Elementary Education Online, 2018; 17(2): pp. 1066-1076
İlköğretim Online, 2018; 17(2): s.1066-1076. [Online]: http://ilkogretim-online.org.tr
doi 10.17051/ilkonline.2018.419355

Dereceli Puanlama Anahtarlarının Güvenirliğinin Farklı Deneyim Yıllarına Sahip Puanlayıcıların Kullanıldığı Durumlarda İncelenmesi[1]

The Examination of Realiability of Scoring Rubrics Regarding Raters with Different Experience Years

Hatice Özlem ANADOL, TOBB ETU Yabancı Diller Bölümü, hanadol@etu.edu.tr
Celal Deha DOĞAN, Ankara Üniversitesi Eğitim Bilimleri Fakültesi, ddogan@ankara.edu.tr

Öz. Bu araştırmanın temel amacı; dereceli puanlama anahtarı (DPA) kullanmaya ilişkin deneyim yılının puanlayıcı güvenirliğine etkisini belirlemektir. Bu amaçla DPA kullanmaya ilişkin farklı deneyim yılına sahip üç grup puanlayıcının bulunduğu durumlarda elde edilen G ve Phi çalışması sonuçları karşılaştırılmıştır. Birinci grupta DPA kullanmaya ilişkin deneyimi az olan (1 yıl ve daha az) ikinci grupta DPA kullanmaya ilişkin deneyimi çok olan(5 yıl ve daha fazla) puanlayıcılar yer almaktadır. Üçüncü grupta ise deneyimi az ve çok olan puanlayıcılar bir arada yer almışlardır. Araştırmada üç farklı grupta yer alan puanlayıcılar İngilizce yazma becerisini ölçmeye yönelik geliştirilmiş açık uçlu bir başarı testini aynı DPA'yı kullanarak puanlamışlardır. Çalışmaya özel bir üniversitede hazırlık eğitimi alan 120 öğrenci ve aynı okulda çalışan 12 okutman dâhil edilmiştir. Araştırmada, birey ve puanlayıcıların maddeler ile çaprazlandığı ancak bireylerin puanlayıcılara yuvalandığı desenden ((b:p)xm) faydalanılmıştır. Araştırma sonucunda nitelikli bir DPA kullanıldığında DPA kullanmaya ilişkin deneyim yılının puanlayıcı güvenirliği üzerinde etkili olmadığı belirlenmiştir.

Anahtar Sözcükler: Genellenebilirlik kuramı, dereceli puanlama anahtarı, puanlayıcı güvenirliği, yuvalanmış desen

Abstract. The main goal of this study is to determine the role of experience years in using scoring rubric on inter-rater reliability. In regards to this objective; the results of G and Phi studies that were obtained by three groups of raters in one of which there are raters who have much experience in using scoring rubric; in one of which there are raters who have little experience in using scoring rubric and in one of which there is a rater with much experience and one with little experience in evaluating students using scoring rubric in the process of written performance assessment with (s:r)xi design were compared. Raters in three different groups evaluated an open ended achievement test developed to test students' written English written skills with the same scoring rubric. 120 students who study at the prep-class of a private university and 12 instructors that work at the same school participated to the study. In the study, a nested ((b:p)xm) design in which individuals are nested in raters but individuals and raters are crossed with items and raters was utilized. Based on the findings, it is found out that experience year in using scoring rubric doesn't affect reliability when a qualified scoring rubric is employed.

Keywords: Generalizability theory, scoring rubric, english writing skill, nested design

[1] Bu makale, birinci yazarın Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ölçme ve Değerlendirme Anabilim Dalında tamamladığı yüksek lisans tezinden üretilmiştir.

Geliş tarihi: 20.09.2017 Kabul tarihi: 07.02.2018 Yayımlanma tarihi: 28.04.2018

SUMMARY

Introduction. Recently, open-ended questions and performance tasks are attempted to be used in countrywide exams and a variety of raters take part in rating processes. In addition to this, open-ended questions are used in in-class evaluations in which at least two raters take part. In cases where more than one rater is used, rater characteristics are one of the factors affecting reliability of the evaluation. It may not be possible to group raters with similar experience years in rating process all the time; therefore, it is inevitable to group raters with different experience years in rating written performances and open-ended questions. The thing to take into consideration is a qualified scoring rubric. A qualified scoring rubric is supposed to yield similar results in different cases. In literature, there are many studies in which open-ended questions and performance tasks were evaluated using scoring rubrics. Yet, there are not any studies in which reliability coefficients of scoring rubrics obtained from G studies are compared based on different experience years of raters in using scoring rubrics. In this sense, the findings of this study can fill a gap in the field. In this context, the main goal of this study is to compare the results of G studies that were obtained by groups formed by three groups of raters in using scoring rubric in the process of English written skill assessment in a private university in Ankara. It attempts to compare the results in regards to raters' experience years.

Method. This is a quantitative study in which three separate G studies were carried out for (s:r)xi design (s: student, i: item, r:rater) by three groups of raters in one of which there are raters who have much experience in using scoring rubric; in one of which there are raters who have little experience in using scoring rubric and in one of which there is a rater with much experience and one with little experience in evaluating students using scoring rubric in accordance with 6 criteria. The study sample consists of 12 raters who work as instructors and 120 students who attend to foreign languages department in a private university in Ankara. Nested design (s:r)xi was used in rating. Based on the findings, the results of G studies were compared.

Results. In accordance with the objective of the study, written skills of 120 prep-class students in a private university in Ankara were evaluated by 12 raters who formed 3 different groups by a scoring rubric. As a result of the study, it is observed that variance rates that were estimated for variables in three groups, G and Phi coefficients, absolute and relative error variances are parallel to each other and they are significantly high. Therefore, it is found that there is no difference between raters with different experience years in using English written skill scoring rubric in terms of reliability and with a nested design, it is possible to obtain significantly high reliability coefficents.

Discussion and Conclusion. It is deduced from the quantitative data that experience year of using scoring rubrics in written performance evaluation does not affect reliability of rating. When a qualified scoring rubric is used, it can yield similar results to raters with different experience years in evaluating written performances or open-ended questions. When there are many written performance tasks are to be evaluated, there is no need to form groups of raters with similar experience years in rating. In addition; there is no need to employ a fully crossed design which requires more time and energy than a nested design since nested designs can also yield reliable results when used with qualified scoring rubrics.

GİRİŞ

Öğrencilerin akademik başarısını belirleyebilmek için, program hedefleri esas alınarak çeşitli başarı testleri kullanılır. Başarı testleri, sonuçlarının benzer gruplar ile karşılaştırabilirliği açısından öğretmen yapımı başarı testleri ve standart başarı testleri diye ikiye ayrılır. Gronlund (1977) standart başarı testlerini tanımlanmış davranış örneklemini ölçmek üzere düzenlenmiş bir dizi test maddesini içeren, uygulaması ve puanlaması el kitabı esas alınarak yapılan, testi hazırlayan kişilerin alanlarında uzman kişiler olduğu, geçerlilik ve güvenilirlik çalışmaları yapılmış olan testler olarak tanımlar. Öğretmen yapımı başarı testleri, sınıf içi başarıyı ölçmek için geliştirilen, küçük gruplarda uygulandığı geçerlilik ve güvenirlik analizleri için sınırlı istatistiksel çalışmadan yararlanılabilen testlerdir.

Öğretmenler, eğitim öğretimin her kademesinde, öğrencilerinin ders ile kazandırılmak istenen kazanımlara ulaşıp ulaşmadıklarını belirlemek isterler. Bu amaçla, çeşitli ölçme yöntemlerine başvururlar. Kendi bilgi ve becerileri dâhilinde hazırladıkları başarı testleri en yaygın kullandıkları ölçme ve değerlendirme araçlarıdır. Bu süreçte öğretmenlerin kullandığı çoktan seçmeli, doğru yanlış gibi yanıtı öğrenci tarafından seçilen testlerin puanlaması nesnel olarak gerçekleştirilebilmektedir. Ancak yanıtını öğrencinin yapılandırdığı açık uçlu maddelerden oluşan testlerin veya öğrencilerin özgün bir ürün oluşturduğu performans görevlerinin puanlama süreci daha öznel olabilmektedir. Bu durum ölçme aracının güvenirliğini olumsuz etkileyebilmektedir. Öznel yargılara açık olan yanıtı öğrencinin yapılandırdığı testlerin veya performans görevlerinin objektif şekilde puanlanması, okullardaki ölçme ve değerlendirmede etkinliklerinin sağlıklı bir şekilde yürütülmesi için önemlidir. Bu süreçte dereceli puanlama anahtarları önemli bir yer tutar.

Dereceli puanlama anahtarı (DPA), bireylerin ürünlerini ve performanslarını detaylı bir şekilde analiz etmek amacıyla kullanılan bir puanlama tasarımıdır (Moskal, 2000). DPA'lar, bütünsel ve analitik olmak üzere ikiye ayrılır. Bütünsel DPA'lar, ürünlerin bir bütün olarak değerlendirildiği puanlama anahtarlarıdır ve öğrencilerin performanslarının bütününe ilişkin tek bir puan vermeyi sağlar. Analitik DPA'lar ise performansın alt bölümlerinin tek tek puanlanmasını ve sonrasında toplam bir puan elde edilmesini gerektirir. Bu yolla, öğrencini hangi aşamada ve hangi konuda eksikliklerinin olduğu belirlenebilir (Moskal, 2000). Bu özellikleri ile DPA'lar puanlayıcı güvenriliğinin artırılmasında önemli bir yer tutar.

Ancak özellikle DPA'ların kullanıldığı durumlarda puanlayıcılar arası güvenirlik ile ilişkili olan bir boyut da puanlayıcı özellikleridir. Puanlayıcıların sahip oldukları bireysel farklılıklar, puanlama davranışı üzerinde önemli bir etkiye sahip olabilir. Özellikle son yıllarda merkezi sınavlarda açık uçlu sorular kullanılmış ve bu süreçte pek çok puanlayıcı görev almıştır. Bunun yanı sıra sınıf içi uygulamalarda da değerlendirme sürecinde birden fazla puanlayıcının yer alması söz konusudur. Bunun gibi birden fazla puanlayıcının yer alması gereken durumlarda belirli kriterlere sahip puanlayıcıların seçilmesi önemlidir.

Bu süreçte dikkate alınabilecek puanlayıcı özelliklerin başında puanlayıcıların deneyim yılı düşünülebilir. Dereceli puanlama anahtarlarının kullanımı belli bir donanım gerektirmekle beraber iyi bir DPA'nın deneyim yılı çok ve az olan puanlayıcıların benzer bir şekilde puanlama yapmalarını sağlaması beklenir.

Birden fazla puanlayıcının olduğu durumlarda pratik nedenlerden dolayı DPA kullanımına ilişkin benzer deneyim yılına sahip puanlayıcılara ulaşmak herzaman mümkün olamayabilir. Dolayısıyla, DPA kullanmaya ilişkin farklı deneyim yıllarına sahip puanlayıcıların, aynı grubu puanlaması kaçınılmazdır. Bu bağlamda farklı deneyim yıllarına sahip puanlayıcıların kullanıldığı durumlarda puanlayıcı güvenirliğinin belirlenmesi ve karşılaştırılmasının alana katkı sağlayacağı düşünülmektedir.

İlgili literatürde farklı kuramlara dayalı yöntemler ile puanlayıcılar arası güvenirliğin test edildiği çalışmalar yer almaktadır (Kasap, 2008; Parlak ve Doğan, 2014; Özel ve Acar, 2014; Büyükkıdık ve Anıl, 2015)**.**Tüm çalışmaların ortak sonucu olarak, DPA'nın değerlendirme yöntemi olarak kullanılması ile daha güvenilir sonuçların ortaya çıktığı belirtilebilir. Ancak, puanlayıcıların deneyim yıllarına dayalı olarak puanlama aracının güvenirliğine ilişkin karşılaştırmaların yapıldığı veya puanlayıcıların deneyim yıllarının güvenirlik katsayısına olan etkisinin incelendiği bir çalışmaya rastlanmamıştır. DPA kullanmaya yönelik deneyim yılı az ve çok olan puanlayıcıların kullanıldığı durumlarda puanlayıcı güvenirliğinin genellenebilirlik kuramına dayalı olarak belirlenmesinin ve elde edilen sonuçların karşılaştırılmasının alandaki bu eksikliğin giderilmesine katkı sağlaması ve göz ardı edilen bu konuya dikkat çekmesi beklenmektedir.

Bu araştırmanın amacı, Ankara'da bulunan bir vakıf üniversitesinde yabancı diller bölümünde verilen yazma dersi ara sınavının, DPA kullanmaya yönelik farklı deneyim yıllarına sahip puanlayıcılar tarafından puanlandığı durumlarda puanlayıcı güvenirliğini belirlemek ve karşılaştırmaktır. Bu doğrultuda aşağıdaki sorular cevaplanmıştır:

İngilizce yazılı anlatım becerisi ara sınavı, DPA kullanımına yönelik

- deneyim yılı az olan (1 yıl ve daha az) puanlayıcıların bulunduğu grup (grup1)
- deneyim yılı çok olan(5 yıl ve daha fazla) puanlayıcıların bulunduğu grup (grup2)
- deneyim yılı az ve çok olan puanlayıcıların birlikte yer bulunduğu karışık grup(grup 3) tarafından puanlandığında elde edilen
    - varyans bileşenleri
    - bağıl ve mutlak hata varyansları
    - G ve Phi katsayısı nasıldır?

YÖNTEM

Araştırma Modeli

Bu çalışmada, analitik dereceli puanlama anahtarlarının güvenirlikleri farklı deneyim yıllarına sahip puanlayıcıların kullanıldığı durumlarda genellenebilirlik kuramına dayalı olarak belirlenmiş ve karşılaştırılmıştır. Bu nedenle çalışma, temel araştırma türündedir. Temel araştırmalar, kuram geliştirmeyi ya da var olan kuramları sınamayı amaçlayan araştırmalardır (Karasar, 2005).

Çalışma Grubu

Puanlayıcılar amaçlı örneklem ile seçilmiştir. Araştırma kapsamında DPA kullanmaya yönelik 1 yıl deneyime sahip 6 okutman deneyimi az, DPA kullanmaya yönelik 5 yıl ve üzeri deneyime sahip 6 okutman ise deneyimi çok, son olarak dereceli puanlama anahtarı kullanmaya ilişkin deneyimi az ve çok olan puanlayıcıların aynı grupta yer alması ile (5 yıl ve üzeri deneyime sahip 3 okutman; 1 yıl deneyime sahip 3 okutman) oluşan grup karışık grup olarak kabul edilmiştir. Araştırmanın yürütüldüğü üniversitede istisnasız tüm okutmanlar, her dönem uygulanan altı sınavın puanlamasında görev almak durumundadır. Dolayısıyla üç dönem eğitim verilen okulda, bir sene içinde 18 kez puanlama yapılmaktadır. Dolayısıyla, "puanlayıcıların deneyim yılı" kavramı ile öğrenci çalışmalarını puanlama sürecinde puanlayıcıların dereceli puanlama anahtarını kullanma sıklıkları ifade edilmiştir. Deneyimi az olan puanlayıcılar uygulamanın yapıldığı kurumda en fazla 1 yıldır görev yapmaktadırlar. Bu bağlamda deneyimi az olarak nitelendirilen puanlayıcılar en fazla 18 sınavda öğrenci çalışmalarının puanlamasında dereceli puanlama anahtarını kullanmışlardır ve önceki çalıştıkları kurumlarda da dereceli puanlama anahtarı ile öğrenci çalışmalarını puanlamamışlardır. Deneyimi çok olarak

nitelendirilen puanlayıcılar ise ilgili kurumda en az 5 yıldır görev yapmaktadırlar ve bu süreçte en az (18 x 5) 90 sınavda öğrenci çalışmalarını dereceli puanlama anahtarı kullanarak puanlamışlardır. Bu bağlamda sözcük sınırı nedeni araştırmanın başlığında "deneyim yılı" kavramı kullanılmıştır.

Veri Toplama Araçları

İngilizce Yazılı Anlatım Becerisi Değerlendirme Sınavı

Araştırmada, öğrencilerin İngilizce yazılı anlatım becerilerini değerlendirmek için, verilen üç konudan birini seçmeleri ve seçilen konu hakkında bir görüş kompozisyonu (opinion essay) yazmaları istenmiştir. Öğrencilerin bu kompozisyon türünde, İngilizce olarak fikirlerini savunmaları ve savundukları fikirleri örneklerle desteklemeleri gerekmektedir. Yazılan kompozisyonlar, "Dilbilgisi Kullanımı, Kelime Kullanımı, Konu Bütünlüğü ve Bağlantı, Cümleler Arası Geçiş, Örneklendirme, Sonuç Cümlesi" olmak üzere 6 ölçüte göre değerlendirilmiştir.

Dereceli Puanlama Anahtarı (Rubrik)

Bu çalışma kapsamında, öğrenci çalışmalarının puanlanması için, uygulamanın yapılacağı vakıf üniversitesinin hazırlık okulunda hali hazırda kullanılanıma devam eden analitik dereceli puanlama anahtarı kullanılmıştır. Çalımada kullanılan bu DPA hazırlık okulunda görev yapan okutmanlarca daha önceki yıllarda geliştirilmiştir ve geliştirme aşamasında ne amaçla geliştirileceği, yine hazırlık okulunda görev yapan okutmanlarca belirlenmiştir. Çalışma kapsamında kullanılan DPA'nın geliştirilme amacı, bir eğitim- öğretim yılı boyunca, öğrencilerinin İngilizce yazılı anlatım becerilerini değerlendirmektir. Amaç belirlendikten sonra, İngilizce yazılı anlatım becerisinin hangi ölçütler kullanılarak değerlendirileceği, alan yazın incelenerek belirlenmiştir. İlgili kaynaklar tarandıktan sonra "Dilbilgisi Kullanımı, Kelime Kullanımı, Konu Bütünlüğü ve Bağlantı, Cümleler Arası Geçiş, Örneklendirme, Sonuç Cümlesi" olmak üzere 6 ölçüt belirlenmiştir. Ölçütlerin yeterlilik düzeyleri ayrıntılı bir şekilde tanımlanmıştır. DPA'daki her bir ölçüt, alan yazındaki çalışmalar göz önünde bulundurularak belirlenmiştir. Hazırlanan taslak ölçek, uzman görüşüne sunulup gereken düzeltmeler yapıldıktan sonra, okutmanlar tarafından sınıf içi yazma etkinliklerinin değerlendirilmesinde kullanılmış ve gelen dönütlerden yararlanarak, gerekli düzeltmelerden sonra, son halini almıştır.

Verilerin Toplanması

Araştırmanın verileri, bir vakıf üniversitesinin üçüncü dönem ilk vize sınavı kapsamındaki 40 dakikalık yazılı yoklamadan, gerekli izinler alınarak toplanmış ve DPA ile puanlanmıştır. 120 kişilik bir grup, 6 deneyimi çok ve 6 deneyimi az ve 1 deneyimi çok ve 1 deneyimi az puanlayıcının oluşturduğu karışık grup tarafından puanlanmış ve elde edilen güvenirlik katsayıları genellenebilirlik kuramına dayalı olarak karşılaştırılmıştır. Araştırmada, birey ve puanlayıcıların maddeler ile çaprazlandığı ancak bireylerin puanlayıcılara yuvalandığı desenden ((b:p)xm) faydalanılmıştır. Bu desende her birey bir maddeyi yanıtlarken her puanlayıcı farklı bireyleri puanlamıştır. Araştırmaya ilişkin veriler (Nisan 2015) 40 dakikalık yazılı yoklamadan, gerekli izinler alınarak toplanmıştır. Uygulama, hâlihazırda yazma dersi veren okutmanlarca yapılmıştır. İlgili yazılı yoklama kağıtları rastgele toplanmış ve puanlayıcılar tarafından birbirlerinden bağımsız olarak 1 hafta içerisinde DPA kullanarak değerlendirilmiştir.

Verilerin Çözümlenmesi

Araştırmada deneyimi çok puanlayıcıların yer aldığı "deneyimi çok grup", deneyimi az puanlayıcıların yer aldığı "deneyimi az grup" ve deneyimi çok ve deneyimi az puanlayıcıların birlikte yer aldığı "karışık grup" için varyans bileşenleri ve her değişkenlik kaynağının göreli etkisi

incelenmiştir. Varyans bileşenlerine dayalı olarak, mutlak hata varyansı, göreli hata varyansı hesaplanmıştır. Akabinde göreli hata varyansına dayalı olarak G katsayıları ve mutlak hata varyansına dayalı olarak da Phi katsayıları hesaplanmıştır. Sonuç olarak 6 deneyimi çok ve 6 deneyimi az puanlayıcının ayrı ayrı, 1 deneyimi çok 1 deneyimi az puanlayıcının beraber yer aldığı üç durum için hesaplanan G ve Phi katsayıları karşılaştırılmıştır. Çalışmada elde edilen nicel veriler genellenebilirlik kuramı kapsamında, varyans bileşenlerinin ve değişkenlerin toplam varyansı açıklama oranlarının hesaplandığı Edu G programı kullanılarak analiz edilmiştir. (B:p)xm deseni için mutlak hata varyansının, bağıl hata varyansının, G ve Phi katsayılarının hesaplanmasında kullanılan formüller aşağıda verilmiştir:

Göreli hata varyansı formülü: $\sigma^2_{Göreli} = \frac{\sigma^2_{pm}}{n_p} + \frac{\sigma^2_{mp:b}}{n_b + n_p}$

Mutlak hata varyansı formülü: $\sigma^2_{Mutlak} = \frac{\sigma^2_{m}}{n_m} + \frac{\sigma^2_{pm}}{n_m} + \frac{\sigma^2_{b,pb}}{n_b n_p} + \frac{\sigma^2_{mb:p}}{n_b n_p}$

G katsayısı hesaplama formülü: $(G) = \frac{\sigma^2_m}{\sigma^2_m + \sigma^2_{Göreli}}$

Phi katsayısı hesaplama formülü: $\phi = \frac{\sigma^2_m}{\sigma^2_m + \sigma^2_{Mutlak}}$

BULGULAR

Bu bölümde, öğrenci çalışmalarının farklı deneyim yılına sahip üç puanlayıcı grup tarafından puanlaması sonucunda (b:p)xm deseni ile elde edilen bulgular sunulmuştur. Çizelge 1'de her bir değişkenlik kaynağı için varyans bileşenleri ve varyans açıklanma oranları, mutlak ve bağıl hata varyansları ve son olarak G ve Phi katsayıları sunulmuştur. Çizelgede her bir değişkenlik kaynağı için parantez içerisinde yer alan değer varyans bileşeni parantez dışında yer alan değer ise varyans açıklanma yüzdesini göstermektedir.

Çizelge 1: *G Çalışması Sonuçlarının Karşılaştırılması*

		Deneyim Yılı Az olan Grup	**Deneyimi Yılı Çok Grup**	**Karışık Grup**
Değişkenlik Katsayısı	**M**	%0.9 (0.00532)	%0.7 (0.00400)	%0 (-0.00163)
	P	%0.2 (0.00141)	%0 (-0.00293)	%1 (0.00491)
	B:P	%47.1 (0.27548)	%51.1 (0.27794)	%50.1 (0.30670)
	MP	%3.2 (0.01877)	%2.7 (0.01459)	%1.2 (0.00761)
	MB:P	%48.5 (0.28396)	%45.5 (0.24789)	%47.8 (0.29235)

Hata Varyansı	**Mutlak hata varyansı**	0.22659	0.21075	0.22359
	Bağıl hata varyansı	0.22462	0.20916	0.22359
Güvenirlik Katsayısı	**G katsayısı**	0.85	0.86	0.86
	Phi katsayısı	0.84	0.86	0.86

İngilizce yazılı anlatım becerisi, puanlama sürecine yönelik hesaplanan varyans bileşenleri incelendiğinde, madde ana etkisine ilişkin hesaplanan varyans bileşeninin (M) en küçük karışık grupta (-0.00163 - %0) kestirildiği görülmektedir. Deneyimi çok (0.00400 - %0.7) ve deneyimi (0.00532 - %0.9) az gruplarda madde ana etkisine ilişkin hesaplanan varyans bileşeni de, karışık gruptakinden çok büyük farklılık göstermemektedir. Bu sonuç, her üç durumda da maddelerin birbirinden farklılaşmadığını ve benzer şekilde puanlandığını göstermektedir.

Puanlayıcılara ilişkin hesaplanan varyans değerinin (P) en büyük karışık grupta (0.00491 - %1) kestirildiği görülmektedir. Puanlayıcılara ilişkin hesaplanan varyans değerleri, puanlama sürecine ilişkin deneyimi çok ((-0.00293)- %0) ve deneyimi az olan (0.00141 %0.2) gruplar için hemen hemen aynı düzeydedir. Sonuç incelendiğinde, puanlayıcı ana etkisine ilişkin hesaplanan varyans bileşeninin gruplar arası önemli bir fark yaratmadığı ve deneyim yılının puanlama güvenirliğini etkilemediği sonucuna varılabilir. Puanlayıcılar birbirleri ile tutarlı puanlamalar yapmışlardır.

(B:p) için hesaplanan varyans bileşeni, en büyük deneyimi çok olan grupta (0.27794 - %51.1) kestirilmiştir. Deneyimi az grupta (0.27548 - %47.1) ve karışık grupta (0.30670 - %50.1) hesaplanan varyans değerleri de deneyimi çok grupta hesaplanan varyans değerinden önemli düzeyde fark göstermemektedir. Bu durum, birey-puanlayıcı ortak etkileşiminden kaynaklı farklılıkların, puanlayıcı davranışlarının bir puanlayıcı grubundan diğerine farklılaşmadığı fakat bir bireyden diğerine değiştiği şeklinde yorumlanabilir.

Madde puanlayıcı ortak etkişimi (MP) için hesaplanan varyans bileşene ilişkin en düşük değer(0.00761 - % 1.2) karışık grupta elde edilmiştir. Deneyimi çok olan (0.01459 - % 2.7) ve deneyimi az olan (0.01877 - % 3.2) biraz daha büyük varyans bileşenleri elde edilmiştir. Ancak he üç grupta da ilgili varyans bileşeinin toplam varyansın küçük bir kısmını açıklaması nedeniyle puanlayıcıların bireyleri bir maddeden diğerine kararlı bir şekilde puanladıkları ve verilen puanın, puanlayıcıdan puanlayıcıya değişmediği şeklinde yorumlanabilir.

Hesaplanan en düşük artık varyans(MP:B) değerleri deneyimi çok olan grupta (0.24789 - % 45.5) elde edilmiştir. Artık varyans karışık grup (0.29235 - 47.8) ve deneyimi az olan grupta (0.28396 – 48.5) bir miktar daha yüksektir. Artık varyans bileşeninin büyük çıkması birey puanlayıcı ve madde ortak etkileşiminin tesadüfî hata kaynaklarından etkilendiğinin göstergesi olabilir.Ancak birey-görev madde etkileşiminin, üç yönlü etkileşimin ve diğer değişkenlik kaynaklarından kaynaklanan etkilerin, gruplar arası büyük bir fark göstermediği şeklinde yorumlanabilir.

(B:p)xm deseni kullanılarak deneyimi az grup için elde edilen $\sigma^2(\delta)$ = 0.22462, $\sigma^2(\Delta)$ = 0.22659 deneyimi çok grup için elde edilen $\sigma^2(\delta)$ = 0.20916, $\sigma^2(\Delta)$=0.21075 ve karışık grup için

elde edilen $\sigma^2(\delta)$ = 0.22359, $\sigma^2(\Delta)$= 0.22359 olarak hesaplanmıştır. Farklı verilerle aynı senaryo durumuna göre oluşturulmuş üç desenden, deneyimi çok puanlayıcıların kullanıldığı durumda bağıl ve mutlak hata varyansları en küçük kestirilmiştir ancak üç grupta hesaplanan değerler arasında çok büyük farklar bulunmamıştır.

(B:p)xm deseni ile araştırmada kullanılan veriler ışığında deneyimi az gruptan elde edilen G katsayısı 0.85, Phi katsayısı ise 0.84 olarak, deneyimi çok puanlayıcıların kullanıldığı durumda G katsayısı 0.86 ve Phi katsayısı 0.86, karışık grupta, G katsayısı 0.86 ve Phi katsayısı 0.86 olarak kestirilmiştir. Böylelikle; farklı veriler ve aynı desen kullanılarak oluşturulmuş üç durumda, G ve Phi katsayılarının, çok yüksek kestirildiği, puanlama güvenirliği üzerinde deneyim yılının önemli bir etkiye sahip olmadığı sonucuna varılabilir.

TARTIŞMA ve SONUÇ

Bu araştırmada, genellenebilirlik kuramına dayalı olarak İngilizce dersi yazılı anlatım becerisinin, puanlamaya yönelik deneyimi az, puanlamaya yönelik deneyimi çok ve karışık grup puanlayıcılar olmak üzere üç puanlayıcı grubu tarafından puanlanmasıyla oluşturulan desenlerin genellenebilirlik (G) çalışması sonuçları karşılaştırılmıştır. Sonuçlar aşağıda özetlenmiştir:

Nitelikli bir DPA kullanılarak yapılan puanlamalarda, puanlayıcı özelliklerinden olan DPA kullanmaya ilişkin deneyim yılının güvenirlik üzerinde etkili olmadığı ve deneyim yılları farklı puanlayıcıların yer aldığı durumlar için hesaplanan güvenirlik katsayılarının önemli bir farklılık yaratmadığı, sonucuna varılmıştır. Tüm gruplarda hesaplanan G ve Phi katsayıları ve bağıl ve mutlak hata varyansları arasında tutarlılık olduğu görülmüştür. Bu sonuçlardan hareketle, puanlayıcıların deneyim yılının iyi hazırlanmış bir DPA'nın kullanıldığı durumlarda güvenirliği etkilemediği sonucuna varılmıştır.

İngilizce yazılı anlatım becerisini ölçmek için hazırlanan yazılı yoklama (b:p)xm yuvalanmış deseni ile ile yapılan G çalışmaları sonucunda hesaplanan varyans bileşenleri ve toplam varyansı açıklama oranları her üç durumda da birbirine yakın ve yüksek sonuçlar üretmiştir. Bu sonuçlar, yuvalanmış desenler ile de yüksek güvenirlik katsayıları elde edilebileceği sonucuna varılabileceğini göstermiştir. Bu sonuç Nalbantoğlu (2009) tarafından yürütülen çalışmanın sonuçlarını destekler niteliktedir. Nalbantoğlu'nun çalışmasında (2009), öxgxp (ö:öğrenci, g:görev ve p:puanlayıcı olmak üzere) deseni ve(ö:p)xg deseni kullanılarak 48 öğrenciden her biri 3 puanlayıcı tarafından 15 görev doğrultusunda puanlanmış ve elde edilen G ve Phi katsayısıları karşılaştırılmıştır. Bu iki desenden, (ö:p)xg deseninde G ve Phi katsayılarının daha yüksek hesaplandığı sonucuna varılmıştır.

İngilizce yazılı anlatım becerisinin değerlendirilmesine yönelik hazırlanan analitik DPA ile yapılan puanlamalara ilişkin G çalışmalarının, çalışmaya katılan tüm gruplar için benzer sonuçlar verdiği görülmüştür. Bu sonuçlar, Andrade ve Du (2005) tarafından yürütülen ve öğrencilerin DPA'yı kullanılması durumunda daha güvenilir puanlar aldıkları sonucuna varan çalışmayı destekler niteliktedir. Ayrıca bu sonuçlar, analitik DPA ile elde edilen puanların yüksek güvenirlik gösterdiği ile sonuçlanan, alan yazındaki diğer birçok çalışmayla örtüşmektedir (Kan, 2005; Kasap, 2008; Özel ve Acar, 20014; Büyükkıdık ve Anıl, 2015).

Sınıf içi değerlendirmelerde kullanılan ve geniş ölçekli sınavlarda kullanılması planlanan açık uçlu soruların güvenilir bir şekilde puanlanması sorunu gündemdedir. Özellikle açık uçlu soruların kullanıldığı geniş ölçekli sınavlarda DPA kullanımına ilişkin farklı deneyime sahip puanlayıcıların bir arada yer alması kaçınılmaz olacaktır. Bu bağlamda geniş ölçekli sınavlarda bulunan açık uçlu soruları değerlendirmede görev alacak puanlayıcı seçiminde DPA kullanımına

ilişkin deneyim yılı göz ardı edilebilir. Burada kastedilen hiç DPA kullanmamış ve puanlama sürecini hiç deneyim etmemiş puanlayıcıların seçilmesi değildir. Ancak DPA kullanmaya ilişkin eğitim almış ve bu konuda deneyimi olan puanlayıcılar arasından deneyimi çok veya az olanların seçilmesinin güvenirlik üzerinde önemli bir etki yaratmadığı vurgulanmak istenmektedir.

Altını çizilmesi gereken diğer husus ise bahsedilen bu durumun nitelikli bir DPA kullanıldığı durumda gerçekleşebileceğidir. Yani kullanılan DPA nitelikli değil ise puanlayıcıların deneyimi önemli bir rol oynayabilir. Ancak nitelikli bir DPA (ölçütleri iyi belirlenmiş, performans tanımları açık ve anlaşılır olan, gereğinden fazla uzun ve karmaşık olmayan vb.) kullanıldığında deneyimi az olan puanlayıcılar da deneyimi çok olan puanlayıcılar kadar doğru ve güvenilir puanlama yapabilmektedirler. Daha öncede belirtildiği gibi iyi bir DPA'nın deneyim yılı çok ve az olan puanlayıcıların benzer bir şekilde puanlama yapmalarını sağlaması beklenir.

Çok sayıda sınav kâğıdının kısa bir süre içinde puanlanması gerektiğinde ve puanlayıcılar arası tutarlılık sağlandığında, tüm puanlayıcıların tüm öğrencileri puanlaması yerine bazı öğrencileri puanlaması, diğer bir deyişle çaprazlanmış desenler yerine yuvalanmış desenler tercih edilebilir. Benzer konuda çalışmayı planlayan araştırmacılara şunlar önerilebilir:

Bu araştırma, 6 puanlayıcılı örneklemlerle gerçekleştirilmiştir. Benzer çalışmalar yapacak olan araştırmacılar, farklı puanlayıcı sayıları kullanarak araştırmalarını düzenleyebilirler. Bu araştırma kapsamında güvenirliğin değerlendirilmesinde, G kuramının yuvalanmış (b:m)xp deseni ile analizi yapılmıştır. Benzer çalışmalarda farklı yuvalanmış desenler ya da tümüyle çaprazlanmış desen kullanılabilir. Benzer çalışmalarda veri çözümleme sürecinde Çok Değişkenli Rash modeli kullanılabilir. Puanlayıcılara ilişkin farklı bireysel özellikler dikkate alınarak güvenirlik üzerindeki etkisi incelenebilir.

KAYNAKÇA

Andrade, H. G. & Du, Y., (2005).*Student perspectives on rubric-referenced assessment. Practical Assessment, Research&Evaluation*.Vol.10 Number 3, 6-7.

Aslanoğlu, A. E. ve Kutlu, Ö. (2003). Öğretimde sunu becerilerinin değerlendirilmesinde dereceli puanlama anahtarı (rubric) kullanılmasına ilişkin bir araştırma. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, Sayı 36 (1-2), 25-36.

Büyükkıdık S. ve Anıl, D., (2015). Performansa Dayalı Durum Belirlemede Güvenirliğin Genellenebilirlik Kuramında Farklı Desenlerle İncelenmesi. *Eğitim ve Bilim Dergisi*: 40 (177), 285-296.

Crocker, L., & Algina, J. (1986). *Introduction to Classicaland Modern Test Theory*. New York: CBS College Publishing.

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.

Goodwin, L.D. (2001). Interrater agreement and reliability*. Measurement in Psychical Educationand Exercises Science*, 5 (1), 13-14. http://dx.doi.org/10.1207/S15327841MPEE0501_2.

Kan, A. (2005). Yazılı Yoklamaların Puanlanmasında Puanlama Cetveli ve Yanıt Anahtarı Kullanımının Puanlayıcı Güvenirliğine Etkisi.*Eurasian Journal of Educational Research*, 5(20), 166-177.

Kan, A. (2007). Performans değerlendirme sürecine katkıları açısından yeni programanlayışı içerisinde kullanılabilecek bir değerlendirme yaklasımı: Dereceli puanlama anahtarı puanlama yönergeleri. *Kuram ve Uygulamada Eğitim Bilimleri,* 7(1), 129-152.

Kasap, Y. (2008). *Dereceli puanlama anahtarı ve puanlama anahtarından elde edilen puanların karşılaştırılması*. Yüksek lisans tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Kutlu, Ö., Doğan, C. D. ve Karakaya, İ. (2008). *Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme*. (3. Baskı), Ankara: Pegem A Yayıncılık.

Nalbantoğlu, F. (2009). *Performans Ölçümlerinde Genellenebilirlik Kuramıyla Farklı Desenlerin Karşılaştırılması.* Yüksek Lisans Tezi. Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

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

Moskal, B. M. (2000). Scoring rubrics: what, when and how? *Practical Assessment Research and Evaluation,* 7 (3), 1-5. (http://pareonline.net/getvn.asp?v=7ven=3.
Moskal, B. M. (2003). Recommendations for Developing Classroom Performance Assessments and Scoring Rubrics. *Practical Assessment, Research & Evaluation*, 8(14). http://PAREonline.net/getvn.asp?v=8&n=14.
Özel, S. ve Acar, T. (2014).*Okullarda Sınıf İçi Ölçmelerde G Katsayısı*. IV. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde Sözlü Bildiri olarak sunulmuştur. Hacettepe Üniversitesi.
Parlak, B. ve Doğan, N. (2014). Dereceli puanlama anahtarı ve puanlama anahtarından elde edilen puanların uyum düzeyleri. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 29(2), 189-197.
Popham, W. J. (1997). What's stil wrong- and what's stil rightwithrubric. *Educational Leadership,* 55(2), 72-75.
Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

EK ***:*** *Yazma Dersi Analitik Dereceli Puanlama Anahtarı*

	1. Dilbilgisi	**2. Kelime**	**3. Tutarlılık ve Bağlantı**	**4. Geçişler**	**5. Örnekler**	**6. Sonuç Cümlesi**
4 Puan	Anlamada hiçbir probleme sebep olmayan sıfıra yakın dil bilgisi hatası ile kurulan, değişik yapılar kullanılan cümleler.	Çeşitli ve doğru yerde kullanılan kelimeler	Paragraf ya da metin türünün gerektirdiği mesajı, anlaşılır, tutarlı ve mantıklı bir şekilde verme.	Cümleler ve fikirler arası geçişi sağlayan, uygun ve çeşitli bağlaçlar	Ana fikir ve metnin akışını destekleyen detaylar ve örnekler	Paragrafın ya da metnin ana fikrini son bir düşünce ile tekrar vurgulayan son cümle
3 Puan	Anlamada ufak problemlere sebep olan dil bilgisi hataları ile kurulan, değişik yapılar kullanılan cümleler.	Yeterli fakat ufak hatalarla kullanılan kelimeler.	Paragraf ya da metin türünün gerektirdiği mesajı, ilgisiz birkaç örnek ile, anlaşılır bir şekilde verme	Cümleler ve fikirler arası geçişi sağlayan, az sayıda ve bazen yanlış yerde kullanılan bağlaçlar	Ana fikir ve metnin akışını kısmen destekleyen detaylar ve örnekler	Paragrafın ya da metnin ana fikrini, ana fikirle aynı kelimelerle vurgulayan son cümle.
2 Puan	Öğrenilen dilbilgisi yapıların birkaçı kullanılarak kurulan, birçok dilbilgisi hatası barındıran cümleler.	Sınırlı ve zaman zaman yanlış anlaşılmalara sebep olabilecek kelimeler.	Paragraf ya da metin türünün gerektirdiği mesajı, ilgisiz birçok örnek ile, anlaşılması zor bir halde verme	Çok az sayıda, çok basit düzeyde ve çoğu kez yanlış kullanılan bağlaçlar	Ana fikir ve metnin akışını desteklemeyen, ilgisiz detaylar ve örnekler	Paragrafta ya da metinde sonuç cümlesinin olmaması
1 Puan	Öğrenilen dilbilgisi yapılarını neredeyse	Çok sınırlı ve çoğu zaman yanlış anlaşılmanla	Paragraf ya da metin türünün gerektirdiği			

	hiçbirini kullanmadan kurulan, neredeyse tüm cümlelerde hata barındıran, anlatılmak istenen mesajın anlaşılmadığı cümleler.	sebep olan kelimeler	mesajı, tamamen ilgisiz örnekler ile, anlaşılması imkansız bir halde verme			