GENELLENEBİLİRLİK KURAMI VE LOJİSTİK REGRESYONA DAYALI HESAPLANAN PUANLAYICILAR ARASI TUTARLIĞIN KARŞILAŞTIRILMASI[1]

Derya ÇAKICI ESER
Kırıkkale Üniversitesi, Eğitim Bilimleri Bölümü, Kırıkkale

Selahattin GELBAL
Hacettepe Üniversitesi,Eğitim Bilimleri Bölümü, Ankara

İlk Kayıt Tarihi: 05.12.2011 ***Yayına Kabul Tarihi: 10.07.2012***

Özet

Araştırmanın amacı G Kuramı ve LR analizinden yararlanarak gerçekleştirilen bir performans puanlamada ortaya çıkan puanlayıcı tutarlığını belirlemek ve karşılaştırmaktır. Araştırmada 106 öğrenciye 15 maddelik bir ölçme aracı verilmiş, öğrencilerin verdikleri cevaplar üç puanlayıcı tarafından dereceli puanlama anahtarı kullanılarak puanlanmıştır. Puanlama ile elde edilen veri seti, G kuramı ve lojistik regresyon analizi ile madde bazında ve testin tamamına dayalı olarak analiz edilmiştir. Analiz sonucunda G kuramı ile elde edilen puanlayıcıvaryans bileşenleri ve toplam varyansı açıklama yüzdeleri ile LR analizi ile elde edilen sınıflama yüzdeleri yorumlanmıştır. Elde edilen bulgulara dayalı olarak G Kuramı ve LR analizinin puanlayıcılar arası tutarlığı belirlemede paralel sonuçlar ürettiği, ancak lojistik regresyon analizinin G kuramı kadar hassas çıktılar vermediği ve G kuramına göre daha yüzeysel bir istatistik olduğu sonucuna varılmıştır.

Anahtar Sözcükler: *Puanlayıcılar arası tutarlık, Genellenebilirlik Kuramı, Lojistik Regresyon*

COMPARISON OF INTERRATER AGREEMENT CALCULATED WITH GENERALIZABILITY THEORY AND LOGISTIC REGRESSION

Abstract

The purpose of this study is to determine the rater agreement and interrater agreement of a performance rating using G theory and LR analysis. In this research 106 students are asked to answer a 15 item scale and the answers have been rated by three raters with an analytic

1. Bu çalışma Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Eğitimde Ölçme ve Değerlendirme Bilim Dalı'nda hazırlanan yüksek lisans tezinden özetlenmiştir.

rubric. Data set has beenbeen analyzed with G theory and LR analysis based on the items of the test and the whole test. According to analyses; rater variance component and total variance explaining percentage of G Study and classification percentage of LR analysis have been interpreted. According to the results; generalizability theory and logisticregression analysis both give parallel results; but logistic regression analysis do not generate as sensitive outputs as G theory and is a superficial statistics in referenceto G theory.

KeyWords: *InterraterAgreement, GeneralizabilityTheory, LogisticRegression*

1. Giriş

Yapılan ölçmelerin sonucunda değerlendirmeye geçebilmek için ölçmenin geçerliğinin ve güvenirliğinin sağlanmış olması gerekmektedir. Geçerlik ölçme aracının istenen amaca hizmet etme derecesi, güvenirlik ise amaca hizmet edebilirliği ne olursa olsun, aracın tutarlı bilgiyi verme kapasitesidir (Ahmann ve Glock 1971:290). Crocker ve Algina (1986:105)'ya göre ise, güvenirlik ölçme sonuçlarının sahip olması istenen tutarlığı olarak tanımlamıştır. Her ölçmenin güvenilir olması istenmekle birlikte, ölçmeye karışan sabit, sistematik veya tesadüfî hatalar güvenirliğin düşmesine sebep olmaktadır. Sabit ve sistematik hata kaynakları ölçmeden uzaklaştırılabilirken, tesadüfî hata kaynağı ölçmeden tam olarak arındırılamamaktadır. Bu durum ölçmenin güvenirliğinin arttırılması için tesadüfî hata kaynaklarından biri olan puanlayıcı güvenirliğinin arttırılması gerektiğine işaret eder. Ölçmeler, bütün puanlayıcıların aynı cevaba aynı puanı vermesi sağlandığı müddetçe, objektiftir (Turgut ve Baykul, 2010:132) ve puanlayıcı güvenirliği sağlanmış olur. Bu sebeple yapılan ölçme işleminde güvenirliği arttırmak amacıyla puanlayıcı sayısı arttırılabilir.

Ancak ölçmenin güvenirliğini arttırmak için puanlayıcı güvenirliğini arttırmaya çalışmak, ölçmenin konusunu duyuşsal, bilişsel ve psiko-motor davranış alanlarının birkaçını içinde barındırabilen performans ölçmelerde daha büyük önem kazanmaktadır. Eğitimdeki öğrenme çıktılarını kazanma başarısı olarak tanımlanan performans; bilişsel, duyuşsal ve psiko-motor davranışlarla üst düzey zihinsel beceriler olmak üzere bütün öğrenme ürünlerini içerir (Turgut ve Baykul, 2010: 259). Performans değerlendirilmesi, performans ölçülmesi veya performansa dayalı ölçme olarak çeşitli şekillerde tanımlanan performans ölçülmesi ise, gözlem ve kanıya dayalı bir değerlendirmedir (Stiggins, 1997 Akt: Palm, 2008). Performans ölçülmesinde öğrencilerin bir aktivite yapması (ör: bir model kurması) ya da orijinal bir yanıt oluşturması, yüksek düzeyde düşünme ve problem çözme becerileri geliştirmesi, problem çözme yeteneklerini gerçek dünyaya uyarlaması, birden fazla çözüm ve stratejiyi göz önünde bulundurması, bilgiye erişebilmesive bunların yanında birkaç dakikadan birkaç güne genişleyebilen zaman periyoduna gereksinim vardır (Lane, 2006: 387). Bu yönüyle performans değerlendirme klasik değerlendirme yaklaşımlarından ayrılır. Bu ayrılışı yaratan fark beraberinde performans değerlendirmede karşılaşılan bazı güçlükleri getirir. Bu güçlüklerden birisi öğrencilerin nasıl puanlanacağına karar vermek, diğeri ise ölçmenin güvenirliğini sağlamaktır.

Performans değerlendirmelerde öğrencilerin nasıl puanlanacağını belirlemek için dereceleme ölçeklerinden, diğer ifadeyle dereceli puanlama anahtarlarından (rubrik) faydalanılabilir. Dereceli puanlama anahtarları öğrencilerin performanslarını ve cevaplarını belirlenen ölçülere göre puanlamada kullanılan kılavuzlardır (Turgut ve Baykul, 2010:266). Değerlendirmenin amacına paralel olarak bütüncül (holistik) veya analitik puanlama anahtarı şeklinde hazırlanabilir. Bütüncül puanlama anahtarı performans sürecine veya ürüne yönelik tek bir puanlama yapılacağı durumlarda kullanılır. Analitik puanlama anahtarı ise performans parçalarının nasıl puanlanacağına ilişkin yönergeler içerir. Bu yönüyle analitik puanlama anahtarı ile öğrencinin hangi noktalarda istenen performansı sergilediği, hangi noktalarda performansında zayıf noktaların ve eksikliklerin bulunduğu ortaya çıkarılabilir.

Performans değerlendirmede öne çıkan bir diğer nokta güvenirliği sağlamaktır. Çünkü performans değerlendirme sürecinde öğrenciden beklenen davranışlar dereceli puanlama anahtarları veya kontrol listeleri ile belirlenmesine rağmen, puanlayıcılar, puanlayıcıların görüş ayrılıkları, objektiflik eksikliği, belirli olmayan puanlama anahtarı ve çevredeki değişiklikler gibi muhtemel hataları ile karşılaşılır (Kim, 2000). Bu hata kaynakları puanlayıcılar arası tutarlığı diğer bir ifadeyle puanlayıcı güvenirliğini düşürür. Bu sebeplerden dolayı, yapılan performans ölçümlerinde ölçme sonuçları değerlendirilmeden önce, puanlayıcılar arası tutarlığın incelenmesi gereklidir. Bunun için Pearson Momentler Çarpımı, Spearman-Brown düzeltme formülü, uyum oranı, uzlaşma katsayısı, Kappa istatistiği gibi farklı yöntemlerden yararlanılabilir. Bu araştırmada puanlayıcılar arası tutarlığın incelenmesi amacıyla Genellenebilirlik Kuramı ve Lojistik Regresyon Analizi kullanılacağından bu iki yöntem ele alınmıştır.

Genellenebilirlik Kuramı

Genellenebilirlik Kuramı (G Kuramı), gözlenen puanlarda var olan veya var olabilecek olan tutarsızlıkların güçlü istatistiksel yöntemler ile belirlenmesini ve ölçülmesini sağlayan, varyans analizi (ANOVA) üzerine kurulu bir kuramdır (Brennan, 2001:4). Crocker ve Algina (1986)'ya göre ise, G Kuramı, bireylere ait ölçme sonuçlarının daha geniş ölçümlere genellenme derecesini veren teknikler ile ilgilidir. Bu kuramın altında yatan temel felsefe, araştırmacının elinde var olan ölçmeleri ait olduğu gözlemler evrenine genellemek için ölçmenin hassasiyet ve güvenirliğini sağlamayı istemesidir (Cronbach, ve ark.,1963 akt: Rentz, 1987). Genellenebilirlik ise, toplam varyansın desendeki bağımsız değişkenlere bölünüp, farklı varyans kaynaklarına ayrılmış gözlenen puanların evren puanlarına genellenmesi ile sağlanır.

Genellenebilirlik kuramında adı geçen bazı önemli kavramlar vardır. Bu kavramlar değişkenlik kaynağı (facet), koşul (condition), gözlemler evreni (universe of admissible observation), genelleme evreni (universe of generalisation) ve evren puanı (universe score)'dır. Değişkenlik kaynağı, ölçmede potansiyel hata kaynağı olan ölçme sürecinin karakteristikleridir. Koşul ise bu değişkenlik kaynaklarının düzeyleridir. Bu tanımlara göre öğrencilerin yerine getirmesi gereken on görevin yer aldığı

bir ölçmede öğrencilerin yerine getirmesi gereken işler görev değişkenlik kaynağını, her bir görev ise ayrı ayrı koşulları oluşturmaktadır. Elde var olan örneklemin yerine geçebilecek olası gözlemler evrenine kabul edilebilir gözlemler evreni (universe of admissible observation), karar veren kişinin genellemek istediği değişkenlik kaynağı koşullarına ise genelleme evreni (universe of generalisation) adı verilir. Kişinin genellenme evreninde beklenen puanının gözlenen değeri ise evren puanını (universe score) oluşturur ve klasik test teorisindeki gerçek puana karşılık gelir (Brennan, 2001:7;Crocker ve Algina, 1986:159).

Genellenebilirlik kuramı genellenebilme çalışması (G Çalışması) ve karar çalışması (D Çalışması) olmak üzere iki çalışma içerir. G çalışması varyans bileşenlerinin tahmin edilmesi amacı ile veri toplama aşamasıdır. Bu aşamada potansiyel hata kaynakları veya değişkenlik kaynakları kullanılarak varyans bileşenleri ve bu bileşenler arasındaki etkileşimler ANOVA ile kestirilmeye çalışılır. Kestirilen varyans bileşenlerinin genellenebileceği varsayılır. G çalışmasında elde edilen veriler D çalışmasında kullanılır. Bu çalışma karar verme ve anlamlandırma sürecidir ve daha yüksek güvenirlik ile düşük hataların olduğu durumlar elde edilmeye çalışılır. (Gleser ve ark, 1965;Smith ve Kulikowich: 2004;Tobar, Stegner ve Kane, 1999).

Değişkenlik kaynaklarının ele alınış biçimine göre genellenebilirlik kuramı desenleri değişmektedir. Buna göre tüm değişkenlik kaynaklarının tüm koşullarının birbiri ile örtüştüğü desen çaprazlanmış, değişkenlik kaynaklarının koşulları diğer değişkenlik kaynakları örtüşmediği desen yuvalanmıştır. Çapraz desen değişkenlik kaynakları arasına konan 'x' işareti ile gösterilir ve yuvalanmış desen değişkenlik kaynaklarının arasına konan ':' işareti ile gösterilir (Crocker ve Algina,1986:160).

Genellenebilirlik kuramında evren sahip olduğu değişkenlik kaynaklarına göre üç başlık altında incelenebilir. Bir potansiyel hata kaynağına sahip evrenler tek değişkenlik kaynaklı evren, ölçmeye karışan iki hata kaynağı içeren iki değişkenlik kaynaklı evren ve üç veya daha fazla değişkenlik kaynağı içeren üç veya daha çok değişkenlik kaynaklı evren şeklinde adlandırılır. Bireylerin yerine getirmeleri gereken görevlerin olduğu ve birden fazla puanlayıcının yer aldığı bir ölçmede görevler ve puanlayıcılar potansiyel değişkenlik kaynaklarını diğer bir ifade ile potansiyel hata kaynaklarını oluşturmaktadır. Söz konusu ölçmede bireylerin değişkenliğe sahip olması ve başarılarının değişkenlik göstermesi beklendiğinden bireyler bir hata kaynağı olarak ele alınmamaktadır. Buna göre bu ölçme deseni iki değişkenlik kaynaklı bir ölçmedir. İki değişkenlik kaynaklı evrenden elde edilen varyans bileşenleri değişkenlik kaynaklarının her birini ayrı ayrı, ikili ve üçlü etkileşim şeklinde ele almaktadır. Birey, madde ve puanlayıcı değişkenlik kaynaklarını içeren iki değişkenlik kaynaklı bir ölçmede var olan varyans kaynakları, varyans tipleri ve varyans gösterimleri tablo 1.1'de yer almaktadır.

Tablo 1: İki Değişkenlik Kaynaklı Ölçmelerde Değişkenlik Kaynakları

	Varyans Tipi	Varyans Gösterimi
Birey (b)	Evren puanı varyansı (ölçmenin nesnesi)	σ_b^2
Madde (m)	Bireylerin bir maddeden diğerine değişen davranışlarındaki tutarsızlıkları (tüm bireyler için sabit bir etkidir)	σ_m^2
Puanlayıcı (p)	Puanlayıcıların birey puanlamadaki katılıkları/ cömertlikleri (tüm bireyler için sabit bir etkidir)	σ_p^2
b x m	Bir bireye ait bir maddeden diğerine davranışlarında gözlenen tutarsızlık	σ_{bm}^2
b x p	Puanlayıcının bireyleri değerlendirmedeki tutarsızlığı	σ_{bp}^2
m x p	Puanlayıcıların bir maddeden diğerine kararlılıklarındaki değişim (tüm bireyler için sabit bir etkidir)	σ_{mp}^2
b x m x p	Birey, madde ve puanlayıcı değişkenlik kaynaklarının ölçülemeyen kombinasyonlarından oluşan artık etki ve/ veya tesadüfi hatalar	$\sigma_{bmp,e}^2$

Bu desende hesaplanan gözlenen puanlara ilişkin varyans aşağıdaki yedi varyans bileşenini içermektedir:

$$\sigma^2\left(X_{bmp}\right) = \sigma_b^2\,\sigma^2\left(X_{bmp}\right) = \sigma_b^2 + \sigma_m^2 + \sigma_p^2 + \sigma_{bm}^2\sigma_m^2 + \sigma_p^2 + \sigma_{bm}^2 + \sigma_{bp}^2 + \sigma_{pm}^2 + \sigma_{bmp}^2\sigma_{bp}^2 + \sigma_{pm}^2 + \sigma_{bmp}^2$$

Bu araştırmada puanlayıcılar arası tutarlığın hesaplanmasında, G Kuramının yanı sıra Lojistik Regresyon Analizi kullanılacaktır.

Lojistik Regresyon Analizi

Regresyon analizi bağımsız değişken (yordayan değişken) ile bağımlı değişken (yordanan değişken) arasındaki ilişkinin istatistiksel yöntemlerle belirlenmesi şeklinde tanımlanır (Kayri ve Çokluk, 2010). Regresyon analizinde amaç verileri; değişkenlerin gözlenen değerlerini, aranılan ilişkinin şeklini tahmin için kullanmaktır (Öztürkcan, 2009:2). Ancak Çokluk, Şekercioğlu ve Büyüköztürk (2010: 49)'ün de belirttiği gibi veri setinin yapısına göre uygulanacak regresyon modellemeleri farklılık göstermektedir. Basit doğrusal regresyon, çoklu doğrusal regresyon ve çok değişkenli doğrusal regresyon modelleri ile çalışma yapmak için bağımlı ve bağımsız değişkenlerin sürekli değişken olması ve değişken setlerinin dağılımları ile hata değerlerine ait dağılımların normalliği gibi sayıtlıların karşılanmasını gerektirir. Ancak sosyal bilimlerde elde edilen sonuçlar genellikle kategoriktir ve doğrusal regresyon modelleri ile incelenmesi mümkün değildir. Bu durumda lojistik regresyon modellemesine gidilir.

Lojistik regresyon(LR) analizini diğer regresyon tekniklerinden farklı kılan özelliği, sürekli, ayrık, ikili veya karışık değişkenlerden oluşan bağımsız değişken setinden sonuç değişkenini tahmin etmemize izin vermesidir. Bu özelliğinden dolayı lojistik regresyon sağlık bilimlerinden sosyal bilimlere kadar pek çok alanda kullanılan bir tekniktir (Tabachnick ve Fidell, 2007: 437). Lojistik regresyon sağlık bilimlerinde hasta/hasta değil çıktısını elde etmek amacı ile kullanılırken, davranış bilimlerinde davranış bozukluğuna sahip/sahip değil, eğitim bilimlerinde başarılı/başarılı değil çıktılarının elde edilmesi amacı ile kullanılabilir. Ayrıca lojistik regresyonda diğer regresyon analizlerinden farlı olarak sonuç değişkeni (yordanan veya bağımlı değişkenin) iki değer alır ve ikili (binary veya dichotomous) olarak tanımlanır, bağımsız değişkenlerin normalliği, doğrusallığı, gruplarda varyansın eşitliği ve yordayıcıların kesikli olması gibi sayıtlılara ihtiyaç duyulmaz ve kategorik bir kriter tarafından üretilen doğrusal olmayan ilişkiler sürekli ve kategorik değişkenlere modellenebilir, kabul edilebilir olasılık tahminlerini hesaplanabilir (Hosmer ve Lemeshow, 2000: 1; Tabachnick ve Fidell, 2007: 437; King, 2003: 393).

Lojistik regresyonun matematiksel alt yapısı, olasılıklar oranı (odds ratio, OR) ve lojit (logit) şeklinde ele alınır. OR bir olayın gerçekleşme olasılığının gerçekleşmeme olasılığına oranıdır. Buna göre p bir çıktı değişkeninin gerçekleşme olasılığı ise, bu değişkenin gerçekleşmeme olasılığı olacaktır. Buna göre odds:

$$odds = \frac{p}{1-p}$$

şeklinde hesaplanır. Olasılıklar oranının doğal logaritması lojit (logit) olarak adlandırılır. (Hosmer ve Lemeshow, 2000: 49; Tabachnick ve Fidell, 2007:438). Lojit:

$$\log(Odds) = logit(P) = \ln\left(\frac{P}{1-P}\right)$$

şeklinde gösterilir. Lojit doğrusal regresyonun arzu edilen özelliklerinin çoğuna sahiptir ve doğrusal regresyondan daha fazla avantaj sağlar. Buna göre sürekli ve doğrusal parametreli olan lojit matematiksel olarak simetriktir. OR 0 ile +∞ arasında asimetrik değerler alırken, lojit P'ye bağlı olarak -∞ ile +∞ arasında değerler almaktadır. OR'nin 1'e eşit olması farklılığın diğer bir ifade ile ilişkinin olmadığına işaret etmektedir. Bu sebeple, zıt yönlü fakat aynı OR'yeişaret eden farklı değerler görülebilmektedir. Örneğin 5.0 (5/1) kazanma olasılığına karşılık kazanmama olasılığı 0.2 (1/5)'tir. Ancak aynı durum lojit ile ele alındığında kazanma olasılığı (ln(5)) 1,609 iken bu olasılığın tersi olan kazanmama olasılığı (ln (0.2)) kazanma olasılığının ters işaretlisi olan -1.609'a eşittir (Hosmer ve Lemeshow, 2000:6 ; Pedhazur,1997:716).

Lojistik regresyonda kullanılan istatistikler yukarıda verilenlerle sınırlı değildir. Bu istatistikler dışında model ve veri uyumunu belirlemek amacıyla -2 LogLikelihood (-2 Log Olabilirlik, -2LL), ki-kare ve sınıflama yüzdeleri kullanılır.

- -2LL istatistiği maksimum olabilirlik (maximumlike lihood) istatistiğinden türe-

tilmiştir ve verilerin modele uyumu hakkında bilgi verir. Maksimum olabilirlik 0 ile 1 arasında küçük değerler alır. Bu sınırlılığı aşmak için bu değerlerin doğal logaritması alınır ve -2 ile çarpılarak -2 loglikelihood değeri elde edilir. Bu şekilde 0 ile 1 arasındaki maksimum olabilirlik değerlerinin daha büyük pozitif değerler alması sağlanır. Mükemmel bir model için maksimum olabilirlik 1 olacağından -2LL, 0 (ln 1 = 0) değerini alır. Diğer bir ifade ile -2LL değeri 0'a yaklaştıkça model uyumu artmaktadır. Son olarak -2LL istatistiği normal dağılım sayıtlısına ihtiyaç duymamaktadır (Pedhazur, 1997: 721).

• Ki- kare istatistiği (χ^2) LR'de hipotez testlerinin reddedilmesi veya kabul edilmesi aşamasında önem derecesi değeri (p) ile birlikte kullanılır. Buna göre, α = 0,05 olduğu durumda, $p \leq$ 0,05 ise null hipotez reddedilir ve bağımsız değişkenlerin bağımlı değişkeni iyi yordadığı yorumu yapılır (Menard, 2002: 22 ; Pedhazur, 1997; 722).

• LR'de karşımıza çıkan bir diğer istatistik, sınıflama yüzdeleridir. Sınıflama yüzdeleri LR sonuçlarında sınıflama tablosu biçiminde çıkar ve model uyum göstergesi olarak yorumlanır (Çokluk, Şekercioğlu ve Büyüköztürk, 2010: 85). Bunun dışında birden fazla puanlayıcının yer aldığı ölçmelerde hesaplanan sınıflama yüzdeleri puanlayıcılar arası tutarlılığın yorumlanmasını sağlayan değerler verir.

PISA

Çalışmada kullanılan ölçme aracı, ülkemizde 2003 ve 2006 yıllarında asıl ve pilot uygulamaları gerçekleştirilen PISA Fen Bilimleri testinin açıklanan sorularından yararlanarak oluşturulmuştur. PISA, Uluslararası Öğrenci Değerlendirme Programı -(Programme for International Student Assessment), Ekonomik İşbirliği ve Kalkınma Örgütü OECD'nin üç yıllık aralıklarla düzenlemekte olduğu ve 15 yaş grubu öğrencilerin kazandıkları bilgi ve becerilerin değerlendirilmesine yönelik yapılan bir tarama araştırmasıdır. Bu proje kapsamında Türkiye de dahil olmak üzere 52 ülkedeki 15 yaş grubu öğrencilerinin Okuma Becerileri, Matematik okuryazarlığı ve Fen Bilimleri okuryazarlığı konu alanları ölçülmekte; ayrıca öğrencilerin motivasyonları, kendileri hakkındaki görüşleri, öğrenme biçimleri, okul ortamları ve aileleri ile ilgili bilgiler toplanmaktadır. Buna göre bu araştırmalarda, öğrencilerin öğretim stratejileri hakkındaki düşünceleri gibi konuları içeren geniş ölçekli eğitim çıktılarının elde edilmesi ve temel konu alanlarındaki performanslarının incelenmesi sağlanmaktadır (MEB, 2009). Bu bakımdan PISA bir performans sınavı şeklinde ele alınabilir.

PISA projesinde kullanılan «okuryazarlık» kavramı öğrencinin bilgi ve potansiyelini geliştirip, topluma daha etkili bir şekilde katılmasını ve katkıda bulunmasını sağlamak için yazılı kaynakları bulması, kullanması, kabul etmesi ve değerlendirmesi olarak tanımlanmakta ve bu doğrultuda ölçmeler yapılmaktadır. Bu proje kapsamında öğrencilere çoktan seçmeli, karmaşık çoktan seçmeli, açık uçlu, kapalı uçlu gibi değişik soru türleri yöneltilmektedir. (MEB, 2010). Uygulaması yapılan sorular dereceli

puanlama anahtarları (rubrikler) ile puanlayıcılar tarafından puanlanmaktadır.

Ancak PISA gibi öğrencilerin performansını yazılı olarak ortaya koyduğu sınavlarda öğrenci başarısı puanlayıcı güvenirliğinden etkilenebilmektedir. Puanlayıcının puanlamadaki katılığı öğrencinin başarısının düşük çıkmasına, puanlamadaki cömertliği ise öğrencinin hak ettiğinden yüksek başarı elde etmesine sebep olabilmektedir. Bu nedenle bu tip sınavlarda bir puanlayıcı yerine birden fazla puanlayıcıya yer vermek tercih edilebilir. Ancak bu durumda puanlayıcıların birbirleriyle tutarlı puanlama yapması sağlanması gereken bir koşul olarak karşımıza çıkmaktadır. Bu koşulun yerine getirilmesi için PISA sınavında olduğu gibi dereceli puanlama anahtarlarından yararlanmaya gidilebilir. Bunun yanında puanlayıcılar arası tutarlık çeşitli istatistiksel yöntemlerle araştırılabilir. Bu amaçla bu çalışmada, PISA sınavından elde dilen form öğrencilere uygulanıp puanlayıcılar tarafından dereceli puanlama anahtarları ile puanlanarak G kuramı ve LR analizi yöntemi ile puanlayıcılar arası tutarlık belirlenmeye yönelik olarak analiz edilmiştir. Puanlayıcı tutarlıklarının ve hata kaynaklarının belirlenmesi amacıyla aşağıdaki problem cümleleri araştırılmıştır.

Problem Cümlesi

Genellenebilirlik kuramı ve lojistik regresyon analizine göre birden fazla puanlayıcının yer aldığı performans ölçmelerinde puanlayıcılar arası tutarlık nasıldır?

Alt Problemler

1. Bilgi, kavrama, uygulama ve analiz basamaklarında yer alan maddelerden elde edilen puanlara dayalı olarak G kuramı ve lojistik regresyon analizi ile hesaplanan puanlayıcılar arası tutarlık nasıldır?

2. Testin tamamından elde edilen puanlara dayalı olarak G kuramının ö x m x p desenine göre ve başarılı-başarısız sınıflamasına dayalı olarak lojistik regresyon analizine göre puanlayıcılar arası tutarlık nasıldır?

Amaç

Bu araştırmanın amacı; birden fazla puanlayıcının yer aldığı yazılı bir performans sınavında puanlayıcılar arası tutarlığı genellenebilirlik kuramı ve lojistik regresyon analizi ile belirlemek ve karşılaştırmaktır. Bu amaçla araştırmada kullanılan yöntemlerden hangisinin puanlayıcılar arası tutarlığı belirlemede daha hassas sonuçlar sağladığı sorusuna cevap verilmesi beklenmektedir.

Yapılan benzer çalışmalarda psiko-motor becerileri ölçen performans sınavlarından yararlanılmıştır. Bu çalışmalarda puanlayıcılar öğrenci davranışlarını en fazla bir defa doğrudan veya kamera aracılığı ile birkaç kez gözlemleyebilmektedir. Ancak öğrenci davranışının bir kez gözlemlenmesi puanlayıcıların öğrenci performansına ilişkin tek seferde karar vermesini zorunlu kılmaktadır. Kamera aracılığı ile yapılan gözlemlerde de sahip olunan teknik donanımdan doğabilecek sıkıntılar, puanlayıcının

sınav ortamından uzak olması, istediği açıdan görüş alanına sahip olamaması gibi kısıtlar söz konusu olabilmektedir. Ancak bu araştırma, puanlayıcıların öğrenci davranışlarını sınav kağıtları aracılığıyla istediği kadar gözlemleyebilmesi yönünden diğer çalışmalardan ayrılmaktadır. Çalışma ayrıca bilişsel becerileri ölçen ve yazılı bir performans sınavı olan PISA sınavından elde edilen bir form ile performans ölçümüne odaklanması ile özgündür.

Türkiye'de yapılan puanlayıcılar arası tutarlık araştırmalarında farklı yöntemlerin birlikte kullanılarak karşılaştırılmasına gidilmemiştir. Bu araştırma, puanlayıcılar arası tutarlığı belirlemek üzere daha önceki çalışmalarda birlikte kullanılmamış olan iki farklı istatistiksel yöntemden yararlanması ve sonuçlarını karşılaştırması yönünden önem taşımaktadır.

Sayıltılar

1. Araştırmaya katılan puanlayıcılar birbirinden bağımsız puanlama yapmıştır.

Sınırlılıklar

1. Bu araştırma kullanılan ölçme aracı ile sınırlıdır.

2. Puanlamada yaşanan güçlüklerden dolayı, öğrenci sayısı 106 ve soru sayısı 15 ile sınırlandırılmıştır.

2. Yöntem

Araştırmanın Türü ve Grubu

Araştırma uygulaması yapılan performans sınavında ortaya konan puanlayıcılar arası tutarlığın genellenebilirlik kuramı ve lojistik regresyon analizi ile belirlenmesine, kullanılan yöntemlerin ürettiği çıktıların karşılaştırılmasına ve güçlü ve zayıf yönlerinin belirlenmesine yönelik olması açısından betimseldir.

Araştırmanın çalışma grubunu, Ankara ili Çankaya ilçesinde bulunan bir Anadolu Lisesinde 2010-2011 eğitim-öğretim yılında dokuzuncu sınıfta öğrenim gören 106 kişilik öğrenci grubu oluşturmaktadır.

Araştırmanın Verileri

Araştırmanın verileri, 2010-2011 eğitim-öğretim yılında dokuzuncu sınıfta öğrenim gören 106 öğrenciye uygulanan ve MEB EARGED tarafından açıklanan PISA Fen Bilimleri testinden yararlanarak oluşturulan ölçme aracı ile elde edilmiştir. Kullanılan ölçme aracı öğrencilerin fen bilimleri okuryazarlığını ölçmeyi amaçlayan 15 adet açık uçlu ve kısa cevaplı sorudan oluşmaktadır.

Ölçme aracının öğrencilere uygulanması ile elde edilen cevap kağıtlarının birer kopyası en az yüksek lisans düzeyinde öğrenime sahip ve sırasıyla fizik, kimya ve

biyoloji alanlarında branşlaşmış puanlayıcılara verilmiştir. Puanlayıcılar öğrencilerin cevaplarını dereceli puanlama anahtarı doğrultusunda birbirinden bağımsız şekilde puanlamıştır. Araştırmacı tarafından PISA Fen Bilimleri Dereceli Puanlama Anahtarı'ndan yaralanarak geliştirilen dereceli puanlama anahtarında sorular gözlenmedi (0), kısmen gözlendi (1) ve gözlendi (2) şeklinde ağırlıklandırılmıştır.

Öğrencilerin kullanılan fen bilimleri testinden alabileceği en yüksek puan 30 (15x2)'dur. Uzman görüşüne dayalı olarak testten başarılı olan öğrencilerin 30 tam puan üzerinden 15 alması gerektiği kararına varılmıştır. Buna göre testten elde ettiği puan 15 ve üzeri olan öğrenciler başarılı (1), testten elde ettiği puan 15'in altında olan öğrenciler başarısız (0) olarak sınıflanmıştır.

Verilerin Analizi

Verilerin analizi madde bazında ve test bazında olma üzere iki aşamada gerçekleştirilmiştir. Veriler alt problemler doğrultusunda genellenebilirlik kuramı ve lojistik regresyon ile analiz edilmiştir.

Genellenebilirlik kuramı doğrultusunda yapılan madde bazındaki analizlerde her bir madde için 106 öğrencinin 3 puanlayıcı tarafından puanlanması ile oluşturulan ö x p çapraz deseni kullanılmıştır. Bu desen için gerçekleştirilen genellenebilme çalışması sonunda kestirilen varyans bileşenlerinden puanlayıcı varyans bileşeni ve puanlayıcı varyans bileşeninin toplam varyansiçersindeki yüzdesi elde edilmiştir.

Lojistik regresyon ile madde bazında yapılan analizlerde puanlayıcıların her bir maddeye verdikleri puan (yordayıcı) başarılı-başarısız sınıflamasına (yordanan) dayalı olarak analiz edilmiştir. Analiz sonunda verilere ilişkin model uyum değerlerinden ki-kare ve -2LL incelenmiş, daha sonra elde edilen sınıflama tablolarında puanlayıcıların uzlaşmalarını veren sınıflama yüzdeleri puanlayıcılar arası tutarlığın bir ölçüsü olarak yorumlanmıştır.

Testin tamamı için gerçekleştirilen genellenebilirlik kuramı analizinde, 106 öğrencinin 15 madde doğrultusunda 3 puanlayıcı tarafından puanlanması ile elde edilen ö x g x p çapraz deseni kullanılmıştır. Kullanılan desen ile yapılan analiz sonunda kestirilen varyans bileşenlerinden puanlayıcıvaryans bileşeni ve toplam varyans içindeki yüzdesi testi puanlama puanlayıcıların göstermiş olduğu puanlayıcı tutarlığını yorumlamada kullanılmıştır.

Lojistik regresyon analizi yöntemi ile başarılı-başarısız sınıflamasına (yordanan) dayalı olarak gerçekleştirilen toplam puan (yordayıcı) analizinde verilere ilişkin model uyum değerlerinden ki-kare ve -2LL incelenmiş, verilerin modele uygun olduğu tespit edildikten sonra elde edilen sınıflama tablolarında puanlayıcıların uzlaşmalarını veren sınıflama yüzdeleri puanlayıcılar arası tutarlığın bir ölçüsü olarak yorumlanmıştır.

3. Bulgular ve Yorumlar

Madde bazında yapılan analizlerde taksonomik sınıflandırmaya göre, bilgi, kavrama, uygulama ve analiz basamaklarında yer alan maddelerin her biri puanlayıcılar arası tutarlık açısından genellenebilirlik çalışmaları ve lojistik regresyon ile analiz edilmiştir. Analiz genellenebilirlik kuramı doğrultusunda ö x p desenine göre yapılmıştır. Lojistik regresyon analizinden faydalanarak yapılan analizlerde öğrencilerin başarılı başarısız şeklinde sınıflandırılmış ve maddeler üzerinden sınıflama yüzdeleri elde edilmiştir.

Yukarıda açıklanan basamaklar doğrultusunda uygulaması yapılan fen bilimleri sınavında yer alan 15 madde için gerçekleştirilen G çalışması ve LR analizi çıktıları Tablo 2'de yer almaktadır.

Tablo 2. Madde Puanlarına Dayalı Gerçekleştirilen Lojistik Regresyon Analizi Ve G Çalışması Çıktıları

	Madde No	LR İle Elde Edilen Sınıflama Yüzdesi	G Kuramı İle Elde Edilen Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdesi		
			Öğrenci (ö)	Puanlayıcı (p)	Öğrenci-Puanlayıcı (öp)
Bilgi Basamağı	1	%87.7	% 78.5	%1.9	%19.6
	5	%88.7	%77.0	%0,3	% 22.7
	8	%87.7	%47.9	%4.8	% 47.3
	12	%89.6	%28.2	%6.2	%65,6
Kavrama Basamağı	4	%87.7	%41.6	%20.7	%37.8
	7	%87.7	%25.0	%7.2	%67.8
	9	%87.7	%72.1	%2.7	%25.2
	14	%90.6	%8.1	%51.7	%40.2

	Madde No	LR İle Elde Edilen Sınıflama Yüzdesi	G Kuramı İle Elde Edilen Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdesi		
			Öğrenci (ö)	Puanlayıcı (p)	Öğrenci-Puanlayıcı (öp)
Uygulama Basamağı	2	%87.7	%36.9	%21.0	%42.1
	3	%87.7	%82.1	%0.7	%17.2
	11	%87.7	%49.8	%2.6	%47.6
	13	%87.7	%52.9	%0.4	%46.7
Analiz Basamağı	6	%88.7	%44.6	%4.1	%51.2
	10	%87.7	%56.2	%4.7	%39.0
	15	%87.7	%43.2	%8.5	%48.4

Tablo 2'de yer alan bilgilere göre LR analizi ile elde edilen sınıflama yüzdeleri %87,7 ile %90,6 arasında değerler almaktadır. Bu sonuçlara göre puanlayıcılar tüm maddelerde yüksek oranda uzlaşma göstermişlerdir. Madde 14 puanlayıcıların en yüksek oranda uzlaşma gösterdikleri maddedir. Bu durum testin tamamında puanlayıcıların tutarlı davrandıkları, öğrencileri puanlamada benzer şekilde davrandıkları ve öğrencilerin performanslarının bir puanlayıcıdan diğerine değişmediği şeklinde yorumlanabilir.

Madde bazında yapılan G çalışması sonuçlarına göre, öğrencilere ait varyans bileşenleri toplam varyans içinde % 78,5 ile %8,1 arasında değerler almaktadır. Öğrenci varyans bileşeninin maddelerin genelinde yüksek değer alması öğrencilerin başarılarının farklılık gösterdiği anlamına gelir. Bu istenen bir durumdur. Puanlayıcı varyans bileşeni toplam varyansın %0,3'ü ile %51'i arasında değerlere sahiptir. %51'lik varyans değerinin bir uç değer olduğu göz önüne alınırsa, puanlayıcıların öğrencileri puanlamada düşük düzeyde değişkenlik gösterdikleri, testin genelinde tutarlı puanlama yaptıkları yorumu yapılabilir.Öğrenci-puanlayıcı ortak etkileşimi varyans bileşenleri incelendiğinde öğrencilerin durumlarının bir puanlayıcıdan diğerine değişiminin madde 12'de en yüksek (%65.6), madde 3'te ise en düşük (%17.2) düzeyde olduğu görülmektedir. Bu durum öğrencilerin durumlarının bir puanlayıcıdan diğerine madde 12'de en yüksek oranda, madde 3'te en düşük oranda değiştiği şeklinde yorumlanabilir.

İki yöntem birlikte değerlendirildiğinde sınıflama yüzdelerinin yüksek olmasının ve puanlayıcı varyans bileşeninin toplam varyans içinde düşük bir yüzdeye sahip ol-

masının birbirini destekler nitelikte olduğu ve puanlayıcıların tutarlı puanlama yaptıklarına işaret ettiği yorumu yapılabilir.

Testin tamamından elde edilen puanlara dayalı olarak gerçekleştirilen G çalışması ve LR analizi çıktıları aşağıda verilmiştir.

Tablo 3. Testin Tamamı İçin ö x m x p Deseninde Gerçekleştirilen G Çalışması Sonunda Elde Edilen Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri

Varyans Kaynağı	Kareler Toplamı	Serbestlik Derecesi	Kareler Ortalaması	Varyans	%
ö	164.55702	105	1.56721	0.01397	2.4
m	489.59413	14	34.97101	0.09557	16.2
p	**53.15010**	**2**	**26.57505**	**0.01428**	**2.4**
öm	1350.00587	1470	0.91837	0.24360	41.2
öp	43.64990	210	0.20786	0.00135	0.2
mp	107.78071	28	3.84931	0.03454	5.8
ömp	551.41929	2940	0.18756	0.18756	31.7
Toplam	2760.15702	4769			100%

Tablo 3'te yer alan bilgilere göre, öğrenci ve puanlayıcıvaryans bileşenlerinin her biri toplam varyansın %2,4'ünü açıklamaktadır. Bu varyans bileşeni değeri açığa çıkan diğer varyans bileşenleri arasında en küçük değerli bileşenler olup, testin tamamında öğrencilerin birbirine yakın başarı gösterdiği ve elde ettikleri puanlarda değişkenliğin düşük olduğu; puanlayıcıların öğrencileri puanlamada tutarlı davranarak, birbirlerine benzer şekilde puanlama yaptığı diğer bir ifade ile testin tamamında puanlayıcı tutarlığının sağlandığı şeklinde yorumlanabilir.

Madde varyans bileşeni maddelerin zorluk kolaylık bakımından farklılık gösterip göstermediği hakkında bilgi verir (Shavelson ve Webb, 1991: 23). Yapılan G çalışmasına göre madde ana etkisi toplam varyansın % 16,2'sini açıklamaktadır. Elde edilen bu sonuç testin farklı güçlerde maddeler içerdiği şeklinde yorumlanabilir.

Öğrenci-madde ortak etkisi öğrencilerin bir maddeden diğerine performanslarında görülen değişiklik olarak açıklanabilir (Shavelson ve Webb, 1991: 23). Bu etkiye ait varyans bileşeni toplam varyansın %41.2'sini açıklamaktadır. Bu sonuçlar öğrencilerin gösterdikleri performans bakımından maddeden maddeye farklılık gösterdikleri biçiminde yorumlanabilir. Bu durumun ortaya çıkmasında maddelerin farklı güçlüklere sahip olmasının etkili olduğu söylenebilir.

Öğrenci-puanlayıcı ortak etkisine ait varyans bileşeni diğer varyans bileşenleri arasında en küçük değerli varyans bileşenidir. Bu varyans bileşeni öğrencilerin

gözlenen durumlarının bir puanlayıcıdan diğerine değişimi hakkında bilgi verir. Testin tamamından elde edilen verilere dayalı olarak hesaplanan öğrenci-puanlayıcı ortak etkisi toplam varyansın %0.2'sini açıklamaktadır. Kestirilen bu düşük değer puanlayıcıların yaptıkları puanlamalarda öğrenciden öğrenciye farklılık göstermediği şeklinde yorumlanabilir. Ayrıca puanlayıcı değişkenlik kaynağından elde edilen sonuca paralel olarak öğrenci-puanlayıcı değişkenlik kaynağının düşük değer alması puanlayıcıların testin tamamında tutarlı puanlamalar yaptığı görüşünü destekler niteliktedir.

Madde-puanlayıcı ortak etkisine ait varyans bileşeni puanlayıcıların bir maddeden diğerine kararlılıklarını göstermektedir. Testin tamamı için gerçekleştirilen G çalışması doğrultusunda madde-puanlayıcı ortak etkisi toplam varyansın % 5.8 'ini açıklamaktadır. Elde edilen bu sonuç diğer varyans bileşenlerine göre düşüktür ve puanlayıcıların bir maddeden diğerine gösterdikleri cömertlik ve/veya katılığın değişmediği şeklinde yorumlanabilir (Shavelson ve Webb, 1991: 23).

Öğrenci-madde-puanlayıcı ortak etkisi ikinci büyük değere sahip ortak etkidir. Bu ortak etkiye ait varyans bileşeni toplam varyansın %31.7'sini açıklar niteliktedir. Ortak etkinin büyük değere sahip olması öğrenci, madde ve puanlayıcı değişkenlik kaynaklarının etkileşiminin veya tesadüfi hata kaynaklarının büyük olduğu şeklinde yorumlanabilir (Shavelson ve Webb, 1991: 23).

Öğrencilerin testten elde ettikleri toplam puana dayalı olarak gerçekleştirilen lojistik regresyon analizi sonucunda verilerin modele uygun olduğu -2LL istatistiği ve ki-kare testi(-2LL=,000; χ^2=,000; p=1.000>0.05) ile belirlenmiştir. Model veri uyumunun sağlanmış olması sınıflama yüzdeleri tablolarının yorumlanabileceği anlamına gelir. Tablo 4'te öğrencilerin testten elde ettikleri ortalama toplam puana dayalı elde edilen sınıflama yüzdeleri yer almaktadır.

Tablo 4. Testten Elde Edilen Ortalama Toplam Puanlara Dayalı Olarak Elde Edilen Sınıflama Yüzdeleri

Gözlenen		Beklenen Sonuç		Doğru Sınıflama
		Başarısız	Başarılı	
Sonuç	Başarısız	13	0	100,0
	Başarılı	0	93	100,0
Toplam				100,0

Tablo 4'te puanlayıcıların on beş maddeye verdikleri puanlar üzerinden lojistik regresyon analizi ile ulaşılan uzlaşma ölçüleri yer almaktadır. Analiz sonuçlarına göre puanlayıcıların testin tamamında gösterdikleri uzlaşma ölçüsü %100'dür. Bu uzlaşma yüzdesi puanlayıcıların öğrencilerin tamamını aynı şekilde başarılı veya başarısız

olarak sınıfladığı anlamına gelmektedir. Bu durum puanlayıcıların öğrencileri değerlendirmede %100 tutarlık gösterdiği biçiminde yorumlanabilir. Tutarlığın yüksek olmasında puanlayıcıların birbirine eş puanlamalar yapmasının etkili olduğu, puanlayıcıların eş puanlamalar yapmasında dereceli puanlama anahtarını amacı doğrultusunda kullanmalarının rol oynadığı yorumu yapılabilir.

Testin tamamı için gerçekleştirilen G çalışmasına göre puanlayıcı varyans bileşeni 0.014 değeri ile toplam varyansın %0.2'sini açıklamaktadır. Bu sonuçlar testin tamamında puanlayıcıların oldukça büyük oranda tutarlık gösterdiği ancak testin tamamında tutarlı olmadıkları şeklinde yorumlanabilir. Aynı veriler ile ortalama toplam puana dayalı gerçekleştirilen LR analizlerine göre puanlayıcılar testin tamamında %100 uzlaşma göstermişlerdir. Bu sonuçlar birlikte değerlendirildiğinde; G kuramının daha ayrıntılı analizler ile hassas sonuçlar oluşturduğu, tutarlıkta var olan küçük değişiklikleri çıktılarına yansıttığı; buna karşılık lojistik regresyonun ayrıntılarla, değişikliklerle ve varyanslarla ilgilenmeyen, puanlayıcıların ortaya koydukları tutarlık derecesinde meydana gelen değişiklikleri tam olarak yansıtmayan analiz çıktıları oluşturduğu yorumu yapılabilir.

4. Sonuç ve Öneriler

Araştırmanın birinci alt problemine göre öğrencilerin bilgi, kavrama, uygulama ve analiz maddelerinden elde ettikleri puanlara dayalı olarak gerçekleştirilen G çalışması ve LR analizi çıktıları karşılaştırıldığında; ölçme aracında yer alan her bir maddede G çalışması ve lojistik regresyon analizi ile elde edilen puanlayıcılar arası tutarlık değerleri farklılık göstermektedir. Buna göre lojistik regresyona ait değerler G çalışması sonucunda elde edilen değerler ile birebir tutarlık göstermemesine rağmen madde puanlarından elde edilen LR analizi sonuçları yüksek uzlaşma yüzdesine, G çalışması sonuçları ise yüksek tutarlık değerlerine işaret etmektedir. Bu sebeple tüm maddelerde puanlayıcıların her iki yönteme göre de tutarlı puanlama yaptıkları sonucuna varılmıştır.

Araştırmanın ikinci alt problemine göre testin tamamından elde edilen puanlara dayalı olarak gerçekleştirilen G çalışması ve LR analizi çıktıları karşılaştırıldığında; G çalışması ile hesaplanan puanlayıcıvaryans bileşeni 0.01428 değeri ile toplam varyansın %2.4'ünü açıklamakta, lojistik regresyon analizi ile elde edilen sınıflama yüzdesine göre, puanlayıcılar öğrencileri %100 doğru sınıflamaktadır. Bu değerler puanlayıcıların öğrencileri puanlamada tutarlı davrandığını gösteren paralel sonuçlardır. Ancak sonuçlar içerdiği bilgi bakımından kıyaslandığında, lojistik regresyon analizinin puanlayıcılar arası tutarlığı belirlemede değişiklikler ile ilgilenmeyen yüzeysel bir istatistik olduğu, G kuramının ayrıntıları ve değişiklikleri analiz sürecine ekleyen detaylı bir istatistiksel analiz olduğu sonucuna varılmıştır.

Bu araştırmanın sonucunda araştırmacıların performans ölçümleri gibi birden fazla puanlayıcının yer aldığı ölçmelerde puanlayıcılar arası tutarlığı hassas ve ayrıntılı

olarak belirlemeleri için genellenebilirlik kuramından yararlanmaları, buna karşılık puanlayıcılar arası tutarlık hakkında ayrıntılar ile ilgilenmeyerek genel bilgi sahibi olmak istedikleri, tutarlığın derecesinden çok varlığının veya yokluğunun ortaya konmasını istedikleri durumlarda lojistik regresyon analizi tekniğinden yararlanmaları önerilmektedir. Gerçekleştirilecek performans ölçümlerinde tutarlı puanlamalar yapılabilmesi için öğrenci cevaplarını puanlamada birden fazla puanlayıcıya yer verilmesi, dereceli puanlama anahtarı ile puanlama yapılması ve puanlayıcıların yapacakları puanlamalar hakkında önceden bilgilendirilmesi önerilmektedir.

5. Kaynakça

AHMANN, J. S. and GLOCK, M. D. (1971). Measuring And Evaluating Educational Achievement, Boston : Allyn and Bacon

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

CROCKER, L. ve ALGİNA, J. (1986). Introduction To Classical And Modern Test Theory, New York: Wadsworth.

ÇOKLUK, Ö., ŞEKERCİOĞLU, G. , BÜYÜKÖZTÜRK, Ş. (2010). Sosyal Bilimler İçin Çok Değişkenli İstatistiksel Analiz: SPSS ve LISREL Uygulamaları. Ankara: PegemA Yayınları.

GLESER, G.C., CRONBACH, L. J., RAJARATHAM N. (1965). Generalizability Of Scores Influenced By Multıple Sources Of Varıance. Psychometrika, 30(4),395-418.

HOSMER, D. W., LEMESHOW, S. (2000). Applied Logistic Regression, Second Edition. USA : A Wiley Inter Science Publication.

KAYRİ, M. Ve ÇOKLUK, Ö. (2010). Using Multinominal Logistic Regression Analysis In Artificial Neural Networt: An Application. Ozean Journal Of Applied Sciences, 3(2), 259-268.

KING, J.E. (2003). Running A Best-Subsets Logistic Regression: An Alternative To Stepwise Methods, Educaional And Psychological Measurement, 63, 392-403.

LANE, S. and STONE, C. A. (2006). Educational Measurment, Fourth Editon (Ed: Robert Brennan) USA:American Council On Education Praeger.

MEB (2010). PISA 2006 Nihai Raporu Milli Eğitim Bakanlığı Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı.

MENARD, S. W. (2002). Applied Logistic Regression Analysis, Thous and Oaks. : Sage.

ÖZTÜRKCAN; M. (2009). Regresyon Analizi. İstanbul: Maltepe Üniversitesi Yayınları, 40.

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

PEDHAZUR, E. J. (1997). Multiple Regression In Behavioral Research Explanation And Prediction (3rd. Ed.).Australia : Wadsworth, Thomson Learning.

RENTZ, J. O. (1987). Generalizability Theory: A Comprehensive Method For Assessing And Improving The Dependability of Marketing Measures. Journal Of Marketing Research , 24(1), 19-28.

SHAVELSON,R.J.,WEBB,N.www.stanford.edu/dept/SUSE/SEAL/Reports-Papers/Generalizability%20Theory-ESM-Final.doc. adresinden 19.11.2010 tarihinde alınmıştır.

SMITH, V. E., KULIKOWICH, J. M. (2004). An Application Of Generalizability Theory And Many Facet Rasch Measurement Using A Complex Problem Solving Skills Assesment. Educational and Psychological Measurement, 64(4), 617-639.

TABACHNICK, B. G., FIDELL, L. S. (2007). Using Multivariate Statistics (5th Ed.).USA:Education Pearson Education Inc.

TOBAR, D. A., STEGNER, A. J., and KANE, M. T. (1999). The Use Of G-Theory in Examming The Dependability Of Scores On The Profile Of Mood States. Measurement In Physical Education and Exercise Science, 3 (3), 142-146.

TURGUT, M. F. ve BAYKUL, Y. (2010). Eğitimde Ölçme ve Değerlendirme. Ankara: PegemA Yayınları.

EXTENDED ABSTRACT

Introduction: Measurements are objective if only all raters give the same scores to the same answers (Turgut and Baykul, 2010:132) and so the rater reliability can be ensured. Therefore, in order to increase the reliability of the measurement process, number of raters can be increased. Rater reliability enhancement studies are also important in the performance assessments which can be defined as assessments based on the observations and conclusions (Stiggins, 1997 Akt: Palm, 2008). In performance assessments deciding how students will be scored and ensure the reliability of raters are the problems encountered. The rubrics con be used to lead scoring, in order to ensure interrater reliability of raters consistency must be investigated. In this study, in order to examine interrater consistency, Generalizability Theory and Logistic Regression Analysis are discussed.

Generalizability Theory (G Theory), is a theory that is based on variance analysis (ANOVA); to determine and measure the inconsistencies that occur or will occur upon the observed scores with strong statistical methods (Brennan, 2001:4). With this theory we are able to generalize observed scores to universal scores; by calculating the variance components of independent variables in universe and the interactions between these variables. Rater variance component and interactions of obtained variance components are informative about the consistency of the raters during scoring.

Logictic Regression (LR) Analysis helps us to approximate the output variable from the independent set of continuous, separate, binary and dichotomous variables. Because of that property Logictic Regression (LR) is used in a wide variety of areas from health sciences to social sciences (TabachnickveFidell, 2007: 437). -2 Log Likelihood, chi-square and classification percentages are some of the statistics used in Logictic Regression. -2LL statistics show the compatibility of data to the model. Chi-square statistics (χ^2) is used together with significance value (p) for the rejection or acceptation of hypothetical tests in the logistics regression. Classification percentages are appeared as classification tables in the logistic regression and interpreted as indicator of model compatibility (Çokluk, Şekercioğlu ve Büyüköztürk, 2010: 85). Classification percentages can also be used for interpretation of interrater agreement.

Problem Statement: How is the interater consistency in the performance measurements according to the generalizability theory and logistic regression analysis?

Sub Problems:

1. How is the interrater consistency analyzed by G Theory and LR analysis upon the items in knowledge, comprehension, application and analysis levels?

2. How is the interrater consistency analyzed by G Theory and LR analysis upon whole test items?

Purpose: The purpose of this study is to examine and compare the interrater consistency in the performance assessments through G Theory and LR analysis. At the end of the study researcher expects to answer the following question: Which method provides more sensitive outputs?

Method : The research is descriptive since it is based on determining interrater consistency and comparison of the outputs produced by the G Theory and LR analysis. Research's study group consists of 106 ninth grade students of an Anatolian High School located in Çankaya-Ankara. Data generated by applying the measurement tool to students. The measurement tool consists of 15 questions and obtained from PISA Science test. The responses are scored by 3 raters as in the rubric. Based on the expert opinion, students who take 15 points over full score of 30 are passed. Others are failed (2).

Data set has been analyzed by s x r (s: student, r: rater) and s x i x r (s: student, i: item r: rater) design of G Theory. In the analysis made by LR analysis chi-square and -2LL are investigated and classification percentages interpreted.

Findings and Comments: According to the LR analysis based on items, classification percentages varies between %87,7 and %90,6. This means raters are consistent in scoring students. Also this result can be interpreted as students performances don't differ from one rater to another. According to the G studies based on items, rater variance component explains total variance between %0,3 and %51. If we take into account the value %51 as an outlier, results can be interpreted as; raters show a low level of variability in scoring students and are consistent through the whole items of the test.

According to the LR analysis based on the whole test, classification percentage is %100. The rater variance component explains %2,4 of total variance. The results obtained by both methods can be interpreted as the raters are consistent on the whole test.

Conclusions and Recommendations; Analysis results indicate high level of rater consistency. Therefore, we can conclude that; the interrater reliability is provided for all items and whole test.

If the results are compared in terms of included information, we can conclude that; LR analysis is statistically superficial analysis for determining interrater consistency. On the other hand; G Theory is a more accurate and detailed statistic and considers the changes in analysis process

As a recommendation of this research; researchers who want to determine interrater consistency precisely and detailed can benefit from G theory. If researchers want to have a general idea about interrater consistency they can benefit from LR analysis.