır ve farklı durumlardaki güvenirliğin nasıl olacağı araştırılır.

Genellenebilirlik ve karar çalışması yapabilmek için araştırmacı araştırma amacı ve verisi doğrultusunda çalışma deseni oluşturur. Oluşturulan bu desenler araştırma verisine göre dengelenmiş (balanced) ya da dengelenmemiş (unbalanced) olabilir. Yuvalanmış bir yüzeyin (facet) koşullarına ait gözlem sayısı tüm yüzey boyunca aynı ise veri dengelenmiş, yüzeyin koşullarına ait hücrelerdeki gözlem sayısı farklılaşıyorsa veri dengelenmemiş olmaktadır (Brennan, 2001). Örneğin; farklı sınıflardan öğrencilerin araştırılmaya dahil edildiği, öğrenci (ö) ve sınıf (s) yüzeylerinin yuvalandığı (ö:s) bir durumda araştırmada kullanılan her bir sınıftaki öğrenci sayısı aynı ise veri dengelenmiş, her bir sınıftaki öğrenci sayısı farklı ise veri dengelenmemiş olmaktadır. Bir başka örnekle, becerilere ait farklı görevlerin araştırmaya dahil edildiği, görev (g) ve beceri (b) yüzeylerinin yuvalandığı (g:b) bir durumda her bir beceri alanına ilişkin görev sayısı eşit ise veri dengelenmiş, her bir beceri alanına ilişkin görev sayısı farklı ise veri dengelenmemiş olmaktadır (Cardinet, Johnson ve Pini, 2010). Örneklerden de görülebildiği gibi gerçek yaşam verilerinde dengelenmemiş durumlar ile sıklıkla karşılaşılabilmektedir. Bununla birlikte genellenebilirlik kuramı ile dengelenmiş ve dengelenmemiş veriler için G ve Phi katsayısı olmak üzere güvenirliği veren iki katsayı hesaplanabilmektedir. Bu katsayılar 0 ile 1 arasında değerler almakta ve klasik test kuramında olduğu gibi bire yaklaştığı oranda güvenirlik artmaktadır.

Literatürdeki çalışmalar incelendiğinde, performansın ölçülmesinde güvenirliğin genellenebilirlik kuramıyla belirlendiği çalışmalar (Anıl ve Büyükkıdık, 2012; Atılgan, 2008; Barneveld, 2005; Bergus ve Kreiter, 2007; Çakıcı Eser, 2011; Deliceoğlu ve Çıkrıkçı Demirtaşlı, 2012; Feeley, Manyon, Servoss ve Panzarella, 2003; Güler ve Gelbal, 2010;

Hatala, Marr, Cuncic ve Bacchus, 2011; Kreiter ve Ferguson, 2001; Lane ve Sabers, 1989; Lee ve Kantor, 2007; Ludlow, 2001; Shavelson, Mayberry ve Webb, 1990; Taşdelen Teker, 2014; Wang, 2005; Zibrowski, Myers, Norman, Goldszmidt, 2011; Yelboğa, ve Tavşancıl, 2010; Yılmaz Nalbantoğlu ve Gelbal, 2011; Yılmaz Nalbantoğlu ve Tavşancıl, 2014) bulunmaktadır. Türkiye'de performansın ölçülmesinde güvenirliğin genellenebilirlik kuramıyla belirlendiği çalışmalar incelendiğinde ise kullanılan desenlerin genel olarak iki yüzeyli ve dengeli veri yapısında olduğu görülmektedir. Fakat yukarıda da belirtildiği gibi gerçek uygulama durumlarında her zaman verilerin dengeli yapıda, yani her bir puanlayıcının eşit sayıda öğrenciyi puanlaması, alt testlerdeki madde sayısının aynı olması beklenemez. Bu nedenle; Türkiye'de performansın ölçülmesinde güvenirliğin dengelenmemiş veri yapılarında genellenebilirlik kuramıyla belirlendiği çalışmaların sınırlı olması, gerçek veri yapısı dengelenmemiş olan yapılandırılmış objektif klinik sınavın dikiş atma ve alma becerileri istasyonunda tıp öğrencilerin mesleki becerilerinin ölçülmesine ilişkin puanların güvenirliğinin belirlenebilmesi nedenleriyle bu çalışmanın yapılması bir gereklilik olarak görülmüştür.

Araştırmanın Amacı

Çalışmanın amacı Hacettepe Üniversitesi Tıp Fakültesi öğrencilerinin dikiş atma ve alma becerilerine ait performans puanlarının güvenirliğini öğrenci, puanlayıcı, beceri ve görev etkileşimlerini de dikkate alarak dengelenmemiş verilerde genellenebilirlik kuramıyla incelemektir. Bu amaç doğrultusunda ise araştırma kapsamında "Öğrenci, puanlayıcı, beceri, görev ve bunların etkileşimlerine ait kestirilen varyans bileşenleri nasıldır?","11 puanlayıcının öğrencileri dönüşümlü olarak dikiş atma ve alma becerilerine ait farklı görevler doğrultusunda puanlamasıyla elde edilen güvenirlik katsayıları nasıldır?" sorularına cevap aranmıştır.

YÖNTEM

Araştırmanın Türü

Dikiş atma ve alma becerileri istasyonundan elde edilen performans puanlarının güvenirliğini belirlemeyi içeren bu araştırma betimsel araştırma niteliğindedir.

Çalışma Grubu

Araştırmanın çalışma grubunu Hacettepe Üniversitesi 2011 eğitim öğretim yılı bahar dönemi Yapılandırılmış Objektif Klinik Sınavı (OSCE) dikiş atma ve alma istasyonuna katılan ve değerlendirme formunun tam olarak doldurulduğu öğrencilerden 309 öğrenci oluşturmaktadır. Ayrıca öğrencilerin ilgili istasyondaki becerilere ait performans durumlarının değerlendirilmesinde tıp alanından 11 puanlayıcı kullanılmıştır.

Verilerin Toplanması

OSCE birden fazla istasyondan oluşan ve her bir istasyonda farklı klinik becerilerin değerlendirildiği bir performans sınavıdır. Araştırma kapsamında OSCE sınavı dikiş atma ve alma becerileri istasyonu ele alınmıştır. İlgili istasyon tıp öğrencilerinin maket üzerinde dikiş atma ve almaya ait mesleki becerilerinin değerlendirilmesini içermektedir.

Öğrencilerin ilgili becerilere ait performans durumlarının belirlenmesinde Hacettepe Üniversitesi Tıp Eğitimi ve Bilişimi Anabilim Dalı tarafından hazırlanan "Dikiş Atma ve Alma Becerisi Değerlendirme Formu" kullanılmıştır. Bu değerlendirme formunda dikiş atma becerisine ait 14 görev, dikiş alma becerisine ait ise yedi görev bulunmaktadır.

Değerlendirme formundaki dikiş atma ve alma becerisine ait görevler Tıp Eğitimi alanındaki uzmanlar tarafından hazırlanmıştır. Öğrencilerin dikiş atma becerisine ait elleri yıkama, steril eldiven giyme, yara kenarlarını içten dışa genişleyen daireler şeklinde temizleme, anestetik maddeyi yara kenarına uygulama, dikiş materyalinin iğnesini 2/3'lük kısmından tutulacak şekilde portegüye yerleştirme, yara kenarını (cildi) penset ile tutma, iğneyi yara dudaklarından eşit mesafede ve derinin tüm katlarını alacak şekilde geçirme, bir cm.'den sık olmamak üzere yeterli sayıda dikiş atma gibi ilgili beceriye ait görevler bulunmaktadır. Dikiş alma becerisine ait ise, elleri yıkama, eldivenleri giyme, alınacak dikişi bir penset kullanarak tutup yukarı kaldırma, dikiş halkası ve cilt arasına bistüri ucu ile girip dikişin bir bacağını kesme, pensetle tutulan dikişi yukarı ve dışa doğru çekerek çıkarma gibi görevler bulunmaktadır. Görevler Tıp alanında bulunan puanlayıcılar tarafından öğrencilerin ilgili beceriyi gösterip (1), göstermemesine (0) göre puanlanmaktadır.

İlgili istasyon kapsamında puanlayıcıların her biri dönüşümlü olarak ve birbirlerinden farklı sayıda öğrenciyi puanlamaktadır. 1. puanlayıcı 11, 2. puanlayıcı 13, 3. puanlayıcı ve 4. puanlayıcı 23, 5. puanlayıcı 24, 6. puanlayıcı 25, 7. puanlayıcı 27, 8. puanlayıcı 32, 9. puanlayıcı 35, 10. puanlayıcı 36 ve 11. puanlayıcı 60 öğrenciyi aynı görevler doğrultusunda puanlamıştır. Bu nedenle araştırma kapsamında öğrencilerin puanlayıcılarla yuvalandığı, dikiş atma ve alma becerilerine ait görevlerin farklı olması nedeniyle görev ve beceri yüzeylerinin yuvalandığı ve bu becerilerin tüm öğrenciler için ortak olduğu ve öğrencilerin (ö:p) ölçmenin nesnesi olduğu (ö:p)x(g:b) deseni (ö: öğrenci, p: puanlayıcı, g: görev ve b: beceri olmak üzere) kullanılmıştır. Araştırmada, her bir puanlayıcı birbirinden farklı sayıda öğrenciyi puanladığı ve becerilere (dikiş atma ve alma) ait farklı sayıda görev olduğu için kullanılan desen dengelenmemiş desendir. Ayrıca araştırma kapsamında dikiş atma ve dikiş alma becerileri dikkate alındığından beceri yüzeyi sabit olarak belirlenmiştir. Araştırmacı amacına ve ilgisine bağlı olarak genellemek istediği yüzey koşullarını sabit ya da tesadüfi olarak belirleyebilir. Eğer araştırmacı yüzey koşullarını sabit olarak belirlerse elde ettiği sonuçları evrendeki diğer koşullar için genellemez. Yüzey koşulları tesadüfi olarak belirlenirse koşullar olası gözlemler evreninden seçilir ve araştırmacı örneklemin ötesinde genelleme yapabilir (Brennan, 2001; Shavelson ve Webb, 1991). OSCE sınavında birden fazla beceri ölçülmektedir. Araştırmada, OSCE sınavında bulunan bu beceri setlerinden dikiş atma ve alma becerileri maksatlı olarak seçilmiş ve elde edilen sonuçların diğer becerilere genelleme amacı güdülmemiştir. Bu açıdan araştırma kapsamında beceri sabit olarak ele alınmıştır.

Verilerin Analizi

Araştırma verilerinin analizinde, dengelenmemiş verilerin genellenebilirlik kuramı ile analizinde kullanılan urGENOVA (Brennan, 2001) programı temelli G_String (G-string-IV, Version 6.1.1.; Bloch & Norman, 2011) ara yüz programı kullanılmıştır.

BULGULAR

11 puanlayıcıdan her birinin birbirinden farklı sayıda öğrenciyi dönüşümlü olarak dikiş atma (n_g=14) ve dikiş alma (n_g=7) becerilerine ait görevler doğrultusunda puanlaması durumuna göre dengelenmemiş (ö:p)x(g:b) deseni ile yapılan G çalışması sonucunda kestirilen varyans bileşenleri Tablo 1'de verilmiştir.

Tablo 1'de verildiği gibi puanlayıcıların birbirlerinden farklı sayıda öğrenciyi dönüşümlü olarak iki farklı beceriye ait farklı sayıda görevler doğrultusunda puanlamasıyla oluşturulmuş dengelenmemiş (ö:p)x(g:b) deseni sekiz varyans kaynağına ayrılmaktadır. Bu varyans kaynaklarına ait kestirilen varyans bileşenleri ise aşağıda açıklanmıştır.

Tablo 1. *(ö:p)x(g:b) Desenine Ait Kestirilen Varyans Bileşenleri*

Varyans Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	Kestirilen Varyans Bileşeni	%
p	47.11638	10	4.71164	0.00397	3,49
ö:p	169.42948	298	0.56856	0.00243	2,14
b	29.90669	1	29.90669	0.00943	8,29
g:b	11.87263	19	0.62488	0.00147	1,29
Pxb	19.22670	10	1.92267	0.00544	4,78
gp:b	28.25497	190	0.14871	0.00369	3,24
bö:p	126.17615	298	0.42341	0.04032	35,43
ög : pb,e	266.37241	5662	0.04705	0.04705	41,34
Toplam	698.35540	6488			100

ö:öğrenci, p: puanlayıcı, g:görev, b:beceri

Tablo 1'deki puanlayıcı (p) ana etkisi için kestirilen varyans bileşeni toplam varyansın % 3,49'unu açıklamaktadır. Puanlayıcı ana etkisi için kestirilen varyansın toplam varyansı açıklama yüzdesinin düşük olması öğrencilerin performans puanları üzerinde puanlayıcı farklılığından kaynaklanan bir etkinin olmadığını göstermektedir.

Öğrencilerin puanlayıcılarla yuvalandığı ö:p etkisine ait varyans bileşeni toplam varyansın %2,14'ünü açıklamaktadır. ö:p etkisi öğrenci ana etkisi ve öğrenci-puanlayıcı ortak etkisine ilişkin bilgi vermektedir. ö:p etkisinin toplam varyansı açıklama yüzdesinin düşük olması öğrencilerin ilgili istasyon kapsamındaki toplam performansları bakımından benzeşik, grubun homojen, öğrenci-puanlayıcı etkileşimine ait farklılığın düşük olduğunu göstermektedir.

Becerilere (b) ait kestirilen varyans bileşeni toplam varyansın %8,29'unu açıklamaktadır. Becerilere ait varyans bileşeninin diğer ana etkilere göre daha yüksek olması dikiş atma ve alma becerilerinin birbirlerinden farklılaştığını göstermektedir.

Görevlerin beceri ile yuvalandığı g:b etkisine ait kestirilen varyans bileşeni görev ve görev-beceri ortak etkileşimine ait bilgi içerir. g:b etkisine ait kestirilen varyans bileşeni toplam varyansın % 1,29'unu açıklamaktadır. g:b etkisine ait kestirilen varyans bileşeninin düşük olması her bir beceriye ait görevlerin kendi içlerinde homojen olduğunu göstermektedir.

Puanlayıcı-beceri (pxb) ortak etkileşimine ait kestirilen varyans bileşeni toplam varyansın %4,78'ini açıklamaktadır. Puanlayıcı-beceri ortak etkisine ait varyans bileşenin düşük olması puanlayıcıların dikiş atma ve dikiş alma becerisi olmak üzere her iki beceriyi değerlendirirken farklılığa neden olmadığını göstermektedir.

Puanlayıcı ve görev ortak etkileşiminin becerilerle yuvalandığı gp:b ait varyans bileşeni toplam varyansın %3,24'ünü açıklamaktadır. gp:b ortak etkisine ait varyans bileşeni görev x puanlayıcı x beceri ve görev x puanlayıcı ortak etkilerinden oluşmaktadır. gp:b ortak etkisine ait kestirilen varyans bileşeninin düşük olması her bir beceriye ait görevlerin bir puanlayıcıdan diğerine farklılık göstermediğini vermektedir.

Beceri ve öğrenci etkileşiminin puanlayıcılarla yuvalandığı bö:p ait kestirilen varyans bileşeni toplam varyansın %35,43'ünü açıklamaktadır. bö:p ait kestirilen varyans

bileşeninin toplam varyansı açıklama yüzdesinin yüksek olması her bir puanlayıcının puanladığı öğrenci grubunun performansının beceriden beceriye değiştiğini göstermektedir. Bir başka deyişle öğrencilerin performansları puanlayıcılar bakımından dikiş atma ve alma becerilerine göre farklılık göstermektedir.

Artık etki toplam varyansın % 41,34'ünü açıklamaktadır. (ö:p)x(g:b) deseninde artık etki σ^2(ögpb), σ^2(ögp), σ^2(ögb), σ^2(ög) varyans bileşenlerini ve/veya tesadüfî hata kaynaklarını içermektedir. Artık etkinin yüksek olması, σ^2(ögpb), σ^2(ögp), σ^2(ögb), σ^2(ög) etkileşimlerine ait varyansın artık etkiye dâhil edilmesinden ve/veya öğrencilerin performanslarını göstermeleri esnasında yaşayabilecekleri kaygı, endişe, heyecan vb. durumlardan, puanlayıcı etkisinden karışabilecek çeşitli tesadüfi hata kaynaklarından olabilir.

Tablo 1'de kestirilen varyans bileşenleri kullanılarak becerilerin sabit olarak ele alındığı, öğrencilerin (ö:p) ölçmenin nesnesi olduğu ve her bir puanlayıcının birbirlerinden farklı sayıda öğrenciyi puanladığı dengelenmemiş (ö:p) x (g:b) desenine ait kestirilen güvenirlik katsayıları ise Tablo 2'de verilmiştir.

Tablo 2. (ö:p)x(g:b) Deseni G-Phi Katsayıları

	$n_{puanlayıcı}$=11
	$n_{öğrenci:puanlayıcı}$=11,13,23,23,24,25,27,32,35,36,60
	n_{beceri}=2, $n_{görev:beceri}$=14, 7
G Katsayısı ($E\rho^2$)	0.91
Phi Katsayısı (Φ)	0.72

Tablo 2'de verildiği gibi dikiş atma becerisine ait 14 görev ve dikiş alma becerisine ait 7 görev olmak üzere toplamda 309 öğrencinin 11 puanlayıcı tarafından ve her puanlayıcının birbirlerinden farklı sayıda öğrenciyi performanslarına göre puanlamasıyla elde edilen genellenebilirlik (G) katsayısı 0.91, Phi katsayısı ise 0.72 olarak kestirilmiştir. Görüldüğü gibi G ve Phi katsayıları arasında fark bulunmaktadır. Bu farklılık iki katsayının hesaplanmasında kullanılan hata varyanslarının farklı olmasındandır.

Şöyle ki, G ve Phi katsayılarını veren güvenirlik katsayısı evren puanı varyansının gözlenen puan varyansına oranıdır. Gözlenen puan varyansı ise evren puanı varyansı ve hata varyansından oluşmaktadır. Hata varyansı ise bağıl ve mutlak değerlendirmeler için farklılaşmakta, bağıl değerlendirmelerde bağıl hata, mutlak değerlendirmelerde ise mutlak hata varyansı kullanılmaktadır. Böylece bağıl değerlendirmelerin yapıldığı durumlarda bağıl hata varyansı kullanılarak G katsayısı, mutlak değerlendirmelerin yapıldığı durumlarda mutlak hata varyansı kullanılarak Phi katsayısı hesaplanmaktadır.

Bağıl hata varyansı evren puanının ortak etkilerini içerir (Shavelson ve Webb, 1991). Becerilerin sabit olarak ele alındığı, öğrencilerin (ö:p) ölçmenin nesnesi olduğu dengelenmemiş (ö:p) x (g:b) deseninde bağıl hata varyansı aşağıdaki gibidir.

$$\sigma^2(\delta) = \sigma^2(\ddot{o}G : pB)$$

Mutlak hata varyansı ise evren puanı varyansı hariç desendeki diğer tüm varyansların toplamıdır (Brennan, 2001). Böylece becerilerin sabit olarak ele alındığı, öğrencilerin (ö:p) ölçmenin nesnesi olduğu dengelenmemiş (ö:p) x (g:b) deseninde ait mutlak hata varyansı ise aşağıdaki gibi gösterilmektedir.

$$\sigma^2(\Delta) = \sigma^2(p) + \sigma^2(G:B) + \sigma^2(pB) + \sigma^2(Gp:B) + \sigma^2(\ddot{o}G:pB)$$

Güvenirlik katsayısının 0 ile 1 arasında değiştiği ve 0.70 ve üzeri katsayının kabul edilebilir olduğu düşünüldüğünde genellenebilirlik kuramıyla bağıl ve mutlak değerlendirmeler için ayrı ayrı elde edilen bu güvenirlik katsayılarının (G ve Phi) kabul edilebilir sınırlar içinde olduğu söylenebilir. Öğrencilerin mesleki becerilerinin değerlendirildiği, bu nedenle de mutlak değerlendirmenin önem kazandığı düşünüldüğünde ise mutlak değerlendirmeler için kestirilen güvenirlik katsayısının (Phi) yüksek çıkmaması ise öğrencilerin görevlerdeki performanslarının farklılık göstermemesine bağlanabilir.

SONUÇLAR ve TARTIŞMA

Tıp eğitiminin önemli bir parçası olan ve öğrencilerin mesleki becerilerinden dikiş atma ve alma becerilerine ait performanslarının genellenebilirlik kuramıyla incelenmesi ile aşağıdaki sonuçlara ulaşılmıştır.

OSCE sınavı yapısı itibariyle öğrencilerin dikiş atma ve alma becerilerine ait performansları farklı puanlayıcılar tarafından değerlendirilmektedir. Araştırma kapsamında yapılan genellenebilirlik çalışmasından elde edilen sonuçlardan hareketle puanlayıcı değişkenliğinin puanlama farklılığına neden olmadığı tespit edilmiştir. Ayrıca her iki beceri bakımından puanlayıcı etkisinden kaynaklı farklılıkların olmadığı, beceri ve ilgili beceriye ait görevlerin puanlayıcıdan puanlayıcıya farklılık göstermediği tespit edilmiştir. Bu durum değerlendirme formunun puanlayıcılar açısından anlaşılır olduğunu, formun uygulamada karışıklık yaratmadığının da bir göstergesidir. Puanlayıcı değişkenliğinin öğrencilerin performansını etkilemediği, puanlayıcıların puanlama bakımından farklılığına neden olmadığı bulgusu Hacettepe Üniversitesi OSCE sınavına ait farklı becerilerin değerlendirildiği farklı istasyonlara ait çalışmaların (Yılmaz Nalbantoğlu ve Gelbal, 2011; Yılmaz Nalbantoğlu ve Tavşancıl, 2014) bulguları ile de benzerlik göstermektedir.

Dikiş atma ve alma becerileri ve bu becerilere ait görevlerle ilgili ise araştırmada becerilerin öğrencilere farklı güçlükte geldiği, öğrencilerin performanslarının dikiş atma ve alma becerileri bakımından farklılık gösterdiği sonucuna varılmıştır. Ayrıca ilgili beceriye (dikiş atma ya da dikiş alma) ait görevlerin kendi içinde homojen olduğu tespit edilmiştir. Bu durum becerilere ait görevlerin iç tutarlılığını göstermekte ve genellenebilirliği artırmaktadır.

Tıp eğitiminde önemli bir yere sahip olan mesleki becerilerin değerlendirildiği bu çalışmada, öğrencilerin dikiş atma ve alma becerilerine ait performanslarının ölçülmesi sürecine yönelik olarak güvenirlik katsayılarını veren G ve Phi katsayılarının ise kabul edilebilir düzeyde olduğu tespit edilmiştir. Bu açıdan tıp eğitiminin bir parçası olarak öğrencilerinin mesleki becerilerinden biri olan dikiş atma ve alma becerilerinin ölçülmesi sonucu elde edilen puanlarla alınan kararların isabetli olduğu söylenebilir.

Çalışma sonuçlarından da görülebileceği gibi performansın ölçülmesinde genellenebilirlik kuramı, varyans kaynaklarına ilişkin araştırmacıya bilgiler sunması, birçok hata kaynağı ve bunların etkileşimlerinin de birlikte değerlendirilmesiyle daha gerçekçi güvenirlik kestirimi yapılabilmesi açısından önemlidir. Bu nedenle daha gerçekçi güvenirlik kestirimleri yapabilmek, öğrencilerin performanslarına, görevlere, puanlayıcılara, değerlendirmeye ait bilgi elde edebilmek ve gelecek çalışmalara yol gösterici karar çalışmaları yapabilmek için performansın ölçüldüğü çalışmalarda güvenirlik belirlenirken

genellenebilirlik kuramının kullanılması önerilmektedir. Ayrıca bu çalışmada da olduğu gibi gerçek veri yapısı dengelenmemiş veri yapılarında olan, farklı puanlayıcıların kullanıldığı, farklı koşul ya da zamanda bilgilerin toplandığı durumlarda bu farklılıkları da dikkate alarak analizlere imkân vermesi bakımından genellenebilirlik kuramının kullanılması önerilmektedir.

Çalışmada OSCE sınavına ait dikiş atma ve alma becerileri istasyonu ele alınmıştır. Bir başka çalışmada OSCE sınavına ait farklı istasyonlardan elde edilen puanların güvenirliğini genellenebilirlik kuramıyla belirlemeye yönelik farklı çalışmalar yapılabilir. Ayrıca bu çalışmada Tıp eğitiminde öğrencilerin mesleki becerileri ele alınmış olup Tıp eğitimi dışında farklı alanlara ilişkin farklı performans durumlarına ait güvenirlik genellenebilirlik kuramıyla belirlenebilir.

KAYNAKLAR

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

Atılgan, H. (2008). Using generalizability theory to assess the score reliability of the special ability selection examinations for music education programmes in higher education. *International Journal of Research & Method in Education*, 31(1), 63-76.

Barneveld, C. V. (2005). The dependability of medical students' performance ratings as documented on in-training evaluations. *Academic Medicine*, 80(3),309-312.

Bergus, G. R. ve Kreiter, C. D. (2007). The reliability of summative judgements based on objective structured clinical examination cases distributed across the clinical year. *Medical Education*, 41, 661-666.

Bloch, R., and Norman, G. (2011). G String 4 User Manual. Web: http://fhsperd.mcmaster.ca/g_string/download/g_string_4_manual_611.pdf

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

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG* . New York, NY: Routledg.

Cronbach, L. J., Gleser, G. C., Nanda, H., ve Rajaratnam, N. (1972). *The dependability of behavioral measurements: theory of generalizability for scores and profiles.* New York: Wiley.

Çakıcı Eser, D. (2011). *Genellenebilirlik kuramı ve lojistik regresyona dayalı hesaplanan puanlayıcılar arası tutarlılığın karşılaştırılması*. Yayımlanmamış Yüksek Lisans Tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

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

Gijbels, D. & Dochy, F. (2006). Students' assessment preferences and approaches to learning: can formative assessment make a difference? Educational Studies, Vol. 32, No. 4, pp. 399–409.

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

Feeley, T., H., Manyon, A., T., Servoss, T., J. ve Panzarella, K,.J. (2003). Toward validation of an assessment tool designed to measure medical student' integration of scientific knowledge and clinical communication skills. *Evaluation The Health Professions*, 26(2), 222- 233.

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

Hatala, R., Marr, S., Cuncic, C. ve Bacchus, C. M. (2011). Modification of an OSCE format to enhance patient continuity in a high stakes assessment of clinical performance. *BMC Medical Education*,11,23.

Kreiter, C. D. ve Ferguson, K. J. (2001). Examining the generalizability of ratings across clerkships using a clinical evaluation form. *Evaluation The Health Professions*, 24 (1), 36-46.

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

Lee, Y., W. ve Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G theory.*International Journal of Testing*, 7(4), 353-385.

Ludlow, C., B. (2001). Using running records As a Benchmark reading assessment: reliability in assessing reading progress. The Degree of Doctor, BrighamYoungUniversity, Provo.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, 23, 5-12.

Shavelson, J.R., Mayberry, P., Webb, M. (1990). Generalizability of job performance measurements: marine corps rifleman. *Military Pstchology*, 2(3), 129-144.

Shavelson, J. R., and Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park. CA: Sage Publications.

Suen, H. K., and Lei, P.W. (2007). Classical versus generalizability theory of measurement. *Educational Measurement*, 4, 1-13.

Taşdelen Teker, G. (2014). Madde takımlarının güvenirlik ve değişen madde fonksiyonu üzerine etkisi, Yayımlanmamış doktora tezi, Hacettepe Üniversitesi, Ankara.

Wang, Z. (2005). Estimating reliability under a generalizability theory model for writing scores in c-base. Master Thesis, Universit of Missouri, Columbia.

Yelboğa, A. ve Tavşancıl, E. (2010). Klasik test ve genellenebilirlik kuramı'na göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri,* 10(3), 1825-1854.

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

Yılmaz Nalbantoğlu, F. ve Tavşancıl E. (2014).Intramuskuler enjeksiyon yapma istasyonu verileriyle genellenebilirlik kuramında dengelenmiş ve dengelenmemiş desenlerin karşılaştırılması. Eğitim ve Bilim Dergisi, Cilt 39, Sayı, 175, 285-295.

Zibrowski, E. M., Myers, K., Norman, G., and Goldszmidt, M. A. (2011). Relying on others' reliability: challenges in clinical teaching assessment. *Teaching and Learning in Medicine*, 23 (1), 21 -27.

EXTENDED ABSTRACT

Introduction

In higher education today has gained importance evaluation of learning rather than testing knowledge by traditional methods (Dochy, Gijbels,&Segers, 2006). Information should be dealt with students how they use cases related professions or how they use the knowledge against in the real life problems. From this point no longer to use the thinking skills, to use information to solve problems encountered in real life and to use performance assessments are remarkable. In this respect, scoring consistency and reliability are two important points in performance assessment. Because many unwanted sources of error may affect the reliability of the scores obtained from performance assessment' study. Although reliability is studied with classical test theory, generalizability theory offers a more robust approach to measure performance and to determine the reliability of raters variability.

Reliability should be assessed various ways in classical test theory in case of errors can interfere from different sources. In other words, there are at least a reliability estimate based on each error source in the classical test theory. And this reliability will vary depending on the error sources (Eason, 1989; Suen ve Lei, 2007). One of the advantages of generalizability theory based on classical test theory can be judged together from multiple sources of errors in generalizability theory. Thus, the generalizability theory made more realistic portrayal error.

The studies as determined the reliability via generalizability theory is limited in Turkey. In addition, it is seen that the studies have two facet and balanced design. But in actual practice situations the data is not expected balanced, namely each raters scoring an equal number of students. Therefore the limited application of unbalanced data structure is seen as a requirement for performing this operation. So the aim of the study is to determine the reliability of performance scores of the students in the skills by suturing and remove stitches and to investigate the scoring reliability of unbalanced data regarding students, raters, skills, tasks and their interactions by using generalizability theory.

Method

The study group consisted of 309 students who attended the Objective Structural Clinical Exam (OSCE) at Hacettepe University. 11 raters from the medical field took part in the assessment of the performance of students.

The OSCE is a performance test which consisted multiple stations and evaluated in different clinical skills in each station. Within this research, suturing and remove stitches skills station is discussed. This station includes the evaluation of suturing (14 tasks) and remove stitches (7 tasks) professional skills of medical students. Within relevant station, raters alternately (by turns) rate the different number of students. So, the unbalanced (ö:p)x(g:b) design where facet ö (student) is nested in facet p (rater); facet g (task) is nested in facet b (skill) and those skills are common for all of the students was used in this study. Furthermore, the research covered by suturing skills and sewing skills to take variable is taken into account is fixed. For analysis he computer program G_String (G-string-IV, Version 6.1.1.; Bloch & Norman, 2011) was used to estimate variance components and decision studies for desing.

Results and Discussion

The variance component for the main effect of rater explains 3,49% of total variance. The variance component for students within raters (2,14% of the total variance) is small. This shows that students are homogeneous in terms of overall performance. Similarly, the component for the tasks within skills (1,29% of the total variance) and the component for the rater by skill interaction (4,78% of the total variance) are also small. So, the tasks are homogeneous in each of the skills and there is no difference between raters in scoring the skills. The variance component for skills is (8,29% of the total variance) large according to other main effects. So this result shows that skills differ from each other. The variance component for raters by tasks within skills (3,24% of the total variance) is small. And the variance component for students by skills within raters (35,43% of the total variance) is large. Finally, the variance component for the residuals explains 41,34% of the total variances. In decision study G coefficient was found as 0.91 and Phi coefficient was found as 0.72 for unbalanced (ö:p)x(g:b) design.

As a result, the variability of raters didn't cause any differences in scoring. There is no difference in both skills and at each task of skills originating from rater effect. G and Phi coefficients obtained by generalizability theory for the performance assessment process of the students'skills of suturing and removal of suture are at the acceptable level.