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

Açık Uçlu Maddelerde Farklı Yaklaşımlarla Elde Edilen Puanlayıcılar Arası Güvenirliğin Değerlendirilmesi*

The Evaluation of Rater Reliability of Open Ended Items Obtained from Different Approaches

Neşe GÜLER ** Gülşen TAŞDELEN TEKER ***

Öz
Bu araştırmada puanlayıcılar arası güvenirliğin belirlenmesinde kullanılabilecek dört farklı yaklaşım üzerinde durulmuştur: korelasyon, ortalamaların karşılaştırılması, uyuşma yüzdesi ve genellenebilirlik kuramı. Bu bağlamda 43 öğrencinin on açık uçlu maddeye verdikleri cevapların iki puanlayıcı tarafından puanlanmasıyla oluşturulan veri setine uygulanan yaklaşımlar ile elde edilen güvenirlik değerlerinin, değer aralıkları ve hesaplanma farklılıklarından dolayı farklılaştığı görülmüştür. Çalışma kapsamında ele alınan yaklaşımlar ile elde edilen güvenirlik kestirimlerindeki en yüksek değer 0,90 olarak genellenebilirlik kuramı ile elde edilmiştir. Bu sonucun yanı sıra, hesaplanan basit korelasyonda yüksek düzeyde ve pozitif yönlü (0,74) bir ilişki ortaya çıkmıştır. Puanlayıcılar arası uyuşma yüzdesiyle kestirilen tam uyum % 58,9 olarak belirlenmiştir. Son olarak, maddeler ayrı ayrı incelendiğinde; puanlayıcıların ortalamaları arasında üç maddede anlamlı bir farklılık çıkmakla beraber genel ortalamada anlamlı bir fark çıkmamıştır. Araştırma kapsamında ele alınan yaklaşımlar içerisinde en karmaşık görünen genellenebilirlik kuramı olmasına rağmen, bu yöntemin pek çok hata kaynağını aynı anda ele alabilme özelliği, diğer yöntemlere göre bir avantaj olarak düşünülebilir. Bu sebeple, özellikle puanlayıcı güvenirliğinin belirlenmesinde kullanılması önerilebilir.

Anahtar Kelimeler: puanlayıcılar arası güvenirlik, korelasyon, ortalama karşılaştırması, uyuşma yüzdesi, genellenebilirlik kuramı

Abstract
In this study, four approaches to the estimation of interrater reliability are studied: correlation, comparison of means, percentage of agreement, and generalizability theory. For the data- composed of ratings for 43 students on ten items by two raters- the reliability estimates varied because of the situation that the ranges of the obtained values by used approaches and different calculation processes. The highest estimate was 0.90 which is estimated by G theory. Besides this result, it was obtained that there was positive and high correlation coefficient (0.74). The estimate of percentage of exact matches of agreement between the two raters was found as 58.9 %. Finally, although there were no statistically differences between general mean of scores, there were statistical differences among three of the items by means of rater scoring. Although G theory seems more complex than the other methods illustrated in the study, it yields more information than the other methods because of handling multiple sources of error at the same time. Therefore, it is proposed to be used when estimating interrater reliability.

Key Words: interrater reliability, correlation, comparison of means, percentage of agreement, generalizability theory

GİRİŞ

Güvenirlik, hem eğitim ve psikolojide kullanılan testler için hem de bu testlerin sonucuna dayalı değerlendirmeler yapmak için önemli bir kavramdır. Bu nedenle ölçme araç ve yöntemlerinde bulunması gereken temel bir özelliktir (AERA, APA ve NCME, 2004; Atılgan, Kan ve Doğan, 2011; Haertel, 2006). Güvenirlik genel olarak, ölçme sonuçlarının tesadüfî hatalardan arınıklık derecesi olarak tanımlanmaktadır (Turgut, 1993).

* Bu araştırma III.Ulusal Eğitimde Ölçme ve Değerlendirme Kongresi'nde sözlü bildiri olarak sunulmuştur.
** Doç. Dr., Sakarya Üniversitesi, Eğitim Fakültesi, Sakarya-Türkiye, e-posta: gnguler@gmail.com
*** Öğr. Gör. Dr., Sakarya Üniversitesi, Eğitim Fakültesi, Sakarya-Türkiye, e-posta: gtasdelen@sakarya.edu.tr

Hata varyanslarının kaynağı, maddeler arası tutarsızlık diğer bir ifadeyle testin heterojen maddelere sahip olması, bir dizi paralel madde ya da testler arasındaki performans farklılıkları ve aynı ya da paralel testlerden elde edilen performansın zaman aralıklarına göre farklılık göstermesi olabilir (Atılgan, Kan ve Doğan, 2011). Bir diğer ifadeyle güvenirliği etkileyecek hata, güvenirliğin anlamına göre (kararlılık, tutarlılık, duyarlılık) veya ele alınan hatanın kaynağına göre (zaman, test formu, madde, durum vb.) farklılık gösterir (Baykul, 2000; Crocker ve Algina, 1986; Lord ve Novick, 1968). Bu kaynakların yanı sıra incelenmesi gereken bir diğer önemli hata kaynağı ise puanlayıcılardır.

Özellikle sosyal bilimlerde ve tıp biliminde birçok alanda birden fazla puanlayıcının görev aldığı ve bu puanlayıcıların yaptıkları puanlamalar arasındaki güvenirliğin belirlenmesini gerektiren pek çok ölçme durumu bulunmaktadır. Puanlayıcı güvenirliği, iki ya da daha fazla puanlayıcının farklı bireylere ve farklı maddelere ilişkin yaptıkları puanlamalar arasındaki tutarlılığın derecesi olarak tanımlanmaktadır (Aiken, 2000; Anastasi ve Urbina, 1997). Puanlayıcı hata kaynağı ise, iki veya daha fazla puanlayıcının her bir test maddesini puanlamaları arasındaki farklılıktan ileri gelmektedir.

Puanlayıcı hatalarının olmadığını söyleyebilmek için, iki veya daha fazla puanlayıcı tarafından yapılan puanlamadan elde edilen madde ve test puanları arasında anlamlı bir fark olmaması gerekir. Ancak eğitimde açık uçlu soruların puanlanması sırasında olduğu gibi kimi ölçme araçlarının puanlanmasına öznel etkilerin karışması söz konusudur. Bu nedenle de puanlayıcıların verdiği öznel puanları ve bu puanlara dayalı verilen kararların doğruluğunu değerlendirmek üzere farklı puanlayıcıların verdikleri madde ve test puanlarının birbiriyle ne derece tutarlı olduğunun belirlenmesi gereklidir (Atılgan, Kan ve Doğan, 2011). Bu amaçla kullanılabilecek pek çok yöntem (Cohen's Kappa, Ağırlıklandırılmış Kappa, Kendall Uyuşma Katsayısı, Çok Yüzeyli Rasch Ölçme Modeli vb.) bulunmakla birlikte bu araştırma kapsamında bu yöntemlerden farklı dört yaklaşım üzerinde durulmuştur. Ele alınan bu yöntemler Pearson korelasyon katsayısı, ortalamaların karşılaştırılması, uyuşma yüzdesi ve genellenebilirlik kuramıdır.

Pearson Korelasyon Katsayısı: Puanlayıcılar arası güvenirliğin hesaplanmasında sıklıkla kullanılan bir yöntem olan Pearson korelasyon katsayısı, iki puanlayıcının yaptıkları puanlamanın tutarlılığı olarak tanımlanır. Bir diğer ifadeyle, bu katsayı, iki puanlayıcının puanlarının doğrusal ilişkisini; yani, birlikte değişimini gösterir (Balcı, 2000; Baykul, 2000). Pearson korelasyon katsayısı; veriler en az eşit aralıklı ölçek düzeyindeyse, verilerin büyüklük sırası önemli değilse ve veri sayısı 30'un üzerinde ise ilişkiyi belirlemek amacıyla kullanılır. Ancak bu katsayı, puanlayıcıların yüzde kaç oranında uyuştuklarıyla ilgili bir bilgi vermez. Puanlayıcılar arasındaki varyansı dikkate almadığından puanlayıcılar arasındaki değişkenliğe duyarsızdır (Şencan, 2005).

Ortalamaların karşılaştırılması: Puanlayıcıların puanları arasındaki korelasyon değeri ortalamadan bağımsızdır (Goodwin ve Goodwin, 1991). Bu durumu şu şekilde açıklayabiliriz: İki puanlayıcının puanları mükemmel ya da mükemmele yakın bir ilişki gösterebilir. Ancak bu durum her iki puanlayıcının da puan değerlerinin birbirine eşit ya da çok yakın değerler olduğunu belirtmez. Bu sebeple korelasyon değeri tek başına puanlayıcıların puanları arasındaki uyumun göstergesi olamaz. Aynı zamanda puanlayıcıların puan ortalamaları arasındaki farkın da test edilmesi gerekmektedir. Bu sebeple bu araştırma kapsamında ortalamaların karşılaştırılması yoluna da gidilmiştir.

Uyuşma Yüzdesi (Uyuşma İndeksi): Puanlayıcıların uyuştukları madde sayısının toplam değerlendirme veya gözlem sayısına olan oranı olarak tanımlanan uyuşma yüzdesi, iki puanlayıcının verdikleri puanların ya da aynı puanlayıcının aynı davranışı iki kez puanlaması durumunda verdiği puanların uyumunun basit yüzdesi olarak kullanılır

(Deliceoğlu, 2009; Meyer, 1999). Eğer elde edilen veriler (puanlar) sınıflama ya da sıralama ölçeği düzeyinde ise Pearson korelasyon katsayısının hesaplanması mümkün olmamaktadır. Ancak uyuşma yüzdesi, sınıflama, sıralama, eşit aralıklı ve eşit oranlı ölçek değerindeki tüm veriler üzerinde kullanılabilmektedir. Hesaplanmasının ve elde edilen sonuçların anlaşılmasının kolaylığı, bu yöntemin avantajları olarak değerlendirilebilir. Ancak puanlayıcıların yaptıkları puanlamalar arasındaki tesadüfi uyuşmaları göz önünde bulundurmaması bir sınırlılık olarak karşımıza çıkmaktadır (Goodwin, 2001; Hughes ve Garrett, 1990). Puanlayıcılar arası değerlendirme sonuçlarının güvenilir sayılabilmesi için uyuşma yüzdesinin %75'in üzerinde olması gerekmektedir. Bu ölçütün altında bir oran, puanlayıcıların puanlamalarının farklılaştığı şeklinde yorumlanır. Daha düşük bir oran, değerlendirmede, puanlayıcıların önemli ölçüde farklı düşündükleri anlamına gelir (Şencan, 2005).

Genellenebilirlik (G) kuramı: Puanlayıcılar arası güvenirliği belirlemek için yukarıda kullanılan yöntemlerin yanında, değişkenlik kaynağı olarak puanlayıcıların da ele alındığı (puanlayıcılardan gelen potansiyel hataların da göz önünde bulundurulduğu) bir diğer yol ise genellenebilirlik (G) kuramıdır (Cronbach, Gleser, Nanda ve Rajaratnam, 1972). Bu kuramın en önemli avantajı, tek bir hata kaynağını ele almak yerine tüm hata kaynaklarının hem ayrı ayrı hem de birbirleriyle etkileşimlerinden kaynaklanan hataları göz önünde tutan bir güvenirlik (G) katsayısının elde edilmesidir (Brennan, 2001; Brennan, 1992; Shavelson ve Webb, 1991; Yin ve Shavelson, 2004). Bu avantajından dolayı bu araştırma kapsamında puanlayıcılar arası güvenirliğin belirlenmesinde G kuramı da ele alınmıştır.

Alanyazın tarandığında puanlayıcı güvenirliği üzerine yapılmış pek çok çalışmaya rastlanmaktadır. Kan (2001); aynı ve farklı öğretmenlerin, aynı ve farklı zamanlarda yapmış oldukları puanlamalar arasındaki ilişkinin belirlenmesinde Pearson Momentler Çarpımı Korelasyon katsayısından yararlanarak elde edilen korelasyonları t-testi ile test etmiştir. Farklı öğretmenlerin aynı zamanlarda puanlama ve cevap anahtarı kullanmadan ve kullanarak verdiği yazılı yoklama puanlarının karşılaştırılması sonucunda, puanlamalar arasında pozitif yönde yüksek ve anlamlı bir ilişki tespit edilmiş, ancak puanlar arasında anlamlı bir farklılık olduğu gözlenmiştir. Ayrıca farklı öğretmenlerin farklı zamanlarda puanlama cetveli ve cevap anahtarı kullanarak ve kullanmadan verdiği yazılı yoklama puanlarının karşılaştırılması sonucunda, puanların ortalamaları arasında anlamlı bir fark olduğu ortaya çıkmıştır.

Goodwin (2001), KTK yaklaşımını kullanarak puanlayıcılar arası uyum ve güvenirliği farklı yöntemler kullanarak araştırmıştır. Öncelikle, Pearson korelasyon katsayısı kullanılmış ve 0,90 güvenirlik katsayısı elde edilmiştir. Basit yüzde ile hesaplanan puanlayıcıların farklı zamanlarda verdikleri puanların uyumu %10-70 aralığında bulunurken puanlayıcıların uyumunun %60-100 aralığında oluğu görülmüştür. Çalışma kapsamında kullanılan bir diğer yöntem olan Kappa istatistiği sonucu ise 0 ile 0,63 arasında hesaplanmıştır. Son olarak, G kuramı ile hesaplanan G katsayısı 0,88 bulunmuştur. Çalışma kapsamında kullanılan diğer yöntemlerle kıyaslandığında, aynı anda birçok hata kaynağını ele alan G kuramının bu tür araştırmalarda daha avantajlı bir yöntem olduğu üzerinde durulmuştur.

Goodwin ve Goodwin (1991) okul öncesinde özel eğitim alanında yürüttükleri çalışmada, iki puanlayıcının on öğrencinin altı maddeye verdikleri cevaplar üzerinde yaptıkları puanlamalara ilişkin hipotetik bir veri seti üzerinde, dört farklı yöntemi kullanarak puanlayıcılar arası güvenirlik sonuçlarını karşılaştırmışlardır. Çalışma kapsamında ele alınan yöntemler, korelasyon, ortalamaların karşılaştırılması, uyuşma yüzdesi ve genellenebilirlik kuramıdır. Elde edilen en yüksek kestirim 0,90 olan korelasyon katsayısı iken en düşük

kestirim, uyuşma yüzdesinde (% 26,67) bulunmuştur. Çalışmada ele alınan diğer yöntemlere kıyasla daha karmaşık olmasına rağmen, genellenebilirlik kuramının, okul öncesinde özel eğitimde karşılaşılabilecek pek çok durumda güvenirlik anlamında çok yararlı bilgiler vereceği üzerinde durulmuştur.

Goodwin, Sands ve Kozleski (1991) özel eğitim alanında yürüttükleri çalışmada, röportaj yapan bireyler arasındaki güvenirliği belirlemede basit yüzde uyumu, Kappa, ağırlıklandırılmış Kappa, Pearson korelasyon katsayısı, t testi ve genellenebilirlik kuramını kullanmışlardır. Çalışmada, ele alınan yaklaşımların benzerlik ve farklılıkları tartışılmış ve belirli durumlar için hangi yaklaşımın kullanılabileceğine ilişkin öneriler sunulmuştur.

Bu çalışma kapsamında ele alınan yaklaşımlarla ilgili yapılan çalışmalar gözden geçirildiğinde uyuşma yüzdesinin geçmişte diğer yöntemlere kıyasla daha sık kullanıldığı Hughes ve Garret'in (1990) yürüttüğü bir çalışmada vurgulanmıştır. Çalışma kapsamında elde edilen sonuçlara göre, Hughes ve Garrett'in gözden geçirdiği 86 makalenin 56'sında (% 65,1) uyuşma yüzdesi puanlayıcılar arası güvenirliğin belirlenmesinde kullanılırken sadece altı tanesinde korelasyon kullanılmıştır. Buna karşın korelasyon (Atmaz, 2009; Deliceoğlu, 2009; Goodwin, 2001; Goodwin ve diğerleri, 1991; Goodwin ve Goodwin, 1991; Kan, 2001), ortalamalar arası fark (Goodwin ve diğerleri, 1991; Goodwin ve Goodwin, 1991) ve özellikle daha önce de üzerinde durulan avantajlarından dolayı genellenebilirlik kuramının kullanıldığı çalışmalara (Atılgan, 2005; Atılgan, 2008; Çakıcı Eser ve Gelbal, 2012; Deliceoğlu, 2009; Gage, Prykanowski ve Hirn, 2014; Goodwin, 2001; Goodwin ve diğerleri, 1991; Goodwin ve Goodwin, 1991; Gugiu, Gugiu ve Baldus, 2012; Güler ve Gelbal, 2010; Hill, Charalambous ve Kraft, 2012; Martinez, Goldschmidt, Niemi, Baker ve Sylvester, 2007; Taşdelen, Kelecioğlu ve Güler, 2010) son yıllarda sıklıkla rastlanmaktadır.

Araştırmanın Amacı

Bu araştırmanın amacı korelasyon, ortalamaların karşılaştırılması, uyuşma yüzdesi ve genellenebilirlik kuramı ile puanlayıcılar arası güvenirliğin kestirilmesi ve bu yaklaşımlardan elde edilen değerlerin incelenmesidir.

YÖNTEM

Araştırmanın Modeli

Araştırma, açık uçlu soruların puanlanmasında, puanlayıcılar arası güvenirliğin hesaplanmasında çeşitli yöntemlerin avantaj ve dezavantajlarının ortaya konmasına, yöntemlerden hangisinin daha çok bilgi verdiğinin saptanmasına ve yöntemlere dayalı olarak elde edilen güvenirlik katsayılarının incelenmesine dayalıdır. Bu yönüyle kuramsal bir araştırmadır.

Çalışma Grubu

Bu araştırmanın çalışma grubunu, 2011-2012 öğretim döneminde Sakarya Üniversitesi Eğitim Fakültesi İlköğretim Bölümü Okul Öncesi Eğitimi bölümü ikinci sınıfta öğrenim gören ve İstatistik I dersini alan 43 öğrenci oluşturmaktadır.

Verilerin Toplanması

Araştırma kapsamında görev yapan iki puanlayıcıdan, İstatistik I dersi vize sınavı kapsamında hazırlanan her biri on puan değerinde olan on açık uçlu maddeye, çalışma grubunda yer alan 43 öğrencinin verdikleri cevapları puanlamaları istenmiştir. Bu amaçla maddelerin puanlamasında kullanılacak bütüncül bir dereceli puanlama anahtarı araştırmacılar tarafından oluşturulmuş ve puanlayıcılar bu aracı kullanarak birbirlerinden

bağımsız bir şekilde her bir öğrencinin on maddeye verdikleri cevapları puanlamışlardır. Bu şekilde iki puanlayıcının on madde için 43 öğrenciye verdikleri puanlardan oluşan veri seti elde edilmiştir.

Verilerin Analizi

Her bir öğrencinin on maddeye verdikleri cevapları puanlayan puanlayıcılar arasındaki puanlama güvenirliğini belirlemek amacıyla korelasyon, ortalamaların karşılaştırılması, uyuşma yüzdesi ve genellenebilirlik kuramından yararlanılmıştır. Bu yöntemlerden korelasyon hesaplaması ve ortalamaların karşılaştırılması yöntemleri için SPSS 18, uyuşma yüzdesinin hesaplanmasında Excel programı, genellenebilirlik kuramı hesaplamalarında ise EduG 6.1 programları kullanılmıştır.

Puanlayıcılar arası güvenirliğin hesaplanmasında sıklıkla kullanılan bir yöntem, puanlayıcılar arası korelasyon değerinin hesaplanmasıdır. Araştırma kapsamında ele alınan veriler sürekli olduğu için Pearson korelasyon katsayısının hesaplanması uygun görülmüştür. Araştırma kapsamında puanlayıcı olarak görev yapan puanlayıcıların puanladıkları on madde için Pearson korelasyon katsayısı hem ayrı ayrı hem de toplam puan için hesaplanmıştır. Ayrıca Fisher'in Z dönüşümü kullanılarak da on madde için ortalama korelasyon hesaplanmıştır (Glass ve Hopkins, 1984).

Çalışma kapsamında ele alınan bir diğer yaklaşım olan puanlayıcıların yaptıkları puanlamalar arasındaki farkların karşılaştırılması amacıyla hem her bir madde için ayrı ayrı hem de tüm maddelerin toplam puanlarının ortalamaları arasındaki farklara ilişkin t testi yapılmıştır.

Uyuşma yüzdesi, iki farklı ölçüte göre hesaplanmıştır. Öncelikle "tam uyuşma yüzdesi" her iki puanlayıcının da her bir maddeye aynı puanı verme yüzdeleri bazında hesaplanmıştır. Ardından daha esnek bir ölçüt olarak puanlayıcıların her bir madde için ±2 puan farklılığına kadar aynı puanı verme yüzdeleri elde edilmiştir. Bu iki ölçüt hem maddeler bazında ayrı ayrı hem de toplam test puanı üzerinden kestirilmiştir.

Çalışmada son olarak puanlayıcılar arası güvenirliği belirlemek için genellenebilirlik kuramı analizleri yürütülmüştür. Bu bağlamda çalışmanın ölçme objesi öğrenciler (*ö*) ve yüzeyler maddeler (*m*) ve puanlayıcılar (*p*) olarak ele alınarak tümüyle çaprazlanmış tesadüfî desen (*öxmxp*) üzerinden analizler yürütülmüştür.

BULGULAR

Çalışma kapsamında ele alınan dört yaklaşıma ilişkin yürütülen analizler sonucu elde edilen bulgular sırayla aşağıda verilmektedir.

Pearson Korelasyon Katsayısı

İki puanlayıcının on maddeye ilişkin öğrenci cevaplarına verdikleri puanlar arasında hesaplanan Pearson korelasyon katsayısı sonuçları Tablo 1'de verilmiştir.

Tablo 1 incelendiğinde, iki puanlayıcının puanları arasındaki en yüksek korelasyon değerinin 0,80 ile 2. maddeye, en düşük korelasyon değerinin ise 0,32 ile 6. maddeye ilişkin olduğu görülmektedir. Korelasyon değerleri için, 0,30'dan küçük ise ilişkinin düşük, 0,30 ile 0,70 arasında ise orta ve 0,70'den büyük olduğunda ise ilişkinin yüksek olduğu (Büyüköztürk, Çokluk Bökeoğlu ve Köklü, 2009) göz önünde bulundurulduğunda; 1., 2. ve 3. maddelerdeki ilişkinin yüksek, diğer tüm maddelerdeki ilişkinin orta düzeyde olduğu söylenebilir. Puanlayıcıların maddelere verdiği puanların ortalamaları arasındaki ilişkiye bakıldığında ise 0,74 ile yüksek bir ilişki değeri gözlenmektedir. Bu değer, her bir maddeye ilişkin korelasyon değerleri arasındaki farklılığı içermediğinden, maddelere ilişkin korelasyon değerlerinin ortalamasından (0,69) daha yüksek bir değere sahip olmaktadır.

Tablo 1. *Pearson Korelasyon Katsayısı Aracılığıyla Hesaplanan Puanlayıcılar Arası Güvenirlikler*

Madde	Puanlayıcılar arası korelasyon	Fisher's Z	Ortalama[a]	Tek bir puanlayıcı için güvenirlik kestirimi
1	0,73*	0,92	0,69	0,53*
2	0,80*	1,09		
3	0,76*	0,99		
4	0,43*	0,47		
5	0,54*	0,60		
6	0,32*	0,33		
7	0,65*	0,77		
8	0,62*	0,72		
9	0,49*	0,53		
10	0,44*	0,48		
Testin tamamı	0,74*			0,59*

[a] On madde için hesaplanan puanlayıcılar arası ortalama güvenirlik değeri için Fisher'in Z dönüşümü kullanılmıştır.
* p<0,01 düzeyinde anlamlı

Tablo 1'de ayrıca Spearman Brown formülü kullanılarak, yalnızca bir puanlayıcı olduğunda hem ortalama korelasyon değeri hem de toplam puanlar arası korelasyon değeri kullanılarak güvenirlik kestirimi, Eşitlik 1 ve Eşitlik 2'de verilen formüller yardımıyla hesaplanarak elde edilmiştir. Açıktır ki, puanlayıcı sayısı azaldığında, toplam varyans içindeki puanlayıcı varyansı azaldığından elde edilecek güvenirlik değeri de azalmaktadır.

$$r'_{xx} = \frac{nr_{xx}}{1+(n-1)r_{xx}} = \frac{(0,5)(0,69)}{1+(0,5-1)0,69} = \frac{0,345}{1-0,345} = 0,53 \qquad \text{Eşitlik (1)}$$

$$r'_{xx} = \frac{nr_{xx}}{1+(n-1)r_{xx}} = \frac{(0,5)(0,74)}{1+(0,5-1)0,74} = \frac{0,37}{1-0,37} = 0,59 \qquad \text{Eşitlik (2)}$$

Eşitlik 2 ve 3'te kullanılan sembollere ilişkin açıklamalar aşağıdaki gibidir:

r_{xx}: Kestirilen orijinal güvenirlik değeri

r'_{xx}: Puanlayıcı sayısının değişmesi sonucu kestirilecek güvenirlik değeri

n: puanlayıcı sayısının artırılma/azaltılma oranı

Ortalamaların Karşılaştırılması

Puanlayıcılar arası güvenirliğin incelenmesinde kullanılabilecek bir diğer yaklaşım ortalamaların karşılaştırılmasıdır. Tablo 2'de puanlayıcıların puanları arasındaki farkların karşılaştırılmasına ilişkin elde edilen sonuçlar verilmiştir. Hem her bir madde için ayrı ayrı hem de tüm maddelerin toplam puan ortalamaları arasındaki farklara ilişkin t testi sonuçları ve anlamlılık değerleri elde edilmiştir.

Tablo 2. *Puanlayıcıların Ortalama Puanlarının Karşılaştırılması*

Madde	1.Puanlayıcının Ortalaması ($\bar{P}_1$)	2.Puanlayıcının Ortalaması ($\bar{P}_2$)	Fark ($\bar{P}_1 - \bar{P}_2$)	t Değeri	p
1	8,93	8,04	0,89	2,476	0,017*
2	8,69	8,58	0,11	0,413	0,684
3	8,20	7,91	0,29	0,818	0,417
4	6,98	7,62	-0,64	-1,182	0,242
5	7,47	8,73	-1,25	-2,936	0,005*
6	7,15	7,44	-0,29	-0,477	0,635
7	7,96	8,20	-0,24	-0,592	0,557
8	7,87	7,55	0,33	0,706	0,483
9	6,76	7,64	-0,87	-1,496	0,140
10	6,05	4,73	1,33	2,135	0,037*
Toplam	76,06	76,44	-0,38	-0,149	0,882

*p<0,05

Tablo 2'de görüldüğü üzere 1., 5. ve 10. maddelere verilen puanların ortalamaları anlamlı farklılık göstermektedir (p<0,05). Fakat Tablo 1'e bakıldığında 1. maddeye verilen puanlar arasındaki ilişki 0,73 korelasyon katsayısı ile yüksek diğer iki madde puanları arasındaki ilişki ise sırasıyla 0,54 ve 0,44 korelasyon katsayıları ile orta düzeyde bulunmuştur. Toplam puana göre bakıldığında ise birinci puanlayıcının ortalama puanı, ikinci puanlayıcıdan 0,38 puan daha düşük çıkmıştır. Ancak bu değer istatistiksel olarak anlamlı değildir (p>0,05). Bir başka deyişle, toplam puan ortalamaları arasında anlamlı bir fark yoktur. Aynı zamanda, Tablo 1'de gözlenen toplam puanlar arasındaki korelasyon değeri (0,74) ile yüksek bir ilişki göstermektedir.

Uyuşma Yüzdesi

Tablo 3'te çalışma kapsamında ele alınan iki farklı yaklaşım olan hem tam uyuşma yüzdeleri hem de ±2 puan aralığı uyuşma yüzdeleri değerleri verilmiştir.

Tablo 3. *Uyuşma Yüzdesi Yaklaşımı ile Puanlayıcılar Arası Güvenirlik Değerleri*

Madde	Tam Uyuşma Yüzdesi	±2 Puan Aralığı Uyuşma Yüzdesi
1	70,9	81,8
2	78,2	83,6
3	76,4	83,6
4	45,5	61,8
5	40,0	45,5
6	47,3	54,5
7	65,5	78,2
8	61,8	72,7
9	61,8	72,7
10	41,8	49,1
Ortalama	58,9	68,4
Standart Sapma	13,6	13,8

Tablo 3'te "tam uyuşma yüzdesi" sütununda yer alan değerler her iki puanlayıcının da her bir maddeye aynı puanı verme yüzdelerini göstermektedir. Bir başka ifadeyle örneğin; 1. maddeyi cevaplayan öğrencilerin %70,9'una her iki puanlayıcı da aynı puanı vermiştir. Tablo 3'ün son sütununda yer alan değerler ise biraz daha esnek olup hem her iki puanlayıcının aynı puanı hem de ±2 puan farklılığa kadar aynı puanı verme yüzdelerini göstermektedir. Tablo 3'te görüleceği üzere tam uyuşma yüzdesi ortalaması 58,9 iken daha esnek olarak hesaplanan uyuşma yüzdesi ortalaması 68,4 olmaktadır. Tablo 3'te ayrıca toplam puanlar üzerinden uyuşma yüzdesi verilmemiştir. Özellikle puan aralığının geniş olduğu durumlarda (Çalışmada puan aralığı madde bazında 0-10 iken test bazında 0-100'dür.) toplam puanlar için uyuşma yüzdeleri pek fazla kullanılmamaktadır (Goodwin ve Goodwin, 1991).

Genellenebilirlik Kuramı

Bu çalışma kapsamında ele alınan diğer yaklaşımlarda göz önünde bulundurulan hata, sadece puanlayıcıların puanları arasındaki farklılaşmadan kaynaklanmaktadır. Ancak G kuramı ile madde (*m*), puanlayıcı (*p*), öğrenci-madde etkileşimi (*öxm*), öğrenci-puanlayıcı etkileşimi (*öxp*), madde-puanlayıcı etkileşimi (*mxp*) ve son olarak öğrenci-madde-puanlayıcı etkileşimi (*öxmxp*) hata kaynakları olarak değerlendirilmektedir. Bu avantajından dolayı bu araştırma kapsamında puanlayıcılar arası güvenirliğin belirlenmesinde G kuramı da ele alınmış ve yürütülen analizler sonucu elde edilen varyans değerleri Tablo 4'te verilmiştir.

Tablo 4. *Varyans Analizi Sonuçları ile Öğrenci, Madde ve Puanlayıcı için Kestirilen Varyans Bileşenleri*

Varyans Kaynağı	sd	KO	Kestirilen Varyans Bileşeni	Kestirilen Toplam Varyans Yüzdesi
Öğrenci (*ö*)	42	112,07	5,14	34,2
Madde (*m*)	9	49,47	0,27	1,80
Puanlayıcı (*p*)	1	97,79	0,18	1,20
öxm	378	11,55	2,46	16,4
öxp	42	4,44	-0,22	0,00
mxp	9	21,12	0,34	2,20
öxmxp,e	378	6,63	6,63	44,1

Tablo 4'te her bir değişkenlik kaynağı için varyans analizi tablosunda olduğu gibi serbestlik dereceleri (sd) ve kareler ortalaması (KO) farklı olarak da kestirilen varyans bileşenleri ve toplam varyanstaki yüzdeler yer almaktadır. ANOVA tablosundan farklı olarak kestirilen varyans bileşenlerinin nasıl hesaplandığı öğrenci varyans kaynağı üzerinden bir örnekle aşağıda verilen Eşitlik 3 ile açıklanmaya çalışılmıştır. Diğer varyans bileşenlerinin hesaplanması ile ilgili daha detaylı bilgiye Brennan (2001) kaynağından ulaşılabilir.

$$\hat{\sigma}^2(\ddot{o}) = \frac{KO_{\ddot{o}} - KO_{\ddot{o}m} - KO_{\ddot{o}p} + KO_{\ddot{o}mp}}{n_m n_p} = \frac{112.07 - 11.55 - 4.44 + 6.63}{(10)(2)} = \frac{102.71}{20} = 5.14 \qquad \text{Eşitlik (3)}$$

Tablo 4'te görüleceği üzere kestirilen varyans bileşenlerinin ilk üçü; öğrenci, madde ve puanlayıcı ana etkileri için elde edilen değerlerdir. Öğrenciye ilişkin değişkenlik toplam varyansın %34,2'sini açıklayarak ana etkiler içinde en yüksek değere sahiptir. Bu istenilen

bir durumdur; öğrenciler arası farklılığın olabildiğince yüksek çıkması beklenir. Maddelere ilişkin varyans, toplam varyansın %1,8 gibi oldukça küçük bir kısmını açıklamaktadır. Buradan maddelerin güçlük düzeylerinin birbirlerinden çok farklı olmadığı sonucuna varılabilir. Puanlayıcılar arası değişkenlik istenmeyen bir durum olup varyansın olabildiğince sıfıra yakın çıkması istenir. Tablo 4'te görülüğü üzere puanlayıcı ana etkisine ilişkin varyans toplam varyansın çok küçük bir kısmını açıklamıştır (%1,2). Buradan puanlayıcıların puanları arasındaki değişkenliğin sıfıra yakın bir değere sahip olduğu söylenebilir.

Tablo 4'te ayrıca öğrenci, madde ve puanlayıcı etkileşimlerine ilişkin varyans değerleri de yer almaktadır. Öğrenci-madde (*öxm*) etkileşimine ilişkin varyansın toplam varyanstaki yüzdesi 16,4'tür. Buradan maddelerin güçlük düzeyinin öğrencilere göre değişkenlik gösterdiği söylenebilir. Bu durum özellikle öğrencilerin geçmiş öğrenme yaşantılarının etkili olduğu matematik, istatistik gibi derslerde olasıdır. Bir diğer bileşen olan öğrenci-puanlayıcı (*öxp*) etkileşimine ilişkin varyansın sıfır olduğu görülmektedir. Bu durum öğrencilere verilen puanların puanlayıcıdan puanlayıcıya değişmediğinin bir göstergesidir. Madde-puanlayıcı (*mxp*) etkileşimine ilişkin varyans ise toplam varyansın %2,2'sine karşılık gelmiştir. Bu etkileşim değeri maddelere verilen puanların puanlayıcıdan puanlayıcıya ne ölçüde değiştiğinin bir göstergesidir. Tablo 4'te son olarak öğrenci-madde-puanlayıcı (*öxmxp*,e) ortak etkisi bir başka deyişle "artık (residual)" ya da "hata" varyansı yer almaktadır. Bu değerin olabildiğince sıfıra yakın çıkması istenir. Ancak çalışmanın sonucunda %44,1 olarak bu varyans en yüksek değere sahiptir. Bu durumun pek çok açıklaması olabilir. Örneğin, öğrenci-madde-puanlayıcı arasında sistematik olmayan bir değişim vardır. Bir başka değişle puanlayıcılar sistematik olmayan bir şekilde öğrenciler boyunca ve maddeler boyunca farklı puanlar vermişlerdir. Başka bir durum da ölçmeye bilinmeyen ve sistematik olarak puanlamayı etkilemeyen faktörlerin karışmış olabileceğidir. Bir diğer durum ise yukarıda açıklanan her iki durumunda aynı anda etki göstermiş olabileceğidir. Bu durum yapılan çalışmada istenmeyen durumların ortaya çıktığının sinyalini vermektedir.

Genellenebilirlik kuramı Tablo 4'te verilen değerlerin yanı sıra genellenebilirlik (G) katsayısı olarak ifade edilen bir güvenirlik katsayısı hesaplanmasına da imkân tanır. Bu katsayı 0,0 ile 1,0 arasında değerler almakla birlikte puanların güvenirliğinin veya genellenebilirliğinin düzeyini belirtir (Shavelson ve Webb, 1991). Eşitlik 4'te nasıl hesaplanacağı açıklanmaya çalışılmış ve araştırma kapsamında ele alınan 43 öğrenci, 10 madde ve iki puanlayıcı için hesaplanan G katsayısı 0,90 gibi oldukça yüksek bir değer olarak bulunmuştur.

$$G = \frac{\hat{\sigma}_o^2}{\hat{\sigma}_o^2 + \frac{\hat{\sigma}_{om}^2}{n_m} + \frac{\hat{\sigma}_{op}^2}{n_p} + \frac{\hat{\sigma}_{omp}^2}{n_m n_p}} = \frac{5.14}{5.14 + \frac{2.46}{10} + \frac{0.00}{2} + \frac{6.63}{20}} = \frac{5.14}{5.72} = 0.90 \qquad \text{Eşitlik (4)}$$

SONUÇLAR ve TARTIŞMA

Araştırma kapsamında puanlayıcılar arası güvenirliğin belirlenmesinde kullanılan dört yöntem olan korelasyon, ortalamaların karşılaştırılması, uyuşma yüzdesi ve genellenebilirlik kuramı sonuçları verilmiştir. Elde edilen sonuçlar ve yapılan yorumlar bir miktar farklılaşmakla birlikte benzerlikler de söz konusudur.

Puanlayıcılar arasında hesaplanan basit korelasyonda yüksek düzeyde ve pozitif yönlü (0,74) bir ilişki ortaya çıkmıştır. Bu bulgu, korelasyon ile puanlayıcılar arası uyumun incelendiği alanyazında yapılan bir takım çalışmalar ile benzerlikler göstermektedir

(Goodwin, 2001; Goodwin ve Goodwin, 1991; Goodwin ve diğerleri, 1991). Ancak bu katsayı ortalamadan bağımsız olduğu için, güvenirliğin hesaplanmasında iki puanlayıcıdan elde edilen puanlar arasındaki benzerlik ve farklılıkları göstermez. Bir diğer ifadeyle, birbirlerine göre puanlamada katılık ya da cömertlik açısından farklılık bulunan puanlayıcıların verdikleri puanlar birlikte değişiyorsa, verilen ortalamalar arası fark büyük olmasına rağmen korelasyon katsayısı yüksek çıkacaktır. Bu sebeple korelasyon katsayısının hesaplanması, puanlayıcılar arası uyumun belirlenmesinde yetersiz kalabilir (Goodwin, 2001).

Puanlayıcıların ortalamaları karşılaştırılıp maddeler ayrı ayrı incelendiğinde, üç maddede anlamlı bir farklılık çıkmakla beraber genel ortalamada anlamlı bir fark bulunmamıştır.

Çalışma kapsamına ele alınan bir diğer yaklaşım olan uyum yüzdesi sonuçları incelendiğinde ise tam uyumda oran % 58,9 iken ±2 puan aralığında ise % 68,4 gibi bir oran çıkmıştır. Bu yöntem hesaplanması kolay ve sonuçların yorumlanması görece basit bir yöntem olmasına rağmen, şansla ya da tesadüfen ortaya çıkan uyumun hesaplanamaması en büyük sınırlılık olarak karşımıza çıkmaktadır. Puanlayıcılar arası şansa bağlı ortaya çıkabilecek yapay uyumun önüne geçmek için ise Cohen's Kappa istatistiği (Cohen, 1960) ile güvenirlik katsayısı hesaplanması önerilebilir.

Son olarak işe koşulan G kuramı sonuçlarında elde edilen güvenirlik katsayısı araştırma kapsamındaki en yüksek değerdir (0,90). Puanlayıcılar arası güvenirliğin araştırıldığı bu çalışmada, Tablo 4'te verilen varyans değerleri göz önünde bulunduğunda, puanlayıcılar arasında ciddi bir farklılığın gözlenmemesi de elde edilen diğer bulguları destekler niteliktedir. G kuramının farklı hata kaynaklarını aynı anda ele alarak bir güvenirlik hesaplaması diğer yöntemlere göre üstün yanıdır. Ancak, özellikle çalışma kapsamında ele alınan diğer yaklaşımlarla kıyaslandığında en karmaşık olan yaklaşımın G kuramı olması, bu kuramın bir sınırlılığı olarak karşımıza çıkmaktadır.

Puanlayıcılar arası güvenirliğin kestiriminde pek çok yöntem kullanılabileceği için, en uygun yaklaşımı kestirmek zor olabilir. Bu kararı verirken, araştırmacının bir dizi faktörü göz önünde bulundurması gerekmektedir. Bu faktörlerin ilki, kullanılan ölçme aracının ölçek düzeyi olabilir. Eğer eldeki veri seti, eşit aralıklı ya da eşit oranlı ölçek düzeyinde ise, bu çalışma kapsamında da ele alınan yaklaşımların yanında Krippendorf alfa istatistiği (Krippendorf, 2004) gibi daha birçok yaklaşım kullanılabilir. Eğer eldeki veri seti, sıralama ölçeği düzeyinde ise Pearson korelasyon katsayısı ve genellenebilirlik kuramının kullanılması uygun olmamakla birlikte uyuşma yüzdesi ve Cohen's Kappa istatistiği (Cohen, 1960; Fleiss ve Cohen, 1973; Fleiss, 1971) bu tür veri setlerinde rahatlıkla kullanılabilir. Göz önünde bulundurulması gereken bir diğer nokta ise puanlayıcı sayısı olabilir. Kimi yöntemler yalnızca iki puanlayıcı arasındaki uyumu belirlemede kullanılırken (Cohen's Kappa) kimi yaklaşımlar üç ve daha fazla puanlayıcı arasındaki uyumun belirlenmesinde (Fleiss Kappa) kullanılabilir. Buna bağlı olarak kullanılabilecek yaklaşımlara karar verilebilir. Ele alınabilecek bir diğer faktör ise eldeki veriyi kullanarak nasıl bir karar verileceği ile ilgili olabilir. Örneğin, yerleştirme gibi birey bazında verilecek kararlar alınacaksa, puanlayıcıların verdikleri puanların birbirine yakın olması istenebilir ve bu durumda uyuşma yüzdesi kullanılabilir. Eğer puanlayıcıların verdikleri puanların düzeyleri önemsiz ancak birlikte değişimleri önemli ise korelasyon katsayısının hesaplanması düşünülebilir.

Puanlayıcı güvenirliğinin kestirimi sırasında birden fazla yaklaşımın bir arada kullanılması, gerçek duruma ilişkin daha net bilgi vermesi açısından bir öneri olarak sunulabilir. Bunun sebebi ise her yaklaşımın bir anlamda kendine özgü bilgi sağlamasıdır.

Bir diğer ifadeyle, birden fazla yaklaşımın kullanılması elimizdeki resme farklı açılardan bakmamıza olanak sağlayıp, durumu daha iyi değerlendirmemiz konusunda yardımcı olacaktır.

KAYNAKLAR

Aiken, L. R. (2000). *Psychological Testing and Assessment*. Boston: Allyn and Bacon.

American Educational Research Association (AERA), American Psychological Association (APA) ve National Council on Measurement in Education (NCME). (2004). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Anastasi, A. ve Urbina, S. (1997*). Psychological Testing*. Upper Saddle River, N.J.: Prentice Hall.

Atılgan, H., Kan, A. ve Doğan, N. (2011). *Eğitimde Ölçme ve Değerlendirme*. 5. Baskı, Ankara: Anı Yayıncılık.

Atılgan, H. (2008). Using Generalizability theory to assess the score reliability of the Special Ability Selection Examinations for music education programs in higher education. *International Journal of Research & Method in Education,* Volume 31, Issue 1.

Atılgan, H. (2005). Genellenebilirlik Kuramı ve Puanlayıcılar Arası Güvenirlik için Örnek Bir Uygulama. *Eğitim Bilimleri ve Uygulama*, 4(7), 95-108.

Atmaz, G. (2009). *Puanlama Yönergesi (Rubrik) Kullanılması Durumunda Puanlayıcı Güvenirliğinin İncelenmesi*. Yayınlanmamış yüksek lisans tezi, Mersin Üniversitesi, Mersin.

Balcı, A. (2001) *Sosyal Bilimlerde Araştırma: Yöntem, Yeti ve İlkeler,* Ankara: Pegem Yayıncılık.

Baykul, Y. (2000) *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması*. Ankara: ÖSYM Yayınları.

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

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA. American College Testing.

Büyüköztürk, Ş., Çokluk Bökeoğlu, Ö. ve Köklü, N. (2009). *Sosyal Bilimler için İstatistik*. Ankara: Pegem Akademi.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20, 37-46.

Crocker, L. M. ve Algina, L. (1986). *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart and Winson.

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. ve Gelbal, S. (2012). Genellenebilirlik Kuramı ve Lojistik Regresyona Dayalı Hesaplanan Puanlayıcılar Arası Tutarlığın Karşılaştırılması. *Kastamonu Eğitim Dergisi*. 21 (2),423-438.

Deliceoğlu, G. (2009). *Futbol Yetilerine İlişkin Dereceleme Ölçeğinin Genellenebilirlik ve Klasik Test Kuramına Dayalı Güvenirliklerinin Karşılaştırılması.* Yayınlanmamış Doktora Tezi, Ankara Üniversitesi, Ankara.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*. 76(5), 378-382.

Fleiss, J. L. ve Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33, 613-619.

Gage, N. A., Prykanowski, D. ve Hirn, R. (2014). Increasing Reliability of Direct Observation Measurement Approaches in Emotional and/or Behavioral Disorders Research Using Generalizability Theory. *Behavioral Disorders*, 39(4), 228-244.

Glass, G. V. ve Hopkins, K. D. (1984). Statistical Methods in Education and Psychology. NJ: Prentice-Hall.

Goodwin, L. D. ve Goodwin, W. L. (1991). Using Generalizability Theory in Early Childhood Special Education. *Journal of Early Intervention*, 193-204.

Goodwin, L. D., Sands, D. J. ve Kozleski, E. B. (1991). Estimating Interviewer Reliability for Interview Schedules Used in Special Education Research. *The Journal of Special Education*, Volume 25, Issue1, 73-89.

Goodwin, L. D. (2001). Interrater Agreement and Reliability. *Measurement in Physical education and Exercise Science*, 5 (1), 13-14.

Gugiu,M. R., Gugiu, P. C. ve Baldus, R. (2012). Utilizing Generalizability Theory to Investigate the Reliability of Grades Assigned to Undergraduate Research Papers. *Journal of Multi-Disciplinary Evaluation*, v8 n19 p26-40.

Güler, N. ve Gelbal, S. (2010). Açık Uçlu Matematik Sorularının Güvenirliğinin Klasik Test Kuramı ve Genellenebilirlik Kuramına Göre İncelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri*, 10 (2), 989-1019.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th edn, pp. 65–110). Westport, CT: Praeger Publishers.

Hill, H. C., Charalambous, C. Y. ve Kraft, M. A. (2012). When Rater Reliability Is Not Enough: Teacher Observation and a Case for the Generalizability Study. *Educational Researcher,* Volume 41, Issue 2, 56-64.

Hughes, M. A. ve Garrett, D. E. (1990). Intercoder Reliability Estimation Approaches in Marketing: A Generalizability Theory Framework for Quantitative Data. *Journal of Marketing Research*, Volume 27, 185-195.

Kan, A. (2001). *Yazılı yoklamaların puanlanmasında puanlama cetveli ve yanıt anahtarı kullanımının puanlamaya ve puanlayıcı güvenirliğine etkisi.* Yayınlanmamış yüksek lisans tezi, Hacettepe Üniversitesi, Ankara.

Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Humanities, Social Sciences and Law*, 38(6), 787-800.

Lord, F. M. ve Novick, M. R. (1968) *Statistical Theory of Mental Test Scores*. New Jersey: Addison-Wesley. Co.

Martinez, J. F., Goldschmidt, P., Niemi, D., Baker, E. L. ve Sylvester, R. (2007). Language Arts Performance Assignments: Generalizability Studies of Local and Central Ratings. *Educatıonal Assessment,* 12(3&4), 267–282.

Meyer, G. J. (1999). Simple Procedures to Estimate Chance Agreement and Kappa for the Interrater Reliability of Response Segments Using the Rasch Comprehensive System. *Journal of Personality Assessment*, 72, 230-255.

Shavelson, R. J. ve Webb, N. M. (1991). *Generalizability Theory: A Primer*. USA: SAGE Publications.

Şencan, H. (2005) *Sosyal ve Davranışsal Ölçmelerde Güvenirlik ve Geçerlik*. Ankara: Sözkesen Matbaacılık.

Taşdelen, G., Kelecioğlu, H. ve Güler, N. (2010). Nedelsky ve Angoff Standart Belirleme Yöntemleri ile Elde Edilen Kesme Puanlarının Genellenebilirlik Kuramı ile Karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 1(1), 22-28.

Turgut, M. F. (1993). *Eğitimde Ölçme ve Değerlendirme Metotları*. Ankara: Saydam Matbaacılık.

Yin, Y. ve Shavelson, R. J. (2004). *Generalizability Analysis for Concept Mapping Assessment of Students' Science Achievement.* Paper presented at the annual meeting of the AERA, San Diego, CA.

EXTENDED ABSTRACT

Introduction

There are many measurement situations in which estimates of reliability are needed and also there are many different approaches to estimate interrater reliability. The concept of interrater reliability can be defined as the extent of consistency in scores obtained from two or more raters. It is relevant whenever the determination of scores requires that subjective judgments be made by the raters. In this study, four approaches to the estimation of interrater reliability are studied: correlation, comparison of means, percentage of agreement, and generalizability theory.

The first approach taken into account in this study is correlation. When used to estimate interrater reliability, the Pearson correlation operationally defines consistency as the extent to which two raters agree in the relative placement of persons (objects, events, etc.) in terms of the dimension being rated. The correlation coefficient indicates the extent of linear association between the two raters' scores on persons. Because the Pearson correlation is "mean-free," differences or similarities between the raters in the levels of their ratings are not taken into account when the reliability coefficient is calculated.

The second approach used for examining interrater reliability was comparisons of means. If the level of the ratings assigned by various raters is of concern, then raters' means should be examined. For example, when the data influence important decisions about individuals, the raters' interpretations of the meanings of the scale points need to be congruent to avoid unfair consequences (Goodwin & Goodwin, 1991).

Simple percentage of agreement, which is the third approached of interrater reliability estimates of this study, defines consistency as the relative number of times that two raters agree about the score that should be assigned to the item of an exam, for instance. An advantage of this approach is that it is quite easy to calculate and understand. Also, it can be used with all types of data such as nominal, ordinal, interval, and ratio.

The techniques of generalizability (G) theory, as presented by Cronbach et al. (1972), represent a different way to conceptualize and estimate reliability and lastly used in this study for estimating interrater reliability. G theory allows for the isolation of multiple sources of error, rather than just one source like in many other reliability estimation approaches. Thus, the G-theory allows one to acknowledge and incorporate several different sources of measurement error into the reliability or G study.

Method

The data of the study was collected through the ratings of two raters for the answers of 43 students on ten-item exam. Each of the items was rated out of ten and as a result each of the students gets scores from the exam out of 100. The ratings of the items were made by using a holistic rubric developed by the researchers. SPSS program was used for correlation and comparison of means, Excel was used for percentage of agreement and finally EduG was used for generalizability theory analysis.

Results and Discussion

For the data - composed of ratings for 43 students on ten items by two raters - the reliability estimates varied because of the situation that the ranges of the obtained values by used approaches and different calculation processes. The highest estimate was 0.90 which is estimated by G theory. Besides this result, it was obtained that there was positive and high correlation coefficient (0.74). The estimate of percentage of exact matches of agreement between the two raters was found as 58.9 %. Finally, although there were no statistically differences between general mean of scores, there were statistical differences among three of the items by means of rater scoring. Although G theory seems more complex than the other methods illustrated in the study, it yields more information than the other methods because of handling multiple sources of error at the same time. Therefore, it is proposed to be used when estimating interrater reliability. While deciding which approach is most appropriate for the estimation of interrater reliability in a particular situation, there are number of factors that must consider. Those factors could be the type of the data, the number of raters and type of decision to be made according to the result of interrater reliability. Finally, a suggestion could be given to the researchers who interested in interrater reliability. If more than one approach were used to determine the interrater reliability, there would be a more detailed and clear view of the situation to evaluate the situation since each approach would give unique results by means of their outputs.