ANKARA ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

EĞİTİMDE ÖLÇME VE DEĞERLENDİRME ANABİLİM DALI

EĞİTİMDE ÖLÇME VE DEĞERLENDİRME PROGRAMI

DERECELİ PUANLAMA ANAHTARLARININ GÜVENİRLİĞİNİN FARKLI DENEYİM YILLARINA SAHİP PUANLAYICILARIN KULLANILDIĞI DURUMLARDA İNCELENMESİ

YÜKSEK LİSANS TEZİ

Hatice Özlem ANADOL

Ankara, Şubat 2017

ANKARA ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

EĞİTİMDE ÖLÇME VE DEĞERLENDİRME ANABİLİM DALI

EĞİTİMDE ÖLÇME VE DEĞERLENDİRME PROGRAMI

DERECELİ PUANLAMA ANAHTARLARININ GÜVENİRLİĞİNİN FARKLI DENEYİM YILLARINA SAHİP PUANLAYICILARIN KULLANILDIĞI DURUMLARDA İNCELENMESİ

YÜKSEK LİSANS TEZİ

Hatice Özlem ANADOL

Danışman: Yrd. Doç. Dr. C. Deha DOĞAN

Ankara, Şubat 2017

JÜRİ ÜYELERİNİN İMZA SAYFASI

Eğitim Bilimleri Enstitüsü Müdürlüğü'ne,

Hatice Özlem ANADOL'un hazırladığı "Dereceli puanlama anahtarlarının güvenirliğinin farklı deneyim yıllarına sahip puanlayıcıların kullanıldığı durumlarda incelenmesi" isimli bu çalışma jürimiz tarafından Ölçme ve Değerlendirme Anabilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

İmza

Başkan: Doç. Dr. İsmail KARAKAYA .....

Üye: Yrd. Doç. Dr. Cidde DOĞAN .....
(Danışman)

Üye: Yrd. Doç. Dr. Ergül DEMİR .....

Onay

Bu tez Ankara Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca yukarıdaki öğretim üyeleri tarafından.../.../... tarihinde uygun görülmüş ve Enstitü Yönetim Kurulunca.../.../... tarihinde kabul edilmiştir.

.... / .... / .....

Prof. Dr. İsmail GÜVEN

Enstitü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.



Hatice Özlem ANADOL

ÖZET

DERECELİ PUANLAMA ANAHTARLARININ GÜVENİRLİĞİNİN FARKLI DENEYİM YILLARINA SAHİP PUANLAYICILARIN KULLANILDIĞI DURUMLARDA İNCELENMESİ

ANADOL, Hatice Özlem

Yüksek Lisans, Ölçme ve Değerlendirme Anabilim Dalı
Tez Danışmanı: Yrd. Doç. Dr. C. Deha Doğan
Şubat 2017, xi + 76 Sayfa

Bu araştırmanın temel amacı; İngilizce yazılı anlatım becerisi puanlamada, dereceli puanlama anahtarı (Rubrik) kullanmaya ilişkin üç grup puanlayıcıdan toplanan verilerin G çalışmaları sonuçlarını karşılaştırmaktır.

2015- 2016 öğretim yılında, Ankara'da bulunan bir vakıf üniversitesinin hazırlık okuluna devam eden ve amaçlı örneklem ile seçilen öğrenciler ve bu öğrencilere eğitim veren okutmanlar araştırmanın çalışma grubunu oluşturmaktadır.

Araştırma, dereceli puanlama anahtarı kullanmaya ilişkin deneyimi az, dereceli puanlama anahtarı kullanmaya ilişkin deneyimi çok ve son olarak dereceli puanlama anahtarı kullanmaya ilişkin deneyimi az ve çok olan puanlayıcıların aynı grupta yer alması ile oluşan karışık grup olmak üzere üç ayrı grupta aynı dereceli puanlama anahtarı ile 6 görev doğrultusunda, (b:p)xm deseni ile (b: birey, m: madde,p: puanlayıcı) öğrencileri değerlendirdiği G çalışmaları yapılmıştır.

Araştırmada yapılan G çalışmaları sonuçları karşılaştırılmıştır. Elde edilen sonuçlara göre her üç durumda da değişkenler için hesaplanan varyans değerlerinin, G ve Phi katsayılarının, mutlak ve bağıl hata varyanslarının birbirleriyle paralellik gösterdiği görülmüştür. Böylece nitelikli bir derece puanlama anahtarı kullanarak

yapılan güvenirliği yüksek değerlendirmelerde, dereceli puanlama anahtarı kullanmaya ilişkin deneyim yılının etkili olmadığı anlaşılmıştır.

Verilerin analizinden sonra verilerin toplandığı üniversitede görev yapan 6 okutmanla, dereceli puanlama anahtarı hakkındaki görüşlerini toplamak amacıyla görüşmeler yapılmıştır. Görüşmeler sonucunda elde edilen veriler, içerik analizi ile analiz edilmiştir. Görüşme yapılan okutmanlar, dereceli puanlama anahtarı kullanarak yazılı anlatım becerisi değerlendirmeye karsı olumlu görüş bildirmişlerdir ve puanlayıcı güvenirliğini artırdığını belirtmişlerdir. Buna ek olarak, dereceli puanlama anahtarı kullanma konusunda yasadıkları zorlukların başında yaşanan zaman sıkıntısı ve kalabalık bir grubu puanlama esnasında yaşanan yorgunluk etkisi olarak belirtmişlerdir. Okutmanların dereceli puanlama anahtarı kullanmaya ilişkin görüşlerinin, uygulama aşamasında karşılaşılan güçlüklere çözüm önerileri getirmelerine yardımcı olacağı düşünülmektedir.

Anahtar Sözcükler

Genellenebilirlik Kuramı, İngilizce Yazılı Anlatım Becerisi, Dereceli Puanlama Anahtarı

ABSTRACT

THE EXAMINATION OF REALIABILITY OF SCORING RUBRICS REGARDING RATERS WITH DIFFERENT EXPERIENCE YEARS

ANADOL, Hatice Özlem

Master Thesis, Department of Measurement and Evaluation
Advisor: Yrd. Doç. Dr. C. Deha DOĞAN
February 2017, xi + 76 Page

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 written performance assessment.

Students that were chosen with purposive sampling from the prep-class of a private university in Ankara in 2015-2016 academic year and instructors that teach the students constitute the study group.

In this study, 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.

Based on the findings, the results of G studies were compared. 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. Therefore, it is found that

there is no difference between raters with different experience years in using scoring rubric in terms of reliability.

After the analysis of data, an interview form was applied to 6 instructors who work at the university where data of the G studies were collected to find out their point of views and practice upon rubric use in testing and evaluation of written skills of students. Data derived from the interview was evaluated with content analysis. Interviewees passed positive remarks on scoring rubric using in written skill evaluation and they emphasized high rater reliability. In addition to this, they expressed difficulties they experienced during the use of scoring rubric which are the pressure of time, the crowd of the classrooms to be scored and fatigue impact. Views of instructors on scoring rubric are expected to help propose solutions to the difficulties they experienced during the stage of the application.

Key Words

Generalizability Theory, English Writing Skill, Scoring Rubric

TEŞEKKÜR

Bu çalışma, dereceli puanlama anahtarı kullanılarak yapılan değerlendirmelerin önemini bir kez daha vurgulamak ve bu alanda yapılacak çalışmalara katkı sağlamak amacıyla gerçekleştirilmiştir.

İş hayatı ile beraber yürüttüğüm ve oldukça emek isteyen bu dönemde bana destek olan birçok kişiye teşekkür borçluyum. Öncelikle, tez sürecim boyunca benden samimi desteğini esirgemeyen, tüm düşüncelerimi paylaşabildiğim sayın danışmanım Yrd. Doç. Dr. Celal Deha DOĞAN'a, Doç. Dr. İsmail KARAKAYA'ya ve Yrd. Doç. Dr. Ergül DEMİR'e,

Değerli katkıları için Yrd. Doç. Dr Gülşen TAŞDELEN TEKER ve Doç. Dr. Neşe GÜLER'e,

Çok değerli zamanlarından bana ayırarak veri toplamama ve uygulama yapmama izin veren TOBB ETU Yabancı Diller Bölümü bölüm başkanı sayın Taner YAPAR'a ve çok değerli hocalarına,

Tez yazma sürecim boyunca desteğini benden esirgemeyen Arş. Gör. Ömer KAMIŞ'a ve Muharrem ŞENGÜL'e,

Hayatım boyunca bana inanıp, her zaman arkamda duran aileme,

Hayatımın sonuna kadar hayata benimle aynı pencereden bakacağından emin olduğum eşim Erman'a,

Sonsuz teşekkürler…

Hatice Özlem ANADOL

İÇİNDEKİLER

Sayfa

JÜRİ ÜYELERİNİN İMZA SAYFASI ii
TEZ BİLDİRİMİ iii
ÖZET iv
ABSTRACT vi
TEŞEKKÜR viii
İÇİNDEKİLER ix
ÇİZELGELER LİSTESİ xi
BÖLÜM I 1
GİRİŞ 1
Problem Durumu 1
Amaç 6
Önem 7
Varsayımlar 9
Sınırlılıklar 9
BÖLÜM II 11
KAVRAMSAL VE KURAMSAL ÇERÇEVE 11
Başarı Kavramı ve Başarı Testleri 11
Dereceli Puanlama Anahtarı 14
Klasik Test Kuramı 18
Genellenebilirlik Kuramı 18
Rasch Modeli 20
İlgili Araştırmalar 22
BÖLÜM III 26
YÖNTEM 26
Araştırma Modeli 26

Çalışma Grubu 26
Veri Toplama Araç ve Teknikleri 28
Dereceli Puanlama Anahtarı (Rubrik) 28
Görüşme Formu 28
Verilerin Toplanması 29
Verilerin Çözümlenmesi 29
BÖLÜM IV 31
BULGULAR 31
BÖLÜM V 49
SONUÇLAR VE ÖNERİLER 49
KAYNAKÇA 53
ÖZGEÇMİŞ 76

ÇİZELGELER LİSTESİ

Çizelge **Sayfa**

1. Bütünsel Dereceli Puanlama Anahtarı Örneği....16
2. Analitik Dereceli Puanlama Anahtarı Örneği....17
3. Okutmanlara İlişkin Değişkenler....27
4. Deneyimi Az Grubun Puanladığı Veri Yapısı Örneği....32
5. Deneyimi Az Gruba İlişkin Hesaplanan Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri....32
6. Deneyimi Az Gruba İlişkin Hesaplanan (b:p)xm Desenine Ait Bağıl ve Mutlak Hata Varyansları....34
7. Deneyimi Az Gruba İlişkin Hesaplanan (b:p)xm Desenine Ait G ve Phi Katsayıları....34
8. Deneyimi Çok Grubun Puanladığı Veri Yapısı Örneği....35
9. Deneyimi Çok Gruba İlişkin Hesaplanan Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri....36
10. Deneyimi Çok Gruba İlişkin Hesaplanan (b:p) x m Desenine Ait Bağıl ve Mutlak Hata Varyansları....37
11. Deneyimi Çok Gruba İlişkin Hesaplanan (b:p)xm Desenine Ait G ve Phi Katsayıları....37
12. Karışık Grubun Puanladığı Veri Yapısı Örneği....38
13. Karışık Gruba ilişkin Hesaplanan Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri....39
14. Karışık Gruba İlişkin Hesaplanan (b:p)xm Desenine Ait Bağıl ve Mutlak Hata Varyansları....40
15. Karışık Gruba İlişkin Hesaplanan (b:p)xm Desenine Ait G ve Phi Katsayıları....40
16. G Çalışması Sonuçlarının Karşılaştırılması....41

BÖLÜM I

GİRİŞ

Araştırmanın bu bölümü araştırmanın problemi, amacı, önemi, varsayımları, sınırlılıkları ve kısaltmalardan oluşmaktadır.

Problem Durumu

Başarı kavramının tanımı, çeşitli biçimlerde ve birçok kez yapılmıştır. Genel anlamda başarı, istenilen sonuca ulaşma, istenileni elde etme olarak tanımlanabilir (Good, 1973). Bireyin akademik programlardan yararlanma derecesi ise, akademik başarı olarak tanımlanmaktadır. Koç (1981:7) akademik başarıyı, "Öğrencinin bulunduğu okul, sınıf ve derse göre belirlenmiş sonuçlara ulaşmada göstermiş olduğu ilerleme" olarak tanımlamaktadır. Bir başka deyişle akademik başarı; "program hedefleriyle tutarlı davranışlar bütünüdür" (Demirtaş ve Çınar, 2004). Arıcı'ya göre (2007), bir öğrenci programdaki hedef davranışları sergilediği takdirde başarılı sayılabilir.

Öğrencilerin akademik başarısını belirleyebilmek için, program hedefleri esas alınarak çeşitli başarı testleri kullanılır. Gronlund'a (1977) göre başarı testleri, öğrencinin "ne kadar" öğrendiğini belirlemede kullanılan sistematik araçlardır. Sonuçları değerlendirmeye odaklanıyor gibi görünse de aslında öğrenme sürecinin her aşamasında kullanılır ve öğrenmeye katkı sağlar. Başarı testleri, eğitsel kararların alınmasına yardımcı olmanın yanında, öğrenci motivasyonunun artmasına, öğrenmelerin kalıcılı hale gelmesine, öğrencilerin kendilerini tanımasına ve öğretim hakkında geri bildirim alınmasına yardımcı olur. Tekin (2009)'e göre başarı testleri, kişinin eğitim sürecinde edindiklerini test eder. Bu testlerin en önemli yönü bireyin geçmişte ne kadar öğrendiğini ortaya çıkarmayı hedef olarak belirlemiş olmasıdır. Başarı testi, belirli bir programa bağlı olarak hazırlanan bir öğretim süreci sonunda, öğrencilerin kat ettikleri gelişmeleri belirleme amacı ile hazırlanan ve kullanılan testlerdir. Bu testler, genel bilgi testleri olarak adlandırılır (Yıldırım, 1999). Anastasi ve Urbina'ya (1997) göre başarı testleri, eğitimi ve öğretimi geliştirmek için kullanılır.

Başarı testleri, sonuçlarının benzer gruplar ile karşılaştırabilirliği açısından öğretmen yapımı başarı testleri ve standart başarı testleri ikiye ayrılır.

Standart başarı testleri, kullanımını, uygulama koşullarını ve yorumlanmasını ayrıntılı bir şekilde açıklayan bir el kitabına sahip olan, uzman kişilerce test geliştirme aşamaları takip edilip hazırlanmış testlerdir (Koç, 1985). 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, öğrencilerin belirli bir dersteki gelişimlerini ölçmek amacı ile öğretmenlerce geliştirilmiş, kalitesi tamamen öğretmenlerin bilgi ve becerisine dayalı olan 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.

Yanıtı öğrenci tarafından yapılandırılan öğretmen yapımı açık uçlu maddelerden oluşan testlerin ve performans görevlerinin nesnel değerlendirilebilmesi için dereceli puanlama anahtarlarından faydalanılır.

Dereceli puanlama anahtarı (DPA, öğretmenler tarafından geliştirilen, öğrencilerin ürünlerini ve performanslarını detaylı bir şekilde analiz etmek amacıyla kullanılan bir puanlama tasarımıdır (Moskal, 2000). Dereceli puanlama anahtarları, öğrencilerin ürünlerinin hangi çerçevede değerlendirilmesi gerektiğini ve hangi performansa hangi düzeydeki puanın denk gelmesi gerektiğini gösteren puanlama araçlarıdır. Bu sayede öğrencileri hangi düzeyde olduklarını ve hangi düzeye ulaşmaları gerektiğini, puanlama aracı ile sistematik bir şekilde görebilmektedirler (Kutlu, Doğan ve Karakaya, 2014). Kan (2007) DPA'yı, değişik düzeydeki performanslara ait karakteristik özellikleri tanımlayan ve performanslara ilişkin değerlendirme yapmakta

kullanılan bir rehber olarak tanımlamıştır. Dereceli puanlama anahtarları, bütünsel ve analitik dereceli puanlama anahtarları olmak üzere ikiye ayrılır.

Bütünsel dereceli puanlama anahtarı, ürünlerin bir bütün olarak değerlendirildiği puanlama anahtarlarıdır ve öğrencilerin performanslarının bütününe ilişkin tek bir puan vermeyi sağlar. Puanlama esnasında, performans düzeyleri yüksekten düşüğe doğru sıralanır ve her düzeye bir puan verilir. Bu düzeyler, sözel ifadelerle de detaylı bir şekilde tanımlanabilir. Ancak, bütünsel puanlama anahtarı ile öğrenci performanslarının alt boyutlarının değerlendirilme olanağı olmadığı için, öğrenci hakkında dereceli puanlama anahtarı kullanılan durumlar kadar detaylı bilgi edinilemez (Turgut, 1997).

Analitik dereceli puanlama anahtarı, öğrenci ürünlerini değerlendirmek için kullanılan ve bir takım ölçütlerden oluşan bir puanlama aracıdır (Popham, 1997). Analitik dereceli puanlama anahtarı, 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 nedenle bu çalışmada analitik dereceli puanlama anahtarı kullanılmıştır.

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. Kan (2007), açık uçlu testlerin bir yanıt anahtarı ile puanlanmasının, güvenilir bir puanlama için oldukça önemli bir yere sahip olduğunu belirtmiştir. Dereceli puanlama anahtarlarının kullanımı özellikle puanlayıcılar arası güvenirlik bağlamında önemli katkılar sağlasa da gerçekleştirilen puanlama işleminin güvenirliğinin psikometrik olarak test edilmesi gerekmektedir.

Birden fazla doğru yanıtın olabileceği, öğrencilerin üstü düzey zihinsel süreçlerini ölçmeyi amaçlayan yanıtlarını öğrencilerin yapılandırdığı açık uçlu maddelerden oluşan testlerin veya öğrencilerin özgün bir ürün ortaya koyduğu veya etkinlik gerçekleştirdiği performans görevlerinin güvenirliğinin belirlenmesinde klasik test kuramına, madde tepki kuramına ve genellenebilirlik kuramına dayalı çeşitli yöntemler mevcuttur.

Klasik test kuramı, bireyin psikolojik bir ölçme aracından aldığı puanın gözlenen puan olduğunu ve bu puanın gerçek puan ve hata puanından oluştuğunu varsaymaktadır (Baykul, 2000). Klasik test kuramında bireyler, ölçme aracına verdikleri tepkiler ile değerlendirilirler. Diğer bir ifade ile bireylerin ölçme aracından aldıkları puanlar,

onların ölçülmek istenen özelliğe sahip oluş düzeylerini gösterir. Klasik test kuramında hatanın değişik kaynaklardan geldiği varsayılır ve güvenirlik kestirimleri, hatanın geldiği düşünülen kaynağa göre değişik isimler alır. Eğer hatanın değişik puanlayıcılardan kaynaklandığı düşünülüyor ise, puanlayıcılar arası güvenirlikten bahsedilir. Puanlayıcılar arası güvenirlik farklı teknikler kullanarak test edilebilmektedir (Goodwin, 2001).

KTK ile bir seferde bir hata türüne ilişkin güvenirlik katsayısı hesaplanabilmektedir. Oysa değişik kaynaklardan gelebilecek hataları ayrıştırarak hesaplayabilmek, hata kaynakları hakkında daha ayrıntılı bilgi sağlayacaktır. Özellikle, puanlayıcıların puanlama sürecine katıldığı durumlarda, puanlayıcı hatalarının dikkate alınması, ölçme güvenilirliği açısından önemlidir (Crocker ve Algina, 1986). Genellenebilirlik kuramı ile güvenirlik yüzey (facet) olarak tanımlanan bütün değişkenler (zaman, maddeler, puanlayıcılar) dikkate alınarak hesaplanabilmekte ve birbirleri ile olan etkileşimleri dikkate alınabilmektedir. Tüm değişkenlerin birbirleri ile olan kombinasyonları tanımlanmaktadır (Shavelson & Webb, 1991). Genellenebilirlik (G) kuramı varyans analizini (ANOVA) temel alır. Genellenebilirlik analizi ile farklı varyans kaynaklarının, toplam varyanstaki oranı, dolayısıyla; değişkenlik kaynaklarının, toplam varyansın yüzde kaçını açıkladığı bulunur (Brennan, 2001). Genellenebilirlik kuramında, Genellenebilirlik (G) ve Karar (K) olmak üzere iki farklı çalışma yer almaktadır.

G çalışmaları, ölçme hatalarının hangi değişkenlik kaynağından kaynaklandığını belirlemek için düzenlenmektedir. G çalışması için öncelikle ölçme örnekleminin genelleneceği evren ve değişkenlik kaynakları tespit edilmelidir. G çalışmalarında farklı hata kaynakları kullanılarak oluşturulan farklı desenler bulunmaktadır. Bunlar çaprazlanmış ve yuvalanmış desenlerdir. Değişkenlik kaynaklarının özeliklerine göre ise tesadüfi ve sabit etkiden söz edilebilir (Crocker ve Algına, 1986).

K (karar) çalışmalarında, G çalışmasından elde edilen veriler üzerinde çalışılarak, değişkenlerden kaynaklanan hataların en aza indirilmesi amaçlanır. Böylece, hangi koşulda maksimum güvenirliğin sağlanabileceği bulunmuş olur (Crocker ve Algina, 1986).

G kuramına bağlı olarak bağıl ve mutlak olmak üzere iki tür değerlendirmeden söz edilebilir. Bağıl değerlendirme, verilerin toplandığı ve üzerinde çalışma yapılan grubun normlarına göre yapılırken, mutlak değerlendirmede önceden belirlenen bir ölçüt vardır ve değerlendirmeler bu ölçüte göre yapılır. Dolayısıyla, G çalışmaları ile

incelenen ölçme hataları, bağıl ve mutlak kararlar için ölçme hataları olmak üzere ikiye ayrılır (Güler, Uyanık ve Teker, 2012). Bu çalışmada bahsedilen avantajlarından dolayı genellenebilirlik kuramından faydalanılmıştır.

Madde Tepki Kuramına dayalı olarak puanlayıcılar arası güvenirliği hesaplamak için Rasch Modeli kullanılabilir. Bu model madde güçlüğünün yeterli, madde ayırt ediciliğinin eşit, şans başarısının ise sıfır olduğunu varsayar ve maddelerin güçlük düzeyi ve bireylerin yetenekleri aynı anda tanımlanmaya çalışır. Bireylerin tepkilerinin birden çok puanlayıcı tarafından puanlanması durumunda, farklı değişkenlik kaynaklarının dikkate alınması gerekir. Birincisi bireyin sahip olduğu yetenek, ikincisi bireylerin yeteneklerini yansıtabilecekleri performans görevi, üçüncüsü puanlayıcılar, dördüncüsü maddeler, beşincisi ise puanlama ölçeğidir. Çok Değişkenlik Kaynaklı Rasch Modeli ile bu beş değişkenlik kaynağının etkileşimi ile bireyin performansı kestirilebilir (Crocker ve Algina, 1986).

Yapılandırıcı yaklaşım ile öğrencilerin üst düzey zihinsel becerilerini ölçmeyi amaçlayan, okul öğrenmelerini gerçek yaşam durumları ile ilişkilendiren, yanıtını öğrencilerin yapılandırdığı açık uçlu sorular ve performansa dayalı değerlendirmeler ön plana çıkmıştır. Bu tip değerlendirme süreçlerinde puanlayıcılar önemli bir yer tutmaktadır. Puanlayıcılar arası güvenirlik, bireyler hakkında güvenilir kararlar verme açısından çok önemlidir.

Ancak özellikle dereceli puanlama anahtarlarının kullanıldığı durumlarda puanlayıcılar arası güvenirlik ile ilişkili olan bir boyut da puanlayıcı özellikleridir. Puanlayıcı özellikleri güvenirlik üzerinde önemli bir etkiye sahiptir. Bu noktada deneyimi çok, alanında uzman puanlayıcıların kullanılması güvenirliği artıracak bir etmen olabilmektedir. Özellikle son yıllarda merkezi sınavlarda açık uçlu soruların 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. Ancak pratik nedenlerden dolayı birebir benzer özelliklere sahip puanlayıcılara ulaşmak mümkün olamamaktadır. Bu nedenle puanlama sürecinde kullanılan dereceli puanlama anahtarlarının farklı özelliklere sahip puanlayıcıların benzer şekilde puanlama yapmasına katkı sağlaması gerekmektedir.

Bu süreçte dikkate alınabilecek puanlayıcı özelliklerin arasında ise puanlayıcıların deneyim yılı düşünülebilir. Dereceli puanlama anahtarlarının kullanımı

belli bir donanım gerektirmekle beraber deneyim yılı çok ve az olan puanlayıcıların benzer bir şekilde puanlama yapmalarını sağlamalıdır.

İlgili literatürde farklı kuramlara dayalı yöntemler ile puanlayıcılar arası güvenirliğin veya dereceli puanlama anahtarlarının güvenirliğinin test edildiği çalışmalar yer almaktadır. Tüm çalışmaların ortak sonucu olarak, DPA'nın değerlendirme yöntemi olarak kullanılması ile değerlendirme sonuçları açısından güvenilir sonuçların ortaya çıktığı görülmüştür. Ancak, puanlayıcıların deneyim yıllarına dayalı olarak puanlama aracının güvenirliğine ilişkin karşılaştırmanın yapıldığı veya puanlayıcıların deneyim yıllarının güvenirlik katsayısına olan etkisinin incelendiği bir çalışmaya rastlanmamıştır. Bu bağlamda dereceli puanlama anahtarlarının güvenirliğinin deneyimi az ve çok olan puanlayıcıların kullanıldığı durumlarda genellenebilirlik kuramına dayalı olarak karşılaştırılması alandaki eksikliği giderecek ve göz ardı edilen bu konuya dikkat çekecektir. Bunun yanı sıra alan yazında puanlama sürecine ilişkin puanlayıcıların görüşlerinin incelendiği nitel bir çalışmaya rastlanmamıştır. Bu çalışma ile elde edilen nicel verilerin puanlayıcı görüşlerine dayalı nitel veriler ile desteklenmesi probleme ilişkin kapsamlı bir belirleme yapılmasına katkı sağlayacaktır.

Amaç

Bu araştırmanın amacı, Ankara'da bulunan bir vakıf üniversitesinde yabancı diller bölümünde verilen yazma dersine yönelik hedef davranışlar üzerinden beceri değerlendirilmesinde kullanılan dereceli puanlama anahtarının güvenirlik düzeyinin, puanlama sürecine ilişkin deneyimi az ve çok olan puanlayıcılardan elde edilen sonuçlar bağlamında karşılaştırılmasıdır.

Bu doğrultuda aşağıdaki sorular cevaplanmıştır:

1. İngilizce yazılı anlatım becerisi, puanlama sürecine yönelik deneyimi az olan İngilizce okutmanları tarafından puanlandığında, dereceli puanlama anahtarına ilişkin hesaplanan
    - varyans bileşenleri
    - bağıl ve mutlak hata varyansları
    - G vePhi katsayısı nasıldır?

2. İngilizce dersi yazılı anlatım becerisi, puanlama sürecine yönelik deneyimi çok olan İngilizce okutmanları tarafından puanlandığında dereceli puanlama anahtarına ilişkin hesaplanan
    - varyans bileşenleri
    - bağıl ve mutlak hata varyansları
    - G ve Phi katsayısı nasıldır?

3. İngilizce dersi yazılı anlatım becerisi, puanlama sürecine yönelik deneyimi az ve çok olan İngilizce okutmanlarının birlikte yer aldığı grup tarafından puanlandığı durumlarda,
    - varyans bileşenleri
    - bağıl ve mutlak hata varyansları
    - G ve Phi katsayıları farklılaşmakta mıdır?

4. Çalışmaya katılan üç farklı puanlayıcı grubuna ilişkin hesaplanan
    - varyans bileşenleri
    - bağıl ve mutlak hata varyansları
    - G ve Phi katsayıları karşılaştırıldığında, sonuç nasıldır?

5. İngilizce yazılı anlatım dersi dereceli puanlama anahtarının kullanım sürecine ilişkin çalışmada yer alan puanlayıcıların görüşleri nelerdir?

Önem

Özellikle son yıllarda merkezi sınavlarda açık uçlu soruların yer almasına ilişkin denemeler yapılmaktadır ve puanlama sürecinde çok sayıda puanlayıcı görev almaktadır. Bunun yanı sıra, eğitimin her kademesindeki sınıf içi değerlendirme süreçlerinde de birden fazla puanlayıcının yer alması söz konusudur. Birden fazla puanlayıcının kullanıldığı durumlarda, güvenirliğe etki eden boyutlardan bir tanesi de puanlayıcı özellikleridir. Puanlayıcıların sahip olduğu bireysel farklılıklar, puanlama davranışı üzerinde etkiye sahip olabilir. Bu süreçte dikkate alınabilecek puanlayıcı özelliklerinin başında, puanlayıcıların DPA kullanmaya ilişkin deneyimleri olabilir. Pratik nedenlerden dolayı tecrübe açısından aynı özelliklere sahip puanlayıcılara ulaşmak mümkün olamayabilir. Dolayısıyla, DPA kullanmaya ilişkin farklı deneyim

yıllarına sahip puanlayıcıların, aynı grubu puanlaması kaçınılmazdır. Ancak nitelikli bir puanlama anahtarının, farklı deneyime sahip puanlayıcıların benzer puanlama yapmasına katkı sağlaması beklenir. Alan yazında çeşitli kuramlara dayalı olarak, açık uçlu soruların veya dereceli puanlama anahtarlarının güvenirliğinin hesaplandığı çalışmalar olmakla beraber, DPA kullanımına ilişkin farklı deneyime yıllarına sahip puanlayıcıların kullanıldığı durumlarda güvenirlik katsayılarının karşılaştırıldığı çalışmalara rastlanmamıştır. Bu anlamda, araştırma bulgularının ölçme değerlendirme alan yazınındaki bu eksikliği gidermesi ve puanlayıcı özelliklerinin güvenirlik üzerindeki etkisine dikkat çekmesi beklenmektedir.

Alan yazın incelendiğinde açık uçlu soruların, performans görevlerinin ve dereceli puanlama anahtarlarının güvenirliğinin, genellenebilirlik kuramına dayalı olarak incelendiği çalışmalar yer almakla beraber bu çalışmaların büyük çoğunluğunda tümüyle çaprazlanmış desenlerin kullanıldığı görülmektedir. Bu çalışmada birey ve puanlayıcıların maddeler ile çaprazlandığı ancak bireylerin puanlayıcılara yuvalandığı desenden ((b:p)xm) faydalanılacaktır. Çalışma bu özelliği ile alandaki benzer çalışmalardan farklılaşmaktadır.

Alan yazında yer alan benzer çalışmalar daha çok nicel boyut içermektedir. Bu çalışmada puanlayıcı olarak görev alan okutmanların görüşlerine yer verilmesinin çalışmaya daha geniş bir bakış açısı katması uygulayıcılara ve araştırmacılara önemli bulgular sunması beklenmektedir.

Ülkemizde merkezi sınavlarda kullanılmaya başlayan açık uçlu soruların puanlanması aşamasında, farklı özelliklere sahip puanlayıcıların aynı puanlayıcı grubu içinde yer alıp, beraber puanlama yapması söz konusu olmuştur. Alan yazında, puanlayıcı özelliklerinin puanlayıcı davranışları üzerindeki etkisini inceleyen bir çalışmaya rastlanmamıştır. Bu bağlamda, bu çalışma, ileriki uygulamalara yol gösterecektir.

Bu çalışmanın yürütüleceği üniversitenin hazırlık okulunda, İngilizce yazılı anlatım becerisini değerlendirmek amacıyla kullanılan DPA'nın güvenirliğine yönelik bir çalışma bulunmamaktadır. Her sene başında yapılan toplantılarda tartışmaya açılan ve gerekli görüldüğü takdirde revize edilen DPA'nın geçerlilik ve güvenirlik açısından, sistematik bir değerlendirmesi yapılmamıştır. İlgili üniversitenin hazırlık okulunda kullanılan İngilizce yazılı anlatım becerisini değerlendirmek için kullanılan DPA'nın geçerliliği ve güvenirliğine dair kanıtlar toplamak ve öğretim elemanlarının görüşleri ışığında söz konusu DPA'nın güçlü ve zayıf yönlerini belirlemek İngilizce yazılı

anlatım becerisine ilişkin daha nitelikli değerlendirmeler yapılmasına ve okutmanların mesleki gelişimlerine katkı sağlanacaktır.

Varsayımlar

Yapılacak olan araştırmada aşağıdaki varsayımlardan hareket edilmiştir:

- Puanlayıcı öğretim elemanları, puanlama işlemini birbirlerinden bağımsız olarak yapmışlardır ve bu süreçte puanlama işlemi ile ilgili bilgi paylaşımında bulunmamışlardır.
- Öğretim elemanları görüşme sorularına içtenlikle yanıt vermiştir.

Sınırlılıklar

Çalışma,

- Tek bir görev ile,
- İngilizce yazılı anlatım becerisine yönelik hazırlanan yazılı yoklamada kullanılan metinlerle,
- Çalışmada kullanılan, bireylerin puanlayıcılara yuvalandığı [(b:p)xm]desen ile sınırlıdır.

Kısaltmalar Semboller ve Tanımlar

G Kuramı	: Genellenebilirlik Kuramı
G Çalışması	: Genellenebilirlik Çalışması
K Çalışması	: Karar Çalışması
G Katsayısı	: Genellenebilirlik Katsayısı
Phi Katsayısı	: Güvenirlik Katsayısı
$\sigma^2(\delta)$	: Bağıl Hata Varyansı
$\sigma^2(\Delta)$	: Mutlak Hata Varyansı
“:”	: Yuvalama
b	: Birey
p	: Puanlayıcı
m	: Madde
(b:p)xm	: Birey: puanlayıcı x madde ortak etkisi
Dereceli Puanlama Anahtarı	: Analitik dereceli puanlama anahtarı

Deneyim Yılı:	:Puanlayıcıların dereceli puanlama anahtarı kullanma sıklıkları
Deneyimi az olan puanlayıcılar	:Uygulamanın yapıldığı kurumda en fazla bir yıldır görev yapan ve en fazla 18 kez puanlama yapan puanlayıcılar
Deneyimi çok olan puanlayıcılar	: Uygulamanın yapıldığı kurumda beş yıl ve üzeri yıldır görev yapan ve en az 90 puanlama yapan puanlayıcılar

BÖLÜM II

KAVRAMSAL VE KURAMSAL ÇERÇEVE

Bu bölümde, başarı kavramı, dereceli puanlama anahtarı, KTK, Rasch modeli, G kuramı ve ilgili araştırmalar yer almaktadır.

Başarı Kavramı ve Başarı Testleri

Eğitim sürecinde bireylerin geleceğine ve akademik hayatına dair pek çok karar verilmekte ve bu kararlar bireylerin ileriye yönelik kendi kararlarını almalarına yardımcı olmaktadır. Eğitimciler, bireyler hakkında bilgi toplamak ve toplanan bilgiler ışığında söz konusu kararları verebilmek için, çeşitli ölçme araçları kullanmaktadır.

Okullarda uygulanan ölçme değerlendirme etkinlikleri, araştırmacılar tarafından yıllardır araştırılan ve üzerinde çalışmalar yapılan bir konudur. Yürütülen ölçme değerlendirme etkinlikleri sonucu toplanan bilgiler, öğrencilerin hâlihazırda ne yaptıkları, gelecekte ne yapabilecekleri ya da gelecekte ne yapmaları gerektiği ile ilgili konuları kararlaştırmada kullanılabilen bilgilerdir (Tekin, 2009).

Özçelik (2011)'e göre değerlendirme bir karar verme işlemidir. Diğer bir deyişle, ölçme sonucunu önceden belirlenen bir ölçüt ile karşılaştırma yolu ile ölçülmek istenen özellik hakkında karara varma işlemidir. Değerlendirme yapılırken, öğrenciyi tanıma ve yerleştirme, öğrenme eksiklerini saptama, bunlara yol açan güçlükleri belirleyerek bunları ortadan kaldırma ve öğrenme düzeyini belirleme amaçlanmaktadır. Değerlendirme, Tekin (2009)'e göre ise bir yargılama işlemidir ve ölçme sonucunun ve ölçütün karşılaştırılmasına dayanır. Değerlendirme, ölçme sonuçlarını yorumlayıp, ölçülen özellikler hakkında bir sonuca varma ve öğrencinin başarısı hakkında karar verme işlemidir.

Öğrenci başarısı denince ilk akla gelen şey akademik başarıdır ve bireyin akademik programlardan yararlanma derecesi olarak tanımlanmaktadır. Öğrencilerin akademik başarısını belirleyebilmek için çeşitli başarı testleri kullanılır. Başarı testleri, eğitsel kararların alınmasına yardımcı olmanın yanında, öğrenci motivasyonunun artmasına, öğrenmelerin kalıcı hale gelmesine, öğrencilerin kendilerini tanımasına ve öğretim hakkında geri bildirim alınmasına yardımcı olur. Tekin (2009)'e göre başarı

testleri, kişinin aldığı eğitim sonrası edindiklerini ölçen testlerdir. Bu testlerin en önemli yönü bireyin geçmişte ne kadar öğrendiğini ortaya çıkarmayı hedef olarak belirlemiş olmasıdır. Başarı testleri, öğrencilerin eğitim öğretim süreci sonucunda sağladıkları gelişmeyi belirlemek amacı ile kullanılan testlerdir. Bu testler, genel bilgi testleri olarak adlandırılır (Yıldırım, 1999). Anastasi ve Urbina'ya (1997) göre başarı testleri, eğitimi ve öğretimi geliştirmek için kullanılır. Başarı testleri, sonuçlarının benzer gruplar ile karşılaştırabilirliği açısından öğretmen yapımı başarı testleri ve standart başarı testleri olmak üzere ikiye ayrılır.

Standart başarı testleri bir el kitabına sahip olan, uzman kişilerce hazırlanmış testler iken, öğretmen yapımı başarı testleri, öğretmenlerce hazırlanmış ve el kitabına sahip olmayan testlerdir. Diğer bir ifade ile öğretmen yapımı başarı testleri, öğretmenlerin sınıf içinde uyguladıkları ve kalitesi kendi bilgi ve becerilerine bağlı olan başarı testleridir (Arıkan vd., 2012). Öğrenci başarısı, Turgut ve Baykul'a (2014) göre:

- Yazılı Yoklamalar
- Kısa Cevap Gerektiren Testler
- Sınıflama Gerektiren Testler
- Seçme Gerektiren Testler
- Sözlü Sınavlar
- Performans değerlendirme ile ölçülebilir.

Yazılı yoklama, öğrencilerin az sayıda soruyu, düşüncelerini organize ederek ve belli bir süre içinde açıklayarak cevaplamaları gereken test türüdür. Yazılı yoklamalarda tek bir doğru cevap bulunmaz ve cevapların tasarlanması öğrenciye bırakılır. Bloom'un bilişsel taksonomisi, "bilgi, kavrama, uygulama, analiz, sentez ve değerlendirme" olmak üzere altı bilişsel basamaktan oluşmaktadır (Bloom vd., 1956). Öğrencilerin düşüncelerini organize etmesine olanak veren yazılı yoklamalar ile analiz, sentez gibi üst düzeyde zihinsel işlemleri ölçmek mümkündür. Yazılı yoklamalarda, cevapların organize edilmesi tamamen öğrenciye bırakılabildiği gibi, onlardan, belirlenmiş bir cevap vermeleri de istenebilir. Bu bağlamda yazılı yoklamalarda kullanılan madde türleri, "sınırlı cevap soruları ve serbest cevap soruları" olmak üzere ikiye ayrılır (Turgut,1997).

Sınırlı cevap sorularında, öğrenciler düşüncelerini örgütlerken bazı sınırlamalar ile karşılaşırlar. Örneğin, cevap verirken bazen tanımlama yapmaları, bazen madde madde açıklama yazmaları, bazen de nedenleri sıralamaları gerekebilir. Bu bağlamada,

cevapların puanlanması, daha nesnel olur. Ancak, sınırlandırmalar, en üst bilişsel düzey olan değerlendirme düzeyinde davranışların ölçülmesinde yetersizdir. Serbest cevap sorularında cevapların organize edilmesi, bilgilerin örgütlenmesi, seçilen olgusal bilgiler, kullanılacak sayfa sayısı tamamen öğrenciye bırakılır. Zaman sınırı olsa da, bu bir kısıtlama olarak sayılmaz, içerikteki söz hakkı tamamen öğrenciye aittir. Bu sayede, en üst düzey bilişsel beceriler bile değerlendirilebilir (Turgut, 1997).

Cevapları bir kelime, bir cümle veya birkaç madde gibi uzun açıklamalar gerektirmeyen maddelerden oluşan testlere kısa cevaplı testler denir. Cevap en fazla bir cümle olmalıdır. Kısa cevaplı testlerin yararları şunlardır:

- Test maddelerini cevapla süresi kısa olduğundan, çok maddeli testler hazırlanabilir. Bu da sınavın kapsam geçerliğini yükseltir.
- Puanlaması açık uçlu maddelerden oluşan testlere göre daha objektif olduğundan, geçerlik ve güvenirlik yüksektir.
- Şans faktörü yok denecek kadar azdır.
- Madde sayısı açık uçlu maddelerden oluşan testlere göre daha fazla olduğu için testin genelinde daha fazla sayıda davranış dikkate alınabilmektedir.

Cevaplayıcıların maddeleri belli ölçütlere göre sınıflaması gereken testler, sınıflama gerektiren testler olarak tanımlanır. En yaygın olanı doğru-yanlış testleridir. Bu testlerde maddeler doğru-yanlış ifadeler şeklindedir ve cevaplayıcılardan verilen ifadelerin doğruluğuna ya da yanlışlığına karar vermeleri istenir. Turgut (1997) doğru yanlış testleri ile üst düzey bilişsel davranışların ölçülebilmesinin ve puanlamasının güvenilir bir şekilde yapılabilmesinin, doğru-yanlış türü test maddelerinin kullanılma sebeplerinden olduğunu savunmaktadır. Ancak, doğru-yanlış testlerinin eleştirildiği bir nokta vardır. Öğrenciler, ölçülmek istenen özelliğe sahip olmasalar bile, doğru cevabı şans başarısı ile bulabilmektedirler. Şans başarısının çok yüksek olması, geçerliliği ve güvenirliği zedeleyebilmektedir.

Öğrenciye, her soru ile birlikte sorunun doğru cevabı ve doğru cevap sanılabilecek olan ifadeler verilen ve öğrenciden, seçenekler arasından sorunun doğru cevabını bulması istenen testler, seçme gerektiren (çoktan seçmeli) testlerdir. Çoktan seçmeli testlerde cevapları öğrenciler yazmaz. Belirli bir zaman diliminde, daha fazla hedef davranış değerlendirilebilir. Dolayısıyla, çoktan seçmeli testlerin kapsam geçerliliği, diğer test türlerinden daha yüksektir. Buna karşı, öğrencilerin sentez düzeyindeki davranışlarını yoklamak çok zordur. Soru sayısının artması da, okuma hızı

ve yorgunluk gibi bağımsız değişkenlerin değerlendirmeye karışmasına neden olur. Çoktan seçmeli testlerde, doğru cevabın şansla bulunması olasılığını tamamen ortadan kaldırmak mümkün değildir. Bu özellik çoktan seçmeli testlerin en çok eleştiri alan kısmıdır (Baykul, 2000).

Sözlü sınavlar, öğrencilere sözlü olarak soru sorulan ve öğrencilerin sorunun cevabını sözlü biçiminde ifade ettiği sınavlardır. Sözlü sınavlar, önceki öğrenmeleri değerlendirmek için öğretmen dersi anlatmaya başlamadan önce yapılabildi gibi, ders sonunda, o günkü hedef davranışa ilişkin de yapılabilir, öğrencilerden alınan cevaplara göre şekillenir. Sözlü sınavlarda, bir defada bir öğrenci değerlendirilebilme şansı olduğundan, zaman açısından sıkıntı yaşanabilmektedir. Kısıtlı bir zaman diliminde az sayıda soru sorulması, dersin sadece birkaç kazanımına dair soru sorma imkânı verir. Farklı öğrencilere yöneltilen farklı sorularla, öğrencilerin öğrenme düzeyleri karşılaştıramaz. Bu da sözlü sınavların geçerliliğini ve güvenirliğini etkileyebilir (Özçelik, 2011).

Performans; değerlendirildiği ölçütler çerçevesinde, bir görevin yerine getirilmesi ve amacın gerçekleştirilmesidir (Kutlu, Doğan ve Karakaya, 2008). Performans değerlendirme ise "öğrencinin çeşitli problem durumları karşısında gösterdiği tepkilerin teşhis edilerek eksiklik ve yeterliklerin belirlenmesini sağlayan bir değerlendirme biçimi" (Önal, 2005:15); olarak tanımlanmaktadır. Performans değerlendirilirken hem süreç hem de ürün değerlendirilmektedir. Böylelikle hem öğrenmeler pekiştirilmiş hem de öğrencilerin bilgiyi nasıl kullandıkları görülmüş olur (Brualdi, 1998).

Performans değerlendirmede öğrenciyi güvenilir bir şekilde puanlamak, performans değerlendirmenin en önemli kısmıdır. Öğrenci performanslarının birden çok puanlayıcı tarafından değerlendirilmesi söz konusu olduğunda, puanlayıcı güvenirliği de ön plana çıkmaktadır. Bunun için, öğrenciden beklenen davranışlar öğretmenler tarafından önceden belirlenmeli ve bu davranışlara göre kontrol listesi veya dereceli puanlama anahtarı hazırlanmalıdır (Moskal, 2003).

Dereceli Puanlama Anahtarı

Dereceli puanlama anahtarı, öğretmenler tarafından geliştirilen, öğrencilerin ürünlerini ve performanslarını detaylı bir şekilde analiz etmek amacıyla kullanılan, yatay ve dikey olmak üzere iki eksenden oluşan puanlama tasarımlarıdır. Dikey eksende

performans kriterleri, yatay eksende performans düzeyleri vardır. Yatay ve dikey eksenlerin kesiştiği noktada ise performans düzeyine ilişkin tanımlar ve puanlar bulunur. Performans düzeyleri "3, 2, 1" gibi rakamla belirtileceği gibi, sözel şekilde de ifade edilebilir. (Goodrich, 2001). 5 düzeyli bir DPA'da 1 ve 2 düşük, 3 orta 4 ve 5 ise yüksek başarıyı ifade etmek için kullanılır. Bu sayısal puanlar toplanır ve bir toplam puan elde edilir. Diğer bir ifade ile dereceli puanlama anahtarları, öğrenci performanslarının değerlendirileceği ölçütleri ve hangi performansın kaç puan olması gerektiğini gösteren puanlama araçlarıdır. Bu yolla öğrenciler, aldıkları puanların sonucunda seviyelerini ve hangi seviyeye ulaşmaları gerektiğini görebilir ve gelişimlerini takip edebilirler (Aslanoğlu ve Kutlu, 2003). Goodrich (2001) DPA hazırlanırken izlenmesi gereken yolları şu şekilde belirtmiştir:

- Ölçüt listelerinin belirlenmesi: Ölçütlerin belirlenmesi sayesinde puanlayıcıdan puanlayıcıya değişmeyen nesnel bir değerlendirme yapılabilir ve öğrenci gelişimi hakkında isabetli kararlar verilebilir.
- Dereceli puanlama anahtarının türünün belirlenmesi: Amaca uygun olarak analitik ya da bütünsel DPA'dan yararlanılabilir.
- Performans düzeylerinin belirlenmesi: Her bir performans düzeyine ilişkin ayrıntılı tanımlamalar yapılarak, performans düzeylerinin sınırları belirlenebilir
- DPA ile ilgili uzman görüşlerinin alınması: DPA oluşturulduktan sonra, ölçme değerlendirme uzmanından, ilgili alan öğretmenlerinden ve dilbilgisi uzmanlarından dönüt alınabilir.

DPA hazırlarken öğretmenlerin karşılaştığı problemlerden biri, puanlama anahtarı hazırlama konusunda yeterince bilgi sahibi olmamalarıdır (Sağlam-Arslan, Devecioğlu-Kaymakçı, Arslan, 2009) Montgomery (2000), öğretmenlerin DPA hazırlarken karşılaştıkları problemlerden birinin, öğrencilere dönüt vermeleri gereken durumlarda, öğrencilerin değerlendirme ölçütlerini anlamamaları olduğunu söylemiştir. Karşılaşılan problemlerden bir diğeri ise öğrencilerin toplam puanlarının nasıl hesaplandığı anlamamalarıdır. Bunlara çözüm olarak, ölçütlerin anlaşılabilir bir dille yazılması ve toplam puanın nasıl hesaplandığının öğrencilere detaylı bir şekilde açıklanması verilebilir.

Hazırlarken karşılaşılan problemlerin yanı sıra, DPA kullanmanın, öğretmenler ve öğrenciler açısından pek çok avantajı vardır. Bunlardan biri, öğrencilerin performanslarının nasıl değerlendirileceğine ve ileriki çalışmalarını nasıl biçimlendirmeleri gerektiğine dair bilgi sahibi olmalarıdır. Öğrenciler güçlü ve zayıf yönlerini hakkında bilgi sahibi olurlar ve gelişim süreçlerini izleyebilirler. DPA kullanmanın diğer bir avantajı, öğretmenlerin değerlendirme süresini kısaltması ve öğretmenlere nesnel değerlendirme olanağı sağlamasıdır (Aslanoğlu ve Kutlu, 2003). Popham (1997) dereceli puanlama anahtarlarını, bütünsel ve analitik dereceli puanlama anahtarları olmak üzere ikiye ayırmaktadır:

Bütünsel DPA, performansların bir bütün olarak değerlendirildiği, performanslarının bütününe ilişkin tek bir puan verildiği ve alt boyutlarının ilişkin değerlendirilme olanağı sağlamayan anahtarlardır.

Çizelge 1

Bütünsel Dereceli Puanlama Anahtarı Örneği

Performans Düzeyi/ Performans Tanımları	**Performans Tanımları**
5 (Mükemmel)	Problemin tamamen anlaşıldığını gösterir. Cevaplar ya da ürün performans görevine ilişkin bütün gerekleri içermektedir.
4 (Başarılı)	Problemin önemli ölçüde anlaşıldığını gösterir. Cevaplar ya da ürün performans görevine ilişkin gerekleri içermektedir.
3 (Gelişmekte)	Problemin kısmen anlaşıldığını gösterir. Cevaplar ya da ürün performans görevine ilişkin çoğu gereği karşılamaktadır
2 (Başlangıç)	Problemin çok az anlaşıldığını gösterir. Cevaplar ya da ürün performans görevine ilişkin çoğu gereği karşılamaktan yoksundur
1 (Başarısız)	Problemin tamamıyla anlaşılmadığını gösterir.
0 (Yetersiz)	Cevap ya da çözüm yok/ performans görevi yapılmamış.

(Kan, 2007: 138)

Analitik dereceli puanlama anahtarı ise performansın parçalarının tek tek puanlanmasını en son toplam bir puan elde edilmesini, dolaysıyla alt boyutların tek tek değerlendirilmesini sağlayan puanlama anahtarlarıdır (Moskal, 2000).

Çizelge 2

Analitik Dereceli Puanlama Anahtarı Örneği

KRİTERLER	PERFORMANS DÜZEYLERİ			
	BAŞLANGIÇ	**GELİŞMEKTE**	**BAŞARILI**	**MÜKEMMEL**
ÖLÇÜT 1	Probleme ilişkin tanımlama ve açıklamalar başlangıç düzeyinde bir performansı yansıtıyor. Oldukça yetersiz.	Probleme ilişkin tanımlama ve açıklamalar, problemi çözmeye ilişkin çaba ve girişimi yansıtıyor. Fakat önemli hatalar var.	Probleme ilişkin tanımlama ve açıklamalar, başarılı düzeyde performans yansıtıyor. Göz ardı edilebilecek hatalar var.	Probleme ilişkin tanımlama ve açıklamalar, üst düzeyde performans yansıtıyor. Hata yok.
ÖLÇÜT 2	Probleme ilişkin tanımlama ve açıklamalar başlangıç düzeyinde bir performansı yansıtıyor. Oldukça yetersiz	Probleme ilişkin tanımlama ve açıklamalar, problemi çözmeye ilişkin çaba ve girişimi yansıtıyor. Fakat önemli hatalar var.	Probleme ilişkin tanımlama ve açıklamalar, başarılı düzeyde performans yansıtıyor. Göz ardı edilebilecek hatalar var.	Probleme ilişkin tanımlama ve açıklamalar, üst düzeyde performans yansıtıyor. Hata yok.

(Kan, 2007: 139)

Yanıtı öğrencinin yapılandırdığı testlerin güvenilir bir şekilde değerlendirilmesi, öğrencilere sağlıklı dönüt vermek, dolayısıyla gelecekte sağlıklı karar almalarına yardımcı olmak açısından gereklidir. Bu süreçte, dereceli puanlama anahtarlarının yeri çok önemlidir. Dereceli puanlama anahtarları ayrıca, sübjektif puanlanan açık uçlu testlerin objektif olarak puanlanabilmesi ve puanlayıcılar arası güvenirlik sağlanması açısından yararlanılması gereken puanlama tasarımlarıdır (Kan, 2007).

Yanıtlarını öğrencilerin yapılandırdığı açık uçlu soruların ya da performans görevlerinin güvenirliğinin hesaplanmasında klasik test kuramına, madde tepki kuramına ve genellenebilirlik kuramlarına dayalı kullanılabilecek çeşitli yöntemler mevcuttur.

Klasik Test Kuramı

Klasik test kuramı (KTK), bireyin psikolojik bir ölçme aracından aldığı puanın gözlenen puan olduğu ve bu puanın gerçek puan ve hata puanından oluştuğunu varsaymaktadır. Klasik test kuramında bireyler, ölçme aracına verdikleri tepkiler ile değerlendirilirler. Diğer bir deyiş ile bireylerin ölçme aracından aldıkları puanlar, onların ölçülmek istenen özelliğe sahip oluş düzeylerini gösterir.

Klasik test kuramında, hatanın tesadüfi olduğu varsayılır. Bir yanıtlayıcının bir testteki bir maddeye verdiği cevaplar (X), bireyin madde ile ölçülmek istenen özelliğe sahip olma derecesi (T) ve hata (E) ile ifade edilir. KTK ile ölçülmek istenen özellik gerçek puandır. Fakat ölçmeye karışabilecek çeşitli hatalar yüzünden, gerçek değerler çoğu zaman elde edilemez ve bazı varsayımlar ile gözlenen puanlardan kestirilmeye çalışılır. Hata puanı da gerçek puan ile gözlenen puan arasındaki fark olarak tanımlanır ve hatanın değişik kaynaklardan geldiği varsayılır. Buradan X=T+E bağlantısı elde edilir. (Baykul, 2000). Hata, geldiği düşünülen kaynağa göre değişik isimler alır. Eğer puanlayıcılardan kaynaklandığı düşünülüyor ise, puanlayıcılar arası güvenirlikten bahsedilir. Puanlayıcı güvenirliği hesaplanmasında, puanlayıcının kendi kararlılığının bir göstergesi olan puanlayıcı-içi (intrarater) güvenirlik ve birbirinden bağımsız puanlayıcıların aynı performans görevini puanlaması ve tutarlı sonuçlar vermelerinin bir göstergesi olan puanlayıcılar-arası (interrater) kullanılmaktadır. Puanlayıcılar arası güvenirlik Pearson Momentler Çarpımı Korelasyonu, Uyuşma Yüzdesi, Kappa İstatistiği ve ortalamaların karşılaştırılması gibi farklı teknikler kullanılarak test edilebilmektedir (Goodwin, 2001).

Genellenebilirlik Kuramı

Genellenebilirlik (G) kuramı ile güvenirlik yüzey olarak tanımlanan bütün değişkenler (zaman, maddeler, puanlayıcılar) dikkate alınarak hesaplanabilmekte ve birbirleri ile olan etkileşimleri dikkate alınabilmektedir. Genellenebilirlik kuramında hata kavramı birçok hata kaynağına ayrılır ve farklı kaynaklardan gelen hatalar (puanlayıcı, madde, zaman vb.) ve bu hata kaynaklarının birbirleri ile olan etkileşimleri birlikte değerlendirilir. Böylece klasik test kuramının aksine, farklı hata kaynakları ve bunların etkileşimleri için tek bir güvenirlik katsayısı hesaplanabilir ve farklı hata

kaynaklarının varyansa etkisi göz önünde bulundurularak, maksimum güvenirlik elde edilebilir (Eason, 1989) . Klasik test kuramında değişik hata kaynakları ve bu hata kaynaklarının birbirleri ile olan etkileşimlerini dikkate alınmaz ve hatanın tek bir kaynaktan geldiği varsayılır. Ancak, genellenebilirlik kuramında tüm hata kaynakları ve birbirleri ile olan etkileşimleri hesaplanabilir ve varyans bileşenlerinin varyanstaki etkisi belirlenebilir.

Genellenebilirlik kuramına dayalı olarak performans değerlendirmelerinde bağıl ve mutlak olmak üzere iki farklı değerlendirmeden bahsedilebilir (Brennan, 2001). Bağıl değerlendirme ile bireyler, ölçme sonuçlarının elde edildiği grup normlarına bağlı olarak değerlendirilirken, mutlak değerlendirmede gruptan bağımsız olarak, önceden belirlenmiş mutlak bir ölçüte göre değerlendirilirler. Bu sebep ile G kuramında ölçme hataları bağıl kararlar için ölçme hataları ve mutlak kararlar için ölçme hataları olarak iki farklı yolla hesaplanır. (Shavelson ve Webb, 1991).

Bağıl ve mutlak hatalara göre G ve Phi katsayıları hesaplanmaktadır. Klasik test kuramındaki gerçek puan, Genellenebilirlik kuramındaki evren puanına, Genellenebilirlik (G) katsayısının klasik test kuramındaki karşılığı güvenirlik katsayısıdır (Rentz, 1987). G katsayısı, evren puanının varyansının gözlenen puan varyansına oranıdır ve hesaplanmasında ölçme objesi ile ilişkili olan varyans bileşenleri kullanılır. Phi katsayısı güvenirlik katsayısı olarak adlandırılır. Bu katsayı evren puanının varyansının, evren puanının varyansı ve mutlak hata varyansının toplamına oranıdır ve hesaplanmasında ölçme objesi ana etkisi dışında kalan diğer tüm varyans bileşenleri kullanılır. Dolayısıyla, bağıl değerlendirmelerde G, mutlak değerlendirmelerde ise Phi katsayısından yararlanılır (Brennan, 2001).

Genellenebilirlik çalışmaları, ölçme hatalarının hangi değişkenlik kaynağından kaynaklandığını belirlemek için düzenlenmektedir. Genellenebilirlik çalışması için öncelikle ölçme örnekleminin genelleneceği evren ve değişkenlik kaynakları tespit edilmelidir. Genellenebilirlik çalışmalarında farklı hata kaynaklarına göre farklı desenler kullanılmaktadır. Bunlar çaprazlanmış ve yuvalanmış desenlerdir.

Genellenebilirlik kuramında dayalı olarak bir desen oluşturulmasında veri yapısına göre yararlanılan çaprazlanmış ya da yuvalanmış olmak üzere iki tür desen bulunmaktadır. Çaprazlanmış desenlerde değişkenlik kaynaklarının koşul olarak adlandırılan her bir düzeyi, diğer değişkenlik kaynaklarının bütün koşulları ile etkileşir ve bu desen tüm değişkenlik kaynakların arasına “x” işareti konularak ifade edilir. Yuvalanmış desenlerde değişkenlik kaynaklarının koşul olarak adlandırılan her bir

düzeyi, diğer değişkenlik kaynaklarının bazı koşulları ile etkileşir ve bu desen tüm değişkenlik kaynakların arasına ":" işareti konularak ifade edilir (Shavelson ve Webb, 1991).

Çaprazlanmış veya yuvalanmış desenler, dengelenmiş (balanced) veya dengelenmemiş (unbalanced) olarak tasarlanmaktadır. Tümüyle çaprazlanmış desenlerde, veri kaybı olmadığı sürece, gözlem sayısının her düzeyde eşit olduğu dengelenmiş desenler oluşturulur ve varyans bileşenleri dengelenmiş desen varyans bileşenleri ile hesaplanır. Örneğin eşit sayıdaki puanlayıcının, eşit sayıda öğrenciyi ve görevi puanladığı bir durumda, dengelenmiş desenden bahsedilebilir. Fakat bazen, yuvalanmış desenlerde değişkenlerin tüm koşullarda eşit sayıda olmadığı dengelenmemiş desenler ortaya çıkabilmektedir. Bu durumda, verilerin bir kısmı atılır ve bu da veri kaybına sebep olur. Veri atma durumu söz konusu ise, dengelenmemiş desende varyans analizi kullanmak yerinde olacaktır (Brennan, 2001).

Değişkenlik kaynaklarının özeliklerine göre ise tesadüfi ve sabit etkiden söz edilebilir. Genellenebilirlik kuramında tüm koşulların tesadüfi olarak örneklendiği ve geniş bir puanlayıcı evreninden tesadüfi olarak seçilen bir puanlayıcı örnekleminin puanlama yapması ile mümkün olan tesadüfi; özellikle seçilmiş olan bir örneklem ile çalışılması ve evrene genelleme kaygısı güdülmeyen durumlarda kullanılan sabit olmak üzere iki tür değişkenlik kaynağı bulunmaktadır (Crocker ve Algina, 1986).

Karar çalışmalarında, genellenebilirlik çalışmasından elde edilen veriler üzerinde çalışılarak, değişkenlerden kaynaklanan hataların en aza indirilmesi amaçlanır. Böylece, hangi koşulda maksimum güvenirliğin sağlanabileceği bulunmuş olur (Crocker ve Algina, 1986).G çalışmalarının amacı varyans bileşenlerinin toplam varyans üzerindeki etkisini kestirmekken, K çalışmalarının amacı maksimum güvenirliği sağlamak adına, G çalışmalarında toplanan verileri yorumlamaktır (Güler, Uyanık ve Teker, 2012).

Rasch Modeli

Madde Tepki Kuramına (MTK) dayalı olarak geliştirilen Rasch modeli ile puanlayıcılar arası güvenirlik hesaplanabilir. Performans görevlerinin birden çok puanlayıcı tarafından puanlanması durumunda, farklı değişkenlik kaynaklarının dikkate alınması gerekir. Birincisi bireyin sahip olduğu yetenek, ikincisi bireylerin yeteneklerini yansıtabilecekleri performans görevi, üçüncüsü puanlayıcılar, dördüncüsü maddeler,

beşincisi ise puanlama ölçeğidir. Çok Değişkenli Rasch Modeli ile bu beş değişkenlik kaynağının etkileşimi ile bireyin performansı kestirilebilir (Crocker ve Algina, 1986).

Çok Değişkenlik Kaynaklı Rasch modeli ile hata varyansı değişkenlik kaynaklarına bölünebilmektedir ve her bir değişkenlik kaynağı için grup düzeyinde istatistikler verilebilmektedir. Çok Değişkenli Rasch modeli performans değerlendirmede bazı avantajlar sağlamaktadır:

- Ölçme ortamında bir veya daha fazla ek yüzeyi kapsayacak şekilde Rasch modeli geliştirilebilmektedir.
- Bireyin yeteneği, görevin güçlüğü ve puanlayıcıların katılığını belirlemede ham puanlar yeterli istatistiklerdir.
- Puanlayıcıların puanlamasını kestirmede her kategorideki gözlemlerin sayılması yeterli istatistiklerdir.
- Verilerin modele uyumunu test etmede kullanılan uyum istatistikleri kalite kontrolünü desteklemektedir.
- Uyumsuzluğun birçok şekline karşı güçlüdür/dirençlidir.
- Eksik verilere karşı güçlüdür, eksik verilerin tamamlanması gibi bir gereklilik yoktur.
- Kestirim yaparken, model verilerdeki tesadüfilik uygunmuş gibi işlem yapmaktadır.
- Veriden yapılandırılan ölçümler doğrusal referans çerçevesinde gözlenmektedir (Linacre, 2001).

Yapılandırmacı yaklaşım ile performansa dayalı değerlendirmeler ön plana çıkmış ve dolayısıyla değerlendirme süreçlerinde puanlayıcılar önemli bir yer edinmiştir. Bu açıdan, puanlayıcılar arası güvenirlik, bireyler hakkında güvenilir kararlar verme açısından çok önemlidir. Puanlayıcılar arası güvenirlik ile ilişkili olan bir boyut da puanlayıcı özellikleridir. Puanlayıcı özelliklerinden biri olan deneyim güvenirlik üzerindeki önemli bir etmendir. Pratik nedenlerden dolayı birebir benzer özelliklere sahip puanlayıcılara ulaşmak mümkün olamamaktadır. Bu nedenle DPA kullanmaya ilişkin deneyim yılı çok ve az olan puanlayıcıların benzer bir şekilde puanlama yapmaları sağlanmalıdır. Başka bir ifade ile hazırlanacak DPA, puanlama sürecine ilişkin deneyimi çok ve az olan puanlayıcıların kullanıldığı durumlarda benzer sonuçlar vermelidir.

İlgili Araştırmalar

Alan yazında farklı kuramlara dayalı yöntemler ile puanlayıcılar arası güvenirliğin veya DPA'nın güvenirliğinin test edildiği yurtdışı ve yurtiçi çalışmalar yer almaktadır.

Lane ve Sabers (1989), puanlayıcı sayısı ile güvenirik düzeyi ilişkisini değerlendirmek için genellenebilirlik kuramını kullanmışlardır. Araştırmada öğrencilerden kompozisyon yazmaları istenmiştir. Kompozisyonlar 8 puanlayıcı tarafından ve bir puanlama sistemi kullanılarak değerlendirilmiştir. Araştırmada öğrenciler (s), puanlayıcılar (r) ve puanlama sistemindeki kategoriler (c) ile gösterilmek tüm değişkenlerin çapraz tasarlandığı sxrxc deseniyle G ve K çalışmaları yapılmıştır. Araştırma sonuçlarına göre puanlayıcı sayısının azalmasının, G ve Phi katsayılarının azalmasına yol açtığı tespit edilmiştir.

Barneveld (2005) çalışmasında öğrencilerin puanlayıcılarla yuvalandığı dengelenmemiş (ö:p) x m desen (ö: öğrenci, p:puanlayıcı ve m:madde) kullanarak bir çalışma yapmış ve sonuçları GENOVA programını kullanarak analiz etmiştir. Araştırma sonuçlarına göre öğrenci performansını değerlendirmede güvenirliğin düşük çıktığı, puanlayıcı sayısının artırılmasıyla yükseldiği görülmüştür.

Wang (2005), öğrencilerin İngilizce yazılı anlatım becerilerini ölçmek için hazırlanmış sınav verileri kullanarak, geleneksel yöntem ve genellenebilirlik kuramıyla hesaplanan güvenirlik katsayılarını karşılaştırmıştır. Bu amaçla. Çalışmada öğrenci (p), madde (i) ve beceri (t) olmak üzere px(i:t) deseniyle G çalışması yapılmıştır. Çalışmanın hesaplanan katsayıların birbirine yakın sonuç verdiği sonucuna varılmıştır.

Lee ve Kantor (2007), TOEFL sınavı yazılı anlatım becerilerinin puanlanmasının güvenirliğini genellenebilirlik kuramını kullanarak hesaplamışlardır. Çalışmada ayrıca Karar çalışması yaparak, ideal puanlayıcı ve görev sayısını belirlemeye çalışmışlardır. Araştırma sonuçlarına göre, maksimum puan güvenirliği için görev sayısını artırmanın puanlayıcı sayısını artırmaktan daha etkili olduğu tespit edilmiştir.

Atılgan (2004) İnönü Üniversitesi müzik öğretmenliği seçme sınavının güvenirliğini, birden fazla görevin puanlandığı durumda Genellenebilirlik Kuramı kullanarak hesaplamıştır. Karar çalışması ile güvenirlik katsayılarının artırılmak istenmesi halinde, puanlayıcı sayısının artırılmasındansa, görev sayısının artırılmasının güvenirliği daha çok arttırdığı sonucuna varmıştır.

Atılgan ve Tezbaşaran (2005), aynı değişkenlik kaynağının farklı düzeyleri için karar çalışmaları ve kurgulanan ve gerçek durumlar için hesaplanan G ve Phi katsayılarının tutarlılıklarını karşılaştırmıştır. Araştırmada müzik öğretmenliği bölümüne öğrenci seçmek için geliştirilen özel yetenek sınavından elde edilen veriler kullanılmıştır. Bu sınav, üç ve dört puanlayıcı tarafından puanlanmıştır. Her iki durum için de G ve Phi katsayıları hesaplanmış, ardından kurgulanan durumlardan elde edilen katsayılarla gerçek durumlar için hesaplanan katsayılar karşılaştırılmıştır. Araştırma sonunda, puanlayıcı sayısı arttıkça G ve Phi katsayılarının arttığı, puanlayıcı sayısı azaldıkça G ve Phi katsayılarının azaldığı görülmüştür.

Kan (2005), kâğıt kalem sınavlarının puanlanmasında puanlama anahtarı kullanımının ve iki puanlama arasındaki zamanın güvenirliğine etkisini incelemiştir. Puanlayıcıların puanlama anahtarı kullanarak ve kullanmayarak yaptıkları ve iki puanlama arasından belirli bir zaman geçtikten sonra puanlama anahtarı kullanarak ve kullanmayarak yaptıkları puanlamalardan elde edilen verilerin karşılaştırılması sonucunda puanların arasında anlamlı bir fark bulunmuş ve zaman faktörünün kararlılık üzerinde etkili olduğu sonucuna varılmıştır.

Yelboğa (2007) "İş Performansı Ölçeği" kullanarak elde ettiği verileri klasik test kuramı ve genellenebilirlik kuramı ile analiz etmiş ve güvenirlik katsayılarını karşılaştırmıştır. Araştırmadan sonuçlarına göre, her iki kuram ile hesaplanan katsayının uyumlu olduğu görülmüştür.

Atılgan (2008), müzik öğretmenliği bölümüne öğrenci seçmek için kullanılan özel yetenek sınavının güvenirlik analizini yapmıştır. Çalışmada, tümüyle çaprazlanmış desen kullanılmıştır. Çalışmada, bağıl ve mutlak değerlendirme için G ve Phi katsayıları hesaplanmıştır. Ayrıca puanlayıcı ve görev sayısını değiştirilip, karar çalışması yapılmıştır. Yapılan analizler sonucunda, testin bütününde, G katsayısı için, puanlayıcı sayısının ve görev sayısının arttırılması güvenirliği artırdığı görülmektedir.

Güler (2008), 1999 yılında TIMSS'de kullanılan açık uçlu matematik sorularından 24'ünü, klasik test kuramı, genellenebilirlik kuramı ve rasch modelini kullanarak karşılaştırmış ve öğrencilerin verdikleri yanıtları bütünsel puanlama anahtarı ile puanlamıştır. Klasik test kuramı kullanılarak hesaplanan puanların iç tutarlılığı 0,92 olarak hesaplanmıştır. G katsayısı 0,92 ve phi katsayısı 0,90 bulunmuştur. Matematik başarısının belirlenmesinde puanlayıcıların birbirleriyle uyumlu puanlama yaptıkları sonucuna varılmıştır.

Kasap (2008) tarafından yürütülen yüksek lisans tezinde, 8. Sınıf matematik dersi öğrenci başarısı, cevap anahtarı niteliğinde hazırlanan puanlama anahtarı ve DPA kullanılarak değerlendirilmiştir. DPA ile değerlendirilen öğrencilerin başarısı ve puanlama anahtarıyla değerlendirilen öğrenci başarısı karşılaştırılmıştır. Ayrıca, anket, öz değerlendirme, akran değerlendirme ve öğretmen değerlendirmesi kullanılarak öğretmen ve öğrencilerden düşüncelerini belirtmeleri istenmiştir. Araştırmanın bulgularına göre, puanlama anahtarından ve DPA'dan elde edilen elde edilen puanların ortalamasıyla korelasyonlarının hesaplanması sonucunda elde edilen sonuçlar arasında anlamlı bir fark bulunmuştur.

Yelboğa (2008) otomotiv sektörüne personel seçme sürecinde kullanılmak üzere hazırlanmış mülakat formunu kullanarak elde edilen verilerin analizi sonucu G ve Phi katsayılarını hesaplamış ve ardından K çalışması yapmıştır. Araştırma sonuçlarına göre, puanlayıcıların birbirlerinden oldukça farklı değerlendirmeler yaptıkları sonucuna varılmıştır. K çalışmasında ise daha güvenilir sonuçlar için madde sayısının artırılması gerektiği belirtilmiştir.

Farrokhi, Esfandiari, Schaefer (2012) 6 ölçütten oluşan analitik dereceli puanlama anahtarı kullanarak, İngilizce eğitim gören 188 üniversite öğrencisinin İngilizce yazılı anlatım becerilerini değerlendirmişlerdir. Yazılı anlatım becerisi öz değerlendirme, akran değerlendirmesi ve dersi aldıkları hocanın değerlendirmesinden yararlanarak değerlendirilmiş ve elde edilen veriler Çok Değişkenlik Kaynaklı Rasch Modeli ile analiz edilmiştir. Elde edilen sonuçlara göre kelime seçiminin en zor, yazım denetiminin ise en kolay puanlanan ölçütler olduğu görülmüştür.

Parlak ve Doğan (2014) tarafından yürütülen bir çalışmada bir meslek lisesinde uygulanan bir sınavın puanlanması puanlama anahtarı ve DPA kullanılarak yapılmış ve elde edilen puanlar arası uyumu bakılmıştır. Elde edilen sonuçlara göre, her iki puanlama aracından elde edilen sonuçlar arası korelâsyonun yüksek olduğu görülmüştür. Ayrıca öğrencilerin birinci dönem notları ile her iki ölçme aracından elde edilen puanların korelâsyonuna bakılmış, DPA kullanılarak verilen puanların ders notlarını yordama gücü daha yüksek çıkmıştır.

Özel ve Acar tarafından yapılan bir araştırmada (2014), 5.sınıf öğrencilerinin Matematik performans görevleri, DPA ile üç puanlayıcı tarafından puanlanmış ve güvenirlik katsayısı genellenebilirlik kuramına göre değerlendirilmiştir. Genellenebilirlik katsayısı 0.85 olarak bulunmuştur. Elde edilen bu sonuç ile, DPA kullanarak elde edilen güvenirliğin yüksek olduğu sonucuna varılmıştır.

Büyükkıdık ve Anıl tarafından (2015) yürütülen bir çalışmada, 6. 7. Ve 8. Sınıf öğrencilerinin problem çözme becerileri bütüncül ve analitik dereceli puanlama anahtarı kullanılarak değerlendirilmiş, analitik dereceli puanlama anahtarı kullanılarak elde edilen puanlardan daha güvenilir olduğu bulunmuştur.

Karakaya (2015) ölçme değerlendirme dersi alan üniversite 3. Sınıf öğrencileri ile bir çalışma yürütmüş, öğrencilerin kendi hazırladıkları portfolyoların değerlendirilmesini veri toplama aracı olarak 7 ölçütten oluşan puanlama anahtarı kullanılarak yapmıştır. Portfolyoları öz değerlendirmeden, akran değerlendirmesinden ve dersi aldıkları hocanın değerlendirmesinden yararlanarak değerlendirmiş ve Çok Değişkenlik Kaynaklı Rasch Modeli ile analiz etmiştir. Sonuç olarak üç farklı değerlendirme arasında istatistiksel olarak manidar farklar bulmuştur.

Bu çalışmaların sonucu olarak, G kuramının güvenirlik analizi yaparken sıklıkla kullanılan bir analiz olduğu görülmektedir. Ayrıca, DPA'nın değerlendirme yöntemi olarak kullanıldığı çalışmalarda güvenilir sonuçların ortaya çıktığı görülmüştür. Ancak, puanlayıcıların deneyim yıllarına dayalı olarak puanlama aracının güvenirliğine ilişkin karşılaştırmanın yapıldığı bir G çalışmasına veya puanlayıcıların deneyim yıllarının güvenirlik katsayısına olan etkisinin incelendiği bir çalışmaya rastlanmamıştır.

BÖLÜM III

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).Çalışmadaki nicel veriler, görüşme formu ile toplanacak nitel veriler ile de desteklenmiştir.

Çalışma Grubu

Bu çalışma 2015- 2016 akademik yılının üçüncü döneminde, Ankara'da bulunan bir vakıf üniversitesinde görev yapan ve çalışmaya puanlayıcı olarak katılan 12 okutman ve farklı kurlarda eğitime devam eden ve İngilizce öğrenimi gören 120 öğrenci ile yürütülmüştür.

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 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 çalışmanın başlığında "deneyim yılı" kavramı kullanılmıştır.

Puanlayıcıların mezun oldukları bölüm, mezuniyet ortalamaları ve yurtdışı deneyimlerine dayalı karşılaştırmaları Çizelge 3'te sunulmuştur:

Çizelge 3

Okutmanlara İlişkin Değişkenler

	Okutmanlar	Mezun Olunan Bölüm	Mezuniyet Ortalaması	Erasmus Deneyimi
Deneyimi az	1	İngilizce Öğretmenliği	3.44	Yok
	2	İngilizce Öğretmenliği	3.69	Yok
	3	İngilizce Öğretmenliği	3.32	Var
	4	İngilizce Öğretmenliği	3.66	Yok
	5	İngilizce Öğretmenliği	3.58	Var
	6	İngiliz Dili ve Edebiyatı	3.21	Yok
Deneyimi çok	1	İngilizce Öğretmenliği	3.65	Yok
	2	İngilizce Öğretmenliği	3.28	Yok
	3	İngilizce Öğretmenliği	3.68	Var
	4	İngilizce Öğretmenliği	3.19	Yok
	5	İngilizce Öğretmenliği	3.56	Yok
	6	Mütercim Tercümanlık	3.35	Var

Çizelge 3 incelendiğinde hem deneyimi çok hem de deneyimi az okutmanlar arasında, sadece 1'er okutman İngilizce öğretmenliği dışında bir bölümden mezundur. Geri kalan okutmanların hepsi İngilizce Öğretmenliği mezunudur. Hem deneyimi çok hem de deneyimi az gruptan 2'şer okutman, Erasmus değişim programı ile yurt dışında bulunmuştur. Tüm okutmanların mezuniyet notu 3.00 ve üzeridir. Mezuniyet ortalamaları arasında fark olup olmadığını anlamak için Mann Whitney U testi yapılmış ve çıkan sonucun manidar olmadığı görülmüştür ($U=0.631$, $p>0.05$). Deneyim yılları hariç diğer değişkenler benzerlik göstermediği için her iki grupta yer alan puanlayıcıların birbirine denk olduğu belirtilebilir.

Çalışma grubunun puanlayıcılar dışındaki kısmını ise 3. dönem derslere devam eden öğrenciler oluşmaktadır. Tüm sınıflar, öğrencilerin bir önceki dönem aldıkları akademik ortalamaya göre homojen şekilde oluşturulduğundan, sınıflar arası seviye farklı yoktur.

Veri Toplama Araç ve Teknikleri

Bu bölümde, çalışma kapsamında kullanılacak DPA ve görüşme formu ayrıntılı olarak açıklanmıştır.

Dereceli Puanlama Anahtarı (Rubrik)

Bu çalışma kapsamında, öğrencilerin İngilizce yazılı anlatım başarılarını değerlendirmek için, uygulamanın yapılacağı vakıf üniversitesinin hazırlık okulunda hali hazırda kullanılanıma devam eden analitik dereceli puanlama anahtarı kullanılmıştır. DPA oluşturulmadan önce, ne amaçla geliştirileceği, 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. Ancak ilgili DPA'nın güvenirliğine dair psikometrik bir çalışma yapılmamıştır (EK.A).

Görüşme Formu

Bu çalışmada, yarı yapılandırılmış görüşme formu kullanılarak, okutmanların puanlama sürecindeki deneyimlerine yönelik görüşleri alınmıştır. Yarı yapılandırılmış görüşme formu karşılaştırmalı sonuçlar elde etmek amacıyla tercih edilmiştir. Görüşme formundaki sorular, alan yazındaki bilgiler ve araştırmanın amacı doğrultusunda oluşturulmuştur. Hazırlanacak olan görüşme formunun amaca ne derece hizmet ettiği ve uygulanabilirliği hakkında görüş almak için iki ölçme değerlendirme uzmanından, dil ve anlatımın uygunluğu hakkında görüş almak için ise, iki dil uzmanından yardım alınmıştır. Hazırlanan olan taslak form, okutmanlar ile yapılan görüşmede kullanılmak üzere oluşturulmuştur.

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. (EK B)

Verilerin Toplanması

Araştırmanın verileri, İngilizce yazılı anlatım becerisine yönelik yazılı yoklama uygulanarak toplanmış ve DPA ile puanlanmıştır Araştırmanın birinci aşamasında 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. İkinci aşamada ise puanlayıcılar ile uygulama sürecine yönelik görüşmeler yapılmış ve içerik analizi ile çözümlenmiştir.

Araştırmaya ilişkin nicel veriler, Ankara'da bulunan bir vakıf üniversitesinin üçüncü dönem ilk vize sınavı kapsamındaki 40 dakikalık yazılı yoklamadan, gerekli izinler alınarak toplanmıştır. İlgili yazılı yoklama kağıtları rastgele toplanmış ve puanlayıcılar tarafından birbirlerinden bağımsız olarak 1 hafta içerisinde DPA kullanarak değerlendirilmiştir.

Nitel veriler, puanlayıcı ile yapılacak görüşmeler neticesinde elde edilmiştir. Yarı yapılandırılmış her bir görüşmenin yaklaşık 15 dakika sürmüştür. Katılımcıların onayı alındığı durumlarda görüşme süreci ses kayıt cihazı ile kayıt edilmiştir (EK C).

Verilerin Çözümlenmesi

Çalışmada elde edilen nicel veriler genellenebilirlik kuramı kapsamında Edu G programı kullanılarak analiz edilmiştir. Araştırmada, birey ve puanlayıcıların maddeler ile çaprazlandığı ancak bireylerin puanlayıcılara yuvalandığı desenden ((b:p)xm) faydalanılmıştır. Bu desende her birey ve puanlayıcı her bir maddeyi yanıtlarken her puanlayıcı farklı bireyleri puanlamıştır.

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. (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:

$$\phi = \frac{\sigma^2_m}{\sigma^2_m + \sigma^2_{Mutlak}}$$

Nitel verilerin çözümlenmesinde içerik analizinden faydalanılmıştır.

Yarı yapılandırılmış görüşme tekniği ile 3 deneyimi çok ve 3 deneyimi az olmak üzere 6 puanlayıcıdan elde edilen nitel veriler içerik analizi yöntemi ile çözümlenmiş ve farklı deneyim yıllarına sahip puanlayıcıların görüşleri karşılaştırılmıştır. İçerik analizinde ham veriler önce temalara ayrılmış, bu sayede toplanan verilerin çözümlenmesi sürecinde içerik analizinden faydalanılmıştır. Pratik nedenlerden dolayı kodlama güvenirliği yapılmamıştır.

BÖLÜM IV

BULGULAR

Bu bölümde, yapılan veri analizi sonucunda, araştırma sorularına cevap verilmiştir. Bulguların sunuluşunda problemlerinin sırası izlenmiştir. Her problemin altında o problem ile ilgili bulgular ve yorumlar yer almaktadır.

Puanlamaya Yönelik Deneyimi Az Olan Puanlayıcıların Bulunduğu Grup İçin G Çalışması Sonuçları

Bu kısımda birinci araştırma sorusuna ilişkin elde edilen bulgular sunulmuştur.

Birinci Araştırma Sorusu

1- İngilizce yazılı anlatım becerisi, DPA kullanmaya yönelik deneyimi az olan İngilizce okutmanları tarafından puanlandığında, dereceli puanlama anahtarına ilişkin hesaplanan
   - varyans bileşenleri
   - bağıl ve mutlak hata varyansları
   - G ve Phi katsayısı nasıldır? araştırma sorusuna ilişkin bulgular:

DPA kullanmaya yönelik deneyimi az puanlayıcıların yer aldığı gruba ilişkin, anlatımı yalınlaştırmak için çalışmanın devamında “deneyimi az grup” olarak adlandırılmıştır. Araştırmada kullanılan (b:p)xm deseni için oluşturulan veri yapısı örneği Çizelge 4’te sunulmuştur:

Çizelge 4

Deneyimi Az Grubun Puanladığı Veri Yapısı Örneği

	P1	P2	P3	P4	P5	P6
B1	X	X	X			
:	X	X	X			
:	X	X	X			
B61				X	X	X
:				X	X	X
B120				X	X	X

Çizelge 4'te görüldüğü üzere, İngilizce yazılı anlatım becerisine yönelik hazırlanan yazılı yoklamaya katılan hazırlık sınıfı öğrencilerinden 120 öğrenci 6 puanlayıcı tarafından, her bir puanlayıcının 60'şar öğrenciyi puanlamasıyla tasarlanan (b:p)xm deseni, G çalışması ile analiz edilmiştir. Puanlayıcıların farklı gruptaki öğrencilerin İngilizce yazma becerisini puanladığı, yani bireylerin puanlayıcılara yuvalandığı, maddelerin ise bireyler ve puanlayıcılar ile çaprazlandığı (b:p)xm deseni, 5 varyans kaynağından oluşmaktadır. Hesaplanan G çalışması sonuçları Çizelge 5'te gösterilmiştir.

Çizelge 5

Deneyimi Az Gruba İlişkin Hesaplanan Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri

Varyans Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	Varyans	%
M	16.63565	5	3.32713	0.00532	0.9
P	17.84676	5	3.56935	0.00141	0.2
B:P	685.64167	354	1.93684	0.27548	47.1
MP	35.25602	25	1.41024	0.01877	3.2
MB:P	502.60833	1770	0.28396	0.28396	48.5
Toplam	1257.98843	2159			%100

Çizelge 5'te verilen, G çalışması sonucunda hesaplanan varyans ve toplam varyansı açıklama yüzdeleri incelendiğinde madde ana etkisi için hesaplanan varyans bileşeninin(σ^2(m)= 0.00532), toplam varyansın %1'ini açıkladığı görülmektedir. Madde ana etkisi, bazı görevlerin diğerlerinden daha zor olup olmadığını göstermektedir

(Shavelson ve Webb, 1991). Bu durumda, maddelerin zorluk bakımından farklılaşmadığı sonucuna varılabilir.

Puanlayıcılar için hesaplanan varyans bileşeninin (σ^2(p=0.00141) toplam varyansın %0.2'sini açıkladığı görülmektedir. Puanlayıcılar için hesaplanan varyans bileşeni en düşük varyans değerine sahiptir. Bu durumda, puanlayıcıların birbirleri ile tutarlı puanlamalar yaptıkları sonucuna varılabilir.

Belirli puanlayıcıların belirli öğrencileri puanlaması sonucu oluşan, bireylerin puanlayıcılara yuvalandığı b:p etkisine ait varyans bileşeni (σ^2(b:p)= σ^2(b,bp)), birey ana etkisine (σ^2(b)) ve bireyxpuanlayıcı ortak etkisine (σ^2(bp)) ait varyans bileşenlerinden oluşmaktadır (Brennan, 2001). Çalışmada her bir puanlayıcı farklı 60öğrenciyi puanladığı için birey değişkeni puanlayıcı değişkenine yuvalanmıştır. Ana etki (b:p) için hesaplanan varyans bileşeni toplam varyansın %47'sini açıklamaktadır. Bu varyans bileşeninin yüksek olması, birey-puanlayıcı etkileşiminin farklılık gösterdiği şeklinde yorumlanabilir. Diğer bir ifade ile bireylerin ve puanlayıcıların değişmesi durumunda puanlamanın da değişebileceği sonucuna varılabilir.

Maddexpuanlayıcı ortak etkisine ait varyans bileşeni (mp=0.01877), madde-puanlayıcı etkisinden kaynaklanan varyansın büyüklüğü ve puanlayıcıların farklı maddeleri puanlarken ne düzeyde tutarlı oldukları hakkında bilgi sağlar (Shavelson ve Webb, 1991).Madde ve puanlayıcı ortak etkileşimi için hesaplanan varyans bileşeni toplam varyansın %3'ünü açıklamaktadır. Bu varyansın küçük olması, puanlayıcıların bireyleri bir maddeden diğerine puanlamada farklılaşmadıklarını gösterir.

Bireylerin maddeler ile çaprazlandığı, puanlayıcılara yuvalandığı durum için hesaplanan varyans (σ^2(mbp=0.28396)), toplam varyansın %49'udur. Artık varyansın hesaplanan varyans değerleri içinde, büyüklük bakımından en yüksek varyans değerine sahip olduğu görülmektedir. Bu durum, tesadüfî hata kaynaklarının büyük olabileceğini göstermektedir. Deneyimi az gruba ilişkin hesaplanan bağıl ve mutlak hata varyansları Çizelge 6'da gösterilmiştir.

Çizelge 6

Deneyimi Az Gruba İlişkin Hesaplanan (b:p)xm Desenine Ait Bağıl ve Mutlak Hata Varyansları

(b:p)xm				
Np	Nb	nb:p	$\sigma^2(\delta)$	$\sigma^2(\Delta)$
3	60	120	0.22462	0.22659

b: birey, p:puanlayıcı ve $\sigma^2(\delta)$: Bağıl Hata Varyansı, $\sigma^2(\Delta)$: Mutlak Hata Varyansı

(B:p)xm deseninden elde edilen bağıl ve mutlak hata varyansları, $\sigma^2(\delta)=0.22462$ ve $\sigma^2(\Delta)=0.22659$ olarak hesaplanmıştır. Bağıl ve mutlak hata varyanslarının küçük olması, G ve Phi katsayılarının yüksek hesaplanması, dolayısıyla puanlama güvenirliğinin yüksek olması anlamına gelmektedir. Deneyimi az gruba ilişkin hesaplanan G ve Phi katsayıları Çizelge 7'de verilmiştir.

Çizelge 7

Deneyimi Az Gruba İlişkin Hesaplanan (b:p)xm Desenine Ait G ve Phi Katsayıları

(b:p)xm				
Np	Nb	nb:p	G	Phi
3	60	120	0.85	0.84

b: birey, p:puanlayıcı, m:madde

(B:p)xm deseninden elde edilen G katsayısı 0.85, Phi katsayısı ise 0.84 olarak hesaplanmıştır. Brennan (2001) G ve Phi katsayılarının 0.80'den büyük olması durumunda, puanlamanın güvenirliğinin yüksek olarak kabul edildiğini vurgulamıştır. Bu bulgulardan yola çıkılarak, kullanılan DPA'nın güvenirliğinin yüksek olduğu sonucuna varılabilir.

Puanlamaya Yönelik Deneyimi Çok Olan Puanlayıcıların Bulunduğu Grup İçin G Çalışması Sonuçları

Bu kısımda ikinci araştırma sorusuna ilişkin elde edilen bulgular sunulmuştur.

İkinci Araştırma Sorusu

2- İngilizce yazılı anlatım becerisi, DPA kullanmaya yönelik deneyimi çok olan İngilizce okutmanları tarafından puanlandığında, dereceli puanlama anahtarına ilişkin hesaplanan
   - varyans bileşenleri
   - bağıl ve mutlak hata varyansları
   - G ve Phi katsayısı nasıldır? araştırma sorusuna ilişkin bulgular:

Yazılı yoklamaya katılan hazırlık sınıfı öğrencilerinden 120 öğrenci 6 puanlayıcı tarafından, her bir puanlayıcının 60'şar öğrenciyi puanlamasıyla tasarlanan (b:p)xm deseni Çizelge 8'deki gibidir ve G çalışması ile analiz edilmiştir:

Çizelge 8

Deneyimi Çok Grubun Puanladığı Veri Yapısı Örneği

	P7	P8	P9	P10	P11	P12
B1	X	X	X			
:	X	X	X			
:	X	X	X			
B61				X	X	X
:				X	X	X
B120				X	X	X

Puanlayıcıların farklı öğrencilerin İngilizce yazılı anlatım becerisini puanladığı, yani bireylerin puanlayıcılara yuvalandığı, maddelerin ise bireyler ve punalayıcılar ile çaprazlandığı (b:p)xm deseni, 5 varyans kaynağından oluşmaktadır DPA kullanmaya yönelik deneyimi çok puanlayıcıların olduğu grup anlatımı yalınlaştırmak için çalışmanın ileriki kısımlarında "deneyimi çok grup" olarak adlandırılmıştır. Deneyimi çok grup için hesaplanan G çalışması sonuçları Çizelge 9'da gösterilmiştir.

Çizelge 9

Deneyimi Çok Gruba İlişkin Hesaplanan Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri

Varyans Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	Varyans	%
M	12.80926	5	2.56185	0.00400	0.7
P	8.68148	5	1.73630	-0.00293	0.0
B:P	678.09444	354	1.91552	0.27794	51.1
MP	28.08519	25	1.12341	0.01459	2.7
MB:P	438.77222	1770	0.24789	0.24789	45.5
Toplam	1166.44259	2159			%100

Çizelge 9'da sunulan, G çalışması sonucunda hesaplanan varyans ve toplam varyansı açıklama yüzdeleri incelendiğinde, madde ana etkisi için hesaplanan varyans bileşeninin (σ^2(m)= 0.00400), toplam varyansın %1'ini açıkladığı görülmektedir. Maddeler için hesaplanan varyans bileşeninin düşük olması görevlerin zorluk düzeylerinin öğrenciler için farklılaşmadığının bir göstergesidir (Shavelson ve Webb, 1991).

G çalışması sonucunda hesaplanan varyans ve toplam varyansı açıklama yüzdeleri incelendiğinde, puanlayıcı ana etkisi için hesaplanan varyans bileşeninin toplam varyans üzerinde bir etkiye sahip olmadığı görülmektedir (%0). Puanlayıcı ana etkisi için hesaplanan varyans bileşeni, en küçük varyans değerine sahiptir. Puanlayıcılara ait varyans bileşeninin küçük olması, puanlayıcıların birbirleri ile tutarlı ve benzeşik puanlama yaptıkları şeklinde yorumlanabilir.

Çalışmada her bir puanlayıcı farklı 60 öğrenciyi gözlemlediği için öğrenci değişkeni puanlayıcı değişkenine yuvalanmıştır. Ana etki (σ^2(b:p=0.27794) için hesaplanan varyans bileşeni toplam varyansın %51'ini açıklamaktadır. (B:p) için hesaplanan varyans bileşeni, en yüksek varyans değerine sahiptir. Bu değerin yüksek olması, birey-puanlayıcı etkileşiminin farklılık gösterdiğinin bir göstergesidir.

(Mp) ortak etkisine ilişkin hesaplanan varyans, puanlayıcıların bireyleri tüm maddelerde tutarlı bir şekilde puanlayıp puanlamadıklarını anlamak için hesaplanır. Madde ve puanlayıcı etkileşimi (σ^2(mp=0.01459)) için hesaplanan varyans bileşeni toplam varyansın %3'ünü açıklamaktadır. Bu durum bazı maddelere verilen puanların bir puanlayıcıdan diğerine çok farklılaşmadığı şeklinde yorumlanabilir.

Bireylerin maddeler ile çaprazlandığı, puanlayıcılara yuvalandığı durum için hesaplanan varyans (σ^2(mbp=0.24789)), toplam varyansın %46'sıdır. Bu varyans değeri en büyük ikinci varyanstır. Artık varyans bileşeninin büyük çıkması birey puanlayıcı ve madde ortak etkileşiminin tesadüfî hata kaynaklarından etkilendiğinin göstergesi olabilir. Deneyimi çok gruba ilişkin hesaplanan mutlak ve bağıl hata varyansları Çizelge 10'da sunulmuştur.

Çizelge 10

Deneyimi Çok Gruba İlişkin Hesaplanan (b:p) x m Desenine Ait Bağıl ve Mutlak Hata Varyansları

		(b:p)xm		
Np	**Nb**	**nb:p**	**$\sigma^2(\delta)$**	**$\sigma^2(\Delta)$**
3	60	120	0.20916	0.21075

b: birey, p:puanlayıcı ve $\sigma^2(\delta)$: Bağıl Hata Varyansı, $\sigma^2(\Delta)$: Mutlak Hata Varyansı

(B:p)xm deseninden elde edilen bağıl ve mutlak hata varyansları$\sigma^2(\delta)$ = 0.20916 ve $\sigma^2(\Delta)$ = 0.21075 olarak hesaplanmıştır. Bağıl ve mutlak hata varyanslarının küçük kestirilmesi, puanlamanın güvenilir bir şekilde yapıldığının bir göstergesidir. Deneyimi çok gruba ilişkin hesaplanan G ve Phi katsayıları Çizelge 11'de sunulmuştur.

Çizelge 11

Deneyimi Çok Gruba İlişkin Hesaplanan (b:p)xm Desenine Ait G ve Phi Katsayıları

		(b:p) x m		
Np	Nb	nb:p	G	Phi
3	60	120	0.86	0.86

b: birey, p:puanlayıcı, m:madde

(B:p)xm deseninden elde edilen G katsayısı 0.86, Phi katsayısı ise 0.86 olarak kestirilmiştir. Bu katsayılar ile DPA kullanarak güvenilir bir puanlama yapıldığı sonucuna varılabilir.

Puanlamaya Yönelik Deneyimi Az ve Çok Olan Puanlayıcıların Birlikte Yer Aldığı Grup İçin G Çalışması Sonuçları

Bu kısımda üçüncü araştırma sorusuna ilişkin elde edilen bulgular sunulmuştur.

Üçüncü Araştırma Sorusu

3- İngilizce dersi yazılı anlatım becerisinin puanlama sürecine yönelik deneyimi az ve deneyimi çok olan İngilizce okutmanlarının birlikte yer aldığı gruba ilişkinhesaplanan,
   - varyans bileşenleri
   - bağıl ve mutlak hata varyansları
   - G ve Phi katsayısı nasıldır? araştırma sorusuna ilişkin bulgular:

İngilizce yazılı anlatım becerisinin puanlama sürecine yönelik deneyimi az ve deneyimi çok olan puanlayıcıların birlikte yer aldığı gruptan, çalışmanın devamında karışık grup olarak bahsedilecektir. Karışık grup oluşturulmasındaki amaç deneyimi çok ve deneyimi az puanlayıcıların beraber puanlama yapmaları durumunda hesaplanan güvenirlik katsayılarını, diğer gruplar ile karşılaştırmaktır. Araştırmada kullanılan (b:p)xm deseni için oluşturulmuş veri yapısı örneği Çizelge 12'de gösterilmiştir:

Çizelge 12

Karışık Grubun Puanladığı Veri Yapısı Örneği

	P1	P1	P1	P2	P2	P2
B1...B20	X					
B21…B40		X				
B41…B60			X			
B61…B80				X		
B81…B100					X	
B101…B120						X

Çizelge 12'de görüldüğü üzere, İngilizce yazılı anlatım becerisine yönelik hazırlanan yazılı yoklamaya katılan hazırlık sınıfı öğrencilerinden 120 öğrenci 2 puanlayıcı tarafından, her bir puanlayıcının 60'şar öğrenciyi puanlamasıyla tasarlanan (b:p)xm deseni, G çalışması ile analiz edilmiştir. Puanlayıcıların farklı öğrencilerin İngilizce yazılı anlatım becerisini puanladığı, yani bireylerin puanlayıcılara yuvalandığı, maddelerin ise puanlayıcılar ve bireyler ile çaprazlandığı (b:p)xm deseni, 5 varyans kaynağından oluşmaktadır. Hesaplanan varyans bileşenleri Çizelge 13'te gösterilmiştir.

Çizelge 13

Karışık Gruba ilişkin Hesaplanan Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri

Varyans Kaynağı	Kareler Toplamı	Sd	Kareler Ortalaması	Varyans	%
M	2.76667	5	0.55333	-0.00163	0.0
P	4.35556	1	4.35556	0.00491	0.8
B:P	251.64444	118	2.13258	0.30670	50.1
MP	3.74444	5	0.74889	0.00761	1.2
MB:P	172.48889	590	0.29235	0.29235	47.8
Toplam	435.00000	719			%100

Çizelge 13'te verilen, G çalışması sonucunda hesaplanan varyans ve toplam varyansı açıklama yüzdeleri incelendiğinde madde ana etkisi için hesaplanan varyans bileşeni toplam varyansın %0,2'sini açıkladığı görülmektedir. Bu değer, en düşük varyans değeridir ve madde güçlüğünün değişmediğinin bir göstergesidir.

G çalışması sonucunda hesaplanan varyans ve toplam varyansı açıklama yüzdeleri incelendiğinde puanlayıcı ana etkisi için hesaplanan varyans bileşeninin toplam varyansın %1'ini açıkladığı görülmektedir. Puanlayıcı ana etkisi için hesaplanan varyans bileşeninin sıfıra yakın olması, puanlayıcılar arasında bir tutarlılığın olduğu şeklinde yorumlanabilir

Çalışmada her bir puanlayıcı farklı 60 öğrenciyi puanladığı için öğrenci değişkeni puanlayıcı değişkenine yuvalanmıştır. Ana etki (σ^2(b:p=0.30670) için hesaplanan varyans bileşeni büyük bir varyans değerine sahip olmakla birlikte toplam varyansı %50'siniaçıklamaktadır. Hesaplanan bireyxpuanlayıcı ortak etkisine ait varyans bileşeninin büyük olması, bireylerin bir puanlayıcıdan diğerine farklı puanlandığı şeklinde yorumlanabilir.

Madde ve puanlayıcı ortak etkileşimi (σ^2(mp=0.00761)) için hesaplanan varyans bileşeni toplam varyansın %1'ini açıklamaktadır. Bu durum bazı maddelere verilen puanların bir puanlayıcıdan diğerine farklılaşmadığı, puanlayıcıların bireyleri bir maddeden diğerine kararlı puanladıkları şeklinde yorumlanabilir.

Bireylerin maddeler ile çaprazlandığı, puanlayıcılara yuvalandığı durum için hesaplanan varyans (σ^2(mbp=0.29235)), toplam varyansın %48'idir. Artık varyans bileşeninin büyük çıkması birey, puanlayıcı ve madde ortak etkileşiminin tesadüfî hata

kaynaklarından etkilendiğinin göstergesi olabilir. Hesaplanan bağıl ve mutlak hata varyansları, Çizelge 14'te gösterilmiştir.

Çizelge 14

Karışık Gruba İlişkin Hesaplanan (b:p)xm Desenine Ait Bağıl ve Mutlak Hata Varyansları

		(b:p)xm		
Np	Nb	nb:p	$\sigma^2(\delta)$	$\sigma^2(\Delta)$
1	20	120	0.22359	0.22359

b: birey, p:puanlayıcı ve $\sigma^2(\delta)$: Bağıl Hata Varyansı, $\sigma^2(\Delta)$: Mutlak Hata Varyansı

(B:p)xm deseninde ile elde edilen bağıl ve mutlak hata varyansları $\sigma^2(\delta)=0.22359$ ve $\sigma^2(\Delta)=0.22359$ olarak hesaplanmıştır. Bağıl ve mutlak hata değerlerinin düşük olması, G ve Phi katsayılarının yüksek olmasının dolayısıyla yüksek güvenirliğin bir göstergesidir. Hesaplanan G ve Phi katsayıları Çizelge 15'te gösterilmiştir.

Çizelge 15

Karışık Gruba İlişkin Hesaplanan (b:p)xm Desenine Ait G ve Phi Katsayıları

		(b:p)xm		
Np	Nb	nb:p	G	Phi
1	20	120	0.86	0.86

b: birey, p:puanlayıcı, m:madde

(B:p)xm deseninde elde edilen G katsayısı 0.86, Phi katsayısı ise 0.86 olarak kestirilmiştir. Bu değerler güvenilir bir puanlamanın göstergesidir.

Farklı Puanlayıcı Gruplarına İlişkin Elde Edilen Sonuçların Karşılaştırılması

Bu kısımda dördüncü araştırma sorusuna ilişkin elde edilen bulgular sunulmuştur.

Dördüncü Araştırma Sorusu

4. Çalışmaya katılan üç farklı puanlayıcı grubuna ilişkin hesaplanan
    - varyans bileşenleri

- bağıl ve mutlak hata varyansları
- G ve Phi katsayıları karşılaştırıldığında, sonuç nasıldır?

Bu bölümde, aynı sayıda öğrenci, madde ve DPA kullanmaya ilişkin farklı deneyim yıllarına sahip puanlayıcı grupları için oluşturulmuş (b:p)xm deseni ile yapılan çalışmalardan elde edilmiş G çalışması karşılaştırılmış ve sonuçları Çizelge 16'da gösterilmiştir.

Çizelge 16

G Çalışması Sonuçlarının Karşılaştırılması

	Deneyimi Az Grup	**Deneyimi Çok Grup**	**Karışık Grup**
M	%1	%1	%0
P	%0	%0	%1
B:P	%47	%51	%50
MP	%3	%3	%1
MB:P	%49	%46	%48
Mutlak	0.22659	0.21075	0.22359
Bağıl	0.22462	0.20916	0.22359
G	0.85	0.86	0.86
Phi	0.84	0.86	0.86

İngilizce yazılı anlatım becerisi, puanlama sürecine yönelik hesaplanan varyans bileşenleri incelendiğinde, madde ana etkisine ilişkin hesaplanan varyans bileşeninin en küçük karışık grupta (%0) kestirildiği görülmektedir. Deneyimi çok ve deneyimi az gruplarda madde ana etkisine ilişkin hesaplanan varyans bileşeni de, karışık gruptan fazla farklılık göstermemektedir (%1). Bu sonuç, her üç durumda da maddelerin birbirinden farklılaşmadığını ve ayırt ediciliklerin iyi olduğunu gösterir.

Puanlayıcılara ilişkin hesaplanan varyans değerinin en büyük (%1) karışık grupta kestirildiği görülmektedir. Puanlayıcılara ilişkin hesaplanan varyans değeri, puanlama sürecine ilişkin deneyimi çok ve deneyimi az gruplar için aynı değeri vermiştir (%0). Sonuç incelendiğinde, puanlayıcı ana etkisine ilişkin hesaplanan varyans bileşeninin gruplar arası bir fark yaratmadığı ve deneyim yılının puanlama güvenirliğini etkilemediği sonucuna varılabilir. Puanlayıcılar birbirleri ile tutarlı puanlamalar yaptıkları yapmışlardır.

(B:p) için hesaplanan varyans bileşeni, en büyük deneyimi çok grupta (%51) kestirilmiştir. Deneyimi az grupta (%47) ve karışık grupta (%50) hesaplanan varyans değerleri de deneyimi çok grupta hesaplanan varyans değerinden çok büyük bir 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 için hesaplanan varyans bileşeni, deneyimi çok ve deneyimi az gruplarda toplam varyansın %3'ünü, karışık grupta ise %1'ini açıklamaktadır. Hesaplanan bu varyans bileşeni, gruplar arası fark göstermemektedir. σ^2(mp) varyans bileşeninin toplam varyansın küçük bir kısmını açıklaması nedeniyle puanlayıcıların bireyleri bir maddeden diğerine kararlı bir şekilde puanladıkları, her üç grupta da, her maddeye verilen puanın, puanlayıcıdan puanlayıcıya değişmediği sonucuna varılmıştır.

Maddeler ile çaprazlandığı, puanlayıcılara yuvalandığı durum için hesaplanan varyans (σ^2(mbp)), deneyimi az grupta toplam varyansın %49'udur, deneyimi çok grupta %46'sı, karışık grupta ise %48'idir.Bu durumda, birey-görev madde etkileşiminin, üç yönlü etkileşimin ve diğer değişkenlik kaynaklarından kaynaklanan etkilerin, gruplar arası büyük bir fark göstermediği sonucuna varılmıştır.

(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 etkili olmadığı sonucuna varılabilir.

Puanlayıcıların Puanlama Sürecine İlişkin Görüşleri

Bu kısımda beşinci araştırma sorusuna ilişkin elde edilen bulgular sunulmuştur.

Beşinci araştırma Sorusu

5. İngilizce yazma dersi dereceli puanlama anahtarının kullanım sürecine ilişkin çalışmada yer alan puanlayıcıların görüşleri nelerdir?

Katılımcılarla yapılan görüşmelerden elde edilen verilerle yapılan içerik analizi neticesinde 9 tema belirlenmiştir. Bu temalar deneyim yılı, puanlama deseni, yorgunluk etkisi, puanlama yapılan ortamın güvenirlik üzerindeki etkisi, DPA'nın avantajları, DPA'nın dezavantajları, puanlayıcı grubunda yer alması gereken puanlayıcı sayısı, puanlamaya ilişkin eğitim verilmesi ve puanlayıcı önerileri olup aşağıda sırası ile sunulmuştur. Katılımcılardan A, B ve C DPA kullanmaya ilişkin deneyimi çok olan puanlayıcılar, D,E ve F ise DPA kullanmaya ilişkin deneyimi az olan puanlayıcılardır.

Deneyim Yılı

Katılımcılarla yapılan görüşmeler sonucunda elde edilen bulgular ışığında, DPA kullanmaya ilişkin deneyimi çok olan puanlayıcıların, DPA kullanmaya ilişkin deneyimi az olan puanlayıcılardan daha hızlı ve daha güvenilir puanlama yaptıklarını düşündükleri görülmektedir. Katılımcı A, beraber puanlama yaptığı partnerinin deneyim yılı kendinden az ise, partnerinin kendisini yavaşlattığını belirtmiştir:

> *Eğer partnerim DPA kullanmaya yönelik deneyimi az bir okutman ise, bana çok soru soruyor ve bu da beni daha fazla yoruyor ve daha erken mola veriyorum. Ama partnerim DPA kullanmaya yönelik deneyimi çok bir okutman ise, mola verme sürem uzayabiliyor. Bazen de ofiste tek başıma puanlama yapıyorum. O zaman beni bölecek biri olmuyor. Mola vermeyebiliyorum* (Katılımcı A).

DPA kullanmaya ilişkin deneyimi az olan puanlayıcılar da hız olarak yavaş puanlama yaptıkları belirtmişlerdir. Fakat kendilerinden daha deneyimi çok olan puanlaycıların daha güvenilir bir puanlama yapmadıklarını düşündükleri, deneyimin sadece değerlendirme hızını etkilediğini belirtmişlerdir. Puanlayıcıların değerlendirme hızlarındaki bu değişimin, kendilerinin de belirttiği gibi deneyim yıllarından kaynaklandığı söylenebilir. Katılımcı D bu konudaki fikrini şöyle açıklamıştır:

> *Bence, DPA kullanmak için deneyimden ziyade süreye ihtiyacımız var. DPA'da tüm ölçütler ayrıntıları ile açıklanıyor. Neye kaç puan verilmesi gerektiği yazıyor. İhtiyacımız olan sadece yeteri kadar bir sürede,*

> *değerlendirebileceğimiz sayıda sınav kâğıdı değerlendirmek. Deneyimin önemi olduğunu düşünmüyorum* (Katılımcı D).

Katılımcılardan deneyim yılı az olan D, E ve F puanlama esnasında DPA'da yazan ölçütlere sık sık baktıklarını belirtmişlerdir. Deneyim yılı çok olan katılımcılardan A, B ve C ise, ölçütleri iyi bildikleri için DPA'ya sık sık başvurmayıp, DPA'yı sadece en son puanlama için kullandıklarını belirtmişlerdir. Buradan, deneyim yılı arttıkça, puanlayıcıların DPA'da yer alan ölçütleri içselleştirdiği ve DPA kullanma konusunda kendilerini daha yetkin hissettikleri sonucuna varılabilir.

Katılımcılardan C, E ve F, DPA'yı kullanarak daha güvenilir bir puanlama yapma ile DPA kullanmaya ilişkin deneyimin arasında bir ilişki olduğunu belirtmişlerdir. Katılımcı A, ilk 1 yılın kritik olduğunu ve bu sürenin DPA kullanmayı öğrenmek için yeterli olduğunu şöyle belirtmiştir:

> *Açıkçası, DPA kullanmaya ilişkin deneyimi az ve deneyimi çok okutmanlar arasında güvenilir puanlama yapma açısından bir fark olduğunu düşünüyorum. Okutmanların zaman geçtikçe ve DPA kullanma konusunda kendilerini rahat hissettikçe daha güvenilir puanlamalar yaptıkları kanısındayım* (Katılımcı A).

Katılımcı B ise, deneyim yılından ziyada, DPA kullanarak yapılan değerlendirme sayısının önemini vurgulamıştır:

> *Özellikle işe yeni başlayan okutman arkadaşlarımız hata yapmamak için çok dikkatli değerlendirmeler yapıyorlar. Sık sık DPA'ya başvurup, her bir ölçütü inceliyorlar. Deneyimi çok olan okutmanlar ise, DPA'yı artık çok iyi bildiklerini varsayıp, değerlendirme esnasında bir daha dönüp bakma gereği duymuyorlar ve hata yapabiliyorlar* (Katılımcı B).

Katılımcılar tarafından varılan genel fikir, en az 1 sene deneyimin, DPA kullanarak güvenilir bir puanlama için gerekli olduğu yönündedir. Bu, puanlayıcıların 1 senede DPA kullanmayı kapsamlı bir şekilde öğrendiklerini göstermektedir.

Puanlama Deseni

Bir gruptaki tüm puanlayıcıların tüm bireyleri puanlayıp puanlamamalarına ilişkin sorulan soruya verilen ortak cevap, teoride güzel olan bu fikrin, pratikte imkânsız olduğu yönündedir. Katılımcılar işin içine girecek olan yorgunluk etkisi ile güvenirliğin düşeceğini belirtmişlerdir ve bireylerin puanlayıcılar arasında iş bölümü yapılarak puanlanması gerektiğini vurgulamışlardır. Katılımcılar, ölçütlerin iş bölümü yapılması fikrine ise sıcak bakmamışlardır. Katılımcı C bu konuyu, bir puanlayıcının gözünden

kaçan bir hatanın, diğer puanlayıcı tarafından bulunduğunu ve ölçütlerin paylaşılması durumunda bu hatayı bulmanın imkânsız olduğunu vurgulayarak savunmuştur:

> *Tüm puanlayıcılar tüm bireyleri puanlamamalı. İş bölümü yapılmalı. Aynı zamanda, Ayrıca DPA'da yer alan tüm ölçütler, bir birey için bir puanlayıcı tarafından değerlendirilmeli. Bazen bir puanlayıcının gözünden kaçan bir kelime hatasını, diğer hoca bulabiliyor. Eğer ölçütler puanlayıcılar arasında paylaşılırsa, bu hatayı bulma imkânı kalmaz* (Katılımcı C).

Yorgunluk Etkisi

Yapılan görüşmede, yorgunluk etkisinin karar verme becerisini ve dolayısıyla güvenilir puanlama yapmayı olumsuz etkilediği anlaşılmıştır. Katılımcılar yorgunluk etkisini azaltmak için puanlama esnasında mola verdikleri, açık havaya çıktıkları ya da bir şeyler yedikten sonra puanlamaya devam ettiklerini belirtmişlerdir. Buradan, fiziki ihtiyaçların, güvenilir puanlamayı etkilediği anlaşılmaktadır.

Değerlendirilen sınav kâğıdı sayısının güvenirlik üzerindeki etkisi sorulduğunda, tüm katılımcılar değerlendirilen ilk kağıtların, son kağıtlardan güvenirlik açısından daha yüksek olduğunu belirtmişlerdir. Bunun çözümü olarak okutmanların puanlama yapılan gün içinde derse girmemeleri ve değerlendirilmesi gereken sınav kâğıdı sayısının iki güne bölünmesi gösterilmiştir. İdeal kâğıt sayısı 15 ve 20 arasında değişmektedir. Katılımcı F bunu şöyle özetlemiştir:

> *Bu etkiyi, ara vererek azaltabiliriz. Ya da değerlendirme yaptığımız günler sayısını ikiye çıkarabiliriz Gündüz zaten derse giriyoruz. Ardın da bir sürü kâğıt okuduğumuz bazen güvenilir olmayabilir sonuçlar. Bu çok normal. Belki kâğıt sayısı ikiye bölünüp, iki günde tamamlanabilir* (Katılımcı F).

Değerlendirme Yapılan Ortamın Güvenirlik Üzerindeki Etkisi

Katılımcıların genel olarak vardıkları görüş, topluca kalabalık bir ortamda yapılan değerlendirmelerin, dikkatlerini dağıttığı yönündedir. Ayrıca katılımcı A, tek başına puanlama yaptığında, mola vermediğini belirtmiştir: "*Bazen de ofiste tek başıma puanlama yapıyorum. O zaman beni bölecek bir şey olmuyor. Mola vermeyebiliyorum*." Katılımcı F, katılımcı A'yı şöyle desteklemiştir:

> *En sona kalan hocalardan biri ben oluyorum genelde. Bunun ilk sebebi, dediğim gibi dikkatimin çabuk dağılması. Kalabalık bir ortamda değerlendirme yapmak benim hızımı yavaşlatıyor* (Katılımcı F).

Dereceli Puanlama Anahtarının Avantajları

Elde edilen bir diğer önemli tema DPA'nın sağladığı avantajlarla ilişkilidir. Bu konuda katılımcılar A ve B, bütüncül puanlama anahtarı kullandığı zaman puanlama

güvenirliğinden şüphe ettiklerini belirtmiş, DPA ile güvenilir bir puanlama yaptığını düşündüklerini söylemiştir. Katılımcı E, DPA kullanmanın avantajını şöyle vurgulamıştır:

> *Sene içinde kalabalık bir öğrenci grubuna pek çok sınav uyguluyoruz. Çoktan seçmeli sorular hariç, tüm açık uçlu maddeler ve kompozisyonlar okutmanlarca aynı anda değerlendiriliyor. Bu yoğunlukta ve koşuşturmada yanlış yapma olasılığımız çok yüksek olabiliyor. DPA kullanmak hata yapma olasılığımızı en aza indiriyor. Bir nevi standardizasyon sağlıyor* (Katılımcı E).

Katılımcı A ayrıca DPA'nın her sene gözden geçirilmesi ile öğrencilerden beklentilerini daha net anladıklarını vurgulamıştır: "*DPA bir nevi öğrenciden istediklerimizi yansıtma şeklimiz oluyor. Öğrencilerden istediklerimizi belirttiğimiz için de, öğretimimizi bu şekilde yapıyoruz. Kendimize de dönüt vermiş oluyoruz aslında*."

Bu durum, öğrencilerin hangi ölçütlere göre değerlendirileceklerini bilmelerinin başarılarını artıracağı şeklinde yorumlanabilir. Katılımcı B ise DPA'nın bütüncül puanlama anahtarından üstün noktalarına dikkat çekmiştir: "Öğrenci çok çeşitli kelimeler kullanmış olabilir ama yaptığı dilbilgisi yanlışlarından dolayı çok düşük olabilir. DPA her bir boyut üzerine eğilmemize yardımcı oluyor."

Dereceli Puanlama Anahtarının Dezavantajları

Katılımcıların geneli zaman problemi yaşamaları, programın çok yoğun olması, değerlendirilmeyi bekleyen kâğıt sayısının fazla olması, değerlendirme yapılan gruplarda dikkat dağıtan faktörlerin bulunması nedenleriyle DPA kullanarak puanlama yapma hususunda sıkıntı çektiklerini belirtmişlerdir. Ayrıca katılımcı D, DPA kullanılarak değerlendirilen bir kompozisyonun, bazen hak etmediği puanı alabildiğini belirtmiştir:

> *Örneğin bir kâğıdı puanlarken dilbilgisine 1, kelimesine 1 verdik. Aslında 1 bile almayı hak etmiyor ama dilbilgisi ve kelimede 0 puan olmadığı için 1 alıyor yine de. Kısacası 20 ya da 30 puan almayı hak eden bir öğrenci, DPA ile puanlama yaptığımız için 50 ya da 60 alabiliyor. Tek dezavantajı bu diyebilirim*

Böyle bir durumla karşılaşmamak için, DPA'da yer alan ölçütler daha detaylı bir şekilde gözden geçirilebilir.

Puanlayıcı Grubunda Yer Alması Gereken Puanlayıcı Sayısı

Değerlendirme yapılan gruptaki ideal puanlayıcı sayısı sorulduğunda, katılımcıların hemfikir olduğu konu, puanlayıcı sayısı arttıkça, puanlama güvenirliğinin arttığı yönündedir. Ancak, mevcut puanlayıcı sayısının yeterli olduğu konusunda birleşmişlerdir. Katılımcı A, belli bir süre geçtikten sonra, tüm bireylere aynı puanı veriyormuş gibi hissettiğini belirtmiştir. Bunun nedeni olarak da iş yükünün artmasının yorgunluğa sebep olacağı ve güvenirliği olumsuz etkiyeceği gösterilebilir. Ayrıca Katılımcı C katılımcı A'yı desteklemiştir:

> *Biz mevcut durumda, iki kişiden oluşan bir grupta puanlama yapıyoruz ve eğer uzlaşamadığımız bir şey olursa, ölçme değerlendirme birim görevlilerinden yardım alıyoruz. 3. bir puanlayıcı güvenirliği etkilemez bence. Çünkü ihtiyaç halinde biz 3. puanlayıcıya danışıyoruz. Ayrıca öğrenci beklediğinden düşük not aldığını düşünüyorsa, dilekçe veriyor ve bir kişi daha değerlendirme yapıyor* (Katılımcı C).

Katılımcı D mevcut puanlayıcı sayısının yeterli olduğunu vurgulamıştır: "*2'den fazla puanlayıcı çok fazla iş demektir. İş yükünün artmasının güvenirlik üzerinde olumsuz bir etkisi olabilir*."

DPA Kullanmaya Yönelik Eğitim Verilmesi

Elde edilen bir diğer tema, puanlayıcılara DPA kullanmaya yönelik eğitim verilmesidir. Katılımcı A, DPA kullanarak güvenilir puanlama yapma konusunda, DPA kullanmaya ilişkin verilmesi gereken eğitimin önemini vurgulamıştır:

> *Revizyon çalışmalarına katılmayan bazı hocalar, bazen, DPA'da yer alan açıklamaları yeterli bulmuyor ve bu açıklamamaların ne anlama geldiğini anlamadıklarını belirtiyor. Bu sorunla, her sene başı revizyon toplantıları düzenleyip tüm hocaları davet ederek ve DPA kullanmaya yönelik eğitimler vererek başa çıkabiliriz. Eğitimlerde, DPA kullanmaya yönelik deneyimi az ve çok olan hocalar beraber örnek kağıtlar üzerinde çalışabilirler, tartışıp ortak bir karara varabilirler* (Katılımcı A).

Puanlayıcı Önerileri

Katılımcılara son olarak, DPA'nın güvenilir bir şekilde kullanılması için önerileri sorulmuştur. Katılımcı A, DPA'nın her yıl revize edilmesi gerektiğini ve bu deneyim yılı gözetmeksizin, sene boyunca İngilizce yazılı anlatım becerisini DPA kullanarak değerlendiren tüm puanlayıcıların katılması gerektiğini vurgulamıştır. Ayrıca DPA kullanmaya yönelik eğitimler verilerek, DPA'da kullanırken uzlaşılamayan

ölçütler konusunda fikir birliğine varılabileceğini belirtmiştir. Katılımcı B, ölçütlerin açıklamalarının daha kısa olması gerektiğini söylemiştir.

> *DPA'nın ölçütleri elbette gelişmeye açıktır ve zamanla problem yaşandıkça revize edilmelidir. Fakat mevcut durumda, ölçütlerin çalıştığını görüyoruz. Belki ölçütlerin açıklamaları, kafa karışıklığı yaratmamak için daha kısa olabilir. Zaten bir çok sınav kağıdı değerlendiriyoruz. Bir de ölçütleri okuyarak zaman kaybetmemeliyiz* (Katılımcı B).

Bu fikir, puanlama esnasında zaman kazanmak ile açıklanabilir. Katılımcı C ise, katılımcı B'nin tam aksine, ölçütlerin açıklamalarının daha detaylı olması gerektiğini söylemiştir. Bu, puanlayıcıların puanlama yaparken nelere dikkat etmeleri gerektiğini, kompozisyonun hangi niteliğine hangi puanların verileceğine dair ölçütleri daha detaylı olarak bilmeleri gerektiğini göstermektedir. Üzerinde birleşilen bir diğer fikir, puanlanması gereken sınav kâğıtlarının ikiye bölünüp iki ayrı günde yapılması yönündedir. İş yükünün azalması güvenirliği artıracaktır. Son olarak, katılımcı C işe yeni başlayan okutmanların, o yıl yazma dersi vermeleri halinde, DPA kullanarak puanlama yaparken, daha güvenilir puanlama yapacakları hususunu vurgulamıştır.

BÖLÜM V

SONUÇLAR VE ÖNERİLER

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 ele alınmıştır. Sonuçlar aşağıda özetlenmiştir:

Nicel veri analizi sonucunda elde edilen verilere göre nitelikli bir DPA kullanılarak yapılan puanlamalarda, puanlayıcı özelliklerinden olan DPA kullanmaya ilişkin deneyim yılının güvenirlik üzerinde etkili olmadığı sonucuna varılmıştır. Ayrıca, İngilizce yazılı anlatım becerisine yönelik hazırlanan yazılı yoklamanın değerlendirilmesinde deneyim yılları farklı puanlayıcıların yer aldığı durumlar için hesaplanan güvenirlik katsayılarının farklılık yaratmadığı, tüm gruplarda hesaplanan G ve Phi katsayıları ve bağıl ve mutlak hata varyansları arasında tutarlılık olduğu görülmüştür. Bu sonuçlardan hareketle, deneyim yılının güvenilir puanlama yapmada etkili olmadığı, iyi hazırlanmış bir DPA ile puanlamaya yönelik deneyimi çok ve deneyimi az puanlayıcılardan elde edilen sonuçların güvenirliğinin birbirine çok yakın olduğu sonucuna varılmıştır. Diğer bir deyişle, puanlayıcı özellikleri iyi hazırlanmış bir DPA'nın kullanıldığı durumlarda güvenirliği etkilememiştir.

İngilizce yazılı anlatım becerisini ölçmek için hazırlanan yazılı yoklama (b:p)xm yuvalanmış deseni ile deneyimi çok, deneyimi az ve karışık gruplar için hesaplanan varyans bileşenleri değerlerinden hareketle, her üç durumda da maddelerin bireylere eşit güçlükte geldiği, puanlayıcıların puanlama bakımından farklılığa neden olmadığı, öğrencilerin yazılı anlatım becerileri bakımından farklılaşmadığı tespit edilmiştir. Ancak puanlayıcıların her birinin farklı bireyleri farklı puanladığı sonucuna varılmıştır. Diğer bir ifade ile İngilizce yazılı anlatım becerisinin değerlendirilmesi için farklı deneyim yıllarına sahip puanlayıcıların aynı sayıda öğrenciyi puanlamasıyla oluşturulmuş (b:p)xm yuvalanmış deseni ile yapılan G çalışmaları sonucunda hesaplanan varyans bileşenleri ve toplam varyansı açıklama oranları her üç durumda da

birbirine yakın ve yüksek sonuçlar üretmiştir ve yuvalanmış desenler ile de yüksek güvenirlik katsayıları elde edilebileceği sonucuna varılabileceğini göstermiştir.

Puanlayıcılar ile yapılan görüşmeler sonucu elde edilen nitel verilerden hareketle, yuvalanmış desenlerin kullanıldığı durumlarda yorgun etkisinin azalttığı sonucuna varılmıştır. Dolayısıyla, İngilizce yazılı anlatım becerisi değerlendirilmesi gereken öğrenci sayısının çok olduğu ve güvenirlik açısından birden fazla puanlayıcı tarafından puanlama yapılması gereken durumlarda ve puanlayıcılar arası tutarlık sağlandığında, tüm öğrencilerin tüm puanlayıcılar tarafından puanlandığı desenler yerine, puanlayıcıların bir kısmının, öğrencilerin bir kısmını puanladığı yuvalanmış desenlerinin kullanılmasının daha yerine olacaktır. 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. Puanlayıcılarla yapılan görüşmelerden elde edilen ve nicel verilerle desteklenen bu sonuç Nalbantoğlu (2009) tarafından yürütülen çalışmanın sonuçlarını destekler niteliktedir.

Görüşmeye katılan puanlayıcılar DPA ile yapılan değerlendirmenin birçok açıdan güvenilir olduğunu belirtmişlerdir. DPA ile yapılan değerlendirmenin güvenilir olduğuna ilişkin sonuç, Andrade ve Du (2005) tarafından yürütülen ve öğrencilerin DPA'yı kullanılması durumunda daha güvenilir puanlar aldıkları sonucuna varan çalışmayı destekler niteliktedir. DPA'nın öğretmen beklentilerini açık hale getirdiğine ilişkin sonuç, Aslanoğlu ve Kutlu (2003) tarafından yürütülen ve DPA'nın öğretmen beklentilerini açık hale getirdiğini vurgulayan çalışmayı desteklemektedir.

Görüşmeye katılan puanlayıcılar DPA ile yapılan değerlendirmenin bazı dezavantajlarından da bahsetmişlerdir. Bu dezavantajlardan bir tanesi, DPA kullanmanın çok zaman almasıdır. DPA'nın bu olumsuzluğu Aslanoğlu ve Kutlu'nun (2003) çalışmasında DPA kullanan öğretmenlerin her ölçüt hakkında ayrıntılı bir şekilde düşünmek zorunda kaldıklarını şeklinde dile getirilmiştir. DPA kullanmanın diğer bir dezavantajı, değerlendirilmesi gereken sınav kâğıdı sayısı arttıkça, yorgunluk etkisi ile güvenirliğin düştüğüdür. Bu sonuç, Demir, Kutlu ve Yıldırım (2014) tarafından yürütülen ve sınıf mevcudunun öğretmenlerin DPA'lara yönelik tutumlarını etkileyen bir faktör olduğu sonucuna varılan çalışmayı destekler niteliktedir.

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, analitik DPA ile elde edilen puanların yüksek güvenirlik gösterdiği ile sonuçlanan, alan yazındaki birçok çalışmayla örtüşmektedir (Kan, 2005; Kasap, 2008; Özel ve Acar, 20014; Büyükkıdık ve Anıl, 2015).

Araştırmadan elde edilen sonuçlar doğrultusunda, bazı önerilerde bulunulmuştur:

- Güvenirliği yüksek bir DPA kullanılarak yapılan değerlendirmelerin güvenirliği üzerinde dereceli puanlama anahtarı kullanmaya ilişkin deneyim etkili olmadığından, puanlayıcı grupları oluşturulurken puanlayıcıların deneyim yıllarını göz önünde bulundurmak güvenirliğini artırmamaktadır.
- 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.
- Puanlayıcı gruplarında beraber puanlama yapan puanlayıcıların sayısını artırmak, iş yükünü artıracağından güvenirliği olumsuz etkileyebilir. Puanlayıcı sayısını artırmak yerine, puanlayıcılar arası tutarlılık sağlanamadığı durumlarda başka puanlayıcılardan yardım alınabilir.
- Çok sayıda sınav kâğıdının kısa bir süre içinde puanlanması gerektiğinde, tüm sınav kâğıtlarının aynı gün içinde değerlendirilmeye çalışılması yorgunluk etkisi ile güvenirliği düşürebilir. Böyle bir durumda, değerlendirme görevinin birkaç güne bölünmesi güvenirliği olumlu etkileyebilir.
- Kalabalık bir ortamda tüm puanlayıcılar tarafından yapılan puanlamalar esnasında gürültü gibi kontrol edilemeyen dış etkenler güvenirliği düşürebilir. Puanlayıcıların puanlama yerlerine kendilerinin karar vermesi güvenirliği artırabilir.
- Puanlayıcılara dereceli puanlama anahtarını kullanmak için eğitim verilmesi, puanlayıcıların dereceli puanlama anahtarında yer alan ölçütleri daha iyi anlamasını sağlayacak ve dolayısıyla puanlama güvenirliğini artırabilir.

Yapılan bu çalışmaya benzer çalışmalar yapacak olan araştırmacılar için aşağıda bazı önerilerde bulunulmuştur:

- Bu araştırmada elde edilen sonuçlara bağlı olarak araştırmacılar puanlayıcılar arası güvenirlik çalışmalarında genellenebilirlik kuramına dayalı olarak karar çalışmaları planlayabilirler.
- Bu araştırma, 6 puanlayıcılı örneklemlerle gerçekleştirilmiştir. Benzer çalışmalar yapacak olan araştırmacılar, farklı puanlayıcı sayıları kullanarak araştırmalarını düzenleyebilirler.
- Kullanılan puanlama anahtarlarının 6 ölçütü vardır. Bu araştırmada kullanılan dereceli puanlama anahtarı dışında farklı sayıda ve türde dereceli puanlama anahtarı kullanılabilir.
- Bu araştırma kapsamında güvenirliğin değerlendirilmesinde, G kuramının yuvalanmış (b:m)xp deseni ile analizi yapılmıştır. Benzer çalışmalarda tümüyle yuvalanmış ya da tümüyle çaprazlanmış desen kullanılabilir.
- Bu çalışmada İngilizce yazılı anlatım becerisi değerlendirilmiştir. Benzer çalışmalarda farklı beceriler değerlendirilebilir.

KAYNAKÇA

Anastasi, A. & Urbina, S. (1997). *Psychological Testing*. New York: MacMillan Pub. Co.Inc.

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

Arıcı, İ. (2007). *İlköğretim Din Kültürü ve Ahlak Bilgisi Dersinde Öğrenci Başarısını Etkileyen Faktörler (Ankara Örneği)*. Yayınlanmamış Doktora Tezi. Ankara Üniversitesi Sosyal Bilimler Enstitüsü: Ankara.

Arıkan, S., Çelen, Ü. Gülleroğlu, H. D., Gültekin, S., Kilmen, S., Köse, A. (Ed: Çıkrıkçı-Demirtaşlı, N.) (2012). *Eğitimde Ölçme ve Değerlendirme*. Ankara: Edge Akademi.

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.

Atılgan H. (2004) *Güzel Sanatlar Eğitimi Müzik Öğretmenliği Özel Yetenek Seçme Sınavlarının Geçerlik ve Güvenirliği (G Kuramı Çalışması).* XIII. Ulusal Eğitim Bilimleri Kurultayı. (sf.496-497). Malatya: İnönü Üniversitesi.

Atılgan, H. ve Tezbaşaran, A. A. (2005). Genellenebilirlik Kuramı Alternatif Karar Çalışmaları ile Senaryolar ve Gerçek Durumlar için Elde Edilen G ve Phi Katsayılarının Tutarlılığının İncelenmesi. *Eğitim Araştırmaları Dergisi*, Sayı 18, 28-40.

Atılgan, H. (2008) Using Generalizability Theory to assess the Scorerer liability of the Special Ability Selection Examinations for Music Education Programmes in Higher Education. *International Journal of Researchand Method Education*, 31(1) 63-76.

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

Barneveld, C. (2005). The Dependability of Medical Students' Performance Ratings as Documented on In Training Evaluations. *Academic Medicine*, 80(3), 309-312.

Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Handbook I: Cognitive domain. New York: David McKay Company.

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

Brualdi, A. (1998). Implementing Performance Assessment in The Classroom. *Practical Assessment, Research & Evaluation,* 6(1-3). http://PAREonline.net/getvn.asp?v=6&n=2.

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.

Bilican Demir, S., Kutlu, Ö. ve Yıldırım, Ö. (2014). *İlköğretim Öğretmenlerinin Dereceli Puanlama Anahtarlarına İlişkin Tutumlarını Yordayan Faktörler*. Eğitim ve Bilim. 39(173), 92-102

Demirtaş, H., Çınar. İ. (2004). Yönetici, Öğretmen, Veli ve Öğrencilerin Başarı Algısı ve Eğitime İlişkin Görüşleri (Malatya İli Örneği). *XIII. Ulusal Eğitim Bilimleri Kurultayı*. Malatya: İnönü Üniversitesi.

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.

Good, V. C. (1973). *Dictionary of Education*. New York: McGraw Hill Book Company.

Goodrich, A. H. (2001). Theeffects of instructional rubrics on learning to write. *Current Issues ID Education*, 4(4), 1-4. http://cie.asu.edu/volume4/number4/.

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.

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

Güler, N., Uyanık, K. G. ve Teker, T. G. (2012). *Genellenebilirlik kuramı*. PegemA Yayıncılık. Ankara.

Gronlund, N. E. (1977). *Constructing achievement test*. New Jersey: Prentice– Hall, Inc.

Farrokhi, F., Esfandiari, R. & Schaefer, E. (2012). *A Many-Facet Rasch Measurement of differential rater severity/leniency in sel assessment, peer assessment, and teacher assessment.* Journal of Basic and Applied Scientific Research, 2 (9), 8786-8798.

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.

Karakaya,İ. (2015). Comparison of Self, Peer and Instructor Assessments in the Portfolio Assessment by Using Many Facet Rasch Model . Journal of Education and Human Development, 4(2)(182-192).

Karasar, N. (2005). *Bilimsel araştırma yöntemleri*. 14. Baskı. Ankara: Nobel Yayın Dağıtım.

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

Koç, N. (1981). *Liselerde Öğrencilerin Akademik Başarılarının Değerlendirilmesi Uygulamalarının Etkililiğine İlişkin Bir Araştırma*. Ankara Üniversitesi Eğitim (Bilimleri) Fakültesi Yayınları, No: 104.

Koç, N. (1985). Standart Başarı Testlerinin, Bir Eğitim Sisteminde Verilen Çeşitli Kararlardaki Yeri ve Önemi. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 17 (1-2), 159-172.

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

Lane, S & Sabers, D. (1989). Use of Generalizability Theory for Estimating the Dependability of a Scoring System for sample Essays. *Applied Measurement In Education*, 2(3),195-205.

Lee, Y., W. & Kantor, R. (2007). Evaluating Prototype Tasks and Alternative Rating Schemes for a New ESL Writing Test through G Theory. *International Journal of Testing*, 7(4), 353-385.

Linacre, J.M. (2001). *A user's guide to facets: Rasch measuremen tcompute rprogram*. Chicago: MESA.

Montgomery, K. (2000). Classroom rubrics systematizing what teachers do naturally. *The Clearing House*, 73(6), 324-328.

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.

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.

Özçelik, D. A. (2011). *Eğitimde Ölçme ve Değerlendirme* (4. Baskı). Ankara: Pegem Akademi.

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.

Rentz, J.O. (1987). *Generalizability theory: a comprehensive method for assessing and improving the dependability of marketing measures.* Journal of Marketing Research, Vol. 24 No. 1, pp. 19-28.

Sağlam-Arslan, A., Devecioğlu-Kaymakçı, Y., Arslan, S.,(2009). Alternatif Ölçme Değerlendirme Etkinliklerinde Karşılaşılan Problemler: Fen ve Teknoloji Öğretmenleri Örneği. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi*, 28(1), 1-12.

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

Tekin, H. (2009). *Eğitimde ölçme ve değerlendirme*. Ankara: Yargı yayınları, No: 17.

Turgut, M. F. (1997). *Eğitimde Ölçme ve Değerlendirme Metotları*. Ankara: Saydam Matbaacılık.

Turgut, M. F. ve Baykul, Y. (2014). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi.

Wang, Z. (2005). *Estimating Reliability Under A Generalizability Theory Model For Writing Scores in C-Base*. Master Thesis. University of Missouri, Columbia.

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

Yelboğa, A. (2008). Güvenirliğin Değerlendirilmesinde Genellenebilirlik Kuramı'nın Kullanılması: Endüstri ve Örgüt Psikolojisinde Bir Uygulama. *Psikoloji Çalışmaları Dergisi*. 28, 35-54.

Yıldırım, C. (1999). *Eğitimde Ölçme ve Değerlendirme*. Ankara: ÖSYM Yayınları.

EK A: İngilizce Yazılı Anlatım Becerisi Analitik Dereceli Puanlama Anahtarı

	1. Dilbilgisi	2. Kelime	3. Tutarlılık ve Bağlantı	4. Geçişler	5. Örnekler	6. Sonuç Cümlesi
4	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	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	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	Öğrenilen dilbilgisi yapılarını neredeyse hiçbirini kullanmadan kurulan, neredeyse tüm cümlelerde hata barındıran, anlatılmak istenen mesajın anlaşılmadığı cümleler.	Çok sınırlı ve çoğu zaman yanlış anlaşılmanla sebep olan kelimeler	Paragraf ya da metin türünün gerektirdiği mesajı, tamamen ilgisiz örnekler ile, anlaşılması imkansız bir halde verme			

EK B: Written Skill Evaluation Exam

Content	Grade	Learning Outcomes	Assessment
English Writing Class- Opinion Essay	University Prep Class- B Level	Writing Skill	Analytical Rubric

- We should-shouldn't get a driver's license before 18.
- We should-shouldn't attend to obligatory prep class at university
- We should-shouldn't keep animals at the zoo

Instruction: Write a well- developed opinion essay on one of the topics above in 50 minutes. (250-300 words)

EK C: Puanlayıcıların İngilizce Yazılı Anlatım Dersi Dereceli Puanlama Anahtarına Yönelik Görüşleri

Puanlama sürecine katılan puanlayıcılara kullandıkları dereceli puanlama anahtarına yönelik sorular sorulmuş ve aşağıdaki ifadelere ulaşılmıştır:

Katılımcı A:

DPA kullanarak değerlendirme yapma sürecinizden kısaca bahseder misiniz? (Bir kâğıdı ne kadar sürede okursunuz? Kaç sınav kâğıdı sonra yorulmaya başlarsınız? Değerlendirdiğiniz ilk sınav kağıdı ile son sınav kağıdı arasında güvenirlik açısından bir fark olduğunu düşünüyor musunuz? Düşünüyorsanız bu neden kaynaklanıyor sizce? Değerlendirme yaparken ara verir misiniz? Sizce, diğer okutmanlara göre sınav değerlendirme hızınız nasıl? Önce komposizyonu tamamen okuyup, daha sonra mı değerlendirme yaparsınız yoksa komposizyonu her bir değerlendirme boyutu için tek tek bir daha mı okursunuz?)

Öncelikle, puanlama yapılan sınıfa geçiyorum. Puanlama grup liderlerinin yaptığı açıklama ya da uyarıları dinliyorum. Değerlendirme yapacağımız kompozisyon türü ile ilgili dikkat etmem gereken noktalara dikkat ediyorum ve gerekirse not alıyorum. Aslında, uzun zamandır puanlama anahtarı kullanarak değerlendirme yaptığım için nelerin önemli olduğunu biliyorum, ama yine de bazen gözümden kaçan şeyler olabiliyor. Puanlama yapacağım kompozisyonları aldıktan sonra ilk baktığım ana fikir cümlesi oluyor. Ana fikir cümlesinin kompozisyondaki yeri ve formatı belli olduğu için ilk olarak, DPA'daki ölçüte bakıp ana fikir cümlesini değerlendiriyorum. Sonuç cümlesinin de yeri ve formatı belli. Onun da değerlendirmesini tüm kompozisyonu okumadan önce yapabiliyorum. Daha sonra bütün kompozisyonu okuyup, puanlama işlemini tamamlıyorum.

Değerlendirme esnasındaki hızımın diğer okutman arkadaşlara kıyasla, ortalama olduğunu düşünüyorum. Benden çok daha yavaş puanlama yapan okutman arkadaşlarım var ve bence daha fazla pratik yapmaya ihtiyaçları var. Yavaş olmaları, dereceli puanlama anahtarı konusunda da yeni olmalarından kaynaklanıyor. Kısacası, DPA kullanma konusunda deneyimi az olan okutman arkadaşlarımın yavaş puanlama yaptıklarını düşünüyorum. Diğer bir yandan, bazı okutman arkadaşlarımın aşırı hızlı değerlendirme yaptıklarını görüyorum. Elbette ki yaptıkları değerlendirmelerin güvenirliğine ilişkim şüphelerim var. Yavaş değerlendirme yapmak güvenilir değerlendirme yapmak değil her zaman, bunu biliyorum. Fakat puanlayıcılar arası güvenirliğin düşük çıktığı sınıfların puanlayıcıları genelde çok hızlı puanlama yapan puanlayıcılar oluyor.

Puanlama esnasında mola veriyorum. Mola zamanım, okuduğum kompozisyon türüne, değerlendirdiğim sınıfın seviyesine, değerlendirmem gereken kompozisyon sayına göre değişiyor. Bazen partnerime göre bile değişiyor. Eğer partnerim DPA kullanmaya yönelik deneyimi az bir okutman ise, bana çok soru soruyor ve bu da beni daha fazla yoruyor ve daha erken mola veriyorum. Ama partnerim DPA kullanmaya yönelik deneyimi çok bir okutman ise, mola verme sürem uzayabiliyor. Bazen de ofiste tek başıma puanlama yapıyorum. O zaman beni bölecek bir şey olmuyor. Mola vermeyebiliyorum.

Dereceli Puanlama Anahtarını ne kadar süredir kullanıyorsunuz?

6 senedir kullanıyorum.

DPA kullanmanın avantajları nelerdir?

Değerlendirmemiz gereken kompozisyonu, DPA kullanarak birçok farklı açıdan değerlendirme şansımız oluyor. Bütüncül bir değerlendirme yaptığım zaman, güvenilirliğine dair şüphelerim oluyor. Fakat DPA kullandığım zaman, kendimden emin bir şekilde puanlama yapıyorum ve öğrencinin gerçekten hak ettiği puanı verdiğimi düşünüyorum.

Ayrıca, hâlihazırda kullandığımız DPA'yı her sene yine biz gözden geçiriyoruz. DPA bir nevi öğrenciden istediklerimizi yansıtma şeklimiz oluyor. Öğrencilerden istediklerimizi belirttiğimiz için de, öğretimimizi bu şekilde yapıyoruz. Kendimize de dönüt vermiş oluyoruz aslında.

DPA kullanmanın dezavantajları nelerdir?

Dezavantaj olarak, DPA kullanarak değerlendirme yapmanın gerçekten çok zaman alıcı olduğunu söyleyebilirim. Bütünsel değerlendirmeler de yapıyoruz zaman zaman. Bu tür değerlendirmeler çok az zaman alıyor. Ancak güvenirlik açısından baktığımız zaman, DPA kullanılarak yapılan değerlendirmelerin, bütünsel değerlendirmelerden daha güvenilir sonuçlar verdiğini söyleyebilirim.

DPA'yi güvenilir bir şekilde kullanmak için, DPA kullanma açısından deneyimi çok olmanın etkili düşünüyor musunuz?

Açıkçası, DPA kullanmaya ilişkin deneyimi az ve deneyimi çok okutmanlar arasında güvenilir puanlama yapma açısından bir fark olduğunu düşünüyorum. Okutmanların zaman geçtikçe ve DPA kullanma konusunda kendilerini rahat hissettikçe daha güvenilir puanlamalar yaptıkları kanısındayım. Örneğin bundan 3 yıl önce, şimdiki kadar güvenilir puanlama yaptığımı düşünmüyorum. Zamanla DPA'nın ölçütlerine daha çok hâkim olduğumu söyleyebilirim. Fakat, DPA kullanmaya yönelik benden daha deneyimi çok bir okutmanın şu anda benden daha güvenilir bir değerlendirme yaptığını da söyleyemem. Sorun sadece ilk yıllarda oluyor bence. Belli bir süre geçtikten sonra, okutmanlar DPA kullanmaya alışmış oluyorlar. Yani DPA kullanmaya ilişkin 1 yıllık deneyime sahip olan okutmanın yaptığı değerlendirmeler ile 8 yılık deneyime sahip okutmanın yaptığı değerlendirmeler arasında fark olacaktır. Bu fark yıllar geçtikçe kapanacaktır. 5 yıllık deneyime sahip bir okutman ile 8 yıl deneyime sahip bir okutman arasında bir güvenirlik farkı olacağını söyleyemem.

İdeal bir puanlayıcı grubundaki puanlayıcı deneyim yılı kaç olmalıdır?

Buna kesin bir şey söylememek lazım. Bazen DPA kullanmaya yönelik 2 yıl deneyime sahip bir okutman o kadar çok sınav değerlendirme görevi alıyor ki, 5 yıl deneyime sahip bir okutmandan çok daha güvenilir değerlendirme yapacak duruma geliyor. Burada önemli olan, yaptığımız değerlendirme sayısı, deneyim yılı değil. Bir de, DPA'nın kriterlerini iyi incelemek gerekiyor. Ama yine de ben en az 4-5 yıl deneyimi çok bir okutmanın güvenilir bir puanlama yaptığını düşünüyorum.

Grup kaç puanlayıcıdan oluşmalıdır?

Bence güvenirlik kesinlikle puanlayıcı sayısından etkileniyor. Ne kadar çok puanlayıcı değerlendirme yaparsa, ben güvenirliğin o derece arttığını düşünüyorum. Ama ideal sayı söyleyebilmek için tüm hususları düşünmek gerekiyor. Sonuçta okutmanlara da iş yükü oluyor. O yüzden iki puanlayıcının yeterli olduğunu düşünüyorum. Sadece puanlayıcılar arasında puan farkının çok olduğu durumlarda 3. puanlayıcıya danışılabilir diye düşünüyorum. Bu 3. puanlayıcı belki de deneyimi çok bir puanlayıcı olabilir.

Tüm puanlayıcılar tüm bireyleri puanlamalı mıdır? Cevabınız hayır ise, öneriniz nedir?

Ben her puanlayıcının her bireyi puanlamasının gereksiz ve yorucu bir iş olduğunu düşünüyorum. Bir gruptaki bireylerin puanlanması, puanlayıcılar arasında iş bölümü yapılıp yapılabilir. Aksi takdirde, yorgunluk etkisi olacağından, yine sağlıklı bir değerlendirme yapılacağını düşünmüyorum.

Çok sayıda sınav kâğıdını değerlendirmek puanlama güvenirliğini nasıl etkiler?

İlk değerlendirdiğim kompozisyonların, son değerlendirdiğim kompozisyonlardan güvenirlik açısından daha yüksek olduğunu söyleyebilirim. Belli bir zaman sonra, sanki herkese aynı puanı veriyormuşum gibi geliyor. Bakıyorum kağıtlara. Sanki hep 75-80 veriyormuşum gibi. Bunu yorgunluk etkisine bağlıyorum. Çünkü yazma dersi değerlendirmelerini genellikle Cuma günü ders çıkışı yapıyoruz ve çok yorgun oluyoruz. Yazma dersi değerlendirme işi örneğin sabah erken saatlerde olsa, güvenirlik daha yüksek olur diye düşünüyorum. Yeni hocalar da, eski hocaların hızına yetişmek için çaba göstermek yerine, kendi hızlarında daha güvenilir bir şekilde değerlendirme yapabilirler. Ya da yine sadece yazma dersi puanlamaya bir gün ayrılsa ve o gün derse girmeden sadece puanlama yapsak, daha fazla kompozisyonu daha güvenilir bir şekilde puanlayabiliriz diye düşünüyorum.

Yorgunluk etkisi güvenirliği nasıl etkiler?

İş bölümü yapmanın uygun bu noktada uygun olduğunu düşünüyorum. Ama bu iş bölümü, DPA'da bulunan boyut bazında olmamalı. Yani, bazı okutmanlar sadece dilbilgisi bazıları kelime bazında bir değerlendirme yaparsa, bu daha yorucu bir iş olabilir ve okutmanlarda kafa karışıklığı yaratabilir. Ayrıca iş yükünü daha da artırabilir. Eğer iş yükünü azaltmak istiyorsak, değerlendirmemiz gereken kompozisyon sayısını azaltmalıyız. Güvenirliği artırmak istiyorsak, değerlendirme yapacağımız saatleri çok iyi ayarlamalıyız ki bedenen yorgun olmayalım.

Puanlama anahtarının daha etkili bir şekilde kullanılması için önerileriniz nelerdir?

DPA'ların öncesinde kesinlikle okutmanlara gönderilmesi gerektiğini düşünüyorum. Ve her sene yapılan DPA revizyon çalışmalarında, sadece ölçme değerlendirme birim elemanlarının değil, deneyim yılı gözetmeksizin, sene boyunca yazma dersini DPA kullanarak değerlendiren tüm okutmanların bulunması gerektiğini düşünüyorum. Aksi takdirde yapılan değişiklikler üstten dayatma şeklinde olacaktır ve bu anahtarı kullanan hocalar, neyi neden yaptıklarının bilincinde olmayacaklardır. Revizyon çalışmalarına katılmayan bazı hocalar, bazen, DPA'da yer alan açıklamaları yeterli bulmuyor ve bu açıklamamaların ne anlama geldiğini anlamadıklarını belirtiyor. Bu sorunla, her sene başı revizyon toplantıları düzenleyip tüm hocaları davet ederek ve DPA kullanmaya yönelik eğitimler vererek başa çıkabiliriz. Eğitimlerde, DPA kullanmaya yönelik deneyimi az ve çok olan hocalar beraber örnek kağıtlar üzerinde çalışabilirler, tartışıp ortak bir karara varabilirler.

Katılımcı B:

DPA kullanarak değerlendirme yapma sürecinizden kısaca bahseder misiniz? (Bir kâğıdı ne kadar sürede okursunuz? Kaç sınav kâğıdı sonra yorulmaya başlarsınız? Değerlendirdiğiniz ilk sınav kağıdı ile son sınav kağıdı arasında güvenirlik açısından bir fark olduğunu düşünüyor musunuz? Düşünüyorsanız bu neden kaynaklanıyor sizce? Değerlendirme yaparken ara verir misiniz? Sizce, diğer okutmanlara göre sınav değerlendirme hızınız nasıl? Önce komposizyonu tamamen okuyup, daha sonra mı değerlendirme yaparsınız yoksa komposizyonu her bir değerlendirme boyutu için tek tek bir daha mı okursunuz?)

Sınav kağıdı paylaşımı olduktan sonra, değerlendirmeye başlamadan önce, öncelikle DPA'yı baştan sona okuyorum. ölçütleri gözden geçiriyorum. Örneğin dilbilgisine tam puan vermek için hangi ölçütler karşılanmalı, tek tek okuyorum. Ölçütleri kafamda tam olarak oturttuktan sonra, değerlendireceğim sınav kağıtlarını değerlendirmeye başlıyorum. Değerlendirme esnasında, DPA'ya bir daha geri dönmüyorum. Böyle bir değerlendirme yapmanın daha sağlıklı olduğunu düşünüyorum.

Bazen paragraf bazense kompozisyon değerlendiriyoruz. Kompozisyon değerlendirirken 15. sınav kağıdından sonra epey yorulduğumu söyleyebilirim. Belki okumaya devam edebilirim ama bu sağlıklı olmaz. O yüzden mola veriyorum. Paragraf değerlendirmede bu sayı daha yukarı çıkabiliyor. 30 paragrafa kadar okuyabiliyorum.

Değerlendirdiğim ilk ve son sınav kağıdı arasında, ben farkında olmasam da güvenirlik farkı olduğunu düşünüyorum. İlk kağıtları daha dikkatli değerlendirebiliyoruz. Ama sona yaklaştıkça ister istemez bir dikkat dağınıklığı oluyor. Bunu aşmak için yapmamız gereken şey, değerlendirilmesi gereken kağıt sayısını azaltmaktır. Bir grup içinde değerlendirme yapan puanlayıcılar aralarında iş birliği yapabilirler belki. Herkes her kağıdı değerlendirirse eğer, iş yükü ile beraber dikkat sorunu da ortaya çıkacak ve bu da güvenirliği düşürecektir. Ya da kesin mola saatleri konulabilir. Örneğin her puanlayıcının 10 sınav kağıdı değerlendirdikten sonra molaya çıkması zorunlu olabilir çünkü bazı okutmanlar mola veriyor bazıları ise vermeyip devam ediyor. Bu da biraz da olsa dikkatimi dağıtıyor.

Dereceli Puanlama Anahtarını ne kadar süredir kullanıyorsunuz?

Yaklaşık 5 yıldır kullanıyorum.

DPA kullanmanın avantajları nelerdir?

DPA, kesinlikle güvenilir bir puanlama yapmamızı sağlıyor. Bütüncül bir değerlendirme yapmak, subjektif değerlendirmelere yol açabilir. Ayrıca, dikkat etmemiz gereken alt boyutları gözden kaçırmamıza neden olabilir. Örneğin öğrenci çok çeşitli kelimeler kullanmış olabilir ama yaptığı dilbilgisi yanlışlarından dolayı çok düşük olabilir. DPA her bir boyut üzerine eğilmemize yardımcı oluyor.

DPA kullanmanın dezavantajları nelerdir?

Genel olarak bir dezavantajının olduğunu düşünmüyorum. Sadece biraz zaman alıcı olduğunu söyleyebilirim. Güvenilir bir değerlendirme yapabilmemiz için, DPA kullanmamızın şart olduğunu düşünüyorum.

DPA'yı güvenilir bir şekilde kullanmak için, DPA kullanma açısından deneyimi çok olmanın etkili olduğunu düşünüyor musunuz?

Deneyimi çok olmak elbette etkili ama ben bazen tam tersini düşünebiliyorum. Özellikle işe yeni başlayan okutman arkadaşlarımız hata yapmamak için çok dikkatli değerlendirmeler yapıyorlar. Sık sık DPA'ya başvurup, her bir ölçütü inceliyorlar. Deneyimi çok olan okutmanlar ise, DPA'yı artık çok iyi bildiklerini varsayıp, değerlendirme esnasında bir daha dönüp bakma gereği duymuyorlar ve hata yapabiliyorlar.

İdeal bir puanlayıcı grubundaki puanlayıcı deneyim yılı kaç olmalıdır?

Kendimden örnek verirsem, 2 yıllık bir DPA kullanma sürecimin ardından, güvenilir puanlama yapar hale geldiğimi söyleyebilirim. 2.yıldan itibaren hem daha hızlı hem de daha güvenilir puanlamalar yapmaya başladım diyebilirim.

Grup kaç puanlayıcıdan oluşmalıdır?

Puanlayıcı sayının artırılması, güvenirliği elbette artıracaktır. 2 kişi yerine 3 kişinin aynı sonuca ulaşması, bu 3 kişinin de işini doğru yaptığının bir göstergesidir. Bu dediğim, yorgunluk etkisi yok sayıldığında mümkün olan bir şeydir. Fakat işin içine girecek olan yorgunluk etkisi ile güvenirlik elbette düşecektir.

Tüm puanlayıcılar tüm bireyleri puanlamalı mıdır? Cevabınız hayır ise, öneriniz nedir?

Tüm puanlayıcıların tüm bireyleri puanlaması teoride istenen ama pratikte imkânsız olan bir şey bence. Güvenirlik elbette artar bu şekilde. Fakat bence , puanlayıcılar bireyleri dönüşümlü olarak puanlamalı.

Çok sayıda sınav kâğıdını değerlendirmek puanlama güvenirliğini nasıl etkiler?

20 kişiden sonra, ister istemez dikkat dağınıklığı yaşıyoruz. Bu sayının üzerinde değerlendirme yapmamız gerekiyorsa, bence aynı gün içinde olmamalı. Aksi takdirde güvenirlik düşecektir.

Yorgunluk etkisi güvenirliği nasıl etkiler?

Elbette düşürecektir. Bunun için değerlendirilmesi gereken sınav kağıdı sayısını minimuma indirmemiz gerekir.

Puanlama anahtarının daha etkili bir şekilde kullanılması için önerileriniz nelerdir?

DPA'nın ölçütleri elbette gelişmeye açıktır ve zamanla problem yaşandıkça revize edilmelidir. Fakat mevcut durumda, ölçütlerin çalıştığını görüyoruz. Belki ölçütlerin açıklamaları, kafa karışıklığı yaratmamak için daha kısa olabilir. Zaten bir çok sınav kağıdı değerlendiriyoruz. Bir de ölçütleri okuyarak zaman kaybetmemeliyiz. Bu durumda, güvenirliği daha da artırmak için yorgun olmadığımız bir günde değerlendirme yapmamız gerektiğini düşünüyorum. Bu yüzden öncelikle zaman konusu iyi ayarlanmalı.

Katılımcı C:

DPA kullanarak değerlendirme yapma sürecinizden kısaca bahseder misiniz? (Bir kâğıdı ne kadar sürede okursunuz? Kaç sınav kâğıdı sonra yorulmaya başlarsınız? Değerlendirdiğiniz ilk sınav kağıdı ile son sınav kağıdı

arasında güvenirlik açısından bir fark olduğunu düşünüyor musunuz? Düşünüyorsanız bu neden kaynaklanıyor sizce? Değerlendirme yaparken ara verir misiniz? Sizce, diğer okutmanlara göre sınav değerlendirme hızınız nasıl? Önce komposizyonu tamamen okuyup, daha sonra mı değerlendirme yaparsınız yoksa komposizyonu her bir değerlendirme boyutu için tek tek bir daha mı okursunuz?)

DPA'daki ölçütlere alıştığım için artık tek tek okuma gereği duymuyorum. Fakat, o sene girmediğim kurdaki bir sınıfın İngilizce yazılı anlatım becerisini değerlendirmem gerekiyorsa, ölçme birimince yapılan örnek puanlamayı izleyip notlar alıyorum. Dikkat etmem gereken yerleri kafamda belirliyorum. Daha sonra, bir sınav kağıdını tamamen okuduktan sonra, DPA'ya dönüp tek tek puan veriyorum. Genelde bir sınıfın yazdığı kompozisyonları değerlendirdikten sonra yorulduğumu hissediyorum. Yazma becerisi değerlendirme hızımın ise çok yavaş olduğunu düşünüyorum. Sınıftaki diğer hocalar gittikten sonra bitiyor benim işim. Dikkatli davranıp, hata yapmamaya özen gösterdiğim için hep yavaş puanlama yapıyorum. Değerlendirdiğim ilk kağıt ve son arasında güvenirlik farklı, eğer kur benim ders anlattığım kur ise olmuyor. O sene için konulara ve istenilenlere hâkim olduğum için, sorun olmuyor. Ama eğer o sene ders anlatmadığım bir kursa ve kur sorumlusunun dikkat çekmediği ama önemli olan bir nokta varsa, hata yapabiliyorum. Ya da örneği B kurunda derse giriyorum ve değerlendirdiğim kur C kuru. C kuru B kurundan seviye olarak daha iyi olduğu için, yazdıkları şeyler de ister istemez daha iyi geliyor bana. Aslında dediğim gibi, değerlendirmeye başlamadan önce bir örnek yapıyoruz. Ama farklı kur olduğu zaman olumlu ya da olumsuz etkileniyorum. Üstüne yorgunluk ve diğer partnerler arası konuşmalar da etkilenince, hata yapabiliyorum.

Dereceli Puanlama Anahtarını ne kadar süredir kullanıyorsunuz?

Toplamda 5 yıldır kullanıyorum.

DPA kullanmanın avantajları nelerdir?

DPA'ları ihtiyaca göre sürekli değiştiriyoruz zaten. Örneğin bazen bütün sınıflarda yapılan bir hata çıkıyor ve bunu puanlama anahtarına eklememiz gerekiyor. O zaman da düzenlemelere gidiyoruz. Bu esnek bir süreç ve sürekli yenilenmesi bizim için iyi bir şey. Aslında beni tam anlamıyla etkilemiyor yenilenip yenilenmemesi çünkü çoğu zaman bakma ihtiyacı hissetmiyorum. Ama yeni hocalar için iyi bir şey bu. Ben de yeni geldiğim zaman sürekli bakardım. Her değiştirildiğinde yine bakıyorum ama puanlama sürecinde sık sık bakmam gerekmiyor. Aynı zamanda, iki puanlayıcı arasında çok fazla fark çıkarsa, ölçütleri tek tek gözden geçirip, daha sağlıklı bir karara varma açısından yararlı olduğunu düşünüyorum. DPA olmasa, iki puanlayıcı arasında çok fazla fark çıkabilir. Bunun önüne geçmiş oluyoruz puanlama anahtarı ile.

DPA kullanmanın dezavantajları nelerdir?

DPA'nın bir dezavantajı olduğunu düşünmüyorum. Hatta biraz önce saydığım bir çok avantajı var.

DPA'yı güvenilir bir şekilde kullanmak için, DPA kullanma açısından deneyimi çok olmanın etkili düşünüyor musunuz?

Ben deneyimin güvenirlik üzerinde etkili olduğunu düşünüyorum. Ben yeniyken de başıma geliyordu ve şuan yeni hocaların da başına geliyor. Deneyimi çok puanlayıcılar, öğrencilerin yazması istenen şeyleri çok daha iyi biliyor çünkü o dersi birkaç senedir veriyor oluyorlar. Öğrencilerin ne yazmaları ya da ne yazmamaları gerektiğini daha iyi biliyoruz. Dersi anlatırken de bu doğrultuda anlatıyoruz. DPA da derste anlatılanların bir yansıması aslında. Deneyimi az olan hocalar da elbette hataları bulabilir. Ama mesela hangi hataların göz ardı edilip hangilerinin

edilemeyeceğini bilemez. Bu hususta, DPA'ların çok daha detaylı olması gerekiyor. Ama o zaman puanlama anahtarları aşırı uzun olur ve insanlar okumazlar uzun diye. Bu da iyi bir şey değil.

İdeal bir puanlayıcı grubundaki puanlayıcı deneyim yılı kaç olmalıdır?

Eğer bir puanlayıcı, işe başladığı yıl yazma dersi verir ve puanlama anahtarını kullanarak puanlama yaparsa, bu güvenilir bir puanlama yapması için yeterli olur. Ama yazma dersi vermezse, o puanlayıcı seneler geçse de güvenilir bir puanlama yapamaz. Bence burada yazma dersi verip vermemesi önemli.

Grup kaç puanlayıcıdan oluşmalıdır?

Biz mevcut durumda, iki kişiden oluşan bir grupta puanlama yapıyoruz ve eğer uzlaşamadığımız bir şey olursa, ölçme değerlendirme birim görevlilerinden yardım alıyoruz. 3. bir puanlayıcı güvenirliği etkilemez bence. Çünkü ihtiyaç halinde biz 3. puanlayıcıya danışıyoruz. Ayrıca öğrenci beklediğinden düşük not aldığını düşünüyorsa, dilekçe veriyor ve bir kişi daha değerlendirme yapıyor.

Tüm puanlayıcılar tüm bireyleri puanlamalı mıdır? Cevabınız hayır ise, öneriniz nedir?

Tüm puanlayıcılar tüm bireyleri puanlamamalı. İş bölümü yapılmalı. Aynı zamanda, Ayrıca DPA'da yer alan tüm ölçütler, bir birey için bir puanlayıcı tarafından değerlendirilmeli. Bazen bir puanlayıcının gözünden kaçan bir kelime hatasını, diğer hoca bulabiliyor. Eğer ölçütler puanlayıcılar arasında paylaşılırsa, bu hatayı bulma imkânı kalmaz.

Çok sayıda sınav kâğıdını değerlendirmek puanlama güvenirliğini nasıl etkiler?

Eğer o gün, tüm gün boyunca dersim varsa, elbette yorgunluk etkisi olaya dâhil oluyor. Ama bazen okul, değerlendirilmesi gereken kâğıt sayısını iki güne paylaştırabiliyor. Gerçi bu bir kez oldu. Değerlendirilmeyi bekleyen çok fazla sınav kâğıdı vardı ve iki güne paylaştırıldı. Bu durum bizi olumsuz etkileyecekti. Birim görevlileri de değerlendirme işini iki güne ayırdı. Bence her seferinde böyle olmalı. Eğer olumsuz bir etkilenme istenmiyorsa, iş yükü azaltılmalı. Ben DPA kullanmaya yönelik deneyimi çok olan bir hocanın, çok çok fazla etkilendiğini düşünmüyorum. Bence bu durum deneyimi az olan hocalar için geçerli. Ben bunu kendimden biliyorum. İşe ilk girdiğim zamanlar olumsuz etkileniyordum eğer sınav kâğıdı sayısı fazla ise.

Yorgunluk etkisi güvenirliği nasıl etkiler?

Yorgunluk elbette olumsuz etkiliyor. Bu sorunu aşmak için de 10 dakika ara verme ihtiyacı hissediyorum.

Puanlama anahtarının daha etkili bir şekilde kullanılması için önerileriniz nelerdir?

Puanlamaya başlamadan önce bir kompozisyonu örnek olarak hep beraber puanlıyoruz demiştim. Bu çok güzel bir uygulama. Ama açıklaması daha detaylı olmalı. Örneğin şu şuşu çok önemli, bu bu bunu görmezden gelebiliriz diye. Puanlama anahtarına tüm detayları yazmak mümkün değil ama açıklamalar esnasında daha detaylı bilgi verilebilir.

Katılımcı D:

DPA kullanarak değerlendirme yapma sürecinizden kısaca bahseder misiniz? (Bir kâğıdı ne kadar sürede okursunuz? Kaç sınav kâğıdı sonra yorulmaya başlarsınız? Değerlendirdiğiniz ilk sınav kâğıdı ile son sınav kâğıdı arasında güvenirlik açısından bir fark olduğunu düşünüyor musunuz? Düşünüyorsanız bu neden kaynaklanıyor sizce? Değerlendirme yaparken ara verir

misiniz? Sizce, diğer okutmanlara göre sınav değerlendirme hızınız nasıl? Önce kompozisyonu tamamen okuyup, daha sonra mı değerlendirme yaparsınız yoksa kompozisyonu her bir değerlendirme boyutu için tek tek bir daha mı okursunuz?)

Değerlendirmeye başlamadan önce, DPA'nın ölçütlerini okuyorum ama bölüm bölüm okuyorum. Örneğin dilbilgisine yönelik ölçütleri okuduktan sonra, kompozisyonu dilbilgisi açısından değerlendirip gereken puanı veriyorum. Sonra kelimeye geçiyorum. Kelimeye yönelik ölçütleri okuduktan sonra kompozisyonu kelime açısından değerlendirip gereken puanı veriyorum. Her öğrenci için bu süreci tekrarlıyorum. Geçen seneye göre şuan biraz daha oturttum gibi DPA kullanarak değerlendirme yapmayı.

Değerlendirdiğim ilk sınav kâğıdı ile son sınav kâğıdı arasında elbette güvenirlik farkı olduğunu düşünüyorum. İlk değerlendirdiğim sınav kâğıtlarını daha dinç kafayla değerlendirdiğim için hata yapma olasılığım daha düşük oluyor. Son kâğıtlara yaklaştığım zaman, hem bir kâğıda ayırdığım süre uzuyor, hem de okuduğum cümleleri anlamakta güçlük çekiyorum. Bu yüzden, ara vermem gerekiyor. En az bir ara veriyorum. Yazma becerisi değerlendirme hızımın ise ortalama olduğunu düşünüyorum. Tabi ki geçen sene başına göre çok çok daha hızlıyım diyebilirim.

Dereceli Puanlama Anahtarını ne kadar süredir kullanıyorsunuz?

Yaklaşık 1 yıldır kullanıyorum.

DPA kullanmanın avantajları nelerdir?

Sene içinde kalabalık bir öğrenci grubuna pek çok sınav uyguluyoruz. Çoktan seçmeli sorular hariç, tüm açık uçlu maddeler ve kompozisyonlar okutmanlarca aynı anda değerlendiriliyor. Bu yoğunlukta ve koşuşturmada yanlış yapma olasılığımız çok yüksek olabiliyor. DPA kullanmak hata yapma olasılığımızı en aza indiriyor. Bir nevi standardizasyon sağlıyor.

DPA kullanmanın dezavantajları nelerdir?

DPA kullanarak değerlendirme yapmanın bir dezavantajı olduğunu düşünmüyorum. Ancak, örneğin bir kâğıdı puanlarken dilbilgisine 1, kelimesine 1 verdik. Aslında 1 bile almayı hak etmiyor ama dilbilgisi ve kelimede 0 puan olmadığı için 1 alıyor yine de. Kısacası 20 ya da 30 puan almayı hak eden bir öğrenci, DPA ile puanlama yaptığımız için 50 ya da 60 alabiliyor. Tek dezavantajı bu diyebilirim.

DPA'yı güvenilir bir şekilde kullanmak için, DPA kullanma açısından deneyimi çok olmanın etkili düşünüyor musunuz?

Bence, DPA kullanmak için deneyimden ziyade süreye ihtiyacımız var. DPA'da tüm ölçütler ayrıntıları ile açıklanıyor. Neye kaç puan verilmesi gerektiği yazıyor. İhtiyacımız olan sadece yeteri kadar bir sürede, değerlendirebileceğimiz sayıda sınav kâğıdı değerlendirmek. Deneyimin önemi olduğunu düşünmüyorum.

İdeal bir puanlayıcı grubundaki puanlayıcı deneyim yılı kaç olmalıdır?

1 senenin yeterli olduğunu düşünüyorum. 1 seneden daha fazla süredir DPA kullanan bir puanlayıcı, 1 senedir DPA kullanan puanlayıcılara göre sadece daha hızlı bir puanlama yapabilir. Daha güvenilir bir puanlama yaptıklarını düşünmüyorum.

Grup kaç puanlayıcıdan oluşmalıdır?

Son kompozisyonlara geldiğimiz zaman bazen, dikkatimiz tamamen dağılmış olabiliyor. 2. bir puanlayıcının olması elbette güvenirliği arttırır. Ama 2'den fazla puanlayıcı çok fazla iş demektir. İş yükünün artmasının güvenirlik üzerinde olumsuz bir etkisi olabilir.

Tüm puanlayıcılar tüm bireyleri puanlamalı mıdır? Cevabınız hayır ise, öneriniz nedir?

Böyle bir şeyin mümkün olmadığını düşünüyorum. Tüm puanlayıcıların tüm bireyleri puanladığı bir durumda elde edeceğimiz sonuç hatalarla dolu olur. Bunu yerine iş bölümü daha mantıklı bir tercih olacaktır.

Çok sayıda sınav kâğıdını değerlendirmek puanlama güvenirliğini nasıl etkiler?

Çok fazla sayıda sınav kâğıdı puanladığımız zamanlar da oluyor. O zaman içimde bir kuşku oluyor, acaba tüm kâğıtları değerlendirirken adil davrandım mı diye. O yüzden bazı kâğıtları tekrar tekrar gözden geçirmek zorunda kalıyorum. Çok fazla kâğıt değerlendirdiğim için yoruluyorum zaten. Bir de vicdanen rahat olmak için bazı kâğıtları iki belki üç kez değerlendiriyorum. Elbette bu durumda hata yapma olasılığım çok yüksek. Değerlendirilmesi gereken kâğıt sayısının en fazla 10 ya da 15 olması gerekiyor.

Yorgunluk etkisi güvenirliği nasıl etkiler?

Çok fazla sınav kâğıdı değerlendirmem gereken bir durumda elbette çok yoruluyorum ve yaptığım puanlamayı çok rastgele yaptığımı düşünüyorum. Güvenirliği kötü etkilememek içinse açık havaya çıkıp mola veriyorum. Ya da açsam mutlaka bir şeyler yiyorum.

Puanlama anahtarının daha etkili bir şekilde kullanılması için önerileriniz nelerdir?

Bence, DPA'ları her sene gözden geçirmeliyiz. Kriterler aynı kalmamalı. Örneğin bazen, o zaman kadar fark etmediğimiz ama önemli olan bir nokta oluyor ve bu nokta kendine DPA'da yer bulamayabiliyor. Bunun önüne geçmek için, DPA her sene revize edilmeli diye düşünüyorum.

Katılımcı E:

DPA kullanarak değerlendirme yapma sürecinizden kısaca bahseder misiniz? (Bir kâğıdı ne kadar sürede okursunuz? Kaç sınav kâğıdı sonra yorulmaya başlarsınız? Değerlendirdiğiniz ilk sınav kâğıdı ile son sınav kağıdı arasında güvenirlik açısından bir fark olduğunu düşünüyor musunuz? Düşünüyorsanız bu neden kaynaklanıyor sizce? Değerlendirme yaparken ara verir misiniz? Sizce, diğer okutmanlara göre sınav değerlendirme hızınız nasıl? Önce kompozisyonu tamamen okuyup, daha sonra mı değerlendirme yaparsınız yoksa kompozisyonu her bir değerlendirme boyutu için tek tek bir daha mı okursunuz?)

Sınavları değerlendirmeye başlamadan önce, DPA'yı gözden geçiriyorum. Artık az çok nasıl kullanmam gerektiğini bildiğim için, her bir ölçüt için DPA'ya geri dönmüyorum. Kompozisyonu tamamen okuduktan sonra DPA yardımı ile değerlendirme yapıyorum. Özellikle kafamı karıştıran bir nokta varsa, o zaman kompozisyon ve DPA arasında git gel yapıyorum. Onun dışında, puan verme işini, en sonra bırakıyorum. 20 kâğıt sonra yoruluyorum diyebilirim. Ama genelde ara vermemeyi tercih ediyorum. Genelde tüm kâğıtları tek seferde değerlendiriyorum. Değerlendirdiğim ilk kâğıt ve son

kağıt arasında fark olduğunu düşünüyorum. İlk kâğıtlar daha özenli oluyor bence. Sonlara doğru fiziksel ve zihinsel yorgunluktan dolayı odaklanma problemleri yaşıyorum her ne kadar kendimi zorlasam da. Hızımın ise ortalama olduğunu düşünüyorum. DPA kullanmaya ilişkin deneyim yılı benden fazla olan hocaların daha hızlı değerlendirme yaptıklarını düşünüyorum.

Dereceli Puanlama Anahtarını ne kadar süredir kullanıyorsunuz?

1 yıldır kullanıyorum.

DPA kullanmanın avantajları nelerdir?

DPA ile öğrencinin yazma becerisine yönelik genel bir izlenimden ziyade, daha detaylı bir fikir sahibi olabiliyoruz. Ayrıca, daha objektif bir değerlendirme yapmamızı sağlıyor.

DPA kullanmanın dezavantajları nelerdir?

Bazen bazı öğrencilerin ders içi yazma aktivitelerinde o kadar da başarılı olmadığını ama yazma sınavından yüksek puan aldıklarını görebiliyoruz. Bunun kısmen DPA'dan kaynaklandığını söyleyebilirim. Ölçütler 1 ve 4 arası genelde. Her şeyi yanlış bile yapsa, 1 alabiliyor öğrenci 4 üzerinden. Bu da yanıltıcı oluyor.

DPA'yı güvenilir bir şekilde kullanmak için, DPA kullanma açısından deneyimi çok olmanın etkili düşünüyor musunuz?

Deneyimin güvenilir puanlamaya katkısı olduğunu düşünüyorum.

İdeal bir puanlayıcı grubundaki puanlayıcı deneyim yılı kaç olmalıdır?

En az 1 yıl deneyimin gerekli olduğunu düşünüyorum. Ayrıca, yazma becerisinin değerlendirildiği kurda da ders veriyor olunmalı diye düşünüyorum.

Grup kaç puanlayıcıdan oluşmalıdır?

2 puanlayıcı mevcut durumda iyi bence. 3 olduğu zaman daha güvenilir sonuçlar elde edilebilir ve belki 3.puanlayıcı 2 puanlayıcının göremediği ve gözden kaçırdığı şeyleri fark edebilir. 3 kişi ile yapılan hatalar hakkında daha fazla konuşma imkânımız olabilir.

Tüm puanlayıcılar tüm bireyleri puanlamalı mıdır? Cevabınız hayır ise, öneriniz nedir?

Bence tüm puanlayıcılar mümkünse tüm bireyleri puanlamalı. Herkesin gözünden bir şeyler olabilir ve ne kadar çok puanlayıcı değerlendirme sürecine dâhil olursa, hata yapma ihtimali o kadar azalabilir. Ama bunun pratikte uygulanması mümkün değil. İdeal olan, 2 puanlayıcının anlaşamadığı noktada,3. bir puanlayıcıya danışmaktır.

Çok sayıda sınav kâğıdını değerlendirmek puanlama güvenirliğini nasıl etkiler?

Çok sayıda sınav kâğıdı değerlendirdiğimiz zaman ilk kağıt ve son kağıt arasındaki güvenirlik farkı olmaması için aşırı çaba göstermemiz gerekiyor.

Yorgunluk etkisi güvenirliği nasıl etkiler?

Objektif bir puanlama yapsak, elbette ki puanlama işlemi etkilenmez. Fakat DPA kullanarak bir değerlendirme yaparken, ölçütlerle beraber kendi kişisel kanımızla bir yargıya varıyoruz ve yorgunluk karar verme yetimizi etkiliyor.

Puanlama anahtarının daha etkili bir şekilde kullanılması için önerileriniz nelerdir?

Puanlamanın tek bir günde yapılmasındansa, birkaç güne yayılmasının güvenirlik açısından daha iyi olacağı kanısındayım. Böylece daha etkili ve isteki puanlama yapabiliriz.

Katılımcı F:

DPA kullanarak değerlendirme yapma sürecinizden kısaca bahseder misiniz? (Bir kâğıdı ne kadar sürede okursunuz? Kaç sınav kâğıdı sonra yorulmaya başlarsınız? Değerlendirdiğiniz ilk sınav kâğıdı ile son sınav kağıdı arasında güvenirlik açısından bir fark olduğunu düşünüyor musunuz? Düşünüyorsanız bu neden kaynaklanıyor sizce? Değerlendirme yaparken ara verir misiniz? Sizce, diğer okutmanlara göre sınav değerlendirme hızınız nasıl? Önce kompozisyonu tamamen okuyup, daha sonra mı değerlendirme yaparsınız yoksa kompozisyonu her bir değerlendirme boyutu için tek tek bir daha mı okursunuz?)

DPA kullanırken, anahtar ve puanlayacağım sınav kâğıdı arasında sürekli gel git yapıyorum. Örneğin önce ana fikir cümlesini buluyorum. Sonra puanlama anahtarına gidip, ölçütleri okuyup gereken puanı veriyorum. Önce tüm parçayı okuyup sonra puanlama yapmıyorum çünkü dikkat ettiğim şeyleri unutacağımı düşünüyorum. Ortalama 15 kâğıt sonra yorulduğumu düşünüyorum. Bunun birkaç sebebi var. Değerlendirme yaptığımız odalar çok kalabalık oluyor. Bu da benim dikkatimi dağıtıyor. Ya da değerlendirme yaptığımız sınıfın seviyesi çok düşükse, anlamak daha zor bir hale geldiği için daha çabuk yoruluyorum. Fakat seviyesi yüksek bir öğrencinin kâğıdını okuduğum zaman, odaklanmak da kolay olduğundan, daha geç yoruluyorum. Kâğıt okuma hızımdan bahsedecek olursam, yavaş olduğumu düşünüyorum. En sona kalan hocalardan biri ben oluyorum genelde. Bunun ilk sebebi, dediğim gibi dikkatimin çabuk dağılması. Kalabalık bir ortamda değerlendirme yapmak benim hızımı yavaşlatıyor. Yavaşlığımın deneyim yılımla ilgisi olduğunu düşünüyorum. İşe ilk başladığım zaman ile şuan arasında bir fark olduğunu görüyorum ve ileriki yıllarda daha hızlı değerlendirme yapacağımı düşünüyorum.

Dereceli Puanlama Anahtarını ne kadar süredir kullanıyorsunuz?

Yaklaşık 1 yıldır kullanıyorum.

DPA kullanmanın avantajları nelerdir?

DPA, değerlendirme yaparken daha adil olmamızı sağlıyor bence. Bu anahtarı kullanarak yaptığımız değerlendirmelerde, partnerimizle aramızda çok bir puan farkı çıkmıyor. Hemen hemen aynı puanları verebiliyoruz. Bu bizim güvenilir bir puanlama yaptığımızı gösteriyor diye düşünüyorum.

DPA kullanmanın dezavantajları nelerdir?

DPA'nın dezavantajlarının olduğunu düşünmüyorum.

DPA'yı güvenilir bir şekilde kullanmak için, DPA kullanma açısından deneyimi çok olmanın etkili düşünüyor musunuz?

Yıllar içinde, bir çok sınav kağıdı değerlendirdiğimiz için, DPA'nın da kriterlerini o denli içselleştiyoruz ve belli bir süre sonra bakma ihtiyacı duymuyoruz. DPA'yı ilk kullandığım zamanları hatırlıyorum mesela. Ölçütleri bilmiyordum ve dönüp tek tek bakmam gerekiyordu mesela. Ama şuan ölçütler kafamda daha net. Bir

kompozisyonda neyi aramam gerektiğini biliyorum. Yıllar geçtikçe hem hızımızın hem de yaptığımız puanlamanın güvenirliğinin artığını düşünüyorum.

İdeal bir puanlayıcı grubundaki puanlayıcı deneyim yılı kaç olmalıdır?

Ben 1 senenin yeterli olduğunu düşünüyorum. Daha sonrasında sadece hızımız artıyor.

Grup kaç puanlayıcıdan oluşmalıdır?

Bence 2 puanlayıcı yeterli. Mevcut durumda, 2 puanlayıcı arasında fazla puan farkı çıkmıyor. O yüzden, daha fazla puanlayıcıyı sürece dâhil etmek sadece bizim iş yükümüz artıracaktır. Güvenirliği etkileyeceğini düşünmüyorum.

Tüm puanlayıcılar tüm bireyleri puanlamalı mıdır? Cevabınız hayır ise, öneriniz nedir?

Kesinlikle tüm puanlayıcılar tüm bireyleri puanlamamalı. Bu zaman açısından mümkün değil. Ya da bir değerlendirici grup içinde beraber çalışan puanlayıcılar, değerlendirilmesi gereken kâğıtları paylaşmalı. Tüm puanlayıcıların tüm bireyleri puanladığı bir sistem mümkün değil. Aynı şekilde, DPA'da yer alan ölçütlerin paylaşılması da çok yanıltıcı olabilir.

Çok sayıda sınav kâğıdını değerlendirmek puanlama güvenirliğini nasıl etkiler?

Özellikle, uzun essay tipi bir kompozisyonu puanlarken, özellikle son kâğıtlara doğru, gözümüzden bazı hatalar kaçıyor olabilir.

Çok sayıda sınav kâğıdını değerlendirmek puanlama güvenirliğini nasıl etkiler?

Elbette olumsuz etkiler. En başta zaten o kadar kâğıdı bir arada görünce moralim bozuluyor. 1-0 geride başlıyorum. Tüm kâğıtları bir günde değerlendirmek zorunda olmak güvenirliği zedeliyor bence.

Yorgunluk etkisi güvenirliği nasıl etkiler?

Bu etkiyi, ara vererek azaltabiliriz. Ya da değerlendirme yaptığımız günler sayısını ikiye çıkarabiliriz Gündüz zaten derse giriyoruz. Ardın da bir sürü kâğıt okuduğumuz bazen güvenilir olmayabilir sonuçlar. Bu çok normal. Belki kâğıt sayısı ikiye bölünüp, iki günde tamamlanabilir.

Puanlama anahtarının daha etkili bir şekilde kullanılması için önerileriniz nelerdir?

Ölçütlerin daha detaylı olması gerektiğini düşünüyorum çünkü bir kriterden beklentimiz bazen partnerimizle aynı olmayabiliyor. Kriterler daha açıklayıcı yazılırsa, bence elde edeceğimiz sonuçlar çok daha güvenilir olur.

EK D: Edu G Verileri

File C:\Users\LENOVO\Desktop\tecrubeli\tecrubeli.gen - [2016-10-17 17:51]

TecrubeliAnaliz 1

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
Madde	M	6	INF	
Puanlayici	P	6	INF	
Puanlayici icindebirey	B:P	60	INF	

Observation and Estimation Designs

Source	SS	Df	MS	Components				
				Random	Mixed	Corrected	%	SE
M	12.80926	5	2.56185	0.00400	0.00400	0.00400	0.7	0.00390
P	8.68148	5	1.73630	-0.00293	-0.00293	-0.00293	0.0	0.00274
B:P	678.09444	354	1.91552	0.27794	0.27794	0.27794	51.1	0.02397
MP	28.08519	25	1.12341	0.01459	0.01459	0.01459	2.7	0.00510
MB:P	438.77222	1770	0.24789	0.24789	0.24789	0.24789	45.5	0.00833
Total	1166.44259	2159					100 %	

Analysis of variance

G Study Table
(Measurement design BP/M)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
	.....	M	.....		0.00067	1.5
P	(0.00000)		.....		.....	
B:P	0.27794		.....		.....	
	.....	MP	0.00243	5.6	0.00243	5.5
	.....	MB:P	0.04132	94.4	0.04132	93.0
Sum of variances	0.27794		0.04375	100%	0.04441	100%
Standard deviation	0.52720		Relative SE: 0.20916		Absolute SE: 0.21075	
Coef_G relative	0.86					

Coef_Gabsolute	0.86

Grand mean for levelsused: 1.92685
Variance error of the mean for levelsused: 0.00196
Standard error of the grand mean: 0.04425
File C:\Users\LENOVO\Desktop\tecrubesiz\tecrubesiz.gen - [2016-10-17 17:55]

TecrubesizAnaliz 1

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
Madde	M	6	INF	
Puanlayici	P	6	INF	
Puanlayici icindebirey	B:P	60	INF	

Analysis of variance

Source	SS	df	MS	Components				
				Random	Mixed	Corrected	%	SE
M	16.63565	5	3.32713	0.00532	0.00532	0.00532	0.9	0.00505
P	17.84676	5	3.56935	0.00141	0.00141	0.00141	0.2	0.00542
B:P	685.64167	354	1.93684	0.27548	0.27548	0.27548	47.1	0.02425
MP	35.25602	25	1.41024	0.01877	0.01877	0.01877	3.2	0.00640
MB:P	502.60833	1770	0.28396	0.28396	0.28396	0.28396	48.5	0.00954
Total	1257.98843	2159					100 %	

G Study Table

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
	.....	M	.....		0.00089	1.7
P	0.00141		.....		.....	
B:P	0.27548		.....		.....	
	.....	MP	0.00313	6.2	0.00313	6.1
	.....	MB:P	0.04733	93.8	0.04733	92.2
Sum of variances	0.27689		0.05046	100%	0.05134	100%

(Measurement design BP/M)

Standard deviation	0.52620	Relative SE: 0.22462	Absolute SE: 0.22659
Coef_G relative	0.85		
Coef_Gabsolute	0.84		

Grand mean for levelsused: 2.87269
Variance error of the mean for levelsused: 0.00254
Standard error of the grand mean: 0.05040
File C:\Users\LENOVO\Desktop\tecrubelitecrubesiz\tecrubelitecrubesiz.gen - [2016-10-17 18:01]

TecrubeliTecrubesizAnaliz 1

Observation and Estimation Designs

	Label	Levels	Univ.	Reduction (levels to exclude)
Madde	M	6	INF	
Puanlayici	P	2	INF	
Puanlayici icindebirey	B:P	60	INF	

Analysis of variance

Source	SS	df	MS	Components				
				Random	Mixed	Corrected	%	SE
M	2.76667	5	0.55333	-0.00163	-0.00163	-0.00163	0.0	0.00415
P	4.35556	1	4.35556	0.00491	0.00491	0.00491	0.8	0.00997
B:P	251.64444	118	2.13258	0.30670	0.30670	0.30670	50.1	0.04597
MP	3.74444	5	0.74889	0.00761	0.00761	0.00761	1.2	0.00668
MB:P	172.48889	590	0.29235	0.29235	0.29235	0.29235	47.8	0.01699
Total	435.00000	719					100 %	

G Study Table
(Measurement design BP/M)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
	.....	M	.....		(0.00000)	0.0
P	0.00491		.....		.....	
B:P	0.30670		.....		.....	
	.....	MP	0.00127	2.5	0.00127	2.5
	.....	MB:P	0.04873	97.5	0.04873	97.5
Sum of variances	0.31161		0.04999	100%	0.04999	100%
Standard deviation	0.55822		Relative SE: 0.22359		Absolute SE: 0.22359	
Coef_G relative	0.86					
Coef_Gabsolute	0.86					

Grand mean for levelsused: 2.91667
Variance error of the mean for levelsused: 0.00605
Standard error of the grand mean: 0.07778

ÖZGEÇMİŞ

Hatice Özlem ANADOL

1 Haziran 1988'de Bursa'da doğdu. İlk, orta ve lise öğrenimini Bursa'da tamamladı.2010 yılında Ortadoğu Teknik Üniversitesi Eğitim Fakültesi İngilizce Öğretmenliği bölümünden lisans derecesini aldı. 2010 yılında Milli Eğitim Bakanlığı bünyesinde İngilizce öğretmeni olarak göreve başladı. Görevinden 2013 yılında ayrılıp aynı yıl TOBB Ekonomi ve Teknoloji Üniversitesi'nde okutman olarak göreve başladı. 2015 yılında Karadeniz Teknik Üniversitesi Sosyal Bilimler Enstitüsü Batı Dilleri ve Edebiyatı Uygulamalı Dil Bilim alanında Yüksek Lisans derecesini aldı.

Halen TOBB Ekonomi ve Teknoloji Üniversitesi'nde okutman olarak çalışmaktadır.